

Um pronome com muitas funções: Descrição e resultados da anotação do pronome *-se* em um treebank segundo o esquema *Universal Dependencies* (UD) para Português

Elvis de Souza¹, Cláudia Freitas²

¹Departamento de Letras – PUC-Rio
Lab. Inteligência Computacional Aplicada – PUC-Rio

²Departamento de Letras – PUC-Rio

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

Abstract. *In this work, we conducted a linguistic description and reported the annotation process of the pronoun *-se* in the PetroGold v3 treebank [de Souza 2023]. Special attention to the pronoun *-se* is justified by the need to correctly annotate cases where the pronoun indicates subject indeterminacy, synthetic passive voice, or pronominal verb, recognizing its relevance for various NLP tasks. As a result, we categorized the 1,960 occurrences of “se” in the corpus according to their syntactic class and presented the verbs associated with each type (or more than one type) of the pronoun *-se*.*

Resumo. *Neste trabalho, realizamos uma descrição linguística e relatamos o processo de anotação do pronome *-se* no treebank PetroGold v3 [de Souza 2023]. A atenção especial ao pronome *-se* se justifica pela necessidade de anotar corretamente os casos em que o pronome indica indeterminação do sujeito, voz passiva sintética ou verbo pronominal, reconhecendo sua importância para diversas tarefas de PLN. Como resultados, discriminamos as 1.960 ocorrências do “se” no corpus por classe sintática e apresentamos os verbos que se associam a cada um (ou mais de um) dos tipos do pronome *-se*.*

1. Introdução

Neste trabalho, realizamos uma descrição linguística e relatamos o processo de anotação do pronome *-se* no treebank PetroGold v3 [de Souza 2023] (250.605 tokens). Apresentamos a metodologia – ferramentas e procedimentos – empregada na anotação e na sua avaliação, as decisões linguísticas subjacentes à anotação e discutimos os resultados sob diferentes pontos de vista.

O treebank integra o projeto *Universal Dependencies* (UD) [de Marneffe et al. 2021], e portanto seguimos as diretrivas de UD para conduzir a anotação. Enquanto as diretrivas, por um lado, permitem uma anotação muito simples para o pronome *-se*, agrupando todos sob uma mesma etiqueta sintática, por outro lado há também a possibilidade, igualmente prevista nas diretrivas, de utilizar etiquetas mais específicas para classificar o pronome. Utilizamos gramáticas do português [Bagno 2012, Bechara 2012, Cunha and Cintra 2016] para ajudar a realizar a anotação dos tipos específicos de pronome *-se* de uma maneira que fosse respaldada tanto pelas diretrivas do projeto UD quanto pela tradição gramatical da língua portuguesa.

A atenção especial ao pronome *-se* se justifica, em termos gerais, pela necessidade de garantir a confiança na anotação padrão ouro do *corpus* e pelo desejo de alinhar as diretrivas do projeto UD às gramáticas do português, facilitando o diálogo entre estudos linguístico-descritivos para a língua portuguesa e tarefas linguístico-computacionais. Em termos específicos, há também a necessidade de anotar corretamente os casos em que o pronome *-se* indica indeterminação do sujeito, voz passiva sintética ou verbo pronominal, reconhecendo que é fundamental para a identificação dos argumentos de um verbo, sendo importante para as tarefas de identificação de papéis semânticos e extração de informação, que são dificultadas em frases de sujeito omitido (pelo uso do *-se*, por exemplo), um problema já abordado anteriormente [Duran and Aluísio 2011, Hartmann et al. 2014, Freitas and de Souza 2021].

A identificação de sujeitos ocultos, por exemplo, só pode ser realizada satisfatoriamente quando o pronome *-se* é corretamente identificado porque, em frases de sujeito indeterminado, não há sujeito a ser encontrado; quando o pronome indica voz passiva sintética ou verbo pronominal, porém, sabe-se que a ausência de um sujeito sintático é resultado de elipse (e que este não é o sujeito semântico nas orações de voz passiva), e que deve poder ser encontrado no texto.

Além de discutir as dificuldades e os resultados da anotação, apresentamos também o impacto das revisões do pronome *-se* na geração de um modelo de aprendizado de máquina. Por fim, disponibilizamos três recursos lexicais processáveis computacionalmente, que podem auxiliar em outros projetos de anotação e em tarefas de processamento de linguagem natural (PLN).

2. Metodologia

A anotação do pronome *-se* foi realizada no *corpus* PetroGold, um *treebank* padrão ouro (250.605 *tokens*) composto por 19 teses e dissertações do domínio do petróleo. A revisão da anotação do pronome foi publicada na terceira versão do recurso, disponível tanto nas páginas do projeto Petrolês quanto do projeto *Universal Dependencies*¹.

O *corpus* integra o projeto *Universal Dependencies* (UD) [de Marneffe et al. 2021], uma iniciativa para padronização da anotação morfosintática e para disponibilização de *datasets* e ferramentas para diversas línguas. Como consequência, o mais importante no processo de anotação do PetroGold é garantir que as diretrivas de anotação empregadas na tarefa se alinharam às diretrivas do projeto.

Antes da revisão relatada neste trabalho, o *corpus* continha uma única etiqueta de função sintática para designar todos os pronomes *-se*: a etiqueta *expl*, para pronomes expletivos. Essa anotação tinha origem no modelo que primeiro anotou o recurso, treinado majoritariamente no *corpus* Bosque-UD [Rademaker et al. 2017], um *treebank* composto por textos jornalísticos que também integra o projeto UD. A anotação não é incorreta segundo as diretrivas do projeto, porém há a possibilidade – prevista nas diretrivas – de utilizar etiquetas específicas para diferenciar o tipo de pronome sendo empregado em cada uma das ocorrências no *corpus*. Realizamos essa especificação com a ajuda de gramáticas do português (doravante GT, ou gramáticas tradicionais), uma vez que as diretrivas do

¹Os endereços dos projetos são, respectivamente, <https://petroles.puc-rio.ai> e <https://universaldependencies.org>. Acesso em 15 de jun. 2023.

projeto UD ainda não têm um detalhamento da descrição para casos específicos da língua portuguesa.

Dessa forma, o processo de anotação do pronome *-se* no PetroGold consistiu em (1) revisar a anotação do pronome *-se*, garantindo que todas as ocorrências anotadas como pronome (e com a etiqueta de função *expl*) são de fato os pronomes que queremos anotar, e (2) especificar o tipo de pronome *-se*, diferenciando quando (a) o pronome tem função de indeterminação do sujeito, (b) é utilizado para empregar voz passiva sintética ou (c) indica um verbo sendo usado pronominalmente (cada um dos três usos corresponde a etiquetas de função sintática diferentes para o pronome).

A etapa (1), relativa à revisão da anotação da classe gramatical, função sintática e lema do “se” para garantir que eram de fato pronomes – e não conjunções subordinativas, por exemplo –, foi realizada utilizando a ferramenta [de Souza and Freitas 2021], um ambiente de busca e revisão de *corpora* anotados morfossintaticamente. Realizamos uma busca por todas as formas da palavra “se” e, para facilitar a revisão de todas as ocorrências, pedimos na ferramenta a distribuição dos casos de acordo com a anotação original de classe gramatical, função sintática e lema, tornando menos trabalhosa a identificação humana de erros uma vez que as ocorrências haviam sido agrupadas nas buscas. Identificados os casos errados, foram corrigidos em lote por meio de regras de correção, uma das possibilidades da ferramenta.

A etapa (2), relativa à especificação do tipo de pronome *-se* (índice de indeterminação, pronome apassivador ou partícula integrante do verbo), foi realizada de forma semelhante à etapa (1), agrupando frases semelhantes – dessa vez, tentamos agrupar os casos de indeterminação de sujeito, verbo pronominal e voz passiva sintética pelas características do verbo ao qual o pronome *-se* se associa. Pedimos à ferramenta pela distribuição dos lemas verbais dos quais o *token* “se” é dependente sintaticamente, o que nos permitiu visualizar a lista de verbos, agrupá-los quanto à transitividade, e então utilizamos o agrupamento para facilitar a leitura e a análise de todas as frases para verificar qual seria a anotação correta para o pronome nos contextos respectivos².

Por fim, para verificar o impacto das revisões e da especificação das funções do pronome *-se* no aprendizado automático – ou, na generalização de casos –, comparamos a saída de um modelo treinado no *dataset* antes das revisões e de outro modelo treinado no *dataset* após as revisões. A hipótese em que se baseia essa metodologia é a de que a qualidade do modelo gerado indica a possibilidade de generalização das análises linguísticas e, indiretamente, o nível de consistência dessas mesmas análises. Mas é importante notar que baixa qualidade do modelo (desempenho fraco) não significa, necessariamente, inconsistência de análise, uma vez que análises podem ser consistentes e de difícil generalização. Além disso, no caso da anotação do pronome *-se*, a análise dos resultados do modelo indica também as dificuldades que a implementação de três novas etiquetas impõem ao modelo de anotação automática.

A ferramenta utilizada para treinar o modelo foi o UDPipe [Straka et al. 2016], e as métricas empregadas na avaliação da qualidade do modelo foram as métricas da avaliação conjunta do CoNLL de 2018 [Zeman et al. 2018]. Daremos ênfase aos resulta-

²Para reproduzir a metodologia de anotação, quem lê pode se referir a [de Souza 2023], onde é apresentado o passo a passo utilizado para especificar o pronome *-se*.

dos de UPOS (*Universal Part-Of-Speech Score*, que mede os acertos de classe gramatical), LEMMA (que mede os acertos de lematização) e LAS (*Labeled Attachment Score*, que mede os acertos de encaixe das dependências sintáticas e de tipo de relação sintática).

3. Classes do *–se*

Para especificar o pronome *–se* foram utilizados três subtipos da relação *expl – expl:impers*, *expl:pass* e *expl:pv* –, além das classes para objetos – *obj* e *obj^j*³. Em todos esses casos, o token “*se*” é o que recebe a etiqueta de relação sintática específica, e a sua anotação de classe gramatical é “PRON”.

Embora o termo “*impers*” da classe “*expl:impers*” diga respeito à impessoalização (*impersonal*, em inglês), entendemos, junto com [Bagno 2012], que tanto a indeterminação do sujeito quanto a voz passiva sintética são estratégias utilizadas para impessoalizar a oração. No caso da indeterminação do sujeito, a impessoalização ocorre pela supressão do sujeito, que não pode ser recuperado na oração, ao passo que no caso da voz passiva sintética a impessoalização ocorre pelo deslocamento do objeto direto para a posição de sujeito paciente, sendo que o agente também não é recuperável dentro da oração. Estudos em diferentes abordagens teóricas também defendem a função de indeterminação do *–se* nos casos que gramáticas normativas classificam como partícula apassivadora [Vieira and de Sá 2015, Lopes and Namiuti-Tempone 2017, dos Santos Silva 2021]. No entanto, uma vez que o conjunto de etiquetas UD prevê a tripla diferenciação, optamos por mantê-la. Outro argumento para a separação “tradicional” entre passiva e indeterminação é justamente viabilizar estudos descritivos como os citados. Por fim, e de um ponto de vista prático, se o interesse estiver apenas na distinção entre indeterminação e demais usos – relevante, por exemplo, na anotação de papéis semânticos – é possível dar um tratamento unificado a *expl:impers* e *expl:pass*.

Assim, estamos utilizando a etiqueta *expl:impers* especificamente para os casos de indeterminação de sujeito e a etiqueta *expl:pass* para os casos de voz passiva sintética. A diferenciação entre ambas as classes, portanto, não se dá pela noção de “impessoalização”, mas por critérios gramaticais que serão explicados a seguir⁴.

***expl:impers* – índice de indeterminação do sujeito:** Comumente, o verbo cujo sujeito está indeterminado é intransitivo ou transitivo indireto. Um dos requisitos para a anotação de indeterminação do sujeito é a ausência de um sujeito sintático para o verbo a que o “*se*” está associado.

1. Considerando-se que o ciclo de o motor é realizado a cada duas rotações completas de o eixo de manivelas, **chega-se** a expressão de a massa de combustível (em kg) utilizada em cada ciclo: (3.1)

³A palavra “*se*” pode ainda ser uma conjunção subordinativa ou um nome próprio – abreviação de sudeste ou Sergipe (SE). Como não são casos de pronome *–se*, não serão tratados neste trabalho.

⁴As diretrivas apresentadas nesta seção resumem, devido a limitações de espaço, os critérios utilizados para anotar o pronome *–se* no *corpus* PetroGold. Para um detalhamento maior sobre casos específicos e dúvidas que surgiram durante a anotação, quem lê pode se referir ao trabalho [de Souza 2023], onde as diretrivas empregadas são exploradas mais extensamente.

expl:pass – pronome apassivador: É condição para a ocorrência de voz passiva sintética um verbo com transitividade direta ou transitividade direta e indireta segundo a GT. Em outras palavras, o requisito é a presença de um objeto direto, o qual, na transformação para voz passiva sintética, será anotado como sujeito do ponto de vista sintático, embora seja paciente do ponto de vista semântico.

Diferentemente da oração em que há indeterminação do sujeito, nesta há um sujeito sintático, anotado como *nsubj:pass* – sujeito paciente.

2. Para a análise de as argilas estudadas, **seguiu-se** o seguinte procedimento; **secou-se** as argilas em 39 uma estufa a 80°C durante 18 horas, e após terem sido retiradas de a estufa, foram moídas em um moinho de bolas durante 18 horas.

Na frase 2, o fato de que “argilas”, sendo sujeito da oração, não concorda em número com “secou”, não é um bom critério para definir se houve indeterminação do sujeito ou voz passiva sintética. Entendemos, assim como [Bagno 2012], que em ambos os casos o objetivo é impessoalizar a oração, de tal maneira que já se tornou usual conjugar o verbo na terceira pessoa do singular, mesmo que o fenômeno empregado seja o da voz passiva sintética. Assim, embora segundo a GT tenha ocorrido um erro de concordância verbal, ele é explicado pela intenção do autor, que não distingue entre um sujeito paciente e uma oração de sujeito indeterminado na hora de impessoalizá-la, e o *-se* foi anotado como *expl:pass*.

expl:pv – verbo pronominal: Indica que o *-se* está associado a um verbo pronominal e recebe o nome, na GT, de partícula integrante do verbo. Essa é uma análise que, para nós, assim como as demais categorias, depende da utilização do verbo em cada frase, não sendo característica intrínseca dos verbos exigir ou não o pronome “se” (veja-se o caso dos verbos “pronominais acidentais”, por exemplo). O verbo está sendo usado de forma pronominal quando há um sujeito sintático e ele não é paciente da ação verbal.

3. Dentre os parâmetros reológicos mais usuais, destaca-se a viscosidade, que se refere a a resistência que uma substância apresenta ao fluxo (...)

Em alguns casos, o sujeito não é nem paciente e nem agente, quando um verbo causativo tornou-se incoativo pelo uso do pronome *-se*, como na frase “O esporte popularizou-se”, levantada em [Duran et al. 2013]. Nela, o verbo, originalmente transitivo direto, está sendo empregado no aspecto incoativo – “o esporte ficou popular” – indicando uma mudança de estado e, por isso, “esporte” não é nem agente nem paciente de um verbo de ação, mas “sede” da mudança de estado indicada pelo verbo (de estado), termo empregado por [Cunha and Cintra 2016]. Frases do tipo foram anotadas como pronominais⁵.

obj / iobj – objeto direto ou indireto: A palavra “se” pode ainda ser anotada como objeto direto ou indireto quando a ação do verbo se estende à terceira pessoa do singular ou plural, na forma do pronome “se”. Nesses casos, o sujeito sintático também é agente e paciente da ação ao mesmo tempo, de modo que o verbo precisa ter um objeto, direto ou indireto. O fenômeno recebe o nome de pronome reflexivo na GT.

⁵Mais discussão sobre esse tipo de construção pode ser encontrada em [Cançado and Amaral 2010].

[Bechara 2012] indica que o pronome “se” como objeto exige um sujeito animado para o verbo, pois somente dessa forma o sujeito será agente e paciente da ação ao mesmo tempo. Em uma frase como “João se banha”, o sujeito é um ser animado e funciona como agente e paciente da ação do verbo. Já na frase “O banco só se abre às 10 horas”, para o autor, o sujeito inanimado impede a ocorrência de pronome reflexivo, sendo um caso de voz passiva. No caso de “Ele se chama João”, sabe-se que, embora animado, o sujeito não é agente da ação, restando a anotação de verbo pronominal.

No PetroGold não foram encontradas ocorrências de “se” como pronome reflexivo (objeto direto ou indireto). Para confirmar que a inexistência do pronome reflexivo no PetroGold está correta, verificamos todos os sujeitos de verbos a que se associam o pronome “se”. A análise dos 281 lemas não retornou nenhum sujeito animado, o que justificaria a ausência do pronome reflexivo no *corpus*, sugerindo ser uma característica dos textos do domínio a frequência baixa ou nula de frases em que o sujeito é animado, um dos requisitos elencados por [Bechara 2012].

4. Resultados

O corpus PetroGold v3 conta com 1.960 ocorrências de “se”, sendo 75 conjunções subordinativas e 1.885 pronomes expletivos, conforme tabela 1. Todos os 1.885 usos expletivos passaram por revisão, uma vez que precisaram ter sua etiqueta modificada para se acrescentar a informação relativa à indeterminação do sujeito, passivização ou uso pronominal do verbo.

<i>expl:impers</i>	<i>expl:pass</i>	<i>expl:pv</i>	total de pronomes
278	807	800	1.885

Tabela 1. Frequência das classificações do pronome “se”

Distribuímos os pronomes “se” expletivos pelos verbos aos quais se associam, e alguns dos verbos podem aparecer associados a mais de um tipo de pronome “se” expletivo (os exemplos 4 a 7 a seguir exemplificam esses casos). Como se vê na tabela 2, 25 verbos podem se associar a pronomes “se” de dois tipos diferentes. Desses, 22 (88%) ora são usados pronominalmente e ora na voz passiva sintética, enquanto somente 3 (12%) são usados ora como verbo pronominal, ora como oração de sujeito indeterminado. Nenhum dos verbos se associa a três categorias ao mesmo tempo, e não encontramos no corpus nenhum verbo que se associe ora a um pronome “se” que indique voz passiva sintética, ora índice de indeterminação do sujeito.

depel do “se”	nº de verbos
<i>expl:impers</i>	21
<i>expl:pass</i>	154
<i>expl:pv</i>	142
duas categorias	25
três categorias	0

Tabela 2. Número de verbos que se associam aos tipos de pronome “se”

O maior número de verbos que podem ser utilizados tanto na voz passiva sintética como na forma pronominal já era esperado – [Azeredo 2000], por exemplo, comenta

sobre o fenômeno da cristalização do “se” em verbos de voz passiva, os quais, pela frequência de uso, vão se tornando pronominais. Sintaticamente, o fato de que muitos verbos podem ser empregados das duas formas é explicável pela semelhança estrutural – ambas requerem um sujeito sintático na frase, sendo que a diferenciação é realizada semanticamente ao interpretar se o sujeito é paciente do conteúdo verbal ou não. Por exemplo, nas frases 4 e 5, o verbo em destaque é “ajustar”. O primeiro, porém, tem como sujeito o substantivo “modelos”, sendo que, na interpretação dos autores, a frase não permite inferir que haveria um agente, propositalmente omitido da oração, responsável por ter ajustado os modelos no contexto em que o verbo foi utilizado, diferentemente da segunda oração, onde um agente não identificado ajustou a “frequência”, que é sujeito paciente da oração.

4. *expl:pv*: A partir do coeficiente de correlação, percebe-se que todos os modelos se **ajustaram**.
5. *expl:pass*: A bomba de água foi acionada com uma freqüência de 30 Hz e então **ajustou-se** a freqüência baseando-se na vazão de água desejada.

Já a diferenciação entre orações com sujeito indeterminado e uso pronominal do verbo pode ser explicada em termos puramente sintáticos. Nos exemplos 6 e 7, o verbo em destaque é “chamar”. Na primeira frase, o pronome relativo “que” retoma “pasta oleosa”, sujeito da oração relativa cujo núcleo é “se chama” e cujo complemento verbal é “petróleo”. Por tratar-se de uma oração com sujeito sintático, a anotação é a de verbo pronominal, diferentemente da segunda frase, em que “bentos” é objeto direto (sem a possibilidade da leitura como sujeito paciente) e “organismos” objeto indireto, faltando um sujeito sintático para a oração (que tampouco está elíptico), marcando, portanto, indeterminação do sujeito⁶.

6. *expl:pv*: Admite-se que o petróleo foi formado há milhões de anos pelo acúmulo de diferentes seres vivos como a decomposição de plânctons - seres que são geralmente encontrados na zona costeira, mares, oceanos e estuários - esses seres teriam se acumulados no fundo dos mares, rios e lagos e soterrados pela ação do movimento da crosta terrestre e posteriormente com o passar dos anos transformando-se em uma pasta oleosa que hoje se **chama** petróleo (VAZ, 2011).
7. *expl:impers*: Em biologia marinha e limnologia, **chama-se** bentos aos organismos que vivem no substrato, fixos ou não, em contraposição com os pelágicos, que vivem livremente na coluna de água.

A ausência de verbos compartilhando o *-se* na voz passiva sintética e o *-se* como indeterminação do sujeito pode ser explicada pelo fato de que os dois fenômenos são muito distintos sintaticamente, sendo que no primeiro há um sujeito sintático na oração, que é marcada por um VTD ou VTDI, e no segundo fenômeno não há sujeito, sendo utilizado um verbo VI ou VTI.

Os resultados da avaliação intrínseca do modelo gerado utilizando o *dataset* após as revisões do pronome *-se* como material de treino podem ser verificados na tabela 3. Entre parênteses, é possível verificar a variação, em pontos percentuais, das métricas de avaliação quando comparadas às métricas anteriores às revisões.

⁶A análise seria diferente caso “organismos” não fosse preposicionado: “Os organismos se chamam bentos” (verbo pronominal).

UPOS	LEMMA	LAS
98,40% (+0,05 p.p.)	98,46% (-0,09 p.p.)	88,71% (-0,39 p.p.)

Tabela 3. Avaliação intrínseca após as revisões do pronome “se”

A métrica relativa ao aprendizado de classe grammatical foi a única que obteve uma melhora (+0,05 p.p.). Isso pode ser explicado por alguns motivos: (a) foram realizadas correções sistemáticas relativas a quando o “se” é pronome ou conjunção subordinativa, facilitando o aprendizado automático de POS; (b) foi realizada uma simplificação das informações morfológicas do pronome expletivo “se” – originalmente, eram anotados como tendo atributos morfológicos de um pronome de terceira pessoa, como se fosse um objeto, portanto herdando as características do objeto da oração. Como se trata de um pronome expletivo, que não representa nem um sujeito nem um objeto, removemos completamente suas informações morfológicas, o que pode ter facilitado o aprendizado do etiquetador.

O decréscimo de 0,09 p.p. na avaliação de lematização, por sua vez, pode ser explicado pelo fato de que havia 22 palavras “sudeste” ou “Sergipe”, abreviadas como “SE”, mas que tinham o lema anotado como “se”, em letras minúsculas, o que facilitava o aprendizado uma vez que, independentemente de a palavra estar em caixa alta ou não, o lema era sempre o mesmo. Quando desfizemos essa anotação de lema, diferenciando o “se” pronome do “SE” abreviação para “sudeste” e “Sergipe”, introduzimos um pequeno obstáculo que pode ter refletido no decréscimo.

Já a métrica que diz respeito ao aprendizado de dependências (LAS) teve o desempenho piorado em 0,39 ponto percentual. Esse dado pode ser explicado por termos introduzido uma granularidade previamente inexistente no corpus quando adicionamos três novas classes para o pronome “se” – *expl:impers*, *expl:pass* e *expl:pv*. Antes das revisões, a classe *expl* obtinha 100% de acertos pois era a única para todos os casos de pronome “se”. Nessa nova versão, os resultados de acerto para as três novas classes foram de, respectivamente, 82,3%, 91,3% e 86,8%. Soma-se a isso o fato de que os verbos são polissêmicos e, dependendo do contexto, as orações – e portanto o “se” – podem ser interpretadas de uma forma ou de outra, conforme discutimos.

Além da anotação revista do *corpus* PetroGold v3, estamos disponibilizando também alguns recursos lexicais que podem ser úteis para outros projetos de anotação ou para alimentar tarefas linguístico-computacionais. Em um repositório dedicado a este trabalho⁷, disponibilizamos: (1) três listas com as frases que tiveram o pronome –se classificados como do tipo *expl:impers*, *expl:pass* ou *expl:pv*; (2) uma lista com todos os verbos associados ao pronome –se, organizados pela frequência com que o –se foi anotado usando as diferentes etiquetas e um exemplo de frase para cada tipo, e (3) a lista de verbos associados a pronomes –se (e as respectivas frases) em que os pronomes –se foram anotados de forma diferente, indicando polissemia do verbo, que admite mais de um fenômeno quando associado ao pronome. Os recursos funcionam como apêndices desse trabalho, e a ideia é que possam ser processados computacionalmente com facilidade.

⁷Disponível em: <https://github.com/alvelvis/recursos-lexicais-se>. Acesso em 22 de jun. 2023.

5. Considerações finais

Realizamos uma descrição linguística e discutimos os resultados do processo de anotação do pronome *-se* no *treebank* PetroGold (v3). O recurso integra o projeto *Universal Dependencies*, que prevê diferentes etiquetas para anotar o pronome *-se* mas não apresenta estruturas típicas da língua portuguesa, as quais anotamos tomando como base o estudo de gramáticas do português.

Como resultados, discriminamos as 1.960 ocorrências do “se” no *corpus* por classe sintática e apresentamos o número de verbos que se associam a cada um dos tipos do pronome *-se* expletivos. Com a revisão, foi possível perceber o comportamento de certos verbos com relação ao *-se*, o que por sua vez se reflete em uma constatação importante para o PLN: nem sempre um *corpus* mais bem anotado levará às melhores medidas de avaliação (medida F1). Em nosso caso, o fato de o *corpus* ter sido consistentemente revisto foi justamente o que fez o desempenho do modelo piorar.

Agradecimentos

Os autores agradecem ao CNPq (Conselho Nacional de Desenvolvimento Científico e Tecnológico, processo #130495/2021-2), à FAPERJ (Fundação Carlos Chagas Filho de Amparo à Pesquisa do Estado do Rio de Janeiro, processo #E-26/202.433/2022) e à ANP (Agência Nacional de Petróleo, Gás Natural e Biocombustíveis, Brasil, associada ao investimento de recursos oriundos das Cláusulas de P,D&I, por meio de Termo de Cooperação entre a Petrobras e a PUC-Rio) pelos auxílios concedidos, sem os quais este trabalho não poderia ter sido realizado.

Referências

- Azeredo, J. C. d. (2000). Fundamentos de gramática do português. *Rio de Janeiro: Jorge Zahar*.
- Bagno, M. (2012). *Gramática pedagógica do português brasileiro*. Parábola Ed.
- Bechara, E. (2012). *Moderna gramática portuguesa*. Nova Fronteira.
- Cançado, M. and Amaral, L. (2010). Representação lexical de verbos incoativos e causativos no português brasileiro. *Revista da ABRALIN*, 9(2):123–147.
- Cunha, C. and Cintra, L. (2016). *Nova gramática do português contemporâneo*. LEXIKON Editora Digital ltda.
- de Marneffe, M.-C., Manning, C. D., Nivre, J., and Zeman, D. (2021). Universal dependencies. *Computational linguistics*, 47(2):255–308.
- de Souza, E. (2023). *Construção e avaliação de um treebank padrão ouro*. Mestrado, PUC-Rio.
- de Souza, E. and Freitas, C. (2021). ET: A workstation for querying, editing and evaluating annotated corpora. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 35–41, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- dos Santos Silva, A. (2021). O clítico *se* no português brasileiro como índice de indeterminação do sujeito. *EDUCTE: Revista Científica do Instituto Federal de Alagoas*, 12(1):1683 a 1692.

- Duran, M. S. and Aluísio, S. M. (2011). O tratamento da partícula “se” para fins de anotação de papéis semânticos. *II Jornada de Descrição do Português-Proceedings of 8th STIL-Cuiabá*, pages 24–26.
- Duran, M. S., Scarton, C., Aluísio, S., and Ramisch, C. (2013). Identifying Pronominal Verbs: Towards Automatic Disambiguation of the Clitic ‘se’ in Portuguese. In *Proceedings of the 9th Workshop on Multiword Expressions*, pages 93–100.
- Freitas, C. and de Souza, E. (2021). Sujeito oculto às claras: uma abordagem descritivo-computacional/Omitted subjects revealed: a quantitative-descriptive approach. *Revista de Estudos da Linguagem*, 29(2):1033–1058.
- Hartmann, N. S., Duran, M. S., and Aluisio, S. M. (2014). 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 [...] J. São Carlos: SBC*.
- Lopes, E. M. and Namiuti-Temponi, C. (2017). A ordem e a função do clítico se no português clássico. *Entrepalavras*, 7(2):151–169.
- 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.
- Straka, M., Hajic, J., and Straková, J. (2016). UDPipe: trainable pipeline for processing CoNLL-U files performing tokenization, morphological analysis, pos tagging and parsing. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC’16)*, pages 4290–4297.
- Vieira, C. B. and de Sá, T. M. M. (2015). Pronome apassivador? uma perspectiva cognitiva na análise do pronome se. *Palimpsesto-Revista do Programa de Pós-Graduação em Letras da UERJ*, 14(21):411–426.
- Zeman, D., Hajic, J., Popel, M., Potthast, M., Straka, M., Ginter, F., Nivre, J., and Petrov, S. (2018). CoNLL 2018 shared task: Multilingual parsing from raw text to universal dependencies. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual parsing from raw text to universal dependencies*, pages 1–21.