Inferências baseadas em sintaxe: a anotação de sujeitos implícitos

Magali Sanches Duran, Maria das Graças Volpe Nunes, Thiago Pardo

ICMC - Universidade de São Paulo, São Carlos, SP
Núcleo Interinstitucional de Linguística Computacional - NILC
Center for Artificial Intelligence - C4AI

magali.duran@gmail.com, gracan@icmc.usp.br,taspardo@icmc.usp.br

Abstract. *This paper presents rules for annotating implicit subjects based on five syntactic dependency relations from the Universal Dependencies (UD) approach. Two of these rules were proposed by UD and improved for Portuguese, and the other three were originally developed for Portuguese and can be applied to other languages with minor adaptations.*

Resumo. *Este artigo apresenta regras para a anotação de sujeitos implícitos a partir de cinco relações de dependência sintática da abordagem Universal Dependencies (UD). Duas dessas regras foram propostas pela UD e aperfeiçoadas para o português, e as outras três foram desenvolvidas originalmente para o português, podendo ser aplicadas a outras línguas com pequenas adaptações.*

1. Introdução

Identificar o sujeito de um predicado talvez seja uma das principais motivações para uma aplicação utilizar um *parser* em seu pré-processamento, sendo importante, por exemplo, para aplicações de extração de informação, sumarização e perguntas e respostas. Com exceção dos verbos impessoais (ex: haver, chover), todo verbo tem potencialmente um sujeito. No entanto, atendendo ao princípio pragmático da economia, os mecanismos das línguas permitem suprimir sujeitos que possam ser inferidos pelos seus falantes. Por exemplo, em "João estava tentando acalmar Luíza e convencê-la a não desistir de viajar", sabemos que "João" é sujeito de "tentar", "acalmar" e "convencer" e Luíza é sujeito de "desistir" e "viajar", mas apenas um sujeito está anotado sintaticamente: "João", como sujeito de "tentar". Em suma, temos 5 verbos e 1 sujeito (não contamos "estava", que é verbo auxiliar).

Mecanismos de economia de sujeitos explícitos ocorrem em todas as línguas, mas são ainda mais frequentes em línguas *pro-drop*, como o português, que admitem a elipse do sujeito. Comparando "João disse que chega amanhã e que pretende jantar conosco" com sua tradução para o inglês "João said he will arrive tomorrow and that he wants to have dinner with us", observamos que as duas sentenças têm quatro verbos, mas, enquanto o português tem apenas um sujeito explícito, o inglês tem três.

Como o sujeito elíptico de uma oração quase sempre pode ser identificado em outra oração, dentro da mesma sentença, nos últimos anos desenvolveu-se um tipo de anotação que utiliza regras para, a partir da pré-anotação sintática, inferir e anotar sujeitos não explícitos. Essa tarefa está compreendida dentro do escopo conhecido como *enhanced dependencies* (ED). As ED foram desenvolvidas primeiro no inglês (Schuster

& Manning, 2016), depois foram generalizadas para outras línguas pela abordagem *Universal Dependencies* (UD)[1] (Nivre et al., 2018, de Marneffe et al. 2021) e instanciadas para o português (Pagano et al., 2023).

Na UD foram estabelecidas regras para atribuir sujeitos de orações coordenadas e de orações subordinadas objetivas diretas e indiretas reduzidas, as quais envolvem as relações de dependência conj e xcomp. Assumindo que toda oração tem potencialmente um sujeito, analisamos todos os demais tipos de orações da UD para avaliar a oportunidade de construir novas regras de ED para atribuição de sujeitos (acl, acl:relcl, advcl, ccomp, csubj). As orações csubj (sujeito oracional) e acl:relcl (orações relativas) não apresentam potencial para isso, mas acl (orações adjetivas), advcl (orações adverbiais) e ccomp (orações objetivas diretas e indiretas desenvolvidas) apresentam.

Discutiremos o refinamento das regras das ED universais para inferência de sujeitos (nas relações UD conj e xcomp) e apresentaremos regras para novas ED concebidas (nas relações UD ccomp, acl e advcl). Para obter os resultados aqui discutidos, analisamos sentenças do córpus Porttinari-base (Duran et al. 2023) que não tinham um sujeito próprio anotado (córpus disponível no site do Projeto POeTiSA[2]). Em cada caso, respondemos a uma pergunta: o sujeito da oração pode ser inferido dentro da própria sentença? Os "sim" nos deram regras para inferências corretas, e os "não" nos deram regras para evitar inferências incorretas. A Seção 2 deste artigo apresenta os tipos de ED; a Seção 3 mostra as relações da UD que podem ter sujeitos implícitos e a Seção 4 traz as conclusões do estudo.

2. *Enhanced Dependencies*

Enhanced dependencies (ED), no projeto UD, são relações inferidas a partir das relações dependências sintáticas básicas. Para exemplificar, tomemos a sentença "*João acordou e saiu*". Nas dependências básicas (em preto), "João" é nsubj de "acordou" (Fig. 1a). Já nas ED (em vermelho), "João" é nsubj de "acordou" e de "saiu" (Fig. 1b).

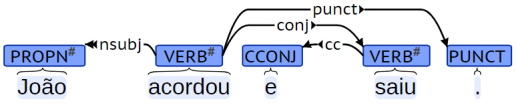

Figura 1a. Dependências básicas da sentença "João acordou e saiu"

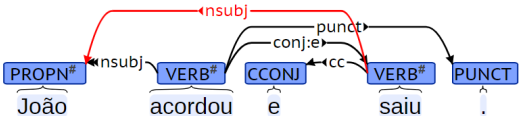

Figura 1b. *Enhanced dependencies* da sentença "João acordou e saiu"

A diferença entre as duas anotações é que, nas ED, foi acrescentada a relação nsubj entre "saiu" e "João", relação não permitida nas dependências básicas porque cada dependente deve ter um único *head*. Além disso, produz um cruzamento de arcos indesejado, pois dificulta o aprendizado automático.

---

[1] São chamadas *Universal Enhanced Dependencies* (EUD). Como nossa proposta inclui, mas extrapola, as EUD, usamos simplesmente *Enhanced Dependencies* (ED).
[2] https://sites.google.com/icmc.usp.br/poetisa

3. Explicitação de sujeito com *enhanced dependencies*

Apresentamos a seguir cinco relações de dependência (xcomp, conj, ccomp, advcl e acl), a partir das quais se pode revelar sujeitos não anotados sintaticamente. As regras que envolvem xcomp e conj foram propostas pela UD e detalhadas neste trabalho para contemplar com maior precisão todos os casos de língua portuguesa. As regras para as outras três são contribuições deste trabalho. Usamos o termo "atribuição" para sujeitos inferidos cuja inserção no texto não é gramaticalmente aceitável (xcomp e acl) e o termo "propagação" para sujeitos inferidos cuja inserção não fere a gramaticalidade.

3.1. Atribuição de sujeito de xcomp

A relação xcomp tem como característica o fato de que seu dependente não admite um sujeito explícito. Essa ausência de sujeito é chamada *null subject* ou "sujeito nulo". Apesar de o dependente de xcomp não ter um sujeito sintático, é possível inferi-lo, pois o sujeito nulo é "controlado" por um *token* presente na oração *head* de xcomp. A seguir são ilustrados diferentes *tokens* assumindo o controle do sujeito da xcomp: o sujeito (nsubj) (Fig. 2); o objeto direto (obj) (Fig. 3) ou, em casos mais raros, o objeto indireto (iobj, obl) (Fig. 4 e 5). As ED são destacadas em vermelho.

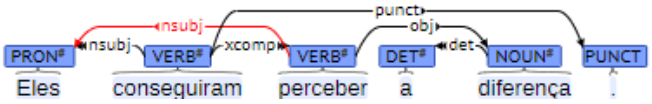

Figura 2. Sujeito da xcomp = sujeito: "Eles conseguiram perceber a diferença"

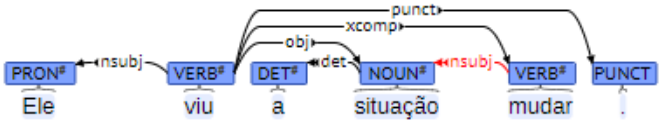

Figura 3. Sujeito da xcomp = obj: "Ele viu a situação mudar"

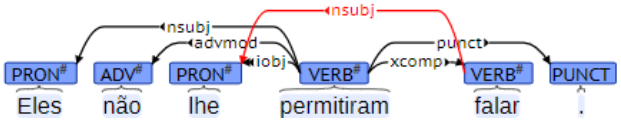

Figura 4. Sujeito da xcomp = iobj: "Eles não lhe permitiram falar"

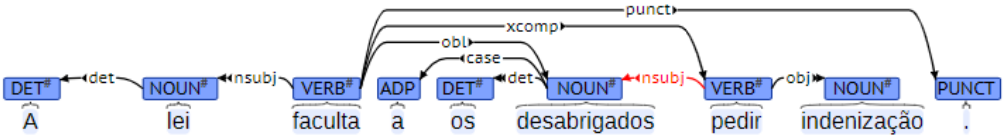

Figura 5. Sujeito da xcomp = obl: "A lei faculta aos desabrigados pedir indenização"

Se o *head* da xcomp está na voz ativa e o dependente da xcomp está na voz passiva, é necessário adequar o nome da relação de dependência, de nsubj para nsubj:pass (ou vice-versa), a fim de refletir isso, como mostra a Figura 6.

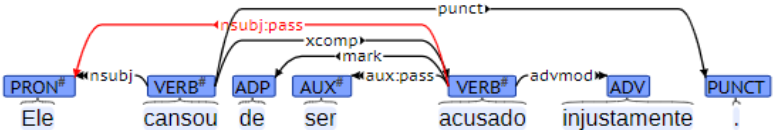

Figura 6. Inversão para voz passiva: "Ele cansou de ser acusado injustamente"

Mesmo quando o dependente de xcomp é um predicado nominal, a ED opera normalmente, como pode ser observado na Fig. 7.

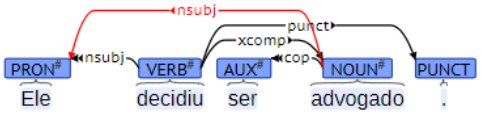

Figura 7. Dependente de xcomp predicado nominal: “Ele decidiu ser advogado”

Se, no entanto, o *token* que deveria controlar o sujeito de xcomp não estiver explícito, o dependente de xcomp não terá um sujeito atribuído, como nos exemplos a seguir, nos quais o *token* controlador elíptico ou indeterminado está indicado por um Ø. Embora na primeira sentença o sujeito esteja marcado na pessoa do verbo (“eu” comecei), não existe um *token* que possa receber a relação de sujeito.

[Ø] Comecei a trabalhar cedinho.
[Ø] Pretende-se chegar até amanhã.
O chefe mandou [Ø] fazer plantão.

Verbos que preveem o objeto da oração matriz como controlador do sujeito do xcomp pertencem a classes restritas (causativos, resultativo, de percepção, com predicativo do objeto), como nos quatro exemplos a seguir.

Ele **nos** mandou/deixou/fez esperar. (“nos”=nós, sujeito de “esperar”)
Ele teve a **casa** invadida. (“casa”, sujeito da passiva de “invadida”)
Ele viu/ouviu/sentiu a **terra** tremer. (“terra”, sujeito de “tremer”)
Ele acha/considera/julga **isso** impossível. (“isso”, sujeito de “impossível”)

3.2 Propagação de sujeito do *head* para dependente de conj

Se dois predicados estão ligados por conj, o sujeito do *head* pode ser propagado para o dependente, observadas as condições (vide Seção 3.6 para condições gerais para propagação de sujeito de conj, ccomp e advcl). A Figura 8 ilustra essa propriedade para o dependente do tipo nsubj na sentença "Essa música é **alegre** e **animada**".

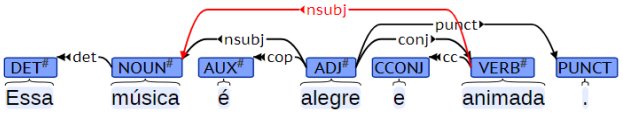

Figura 8. Propagação da relação nsubj para o dependente de conj

As Figuras 9, 10 e 11 ilustram casos de propagação de sujeito da ativa (nsubj), sujeito da passiva (nsubj:pass) e sujeito oracional (csubj), respectivamente.

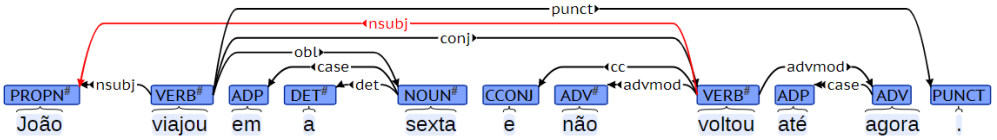

Figura 9. Propagação de nsubj em "João viajou na sexta e não voltou até agora"

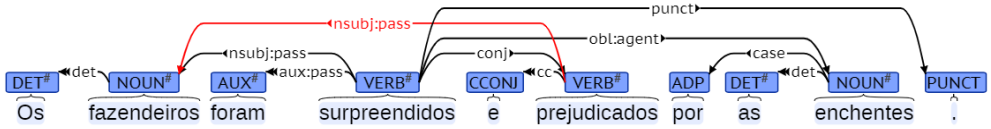

Figura 10. Propagação de nsubj:pass em "Os fazendeiros foram surpreendidos e prejudicados pelas enchentes"

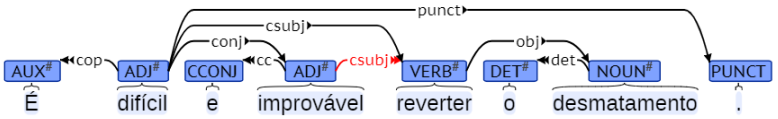

Figura 11. Propagação de csubj em "É difícil e improvável reverter o desmatamento"

Se o *head* de conj não tiver sujeito ("**Acordei** cedo e logo **saí** para caminhar") ou se o dependente de conj já tiver seu próprio sujeito ("A comida era **gratuita** e a bebida era **barata**"), não haverá sujeito a ser propagado.

Quando a relação conj liga duas orações, pode haver inversão de voz ativa para passiva ou vice-versa, conforme ilustra a Fig.12.

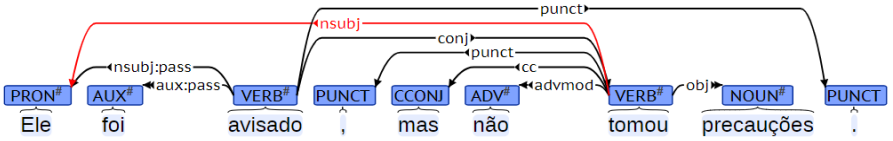

Figura 12. Propagação do sujeito em "Ele foi avisado, mas não tomou precauções"

Também é possível "aproveitar" o sujeito atribuído numa ED para fazer outra ED. Isso ocorre na Figura 13 ("Ele pretende **estudar** engenharia e **trabalhar** na construção de navios"), onde duas orações coordenadas, sem sujeito ("estudar" e "trabalhar"), ocorrem depois de uma oração xcomp que teve seu sujeito (Ele) atribuído em uma outra ED.

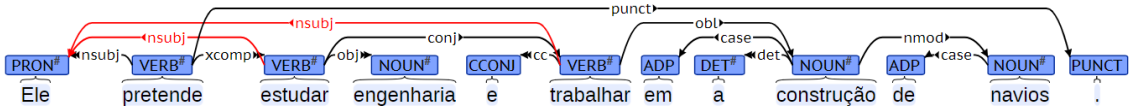

Figura 13. Propagação de sujeito de xcomp para head e dependente de conj

3.3. Propagação de sujeito do *head* para dependente de ccomp

No português, as orações subordinadas do tipo ccomp podem ter o sujeito elíptico e esse sujeito pode coincidir com o sujeito da oração *head* de ccomp, por isso é possível criar uma ED propagando o sujeito do *head* de ccomp para o dependente de ccomp. Vide regras de propagação gerais na Seção 3.6.

Essa propagação de sujeito de ccomp colocaria as línguas *pro drop* em condições de igualdade com as línguas que não admitem a elipse de sujeito. As Figuras 14a e 14b ilustram a propagação de um sujeito de ccomp no português ("**Ele** disse que **[ele]** vai aposentar") e equivalente em inglês (***He*** *said* ***he*** *will retire*), língua na qual o sujeito do dependente de ccomp não pode sofrer elipse e, portanto, não precisa da ED.

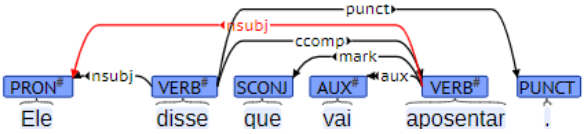

Figura 14a. Propagação de sujeito de ccomp no português

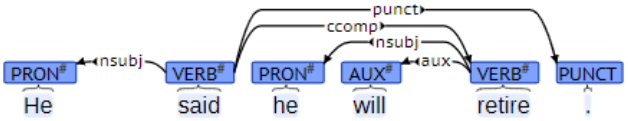

Figura 14b. Sujeito de ccomp explícito no inglês

A propagação do sujeito para preencher o sujeito elíptico do dependente de ccomp tem outra vantagem: não interrompe a anotação de cadeias de sujeitos implícitos. Na sentença "**Ele** disse que **[ele]** vai aposentar e **[ele]** pretende **[ele]** viajar durante um ano" (Figura 15), a anotação do sujeito do dependente de ccomp ("aposentar") torna possível propagar o mesmo sujeito para a oração dependente de conj ("pretende") e anotar o sujeito controlador da oração dependente de xcomp ("viajar").

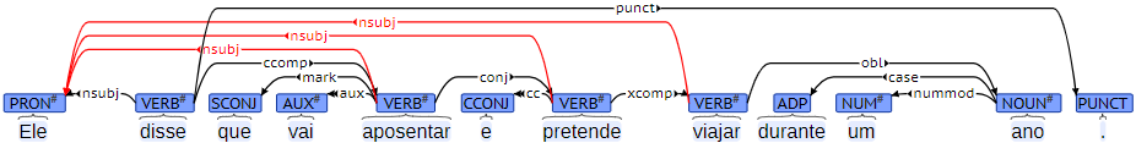

Figura 15. Sentença com ccomp com propagação do sujeito

Se um dos predicados envolvidos na relação ccomp tiver Voice=Pass e o outro não, adapta-se a relação de nsubj:pass para nsubj ou nsubj para nsubj:pass.

3.4 Propagação de sujeito do *head* para dependente de advcl

As orações subordinadas do tipo advcl também podem apresentar elipse de sujeito e, portanto, são candidatas a obter a propagação do sujeito de seus respectivos *heads* (vide regras de propagação gerais na Seção 3.6). Contudo, como há vários tipos de orações adverbiais, muitas delas com várias elipses, as regras para propagação do sujeito nem sempre alcançam uma boa precisão. Nos exemplos a seguir, o sujeito do *head* não propaga para o dependente da advcl, embora as regras básicas tenham sido atendidas.

> Quando **acordou**, o veículo não estava mais **lá**. ("veículo" é sujeito do predicado nominal "lá", mas não é sujeito de "acordou")

> Minha intenção é **contribuir** com isso, **trazendo** dados confiáveis. ("intenção" é sujeito de "contribuir", mas não é sujeito de "trazendo")

Não temos estatísticas para avaliar o percentual de casos que fogem à regra (isso será feito quando as regras forem implementadas computacionalmente e os casos revisados). As Figuras 16 e 17 ilustram respectivamente uma advcl temporal e uma advcl concessiva com propagação do sujeito.

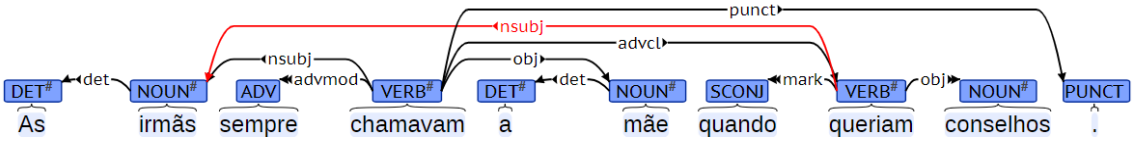

Figura 16. Propagação de sujeito de advcl "As irmãs sempre chamavam a mãe quando queriam conselhos."

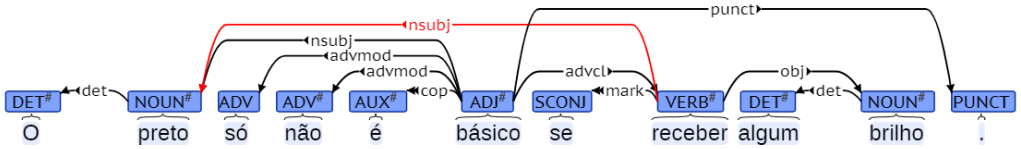

Figura 17. Propagação sujeito de advcl "O preto só não é básico se receber algum brilho."

3.5 Atribuição de sujeito de acl

Os dependentes de acl são orações adjetivas ou completivas nominais e apresentam forma de orações reduzidas, isto é, estão no particípio, gerúndio ou infinitivo. A relação acl une um *token* substantivo a uma oração. Em algumas condições, o *token head* da acl é também o sujeito lógico do dependente da acl e por isso poderia ser atribuído nas ED.

As acl cujo dependente é um **verbo no particípio** são orações adjetivas reduzidas de voz passiva. Por isso, o *token head* da relação acl será nsubj:pass do dependente, como ilustra a Figura 18 na sentença "A certeza de vitória **demonstrada** pelos advogados de defesa é surpreendente" (= "a certeza que foi demonstrada").

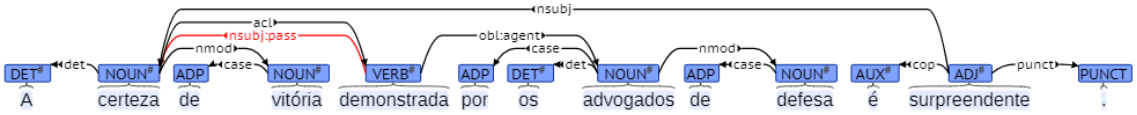

Figura 18. Anotação de nsubj:pass de dependente de acl no particípio

As acl **reduzidas de gerúndio** são sempre adjetivas e podem, seguramente, receber atribuição de sujeito, como na sentença da Figura 19 "Recebi uma mensagem **dizendo** que a carga tombou." (= "mensagem que dizia"). Nesse caso, o sujeito é nsubj.

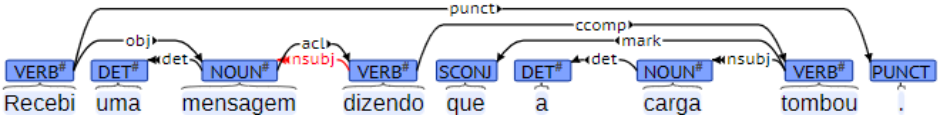

Figura 19. Anotação de nsubj de dependente de acl no gerúndio

As acl **reduzidas de infinitivo** podem constituir orações adjetivas ("máquina de lavar roupa", "preço a combinar") ou orações completivas nominais ("vontade de chorar", "interesse em viajar"). Quando são adjetivas, o *head* da acl pode ser anotado como nsubj ou como nsubj:pass do dependente da acl, como mostram, respectivamente, o primeiro e o segundo exemplo a seguir.

Ela foi a primeira mulher a **assumir** o cargo. (= mulher que assumiu o cargo)
Esse é o preço a **pagar** por um jantar ali. (= "preço que será pago")

As acl do tipo complemento nominal não obedecem essa regra (o *head* da acl não é nsubj ou nsubj:pass do dependente). Algumas dessas orações não admitem atribuição de sujeito, pois denotam uma indeterminação do sujeito ("Propina é a **maneira** mais fácil de [se] **atravessar** um muro na fronteira"); outras têm, como sujeito lógico, o sujeito da oração em que se encontra o *head* da acl (“Eles têm **vontade** de **viajar**”).

3.6 Regras gerais para a propagação de sujeitos de conj, ccomp e advcl

As regras para propagar sujeito do *head* para o dependente de conj, ccomp e advcl são:

- o *head* de conj, ccomp ou advcl precisa ter um sujeito explícito (nsubj, nsubj:pass, csubj, csubj:pass);
- o predicado dependente de conj, ccomp ou advcl não deve ter sujeito próprio;
- o predicado *head* e o predicado dependente de conj, ccomp ou advcl têm que ter a mesma pessoa e número (condição estendida para o auxiliar ou verbo de cópula caso o predicado os tenha como dependentes);
- o predicado dependente de conj, ccomp ou advcl não pode ser um verbo impessoal (por exemplo: “haver”, “chover”);
- o predicado dependente de conj, ccomp ou advcl não pode estar impessoalizado pelo índice de indeterminação do sujeito “se” (expl:impers, na anotação UD).

4. Conclusões

Discutimos o potencial de multiplicação de anotação de sujeitos sintáticos inferíveis por meio de regras construídas a partir das relações de dependência sintática UD, relações chamadas de enhanced dependencies. Todos os tipos de orações da UD foram avaliados e, portanto, não enxergamos possibilidade de novas EDs de sujeito nesse momento. Algumas dessas regras foram definidas no âmbito do projeto UD e aperfeiçoadas para o português neste estudo (sujeitos de xcomp e conj). Outras foram levantadas a partir da observação de contextos do português, mas são igualmente aplicáveis a outras línguas, incluindo o inglês (sujeitos de advcl e acl), ou exclusivamente a línguas que admitem elipse do sujeito, como espanhol e italiano (sujeito de ccomp). As regras descritas serão utilizadas em um programa para automatização da anotação de *enhanced dependencies* no português, seguindo a abordagem simbólica já adotada por outras línguas, conforme relatos de duas *shared tasks* dedicadas ao assunto [Bouma et al., 2020; 2021]. Esse programa será utilizado para anotar o córpus Porttinari-base (Duran et al. 2023).

Agradecimentos

Este trabalho foi realizado no âmbito do Centro de Inteligência Artificial da Universidade de São Paulo (C4AI -http://c4ai.inova.usp.br/), com o apoio da Fundação de Amparo à Pesquisa do Estado de São Paulo (processo FAPESP #2019/07665-4) e da IBM. Este projeto também foi apoiado pelo Ministério da Ciência, Tecnologia e Inovações, com recursos da Lei N. 8.248, de 23 de outubro de 1991, no âmbito do PPI-Softex, coordenado pela Softex e publicado como Residência em TIC 13, DOU 01245.010222/2022-44.

Referências

Bouma, G., Seddah, D., and Zeman, D. (2020). Overview of the iwpt 2020 shared task on parsing into enhanced universal dependencies. In 58th Annual Meeting of the Association for Computational Linguistics.

Bouma, G., Seddah, D., and Zeman, D. (2021). From raw text to enhanced universal dependencies: The parsing shared task at iwpt 2021. In Proceedings of the 17th International Conference on Parsing Technologies and the IWPT 2021 Shared Task on Parsing into Enhanced Universal Dependencies (IWPT 2021), pages 146–157. Association for Computational Linguistics (ACL).

Duran, M. S. (2024). Anotação de Enhanced Dependencies: Orientações para anotação de relações de dependência sintática do tipo enhanced em Língua Portuguesa, seguindo as diretrizes da abordagem Universal Dependencies (UD). Relatório Técnico do Núcleo Interinstitucional de Linguística Computacional (NILC). Disponível em: https://repositorio.usp.br/item/003209188

Duran, M. S.; Lopes, L.; Nunes, M.G.V.; Pardo, T.A.S. (2023). The Dawn of the Porttinari Multigenre Treebank: Introducing its Journalistic Portion. In the Proceedings of the 14th Symposium in Information and Human Language Technology (STIL), pp. 115-124. September, 25-29. Disponível em: https://sol.sbc.org.br/index.php/stil/article/view/25443/25264

de Marneffe, M.-C.; Manning, C.D.; Nivre, J.; Zeman, D. (2021) Universal Dependencies. Computational Linguistics 47(2), 255-308.

Nivre, J.; Marongiu, P.; Ginter, F.; Kanerva, J.; Montemagni, S.; Schuster, S.; Simi, M. (2018) Enhancing Universal Dependency Treebanks: A Case Study. In Proceedings of the Second Workshop on Universal Dependencies, pages 102-107.

Pagano, A. S.; Duran, M. S.; Pardo, T. A. S. (2023) Enhanced dependencies para o português brasileiro. In: Proceedings of the 2nd Edition of the Universal Dependencies Brazilian Festival, p. 461–470, Belo Horizonte, Brasil. Association for Computational Linguistics. Disponível em: https://sol.sbc.org.br/index.php /stil/article/view/25486/25307

Schuster, S.; Manning, C. D. (2016). Enhanced English Universal Dependencies: An Improved Representation for Natural Language Understanding Tasks. In Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16), pages 2371–2378, Portorož, Slovenia. European Language Resources Association (ELRA).