Enhanced dependencies para o português brasileiro

Adriana S. Pagano[1], Magali Sanches Duran [2], Thiago Alexandre Salgueiro Pardo[2]

[1]Faculdade de Letras – Universidade Federal de Minas Gerais (UFMG)
Belo Horizonte – MG – Brasil
[2]Núcleo Interinstitucional de Linguística Computacional (NILC)
Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo (USP)
São Paulo - SP - Brasil

apagano@ufmg.br, magali.duran@uol.com.br, taspardo@icmc.usp.br

Abstract. *This article explores Universal Dependencies' guidelines for annotating the enhanced flavor of dependency relations and presents the main types of enhancement as applied to Brazilian Portuguese. It briefly discusses models for automatically converting basic into enhanced dependency relations with a view to their future implementation to enrich Brazilian Portuguese treebanks.*

Resumo. *Este artigo explora as diretrizes de anotação de "enhanced dependencies" do modelo "Universal Dependencies" e apresenta as principais configurações de anotação em português brasileiro, discutindo a relevância e as aplicações desse tipo de informação. São também discutidos modelos automáticos de conversão das relações de dependência básicas para "enhanced" e tecidas considerações sobre seu uso para o português.*

1. Introdução

As chamadas *enhanced dependencies*[1] (ED) visam o enriquecimento de informações em *treebanks* anotados com relações de dependência básicas da abordagem *Universal Dependencies* – UD (de Marneffe et al., 2021). Esse enriquecimento se mostra útil para aplicações mais avançadas, envolvendo semântica, como a recuperação de informações e a anotação automática de papéis semânticos (Nivre et al., 2018; Ek & Bernady, 2020).

Atualmente, no âmbito da UD, um número ainda limitado, embora crescente, de línguas disponibiliza *treebanks* anotados com ED, como é o caso das línguas inglesa, espanhola, russa, polonesa, finlandesa e holandesa, para citar algumas delas. A língua portuguesa não conta ainda com *treebanks* anotados com ED, lacuna que justifica o esforço aqui apresentado de interpretar as diretrizes da UD quanto à anotação das ED e de ilustrar as explicações com exemplos extraídos de *corpora* de português brasileiro. Busca-se, portanto, estimular e facilitar a implementação das ED em *treebanks* de português.

A seguir, na Seção 2, relatamos sucintamente a evolução da proposta de *enhanced dependencies*. Na Seção 3 discutimos a infraestrutura computacional para anotação e visualização das ED. Na Seção 4 são apresentadas as principais configurações sintáticas passíveis de serem enriquecidas e suas respectivas anotações,

[1] Apesar de ser possível traduzir o termo, optamos pelo termo em inglês, por ser já conhecido no Brasil.

seguindo as diretrizes da UD. Na Seção 5, são discutidos trabalhos que exploram modelos automáticos de conversão de relações UD para relações ED e são tecidas considerações sobre sua possível utilização em *corpora* de português.

2. *Enhanced dependencies*

A UD tem, entre seus precursores, o modelo de dependências de Stanford, que já previa dois tipos de anotação, de acordo com os requisitos das tarefas nas quais eram utilizados: anotação de relações básicas e anotação de relações do tipo *collapsed* (ou *cc processed*) (de Marneffe et al., 2006; de Marneffe & Manning, 2008). As relações *collapsed* permitem anotar cada preposição (classe de palavra relevante enquanto indicadora de papéis semânticos) e cada conjunção coordenativa juntamente com o head das relações de dependência das quais participam (*nmod, obl, conj e advcl*), auxiliando em tarefas como a extração de informação (Schuster et al., 2017).

Schuster & Manning (2016) expandiram as representações do tipo *collapsed* para além de preposições e conjunções, e as adaptaram às diretrizes da UD. As representações ganharam o nome de *enhanced dependencies*. Os autores avaliaram seu uso na extração de informação e detectaram alguns problemas de anotação que não podiam ser resolvidos nem pelas relações básicas, nem pelas relações *enhanced.* Isso os levou a propor um terceiro tipo de anotação, que denominaram *enhanced dependencies* ++, ainda não implementada em português, cuja discussão extrapola o escopo deste artigo..

Para a anotação de ED em arquivos CoNLL-U, que é o formato de arquivo padrão da UD, é reservada a nona coluna, rotulada DEPS. Nesta coluna, temos a relação das arestas que chegam a cada uma das palavras (*tokens*) da sentença. A representação obtida por meio da anotação é um grafo, mas não necessariamente uma árvore, uma vez que podem haver nós vazios, várias arestas chegando a um mesmo nó e ocorrência de ciclos.

Droganova & Zeman (2019) esclarecem que a anotação das ED é opcional em *treebanks*, podendo-se anotar apenas uma, várias ou todas as configurações previstas. Contudo, uma vez decidido incluir um dos tipos de ED, é fundamental que isso seja feito no *corpus* inteiro, por uma questão de consistência.

4. Anotação e Visualização das ED4. Anotação e Visualização das ED

O estudo das ED e o esforço para instanciá-las em língua portuguesa nos levou a perceber que há uma carência muito grande de ferramentas para anotá-las e visualizá-las, embora para anotar e visualizar as relações básicas da UD existam várias ferramentas[2] . A única ferramenta que encontramos para edição das ED foi o CONLL-U Editor[3], utilizado para construir as árvores das figuras que ilustram este artigo. Para visualização, encontramos ainda as ferramentas Grew-Web[4] e Inception[5].

[2] https://universaldependencies.org/tools.html
[3] https://github.com/Orange-OpenSource/conllueditor
[4] https://web.grew.fr/
[5] https://inception-project.github.io/

Há discussões sobre a melhor forma de visualizar o resultado da anotação das ED[6]. Uma opção é visualizar as ED simultaneamente à visualização das dependências básicas, umas sobrepostas às outras. Isso seria simples se as ED só fizessem acréscimos, porém dois tipos de EDs alteram algumas das relações básicas e, portanto, não há como visualizá-las sob um mesmo plano. Uma alternativa é mostrar as ED abaixo da sentença anotada e as relações básicas acima da sentença anotada, o que é feito pela ferramenta CONLLU-Editor, como pode ser observado nas figuras que ilustram este artigo. A outra alternativa, adotada nas diretrizes da UD, é não visualizá-las simultaneamente. Opta-se por visualizar as relações básicas (anotadas na coluna 8 do CONLLU) ou por visualizar as ED (anotadas na coluna 9 do CONLLU). Em qualquer das alternativas, o recurso de "pintar" as relações não compartilhadas pelas colunas 8 e 9 nos pareceu muito bom para fins de visualização. As instruções da UD acerca da anotação das ED apresentam um "antes" (visualização da coluna 8) e um "depois" da anotação das ED (visualização da coluna 9) e pintam de vermelho as relações básicas não compartilhadas com as ED e de verde as relações das ED não compartilhadas com as relações básicas. Simplificando, o *diff* entre as colunas 8 e 9 aparece em vermelho na visualização da coluna 8 e em verde na visualização da coluna 9 e em ambas visualizações as relações compartilhadas aparecem em preto, sem destaque portanto.

Nas Figuras exibidas na Seção 4, por motivo de economia, só replicamos as relações compartilhadas entre as colunas 8 e 9 no caso em que há inserção de token e muitas mudanças de relações decorrentes dessa inserção (4.2.2). Nas demais figuras, as relações compartilhadas só são exibidas na parte superior, o que não significa que não estejam presentes também na parte inferior (apenas foram ocultadas). Utilizamos a cor vermelha para mostrar, na parte superior, o *diff* das relações básicas em relação às ED e a cor azul para mostrar, na parte inferior, o *diff* das ED em relação às relações básicas.

4. Configurações passíveis de serem anotadas com *enhanced dependencies*

De acordo com as orientações fornecidas pela UD[7], há seis casos previstos de anotação de ED. Em linhas gerais, acreditamos que essas seis ED podem ser agrupadas em duas categorias: aquelas que produzem um acréscimo de informações às dependências básicas (exemplificadas na Seção 4.1) e aquelas que produzem uma cópia modificada das dependências básicas (exemplificadas na seção 4.2).

4.1 *Enhanced Dependencies* de acréscimo

As ED que apenas acrescentam informações às dependências básicas são de quatro tipos: as que reproduzem, nos *heads* das relações *nmod*, *obl*, *conj,* e *advcl*, as preposições e as conjunções que introduzem seus dependentes (4.1.1), as que promovem a propagação de sujeitos de *xcomp* (4.1.2), as que propagam *head* compartilhado de elementos coordenados (4.1.3) e as que propagam dependentes compartilhados por elementos coordenados (4.1.4).

[6] Vide discussão sobre o assunto no Fórum da UD em https://github.com/UniversalDependencies.
[7] https://universaldependencies.org/u/overview/enhanced-syntax.html

4.1.1 Acréscimo de preposições e conjunções

Esta configuração enriquece relações de dependência como *nmod, obl, conj* e *advcl,* acrescentando-lhes uma preposição ou uma conjunção com a qual constroem relações semânticas. A anotação pode inclusive conter especificação do significado construído pela preposição ou conjunção. Se o treebank não tiver sub-relações de *nmod*, *obl, cc* e *advcl* não será possível herdar informações semânticas e, portanto, seria necessário anotá-los manualmente caso sejam de interesse do projeto de anotação de ED. No caso das preposições, constroem-se significados relativos aos casos *gen* (genitivo), *loc* (locativo), *tem* (temporal), *ins* (instrumental), *dat* (dativo), *acc* (acusativo) e outros. O exemplo (7) ilustra esta configuração, especificando nas ED que o verbo "proibiu" apresenta um *obl* introduzido pela preposição "por" com o papel semântico de temporal (*tem*), e outro *obl* introduzido pela preposição "em" com papel semântico de locativo (*loc*). O grafo pode ser visualizado na Figura 1.

(1) O governo proibiu por 120 dias as queimadas em todo o Brasil.

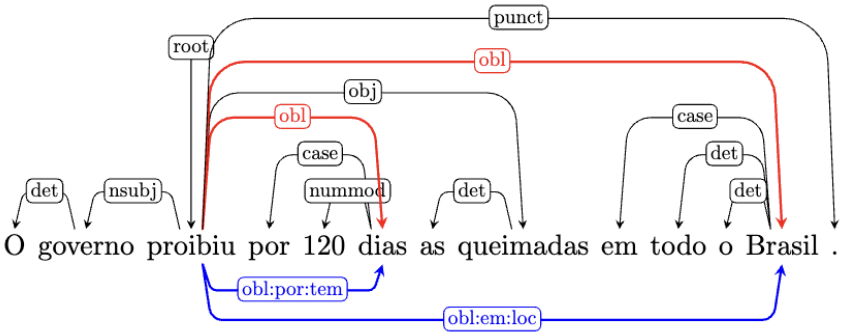

Figura 1. Heads de modificadores enriquecidos com marcadores de caso e papéis semânticos

4.1.2 Propagação de sujeitos de *xcomp*

Esta configuração abrange sentenças com complementos oracionais abertos (*xcomp*), nas quais o sujeito da oração subordinada é nulo (não expresso), mas é controlado pelo sujeito ou pelo objeto da oração matriz. Nas *enhanced dependencies*, anota-se essa relação entre a oração subordinada e o sujeito ou objeto da oração matriz, utilizando a relação *nsubj:xsubj*. O exemplo (2) mostra o sujeito da oração subordinada controlado pelo sujeito da oração matriz, enquanto o exemplo (3) mostra o sujeito da oração subordinada controlado pelo objeto da oração matriz. Os grafos desses dois exemplos podem ser visualizados, respectivamente, nas Figuras 2 e 3.

(2) O governo decidiu proibir as queimadas.

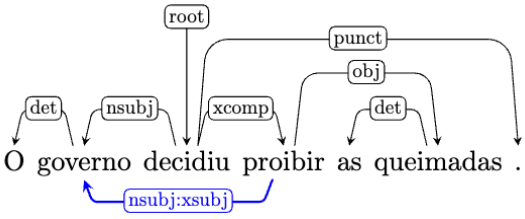

Figura 2. Sujeito da subordinada controlado pelo sujeito da oração matriz

(3) O governo convenceu a oposição a votar no projeto.

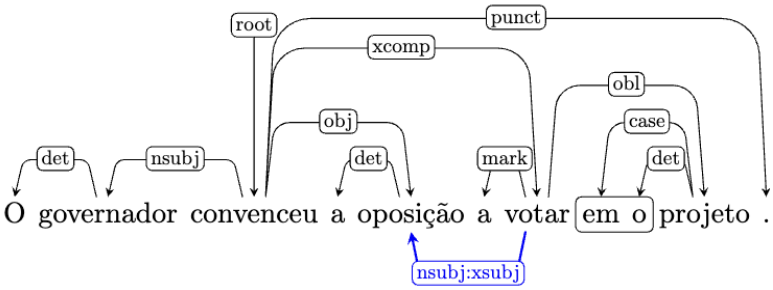

Figura 3. Sujeito da subordinada controlado pelo objeto da oração matriz

4.1.3 Propagação de *head* compartilhado por elementos coordenados

A UD conta separadamente, como dois tipos de ED, o *head* compartilhado por dois elementos coordenados e os dependentes compartilhados por dois elementos coordenados.

A ED de *head* compartilhado ocorre quando há coordenação entre múltiplos elementos que são dependentes de um mesmo *head*, podendo ser vários sujeitos ou objetos de um mesmo predicado ou vários modificadores de um mesmo sintagma nominal. Esses casos são anotados nas dependências básicas com a relação *conj* na direção do *head* para o primeiro elemento coordenado, sendo que os demais elementos coordenados são vinculados ao primeiro. Nas *enhanced dependencies*, estabelecem-se relações entre cada um dos coordenados e o *head* da coordenação. O exemplo (4) ilustra esta configuração, cujo grafo está na Figura 4.

(4) A escolha de Rússia e Qatar para as duas próximas Copas desarranjou a Fifa.

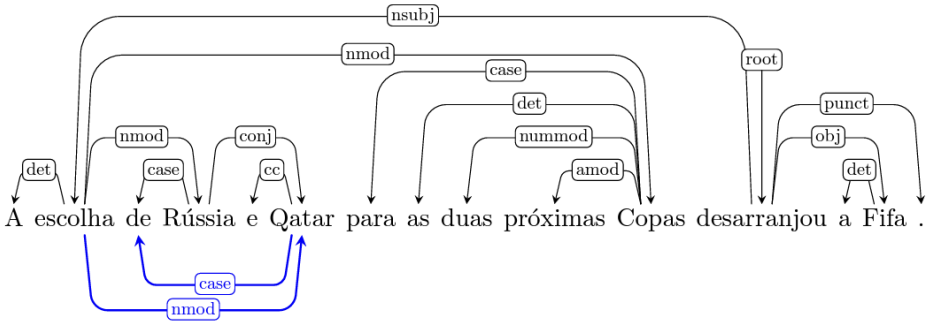

Figura 4. Coordenação de múltiplos elementos dependentes de um mesmo *head*

As relações de dependência básicas na Figura 4 mostram a coordenação de "Rússia" e "Qatar" por meio da relação *conj*, sendo que apenas a primeira está em relação de *nmod* com "escolha". Já nas *enhanced dependencies*, em azul, explicita-se a relação *nmod* entre "Qatar" e "escolha" e a relação de *case* entre "Qatar" e "de".

4.1.3 Propagação de dependentes de elementos coordenados

Já a ED de dependentes compartilhados ocorre quando há dois ou mais elementos coordenados (ligados pela relação *conj*) que possuem um ou mais dependentes em comum. Nas relações de dependência básicas, os dependentes estão ligados a apenas um dos elementos coordenados. Já nas *enhanced dependencies*, os dependentes compartilhados recebem uma ligação para cada um dos elementos coordenados (o que faz com que um mesmo dependente esteja ligado a mais de um *head*). O exemplo (5), cujo grafo pode ser visualizado na Figura 5, ilustra esta configuração, pois o sujeito (*nsubj*) é compartilhado pelos predicados verbais coordenados "pica" e "apresenta".

(5) O inseto pica durante o dia e apresenta fotofobia.

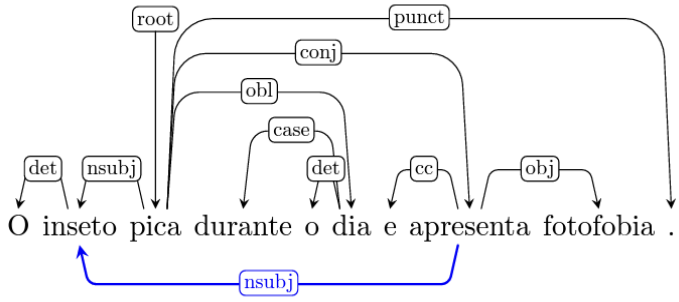

Figura 5. Coordenação de múltiplos predicados com um mesmo dependente

4.2 Enhanced Dependencies de Modificação

As duas ED que modificam as relações da árvore de dependências básicas são: a anotação de referentes de pronomes relativos e a inclusão de predicados elípticos. Essas relações não são tão fáceis de serem anotadas automaticamente e, por isso, demandam maior intervenção humana, seja para anotá-las do zero, seja para revisá-las.

Esses dois tipos de ED justificam algo que vem sendo apresentado por alguns *corpora* que anotam ED e que parece ter se tornado padrão nas diretrizes da UD: a replicação, na coluna 9 (dedicada às ED), de todas as relações compartilhadas com a coluna 8 (dedicada às relações básicas). A seguir explicamos e exemplificamos a anotação das ED de referente de pronomes relativos (4.2.1) e de inserção de token de predicado elíptico (4.2.2).

4.2.1 Anotação do referente dos pronomes relativos

Esta configuração ocorre sempre que há orações relativas, nas quais os pronomes relativos estabelecem uma relação de correferência com o antecedente nominal que

retomam. As dependências básicas anotam a função do pronome relativo em relação ao predicado da oração relativa. Nas ED, o antecedente é vinculado ao pronome relativo que o retoma por meio da relação *ref*. A relação que unia o predicado da oração relativa ao pronome relativo passa ser feita diretamente com o antecedente. O exemplo (6) ilustra esta configuração, cujo grafo pode ser visualizado na Figura 6.

(6) Conheça os 21 livros que o ex-presidente já leu na prisão.

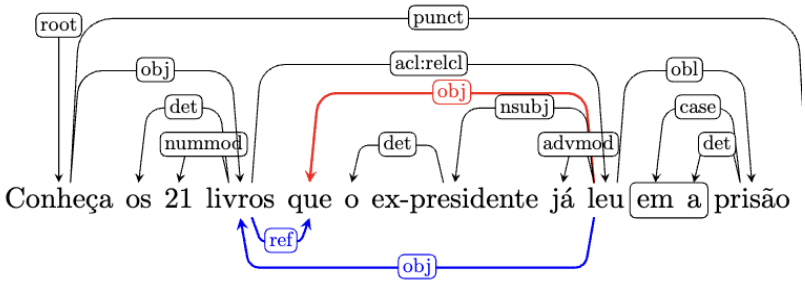

Figura 6. Relação de correferência em construções de orações relativas

4.2.2 Inclusão de predicado elíptico

Esta configuração diz respeito a sentenças nas quais há duas orações, sendo que em uma delas há a realização de um predicado e seu sujeito, enquanto na outra apenas o sujeito é realizado e há elipse do predicado. Nas relações de dependência básicas, sentenças como essas são anotadas com a relação *orphan* para indicar que há um predicado elíptico. Nas *enhanced dependencies*, insere-se um *token* nulo para representar o predicado elíptico e receber as relações a ele devidas. Esse *token* recebe também o lema, a categoria morfossintática e os atributos morfológicos da palavra elíptica. Cumpre destacar que esse tipo de elipse (de predicado) é a única configuração que permite a inserção de um nó vazio na UD. O exemplo (7) ilustra essa configuração de elipse do segundo predicado de uma coordenação. A Figura 7 apresenta uma visualização do grafo dessa sentença, inserindo o token do verbo elíptico "ficam".

(7) A gente fica preso e os bandidos, soltos.

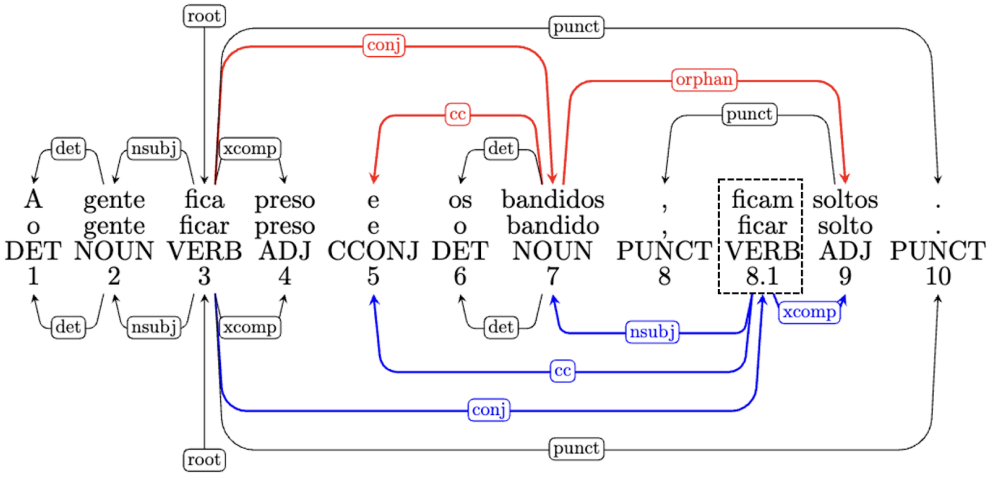

Figura 7. Anotação de predicado elíptico

Na parte de cima da Figura 7, vemos as relações de dependência básicas, nas quais a relação entre "bandidos" e "soltos" é anotada como *orphan*, uma vez que há elipse do verbo. Na parte de baixo, vemos as ED, mostrando a inserção de um *token* no lugar do verbo elíptico, destacado com linha pontilhada na Figura 7. Esse *token* recebe a numeração do *token* anterior (neste caso, 8) juntamente com a indicação de que se trata do primeiro *token* inserido (8.1). Recebe também a forma da palavra que preenche a elipse ("ficam"), além do respectivo lema ("ficar") e a classe de palavra (VERB) com seus atributos morfológicos. As relações que o *token* inserido estabelece com as outras palavras estão destacadas em azul: *head* de *nsubj* com "bandidos", dependente de *conj* com "fica", *head* de *xcomp* com "soltos", e *head* de *cc* com a conjunção "e".
Uma dificuldade adicional a esse tipo de anotação em português é o fato de que, embora o verbo já tenha aparecido na sentença, nem sempre a forma será repetida, pois o preenchimento da elipse pode exigir uma flexão diferente.

5. O enriquecimento de *treebanks* em português com *enhanced dependencies*

Uma vez definidas as configurações de ocorrência de *enhanced dependencies* em português e suas diretrizes de anotação, o próximo passo natural é proceder à anotação das *enhanced dependencies* nos *treebanks* já anotados de acordo com a UD para o português, como o Bosque (Rademaker et al., 2017), o Porttinari (Pardo et al., 2021) e o PetroGold (Souza et al., 2021), entre outros.

A anotação de ED conta com iniciativas automáticas relatadas na literatura da área, valendo-se normalmente da conversão automática de relações de dependência básicas para *enhanced*. Essa alternativa é interessante, pois permite que anotadores humanos avaliem a qualidade da anotação e revisem os casos necessários, como relatado por Nivre et al. (2018). Schuster & Manning (2016) e Grünewald et al. (2021), por exemplo, desenvolveram conversores de relações de dependência básicas para ED para a língua inglesa. Na avaliação deles, embora a conversão tenha gerado resultados com alta acurácia, houve casos nos quais os grafos obtidos construíram significados distintos daqueles construídos pelas sentenças de origem.

Nivre et al. (2018) exploraram duas técnicas de conversão automática aplicadas a múltiplas línguas, uma delas adaptando o conversor de Schuster & Manning (2016) e a outra adaptando o conversor de Nyblom et al. (2013) e avaliaram os resultados como satisfatórios. Heinecke (2020) também relataram bons resultados por meio de uma abordagem híbrida que inclui o *parsing* para a extração de relações básicas e a aplicação de regras linguísticas para a geração de relações *enhanced*.

Portanto, para proceder à anotação de ED para o português, é interessante testar um conversor ou *script* já utilizado por outras línguas e avaliar os resultados. Se forem satisfatórios, passa-se diretamente à revisão manual; se não, procede-se primeiramente à adaptação do conversor utilizando regras específicas para o português, para depois fazer a revisão manual. Como hipotetizamos anteriormente, há relações ED que apresentam maior probabilidade de serem automatizadas do que outras, o que pode indicar a

conveniência de se proceder à anotação de um tipo de ED de cada vez. Uma vez concluído esse trabalho, os *treebanks* anotados com *enhanced dependencies* poderão ser usados para treinar classificadores que realizem essa anotação automaticamente, sem o uso de regras, dedicados à língua portuguesa.

Agradecimentos

Adriana S. Pagano é bolsista de Produtividade em Pesquisa do Conselho Nacional de Desenvolvimento Científico e Tecnológico (Processo CNPq 313103/2021-6). Magali Duran e Thiago Pardo realizaram este trabalho 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. Também receberam apoio do 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

Candito, M.; Guillaume, B.; Perrier, G.; Seddah, D. (2017) Enhanced UD Dependencies with Neutralized Diathesis Alternation. In Proceedings of the Fourth International Conference on Dependency Linguistics, pages 42-53.

Duran, M.S. (2021. Manual de Anotação de PoS tags: Orientações para anotação de etiquetas morfossintáticas em Língua Portuguesa, seguindo as diretrizes da abordagem Universal Dependencies (UD). Relatório Técnico do ICMC 434. Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo.

Duran, M.S. (2022) Manual de Anotação de Relações de Dependência - Versão Revisada e Estendida: Orientações para anotação de relações de dependência sintática em Língua Portuguesa, seguindo as diretrizes da abordagem Universal Dependencies (UD). Relatório Técnico do ICMC 440. Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo.

Droganova, K.; Zeman, D. (2019) Towards Deep Universal Dependencies. In Proceedings of the Fifth International Conference on Dependency Linguistics, pages 144-152.

Ek, Adam; Bernardy, Jean Philippe. (2020) How much of enhanced UD is contained in UD? Proceedings of the16th International Conference on Parsing Technologies and the IWPT 2020 Shared Task, pages 221–226. Virtual Meeting, July 9, 2020.c2020 Association for Computational Linguistics.

Grünewald, S.; Piccirilli, P.; Friedrich, A. (2021) Coordinate Constructions in English Enhanced Universal Dependencies: Analysis and Computational Modeling. In Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, pages 795-809.

Heinecke, J. (2020) Hybrid Enhanced Universal Dependencies Parsing. In Proceedings of the 16th International Conference on Parsing Technologies and the IWPT 2020 Shared Task on Parsing into Enhanced Universal Dependencies, pages 174-180.

de Marneffe, M.-C.; MacCartney, B.; Manning, C.D. (2006. Generating Typed Dependency Parses from Phrase Structure Parses. In Proceedings of the Fifth International Conference on Language Resources and Evaluation, pages 449-454.

de Marneffe, M.-C.; Manning, C.D. (2008) The Stanford Typed Dependencies Representation. In Proceedings of the workshop on Cross-Framework and Cross-Domain Parser Evaluation, pages 1-8.

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

Nyblom, J, Kohonen, S.; Haverinen, K.; Salakoski, T.; and Ginter, F. (2013) Predicting conjunct propagation and other extended stanford dependencies. In Proceedings of theSecond International Conference on Dependency Linguistics (DepLing2013). pages 252–261.

Pardo, T.A.S.; Duran, M.S.; Lopes, L.; Di Felippo, A.; Roman, N.T.; Nunes, M.G.V. (2021) Porttinari - a large multi-genre treebank for Brazilian Portuguese. In Proceedings of the XIII Symposium in Information and Human Language, pages 1-10.

Rademaker, A.; Chalub, F.; Real, L.; Freitas, C.; Bick, E.; Paiva, V. (2017) Universal Dependencies for Portuguese. In Proceedings of the Fourth International Conference on Dependency Linguistics, pages 197-206.

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, pages 2371-2378.

Schuster, S., de La Clergerie, É. V., Candito, M. D., Sagot, B., Manning, C. D., & Seddah, D. (2017) Paris and Stanford at EPE 2017: Downstream Evaluation of Graph-based Dependency Representations. In EPE 2017-The First Shared Task on Extrinsic Parser Evaluation, pages 47-59.

Souza, E.; Silveira, A.; Cavalcanti, T.; Castro, M.; Freitas, C. (2021) Petrogold – corpus padrão ouro para o domínio do petróleo. In Proceedings of the XIII Symposium in Information and Human Language, pages 29-38.