Quantificando (e qualificando) o sujeito oculto em português

Cláudia Freitas, Elvis de Souza, Luisa Rocha

Departamento de Letras
PUC-Rio – Brasil

claudiafreitas@puc-rio.br, elvis.desouza99@gmail.com, l.rocha7@globo.com

Abstract. *In information extraction we seek to find out who does what, when, where, and why. In Portuguese, it is possible to construct sentences with the omission of its subject. This omission – easy for humans but difficult for machines – causes information extraction to be compromised since we cannot fill the gap of who is responsible for the action. In this paper, we present a quantification of the hidden subject in Portuguese in corpora of different textual genres: journalistic, encyclopedic and literary. To do so, we make use of morphosyntactically annotated corpora and use a tool that allows us to interrogate corpora annotated with syntactic dependencies.*

Resumo. *Na extração de informações buscamos descobrir quem faz o quê, quando, onde, e por que. Diferentemente do inglês, na língua portuguesa é possível construir sentenças inteiras com a omissão do sujeito. Essa omissão – de fácil recuperação para humanos, mas difícil para máquinas – faz com que a extração de informações fique comprometida, uma vez que não temos como preencher a lacuna de quem é o responsável pela ação, já que boa parte da função sintática do sujeito corresponde à função semântica de agente. Neste trabalho, apresentamos uma quantificação do sujeito oculto em português em corpora de diferentes gêneros textuais: jornalístico, enciclopédico e literário. Para tanto, fazemos uso de corpora morfossintaticamente anotados e utilizamos uma ferramenta que nos permite interrogar corpora anotados com dependências sintáticas.*

1. Apresentação e motivação

Toda descrição é feita a partir de um ponto de vista. Ainda que, na descrição das línguas, a descrição seja frequentemente assumida como uma área/tarefa em si, lembramos que ela não é, e nem poderia ser, neutra ou desvinculada de uma motivação. Neste trabalho, assumimos uma descrição motivada por uma aplicação: a extração automática de informação em textos, uma das tarefas do PLN.

De maneira geral e muito simplificadamente, na extração de informações buscamos descobrir *quem* faz o *quê* (e *quando*, *onde*, e *por que*, dentre outros). Diferentemente do inglês (que concentra boa parte dos trabalhos de PLN), no entanto, a língua portuguesa permite a omissão do sujeito. Um dos tipos de omissão acontece em contextos em que o sujeito é facilmente recuperável – quer pelas pistas flexionais do verbo, quer pelo nosso conhecimento de mundo [1]. Nesses casos, temos o chamado "sujeito oculto" segundo a tradição gramatical.

[1] No entanto, como bem notou um dos revisores, embora a desinência verbal possa identificar a "pessoa"

A omissão do sujeito – facilmente recuperável pelos humanos, mas difícil para as máquinas – faz com que a extração de informações (e tarefas relacionadas, como a extração de citação (*Quotation Extraction*) e a anotação de papéis semânticos) fique comprometida, uma vez que não temos como preencher a lacuna de quem é o responsável pela ação, já que boa parte da função sintática do sujeito corresponde à função semântica de agente.

Mas qual o tamanho do problema para a língua portuguesa? E como quantificá-lo? Qual a correlação entre o sujeito oculto e gêneros textuais? Em trabalho anterior [Martins and Freitas 2019], realizamos de maneira semi-automática a quantificação de sujeitos ocultos em 6 verbetes do Dicionário Histórico Biográfico Brasileiro (DHBB), uma enciclopédia sobre a história política brasileira, publicada pelo CPDOC/FGV, que já vem sendo objeto de extração de informação [Higuchi et al. 2019]. Especificamente, foram escolhidos 6 verbetes biográficos que, juntos, totalizavam 25 mil palavras. Essa pequena amostra foi lida integralmente para a identificação dos casos de sujeito oculto. Em seguida, o texto foi processado automaticamente pela ferramenta UDPipe [Straka and Straková 2017] para a contabilização total dos verbos. Neste pequeno exercício, obtivemos uma distribuição desigual do sujeito oculto por verbete: o verbete com mais sujeitos ocultos apresentava o fenômeno em 45% das frases, e aquele com menos sujeitos ocultos, 18%. Considerando o tamanho do DHBB (mais de 8 milhões de palavras), e a relevância da explicitação do sujeitos para a extração de informação, decidimos ampliar o estudo, levando a cabo uma análise de todo o DHBB, e comparando os resultados com textos de outros gêneros textuais. Um desafio associado à tarefa é a existência de uma ferramenta que possibilite a contagem, já que no exercício de [Martins and Freitas 2019], todo o trabalho foi feito manualmente.

Neste trabalho, apresentamos uma quantificação do sujeito oculto em português em corpora de diferentes gêneros textuais: jornalístico, enciclopédico e literário. Para tanto, fazemos uso de corpora morfossintaticamente anotados e utilizamos uma ferramenta que nos permite interrogar corpora anotados com dependências sintáticas.

2. O que conta como sujeito oculto?

A língua portuguesa possui alguns fenômenos que envolvem a omissão de sujeito: o sujeito oculto propriamente, o chamado sujeito indeterminado e, ainda, as orações sem sujeito. Considerando nossa motivação principal, tarefas de PLN, não fizemos distinção, em nossa contagem, entre sujeito oculto (1) e sujeito indeterminado (2) - ou seja, nossa busca (e quantificação) pelas frases com "sujeito oculto" considera ambos os casos, já que, em ambos, não há sujeito explícito, ainda que exista um sujeito. A gramática [Cunha and Cintra 2008], por exemplo, considera o primeiro "sujeito oculto (determinado)" e o segundo "sujeito indeterminado", sendo a diferença entre eles a possibilidade de determinação do sujeito pela desinência do verbo. Do mesmo modo, não levamos em conta as chamadas orações sem sujeito, como frases com verbos impessoais (3) e fenômenos da natureza.

---

a que corresponde o sujeito, nem sempre é possível recuperar o sujeito sem resolver a correferência do pronome pessoal. Isto porque as segundas pessoas são conjugadas como as terceiras pessoas (por exemplo, para preencher o sujeito da oração *Conseguiram um grande avanço* temos três possibilidades: *vocês*, *eles* ou *elas*.)

1. CP22-3: “Eu tentei, o senhor Vance tentou, se for respeitado, urrah!”, **comentou**.
2. CP31-3: Sempre que surge um problema, **chamam**-na
3. CP23-8: – **Há**, no ar, uma certa ideia de invasão.

Também não consideramos o sujeito oculto em orações subordinadas ou coordenadas à oração principal. Isto porque, nesses casos, o sujeito deve poder ser retomado no âmbito da frase. Ou seja, os sujeitos ocultos contabilizados foram apenas aqueles das orações principais.

Por fim, sabemos também que, em português, é possível que o *-se* exerça a função de um sujeito indeterminado. Neste caso, temos a seguinte situação: a presença formal de um elemento que conta como sujeito mas que, na prática, funciona como um índice de indeterminação. 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 – porque são aqueles em que seria possível recuperar o sujeito – dos demais casos, não nos preocupamos em dar ao *-se* sujeito, neste momento, um tratamento especial, e também excluímos esses casos de nossa contagem.

Não somos os primeiros a se interessar pela a omissão do sujeito de um ponto de vista quantitativo. Em [Sardinha et al. 2014], por exemplo, a omissão do sujeito – codificada como *subjdrop* –é um dos critérios elencados para a caracterização de gêneros textuais em português. No entanto, não sabemos quais fenômenos são abarcados sob o referido rótulo. Os autores indicam que os critérios tiraram proveito da anotação feita pelo PALAVRAS [Bick 2000], e sabemos que o PALAVRAS reconhece (e distingue) verbos de orações sem sujeito explícito (sujeito oculto) e verbos de orações sem sujeito formal (oração sem sujeito), informação disponível desde 2008 nos corpora da Floresta Sintá(c)tica [Freitas et al. 2008]. Talvez, para a caracterização de gêneros textuais, seja irrelevante a diferença. No PLN, não é, dado que, nos sujeitos ocultos, recurso estilístico, podemos (e queremos) recuperar o sujeito sintático da oração.

3. Método e resultados

A pesquisa foi realizada nos corpora Bosque-UD (versão 2.4), com 9.366 frases; DHBB, com 323.301 frases; e em um subconjunto das obras de Machado de Assis (especificamente todos os contos, crônicas e romances), que totalizam 323.301 das frases do corpus OBras. Todo o material foi anotado pelo UDPipe [Straka and Straková 2017] e está no formato Universal Dependencies (UD) [Nivre et al. 2016] [2]. Apenas o Bosque-UD teve sua anotação gramatical revista [Rademaker et al. 2017], por ser ser o material de treino do parser UDPipe.

Para realizar as pesquisas, utilizamos a ferramenta *Interrogatório*, um dos ambientes da ET, uma Estação de Trabalho para busca, revisão, edição e avaliação de corpora anotados [de Souza and Freitas 2019] – trata-se de uma ferramenta que surgiu motivada exatamente pela tarefa de contar sujeitos ocultos, já que não temos notícia de uma ferramenta de fácil acesso capaz de realizar contagens em material anotado com dependências sintáticas, formato subjacente à abordagem UD.

[2] http://universaldependencies.org

O principal desafio na contabilização dos sujeitos ocultos está no fato de que precisamos contar algo que não está presente. Na sintaxe de dependências, a primeira etapa foi encontrar todos os verbos de orações principais que não têm um sujeito que dele dependa. Tomando o Bosque-UD como exemplo, a pesquisa retornou 2774 frases. No entanto, essa expressão de busca retornou também outros tipos de frases, como construções com o verbo *haver* impessoal (exemplo 3). Criamos então, na ferramenta, um filtro para eliminar essas frases. Em seguida, fizemos mais um filtro, para eliminar as frases em que o verbo indicava um fenômeno da natureza.

Depois de aplicados os filtros, ficamos com 1480 sentenças, ou seja, cerca de 16% das frases do Bosque-UD. No DHBB, utilizando os mesmos critérios e expressões de busca, verificamos que 39.5% do corpus apresenta sujeitos ocultos. Por fim, no material literário (ou, ao menos, na escrita de Machado de Assis), 28.42% das frases continha sujeito oculto (tabela 1). Os resultados indicam que, independentemente do tipo de texto, os números são altos, justificando um tratamento especial para o fenômeno no PLN em português.

Tabela 1. Distribuição de sujeitos ocultos por corpus

Corpus	Frases com sujeito oculto
Bosque-UD (v.2.4)	15.8%
DHBB	39.5%
Machado de Assis	28.42%

Um dado interessante é a diferença na distribuição do fenômeno. O DHBB é, de longe, o corpus com mais sujeitos ocultos. A constatação é facilmente explicada quando sabemos que o material possui dois tipos de verbetes: biográficos e temáticos, sendo os verbetes do primeiro tipo a maioria. E, justamente por se tratar de um artigo biográfico, a omissão do sujeito – na imensa maioria das vezes, correspondente à pessoa verbetada – funciona como um recurso estilístico capaz de trazer fluidez ao texto, evitando repetições desnecessárias. Em seguida, temos o texto literário e o Bosque, composto por textos jornalísticos. O material de Machado de Assis possui o dobro de sujeitos ocultos do Bosque. Por outro lado, o Bosque-UD teve sua anotação revista, e o corpus com as obras de Machado, não. Além disso, vale lembrar que tanto este corpus, como o DHBB, foram anotados por um modelo que foi treinado no Bosque-UD. Deste modo, é possível que a contagem sofra efeitos de uma anotação sintática malfeita.

A fim de verificar o quanto os resultados da tabela 1 são confiáveis, analisamos manualmente uma amostra de 150 frases identificadas como sujeito oculto, 50 de cada corpus/gênero. Destas, 134 (90%) foram avaliadas como corretas, o que nos dá confiança quanto aos números apresentados, sobretudo no que se refere ao DHBB, onde todas as frases identificadas estavam corretas. A tabela 2 apresenta a distribuição dos resultados.

O corpus com textos literários (e diacrônico) é o que mais apresenta erros. A análise dos casos errados (apenas 16 erros) indicou que o principal motivo do erro (9 casos) decorre do processamento automático, e acontece quando temos um sujeito posposto ao verbo (1), sobretudo se estamos diante de um sujeito oracional (2). Em seguida, temos 3 erros que decorrem da nossa forma de pesquisa: buscamos pela ausência de sujeito na oração principal, mas não é raro que orações adverbiais antecedam a oração principal,

Tabela 2. Resultados da análise manual de 150 frases, por

Corpus	Acertos
Bosque-UD (v.2.4)	44 (88%)
DHBB	50 (100%)
Machado de Assis	40 (80%)
Total de acertos	134 (90%)

trazendo com elas o sujeito (3). Isto é algo que precisamos melhorar para números mais precisos, mas não acreditamos que a frequência dessas construções seja capaz de interferir de maneira significativa na contagem final – foram apenas 3 casos desses em 150 frases. Por fim, tivemos também 4 ocorrências, no Bosque-UD, da construção com sujeito indeterminado (construções com "tratar-se de"), e também precisaremos eliminar esses casos da busca.

1. Mas por outro lado, sem a apresentação de Miss Dollar, *seria* o **autor** obrigado a longas digressões, que encheriam o papel sem adiantar a ação
2. Era conveniente ao romance *que o leitor ficasse muito tempo sem saber quem era Miss Dollar*.
3. Se eu dirigisse uma federação, *apresentaria* balanços mensais e liberaria minhas contas bancárias

4. Considerações finais

Apresentamos aqui um primeiro estudo sobre a quantificação do sujeito oculto em português, levando em conta também o gênero textual. Os resultados indicam que o gênero é um aspecto relevante no que se refere à frequência do sujeito oculto. No caso de uma enciclopédia biográfica, por exemplo, e quando pensamos em cadeias de relações como *quem* faz *o quê* (e *quando* e *onde*), em cerca de 40% dos casos não temos como resolver a dimensão *quem*. Por outro lado, em textos jornalísticos, a quantidade de frases sem sujeito cai para cerca de 15%, o que, se não é muito, também não é insignificante.

Outro ponto a ser destacado é a necessidade de ferramentas capazes de auxiliar linguistas na sua tarefa de manipulação de grandes corpora anotados. Em nosso caso, a necessidade de contar de forma simples o sujeito oculto acabou levando ao desenvolvimento de um ambiente complexo para trabalhar com corpus, de uma maneira linguisticamente motivada.

Por fim, reforçamos os ganhos do lado linguístico e do lado computacional ao se pensar uma descrição do português motivadas pelos desafios empíricos do PLN.

Referências

Bick, E. (2000). *The Parsing System "Palavras": Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework*. PhD thesis.

Cunha, C. and Cintra, L. (2008). *Nova gramática do português contemporâneo*. Lexikon.

de Souza, E. and Freitas, C. (2019). 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)*.

Freitas, C., Rocha, P., and Bick, E. (2008). Um mundo novo na floresta sint á(c)tica – o treebank do português. *Calidoscópio*, 6(3):142–148.

Higuchi, S., Santos, D., Freitas, C., and Rademaker, A. (2019). Distant reading brazilian politics. In Navarretta, C., Agirrezabal, M., and Maegaard, B., editors, *Proceedings of the Digital Humanities in the Nordic Countries 4th Conference*, volume 2364, Copenhagen, Denmark. http://ceur-ws.org/Vol-2364/.

Martins, F. and Freitas, C. (2019). Sujeitos ocultos em verbetes biográficos: Contornando dificuldades da extracção automática de informacções. In *XI Congresso Internacional da ABRALIN*.

Nivre, J., De Marneffe, M.-C., Ginter, F., Goldberg, Y., Hajic, J., Manning, C. D., McDonald, R., Petrov, S., Pyysalo, S., Silveira, N., et al. (2016). Universal dependencies v1: A multilingual treebank collection. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, pages 1659–1666.

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

Sardinha, T. B., Kauffmann, C., and Acunzo, C. M. (2014). A multi-dimensional analysis of register variation in brazilian portuguese. *Corpora*, 9(2):239–271.

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