# PropBank e anotação de papéis semânticos para a língua portuguesa: O que há de novo?

**Cláudia Freitas**[1], **Thiago Alexandre Salgueiro Pardo**[1]

[1]Núcleo Interinstitucional de Linguística Computacional (NILC)
Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo

claudiafreitas@usp.br, taspardo@icmc.usp.br

***Abstract.*** *This paper introduces Porttinari-base PropBank (PBP): the Porttinari-base corpus with a semantic role layer. The annotation was performed on syntactic dependencies, using linguistic rules and under human inspection. More than 40,000 arguments were annotated, and the results are discussed in light of works investigating the generalization of PropBank labels.*

***Resumo.*** *O artigo introduz o Porttinari-base PropBank (PBP): o corpus Porttinari-base com uma camada de papéis semânticos. A anotação foi feita sobre dependências sintáticas, usando regras linguísticas e sob inspeção humana. Foram anotados mais de 40 mil argumentos, e os resultados são discutidos à luz de trabalhos que investigam a generalização das classes do PropBank.*

## 1. Introdução

Entre os métodos utilizados para *representar* computacionalmente informação semântica em textos está a anotação de papéis semânticos (ou SRL – *Semantic Role Labeling*). Papéis semânticos são responsáveis por indicar *quem* fez *o quê*, *para quem*, *onde*, *quando*, *como*, *por quê*, *para quê*, *com o quê*, *com quem*, etc., e assim estruturam de maneira *explícita* e *interpretável* a informação contida em enunciados linguísticos. Enquanto tarefa, a anotação de papéis semânticos tem como objetivo atribuir etiquetas a argumentos de predicadores, indicando o papel que estes argumentos exercem em uma frase.

A anotação de papéis semânticos permite criar representações semânticas *estáveis* ao longo de diferentes realizações linguísticas, e as frases 1 a 5 ilustram este ponto. Sintaticamente, "porta" exerce diferentes funções, assim como "chave". Na atribuição de papéis semânticos, "porta" é a "coisa abrindo" em todas as 5 frases, e "chave" é o instrumento de abertura em todas as 5 frases, não importa a função sintática que exerçam. Na frase "O tempo abriu no feriado", entretanto, teríamos uma outra estrutura argumental (e outra representação semântica), já que estaríamos diante de um outro sentido de "abrir".

1. A chave **abriu** a porta.
2. Ela **abriu** a porta com a chave.
3. A porta foi **aberta** com a chave.
4. A porta foi **aberta** por ela com a chave.
5. A porta **abriu** com a chave.

Um PropBank (*Proposition Bank*, ou Banco de Proposições) é um corpus que contém anotação de papéis semânticos, relacionando verbos[1] e seus argumentos diretamente às estruturas sintáticas de um *treebank* e conforme o modelo sugerido por

[1]Atualmente, substantivos e adjetivos também podem ser considerados.

[Palmer et al. 2005], já que, teoricamente, papéis semânticos estariam na interface sintaxe-semântica [Palmer et al. 2005, Levin 1993, Levin and Rappaport Hovav 2005].

Apesar de ser um fenômeno linguístico amplamente estudado, não há consenso relativo ao conjunto de papéis semânticos da língua. No PropBank, a diversidade teórica acerca dos papéis semânticos é contornada com a utilização de i) argumentos numerados, que vão de Arg0 a Arg5; e ii) argumentos modificadores, um conjunto mais amplo de argumentos. A motivação para o conjunto limitado de papéis numerados é facilitar a generalização para o aprendizado de máquina, ainda que alguns estudos mostrem que este objetivo é apenas parcialmente alcançado [Merlo and Van Der Plas 2009, Gung and Palmer 2021, Li et al. 2023]. A diferença entre argumentos numerados e argumentos modificadores está sobretudo na natureza da relação sintática que o argumento mantém com o verbo (*exigência*, nos argumentos numerados, vs *opcionalidade*, nos argumentos modificadores). A distinção entre os argumentos também é motivada pela assistematicidade semântica dos papéis com relação aos verbos, e por isso argumentos numerados são específicos de verbos (o Arg0 do verbo "abrir" é "quem abre", e o Arg0 de "alagar" é "causador do alagamento"). Os ArgM, por outro lado, têm uma semântica específica e previsível (indicada pelo nome da etiqueta, como ArgM-tmp para "tempo" e ArgM-cau para "causa"), e podem se associar a qualquer verbo.

A semântica dos argumentos numerados é revelada com o alinhamento entre a anotação do corpus e o recurso lexical associado ao PropBank, os chamados *Frame Files* – em nosso caso, dispomos do Verbo-Brasil [Sanches Duran and Aluísio 2015]. Assim, a frase (2), segundo o estilo PropBank, é anotada como indicado abaixo. Um PropBank, portanto, não é apenas um corpus anotado, mas a associação entre um corpus anotado e um léxico, que indica como os elementos devem ser anotados e o que eles significam.

- Ela[Arg0] **abriu** a porta[Arg1] com a chave[Arg2].

Neste artigo, apresentamos o *Porttinari-base PropBank* (PBP), que corresponde à adição de uma camada de papéis semânticos a todas as frases do corpus jornalístico *Porttinari-base*, que compõe o *treebank* Porttinari [Pardo et al. 2021, Duran et al. 2023]. O Porttinari-base é padrão ouro na anotação sintática conforme a teoria *Universal Dependencies* (UD) [de Marneffe et al. 2021]. Com o PBP, o Porttinari-base passa a ser duplamente padrão ouro – nas dependências sintáticas e nos papéis semânticos —, contribuindo para a disponibilização de recursos de alto nível para o processamento do português.

## 2. Motivação e trabalhos relacionados

O PropBank foi criado com o objetivo de treinar modelos no aprendizado supervisionado. Levando em conta a onipresença de arquiteturas neurais e grandes modelos de língua, discutimos brevemente a relevância para o PLN da anotação de papéis semânticos.

Entre as críticas ao atual paradigma de IA, estão a falta de transparência e de explicabilidade dos métodos e resultados. Neste contexto, papéis semânticos oferecem maneiras interpretáveis de representar semanticamente enunciados verbais, podendo servir de insumo, por exemplo, para a criação de grafos de conhecimento [Mohebbi et al. 2022], o que torna este tipo de representação semântica relevante para a investigação acerca da articulação entre os grandes modelos de língua (os LLMs – *Large Language Models*) e fontes de conhecimento estruturado [Dong 2023]. Na articulação entre redes neurais e

SRL, [Mohebbi et al. 2022] propõem uma abordagem de aprendizado de grafos profundos para computar similaridade semântica de documentos, usando papéis semânticos.

A anotação de SRL também pode ser usada para avaliação de modelos de língua quanto à capacidade de representar informação semântica estruturada e interpretável [Tenney et al. 2019a, Tenney et al. 2019b, Han and Choi 2020]. Nessa vertente, notamos a escassez de conjuntos de validação criados para o português, que nos faz utilizar conjuntos de dados traduzidos do inglês [Rodrigues et al. 2023]. A utilidade da anotação de papéis semânticos no PLN também pode ser indireta. Considerando o paradigma de avaliação extrínseca, [Evans and Orasan 2019] utilizam SRL para verificar se a simplificação textual é capaz de facilitar o desempenho na tarefa de SRL. Uma vez que a tarefa de SRL pode ser considerada um passo além da análise sintática, diferentes modelos de representação sintática também podem ser avaliados em função do desempenho obtido na anotação SRL, como sugerido em [Freitas 2023].

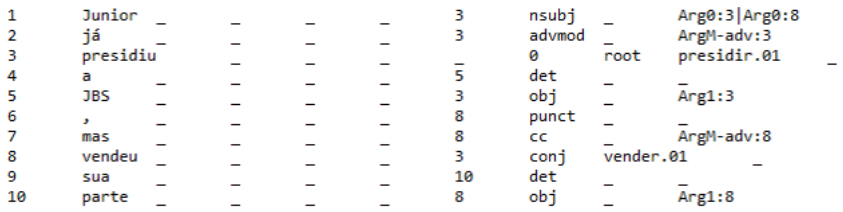
Desde 2011, o português dispõe do PropBank-BR [Duran and Aluísio 2011], que anotou papéis semânticos ao estilo PropBank sobre a parte brasileira do treebank Bosque, em sua versão de sintaxe de constituintes disponibilizada pela Linguateca. Este PropBank levou à formulação (e adaptação do inglês) de diretivas de anotação e permitiu a criação do recurso léxico que codifica os sentidos dos verbos e descreve seus frames sintáticos, o Verbo-Brasil [Sanches Duran and Aluísio 2015]. No entanto, esse recurso é ainda limitado. Tendo em vista a criação de um material maior e mais lexicalmente diversificado, foi criado o PropBank-BR v2 [Duran et al. 2014, Hartmann et al. 2016]. Diferentemente da versão anterior, este material foi construído sobre árvores sintáticas não revistas. Em ambos os casos, o processo de anotação seguiu o PropBank original, com a anotação feita de maneira linear, frase a frase. O português conta ainda com o CINTIL-PropBank [Branco et al. 2012], um corpus de frases anotadas com estrutura de constituintes e papéis semânticos, criado de maneira semiautomática, com algumas etiquetas anotadas automaticamente, e com um conjunto de papéis semânticos que é uma adaptação dos argumentos numerados de [Palmer et al. 2005]. Por fim, em uma abordagem baseada em regras, [Bick 2007] faz SRL seguindo a *Constraint Grammar*.

## 3. O Porttinari-base PropBank

No PBP, a anotação de papéis semânticos foi baseada em dependências, sendo cada argumento um único token. A anotação foi feita em um arquivo no formato CoNLL-U, que consiste em um texto simples com 10 campos separados por caracteres de tabulação[2]. A anotação foi feita nas colunas 10 (nomeada de MISC) para a atribuição dos argumentos, e 9 (coluna DEPS) para a anotação dos frames. Foram anotados argumentos explícitos e implícitos, como sujeitos omitidos. A Figura 1 apresenta a codificação da frase "Júnior já presidiu a JBS, mas vendeu a sua parte". Para facilitar a leitura, omitimos os conteúdos das colunas 3 a 6. A coluna 9 indica os frames dos verbos "presidir" e "vender" (respectivamente, presidir.01 e vender.01). A coluna 10 informa os papéis semânticos e seus predicadores. Por exemplo, o token 1, "Junior", é Arg0 do token 3 ("presidir") e Arg0 do token 8 ("vender"), mesmo que esta última informação não esteja explícita na frase.

A anotação do PBP utilizou 26 etiquetas. Diferentemente do Propbank original, criamos, no PBP, etiquetas que especificam alguns casos da classe mais geral ArgM-adv:

[2]`https://universaldependencies.org/format.html`



| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 1 | Junior | _ | _ | _ | _ | 3 | nsubj | _ | Arg0:3\|Arg0:8 |
| 2 | já | _ | _ | _ | _ | 3 | advmod | _ | ArgM-adv:3 |
| 3 | presidiu | | _ | _ | _ | _ | 0 | root | presidir.01 _ |
| 4 | a | _ | _ | _ | _ | 5 | det | _ | _ |
| 5 | JBS | _ | _ | _ | _ | 3 | obj | _ | Arg1:3 |
| 6 | , | _ | _ | _ | _ | 8 | punct | _ | _ |
| 7 | mas | _ | _ | _ | _ | 8 | cc | _ | ArgM-adv:8 |
| 8 | vendeu | _ | _ | _ | _ | 3 | conj | vender.01 | _ |
| 9 | sua | _ | _ | _ | _ | 10 | det | _ | _ |
| 10 | parte | _ | _ | _ | _ | 8 | obj | _ | Arg1:8 |

**Figura 1. Anotação de papéis semânticos nas colunas 9 e 10 do CoNLL-U**

*ArgM-conseq*, para indicar consequência; *ArgM-cond*, para indicar condições e *ArgM-comp*, para indicar comparações. A divergência com relação à lista de papéis do PropBank original está na classe *ArgM-src* (*source*, fonte da informação), para ocorrências como *De acordo com a polícia, trata-se de uma "prisão significativa"para as investigações*. No PropBank original, este tipo de construção não deve ser anotado, mas as anotamos pela relevância argumentativa/retórica. Em consonância com as versões do PropBank-BR, utilizamos etiquetas específicas para *verbos auxiliares* (de tempo, modo, aspecto e voz) para o pronome *-se*.

O material foi anotado com base no *Manual de anotação do PropBank-BR v2* [Duran 2014] e nos frames verbais elencados no recurso Verbo-Brasil. Ao longo do projeto, as diretivas de anotação foram enriquecidas e atualizadas, dando origem a [Duran and Freitas 2024]. Com relação a seus antecessores brasileiros, a anotação PBP difere quanto à independência da camada sintática no que se refere à identificação/segmentação de argumentos (se a segmentação da análise sintática e a segmentação de argumentos indicada no Verbo-Brasil divergirem, seguimos o Verbo-Brasil).

É interessante destacar as interferências da sintaxe UD na tarefa de SRL. No PBP, as divergências entre anotações sintática e semântica foram motivadas pela impossibilidade, em UD, de cruzar arcos sintáticos (exemplo 1), e em frases com verbos auxiliares, uma vez que a sintaxe UD é bastante econômica quanto ao que deve ser considerado verbo auxiliar. Na anotação UD do Porttinari, estão anotados como auxiliares apenas auxiliares de tempo e voz. Na atribuição de papéis semânticos, porém, esta economia tem como resultado a (falsa) necessidade de atribuir papéis a elementos que, em português, não estão atuando como verbos plenos ("possam"no exemplo 2), e que portanto não deveriam receber papéis semânticos – e o resultado é uma construção sem sentido.

1. O defeito, **que** a Takata demorou a **reconhecer**, foi revelado em...
   (a) Codificação UD: Takata demorou o defeito
   (b) Codificação PBP: Takata reconheceu o defeito
2. O projeto **prevê** que as deduções só possam ocorrer a partir de 2021
   (a) Codificação UD: deduções possam; prevê possam; possam ocorrer
   (b) Codificação PBP: ocorrer deduções; prevê ocorrer

O fato de *verbos de ligação* serem considerados AUX, com relações de dependência "especiais" (os argumentos sintáticos – sujeito e predicativo – ficam dissociados do verbo "ser", e o núcleo do sintagma é o elemento nominal predicativo) também levou à divergência de anotações, já que o verbo "ser" tem papéis semânticos para as posições de sujeito e de predicativo do sujeito.

## 4. Metodologia

Em termos gerais, a anotação de papéis semânticos no PBP seguiu as seguintes etapas, algumas delas concomitantes:

1. Identificação do predicador, que em nosso caso foram apenas os verbos;
2. Consulta ao Verbo-Brasil para seleção do frame adequado;
3. Anotação dos argumentos numerados conforme descritos no Verbo-Brasil;
4. Anotação dos argumentos modificadores conforme descritos nas diretivas;
5. Caso necessário, criação de frames provisórios para verbos novos ou sentidos novos de formas verbais já presentes no Verbo-Brasil;
6. Aplicação de regras de validação para detectar problemas na anotação.

Todo o processo de anotação foi feito com base em regras linguisticamente motivadas e de maneira não-linear [Wallis 2003], seguindo o exemplo de [Freitas et al. 2023]. Nisto, diferimos do processo de anotação do PropBank original, no qual cada frase era anotada inteiramente de uma vez, e do PropBank-BR versões 1 e 2, que seguiu a mesma metodologia. A anotação foi feita com a ferramenta ET [de Souza and Freitas 2021], utilizando o ambiente *Interrogatório*.

A anotação foi feita em 3 fases: (i) anotação de elementos explícitos, sempre que possível usando regras com padrões léxico-sintáticos derivados dos exemplos de Duran (2014) e das frases-exemplo no Verbo-Brasil; (ii) anotação de elementos implícitos (que envolveu sobretudo a propagação de sujeitos, feita com regras) e (iii) aplicação final de regras de validação e detecção de inconsistências. Quase todo o processo foi semi-automático, utilizando regras que associam um padrão de busca a uma regra de anotação, sempre com revisão humana. Na propagação de sujeitos omitidos de verbos na forma infinitiva, explicitamos argumentos apenas se estes fossem recuperáveis (frase 1). Em caso de dúvida ou em caso de argumentos não recuperáveis (frase 2), nada foi feito. Em [Freitas 2024] estão detalhados os procedimentos e regras utilizados.

1. O *presidente* centrista *optou* por **garantir** pela a primeira vez em anos que... (O presidente garantiu)
2. São mecânicas *que pressionam* a **entender** que isso tem custo. (não é possível determinar quem entenderá)

A anotação de elementos implícitos levou a um outro tipo de desalinhamento entre sintaxe e semântica. Na frase "A Folha **pediu** [*contato com o **general Mourão***] , para que **comentasse** suas declarações, mas (...)", o segmento "contato com o general Mourão" é Arg1 do verbo "pedir", mas apenas "general Mourão" é sujeito/Arg0 de "comentar".

A validação final consistiu na aplicação de 4 regras (Figura 2), que buscavam frases com condições suspeitas. Foram encontrados 101 casos suspeitos, e apenas dois deles (derivados da regra 4) eram falsos positivos. Todos os erros foram corrigidos manualmente. Diferentemente das regras utilizadas no processo de anotação, as regras de validação podem ser aplicadas para a verificação final de outros corpora com anotação de papéis semânticos. As regras de anotação, por sua vez, não foram criadas com o objetivo de serem generalizáveis para outros corpora, mas de criar um material padrão ouro de qualidade e no menor tempo possível. No entanto, a elaboração de um anotador baseado em regras, que aproveite estas regras e o corpus já anotado, é algo bastante possível.

1. Um mesmo token que contenha dois ou mais argumentos diferentes que estejam associados a um mesmo head. Esta regra garante que não há um mesmo token com papéis diferentes com relação ao mesmo verbo.
2. Dois tokens com exatamente a mesma etiqueta no que se refere a Args numerados. Esta regra garante que um verbo não terá dois Arg1 associados a ele, por exemplo.
3. Nenhum token cujo deprel é root pode ter papel semântico.
4. Não há token dependente de ArgM-mod com anotação do tipo Arg1_.*xcomp na coluna 10. Esta regra garante que todos os tokens anotados com ArgM-mod precisam ter um dependente com anotação do tipo Arg1.* na coluna 10.

**Figura 2. Regras para detecção de erros**

A anotação foi feita por uma única pessoa, por 7 meses, a partir das informações contidas no Manual de anotação do PropBank v2 e no Verbo-Brasil. A fim de avaliar a qualidade da anotação, foi feita uma concordância inter-anotadores *a posteriori*, tomando como base uma amostra com as 100 frases com a maior quantidade de argumentos anotados no Propbank-BR v2, sobre o qual se baseiam as diretivas de anotação e o Verbo-Brasil. Essa amostra foi reanotada sintática e semanticamente, e os resultados comparados. É importante notar também que, embora tenhamos escolhido as 100 frases com a maior quantidade de argumentos anotados, nem todos os verbos do PropBank-BR v2 têm seus argumentos anotados, apenas aqueles mais frequentes no corpus. A comparação, medida com o índice *kappa*, foi sobre 443 tokens anotados por ambas as anotações, e resultou em uma convergência de .90 (como comparação, no Propbank original, e considerando apenas a classificação de papéis incluindo Arg-M, o kappa foi de .93).

O processo de anotação dos frames dos verbos foi concomitante à anotação dos papéis semânticos. Uma vez que o foco da anotação PBP esteve na atribuição dos papéis semânticos, o processo de atribuição de sentidos aos verbos não foi exaustivo. Além de não exaustiva, a anotação privilegiou o alinhamento com verbos não monossêmicos, uma vez que a anotação de verbos monossêmicos poderia ser feita (e foi) de forma automática.

Apesar de já dispor de um recurso como o Verbo-Brasil, um corpus novo sempre traz novas formas verbais e novos sentidos para verbos já descritos. Para os sentidos (ainda) sem frames, foi feita uma busca por um verbo similar no Verbo-Brasil, e a solução foi indicada em um documento para posterior aprimoramento do Verbo-Brasil. Ao final do processo, foram documentados cerca de 350 sentidos de verbos sem frames, com exemplos do corpus e soluções provisórias em boa parte deles. A anotação de frames monossêmicos foi feita de maneira automática para os casos de verbos monossêmicos (ou seja, que só dispunham de um frame). Este procedimento levou à inclusão de mais de 500 frames no PBP.

A relação estreita entre anotação sintática e anotação de papéis semânticos, por um lado, e as interferências da anotação sintática UD, por outro, levaram à criação de diferentes versões do corpus, e com isso também criamos condições para investigar o papel de diferentes representações linguísticas no aprendizado de SRL. Cada uma das versões é gerada automaticamente a partir de uma versão base.

1. **PBP na versão UD**: Esta versão se caracteriza pela atribuição de papéis semânticos apenas aos elementos considerados verbos plenos na UD. Em consequência: (i) não foram anotados os papéis de argumentos do verbo “ser”; (ii)

foram anotados os papéis de argumentos de verbos considerados plenos pela UD, mas considerados auxiliares no Verbo-Brasil. No entanto, para diferenciá-los dos demais argumentos, receberam a etiqueta Arg1_d e Arg0_d. Se desejável, ambas as etiquetas podem ser substituídas por Arg1 e Arg0, respectivamente.

2. **PBP na versão clássica**: Esta versão prioriza o conceito de *proposição*, e se caracteriza pela atribuição de papéis conforme o modelo PropBank, independentemente do que foi considerado verbo pleno pela UD. Em consequência: (i) foram anotados os papéis de argumentos do verbo "ser"; (ii) não foram anotados os papéis de argumentos de verbos que, apesar de considerados plenos na UD, são considerados auxiliares no Verbo-Brasil e na documentação [Duran and Freitas 2024]; e iii) foram anotados como modificadores auxiliares os verbos que, apesar de considerados plenos pela anotação UD, são considerados auxiliares no PBP (ArgM-mod, ArgM-asp), além daqueles considerados sempre auxiliares (ArgM-tml; ArgM-pas), seguindo a decisão do PropBank-BR v2 [Duran 2014].

## 5. Resultados e conclusões

Foram anotados 45.813 argumentos verbais e 13.395 instâncias verbais contêm anotação de frames, distribuídos em quase 1.018 frames distintos (60,8% dos verbos possui anotação de frame). As versões anteriores do PropBank-BR continham cerca de 7 mil argumentos anotados. A Figura 3 apresenta a distribuição dos argumentos por função sintática e a Figura 4 traz, com mais detalhes, os papéis semânticos mais frequentes para cada relação sintática (deprel). Todos os números se referem à versão clássica, alinhada à versão 1.0 do *treebank* Porttinari-base.

Vemos que a associação mais frequente é entre obj e Arg1, com 92,03%, seguida da associação entre nsubj e Arg0, com 63,49%. O artigo de [Palmer et al. 2005] traz o mesmo tipo de análise para o PropBank original. No entanto, uma vez que cada gramática recorta e define os elementos que lhes parecem relevantes, é necessário algum cuidado nessa comparação. O elemento sentencial/oracional S, presente no PropBank original (derivado da anotação do *Penn Treebank*), não existe em UD, e está distribuído entre alguns casos de *xcomp* e de *ccomp*, por exemplo.

De forma geral, analisando os dados do PBP, extraímos as seguintes informações: (a) Arg1 se distribui principalmente entre as relações *obj*, *nsubj*, *obl*, *xcomp* e *ccomp*; (b) Arg0 se concentra em *nsubj*; (c) Arg2 se distribui principalmente entre *obl* e *xcomp*; (d) ArgM-tmp e ArgM-mnr se distribuem entre *obl*, *advmod* e *advcl*; (e) ArgM-loc se concentra em *obl*. Com a perspectiva da sintaxe: (a) *obj* participa de Arg1 e parece ser a generalização mais fácil: "se é um obj, então é Arg1"; (b) *xcomp* participa igualmente de Arg1 e Arg2; (c) *obl* participa de Arg2, Arg1, tempo, modo e local; (d) *advmod* participa de neg, adv, tempo, dis e modo; (e) *advcl* participa de tempo, finalidade, modo e Arg2.

Os dados apontam para uma regularidade entre funções sintáticas e papéis numerados – especificamente, papéis Arg0 e Arg1. Os demais argumentos numerados, com baixa frequência, são de mais difícil generalização. Estas constatações convergem com resultados para o inglês, que verificam que a anotação ao estilo PropBank captura melhor regularidades sintáticas, sobretudo para argumentos de frequência alta, em oposição ao estilo VerbNet de anotação de papéis semânticos [Merlo and Van Der Plas 2009]. Lidos de outra forma, embora o objetivo de um PropBank seja muitas vezes servir como

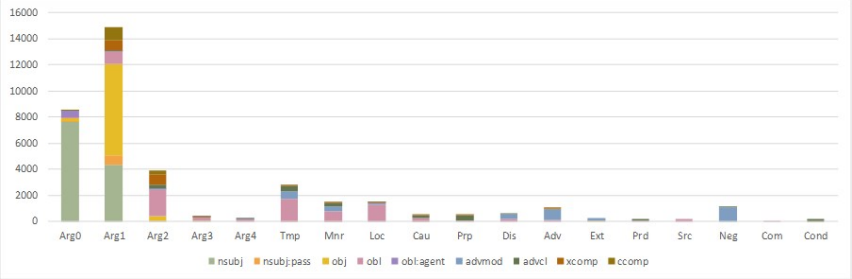



**Figura 3. Distribuição de papéis semânticos por relação sintática**

| *Deprel* | Total de ocorr | Papeis semânticos mais frequentes (%) | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| NSUBJ | 12038 | Arg0 | 63,49 | Arg1 | 35,79 | Arg2 | 0,52 | Arg3 | 0,17 | Mnr | 0,02 | | | | | | | | | | |
| OBL | 6342 | Arg2 | 31,43 | Arg1 | 15,29 | Arg0 | 1,31 | Arg3 | 4,42 | Tmp | 26,70 | Loc | 20,07 | Mnr | 11,89 | Cau | 3,63 | Src | 3,26 | Dis | 2,49 |
| OBJ | 7562 | Arg1 | 92,03 | Arg0 | 2,26 | Arg2 | 4,84 | Arg3 | 0,40 | Ext | 0,21 | Nse | 0,16 | Tmp | 0,07 | | | | | | |
| ADVMOD | 3693 | Neg | 29,62 | Adv | 20,15 | Tmp | 15,52 | Dis | 10,45 | Mnr | 9,83 | Ext | 5,58 | Loc | 4,82 | Arg2 | 2,17 | Arg1 | 1,11 | Arg4 | 0,30 |
| XCOMP | 1790 | Arg1 | 45,20 | Arg2 | 45,92 | Arg3 | 0,28 | Mnr | 0,45 | Cau | 0,89 | | | | | | | | | | |
| ADVCL | 2303 | Tmp | 20,41 | Prp | 16,93 | Mnr | 12,77 | Arg2 | 12,85 | Cau | 9,38 | Cond | 6,82 | Prd | 5,82 | Adv | 5,08 | | | | |

**Figura 4. Papéis semânticos mais frequentes para cada relação sintática**

material de treino para IA, a forma de codificar os papéis talvez seja mais indicada (ou *também* seja indicada) para um anotador baseado em regras e que considere sintaxe. De fato, [Palmer et al. 2005] relatam que um anotador simples baseado em regras tem desempenho de 83%, sendo 85% o estado-da-arte em inglês, considerando cenários difíceis de avaliação, com verbos não vistos na fase de treino [Wang et al. 2022].

Apesar de raros, estudos sobre a capacidade de generalização da anotação ao estilo PropBank têm mostrado que, quando comparada à anotação ao estilo VerbNet, a anotação PropBank leva a resultados inferiores no que se refere a argumentos Arg2 a Arg5 e, em termos gerais, os bons resultados obtidos com a anotação PropBank se referem aos argumentos mais frequentes [Merlo and Van Der Plas 2009, Yi et al. 2007, Gung and Palmer 2021, Li et al. 2023]. No entanto, todos os estudos foram feitos para a língua inglesa, e apenas a disponibilização de recursos similares para o português nos permitiria verificar os resultados para a nossa língua.

Por fim, a criação de duas versões (UD e Clássica) também permite comparar a anotação PBP com outros PropBanks que tenham seguido mais fielmente a anotação UD. As diferentes versões do corpus, bem como documentação linguística detalhada e regras de anotação utilizadas, estão públicas no portal web do projeto POeTiSA[3].

## Agradecimentos

Este trabalho foi realizado no âmbito do Centro de Inteligência Artificial da Universidade de São Paulo (C4AI[4]), 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. Os autores agradecem também a Elvis de Souza pela preparação dos arquivos para disponibilização.

[3] https://sites.google.com/icmc.usp.br/poetisa/resources-and-tools
[4] http://c4ai.inova.usp.br/

## Referências


Bick, E. (2007). Automatic semantic role annotation for portuguese. In *Proceedings of TIL 2007 - 5th Workshop on Information and Human Language Technology*, pages 1713–1716, Rio de Janeiro. Sociedade Brasileira de Computação (SBC).

Branco, A., Carvalheiro, C., Pereira, S., Silveira, S., Silva, J., Castro, S., and Graça, J. (2012). A PropBank for Portuguese: the CINTIL-PropBank. In Calzolari, N., Choukri, K., Declerck, T., Doğan, M. U., Maegaard, B., Mariani, J., Moreno, A., Odijk, J., and Piperidis, S., editors, *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 1516–1521, Istanbul, Turkey. European Language Resources Association (ELRA).

de Marneffe, M.-C., Manning, C. D., Nivre, J., and Zeman, D. (2021). Universal Dependencies. *Computational linguistics*, 47(2):255–308.

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.

Dong, X. L. (2023). Generations of knowledge graphs: The crazy ideas and the business impact. *Proc. VLDB Endow.*, 16(12):4130–4137.

Duran, M., Lopes, L., das Graças Nunes, M., and Pardo, T. (2023). The dawn of the porttinari multigenre treebank: Introducing its journalistic portion. In *Anais do XIV Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana*, pages 115–124, Porto Alegre, RS, Brasil. SBC.

Duran, M. S. (2014). Manual de anotação do PropBank-Br v2. Technical report, ICMC-USP.

Duran, M. S. and Aluísio, S. M. (2011). Propbank-br: a Brazilian Portuguese corpus annotated with semantic role labels. In *Proceedings of the 8th Brazilian Symposium in Information and Human Language Technology*.

Duran, M. S. and Freitas, C. (2024). Guia de anotação de papéis semânticos seguindo o modelo PropBank no corpus Porttinari-base. (no prelo). Technical report, ICMC-USP.

Duran, M. S., Torres, L. S., Viviani, M. C., Hartmann, N., and Aluísio, S. M. (2014). Seleção e preparação de sentenças do corpus PLN-BR para compor o corpus de anotação de papéis semânticos Propbank-Br.v2. Technical report, Núcleo Interinstitucional de Linguística Computacional.

Evans, R. and Orasan, C. (2019). Sentence simplification for semantic role labelling and information extraction. In Mitkov, R. and Angelova, G., editors, *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*, pages 285–294, Varna, Bulgaria. INCOMA Ltd.

Freitas, C. (2023). Dataset e corpus. In Caseli, H. and Volpe Nunes, M. d. G., editors, *Processamento de Linguagem Natural: conceitos, técnicas e aplicações em Português*, pages 1–37. BPLN.

Freitas, C. (2024). Anotação de papéis semânticos no corpus Porttinari-base: Procedimentos, resultados e análise. (no prelo). Technical report, ICMC-USP.

Freitas, C., Souza, E., Castro, M. C., Cavalcanti, T., Ferreira da Silva, P., and Corrêa Cordeiro, F. (2023). Recursos linguísticos para o PLN específico de domínio: o Petrolês. *Linguamática*, 15(2):51–68.

Gung, J. and Palmer, M. (2021). Predicate representations and polysemy in VerbNet semantic parsing. In Zarrieß, S., Bos, J., van Noord, R., and Abzianidze, L., editors, *Proceedings of the 14th International Conference on Computational Semantics (IWCS)*, pages 51–62, Groningen, The Netherlands (online). Association for Computational Linguistics.

Han, H. and Choi, J. (2020). Establishing strong baselines for the new decade: Sequence tagging, syntactic and semantic parsing with bert. In *Proceedings of the Thirty-Third International Florida Artificial Intelligence Research Society Conference (FLAIRS 2020)*.

Hartmann, N. S., Duran, M. S., and Aluísio, S. M. (2016). Automatic semantic role labeling on non-revised syntactic trees of journalistic texts. In Silva, J., Ribeiro, R., Quaresma, P., Adami, A., and Branco, A., editors, *Computational Processing of the Portuguese Language*, pages 202–212, Cham. Springer International Publishing.

Levin, B. (1993). *English Verb Classes and Alternations: a preliminary investigation*. The University of Chicago Press, London.

Levin, B. and Rappaport Hovav, M. (2005). *Argument Realization*. Cambridge University Pres, Cambridge.

Li, T., Kazeminejad, G., Brown, S., Srikumar, V., and Palmer, M. (2023). Learning semantic role labeling from compatible label sequences. In Bouamor, H., Pino, J., and Bali, K., editors, *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 15561–15572, Singapore. Association for Computational Linguistics.

Merlo, P. and Van Der Plas, L. (2009). Abstraction and generalisation in semantic role labels: PropBank, VerbNet or both? In Su, K.-Y., Su, J., Wiebe, J., and Li, H., editors, *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 288–296, Suntec, Singapore. Association for Computational Linguistics.

Mohebbi, M., Razavi, S. N., and Balafar, M. A. (2022). Computing semantic similarity of texts based on deep graph learning with ability to use semantic role label information. *Scientific Reports*, 12(1).

Palmer, M., Gildea, D., and Kingsbury, P. (2005). The proposition bank: An annotated corpus of semantic roles. *Computational linguistics*, 31(1):71–106.

Pardo, T., Duran, M., Lopes, L., Felippo, A., Roman, N., and Nunes, M. (2021). Porttinari - a large multi-genre treebank for brazilian portuguese. In *Anais do XIII Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana*, pages 1–10, Porto Alegre, RS, Brasil. SBC.

Rodrigues, J., Gomes, L., Silva, J., Branco, A., Santos, R., Cardoso, H. L., and Osório, T. (2023). Advancing neural encoding of portuguese with transformer albertina pt-*. In Moniz, N., Vale, Z., Cascalho, J., Silva, C., and Sebastião, R., editors, *Progress in Artificial Intelligence*, pages 441–453, Cham. Springer Nature Switzerland.

Sanches Duran, M. and Aluísio, S. (2015). Automatic generation of a lexical resource to support semantic role labeling in Portuguese. In Palmer, M., Boleda, G., and Rosso, P., editors, *Proceedings of the Fourth Joint Conference on Lexical and Computational Semantics*, pages 216–221, Denver, Colorado. Association for Computational Linguistics.

Tenney, I., Das, D., and Pavlick, E. (2019a). BERT rediscovers the classical NLP pipeline. In Korhonen, A., Traum, D., and Màrquez, L., editors, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4593–4601, Florence, Italy. Association for Computational Linguistics.

Tenney, I., Xia, P., Chen, B., Wang, A., Poliak, A., McCoy, R. T., Kim, N., Durme, B. V., Bowman, S. R., Das, D., and Pavlick, E. (2019b). What do you learn from context? probing for sentence structure in contextualized word representations. In *International Conference on Learning Representations*.

Wallis, S. (2003). Completing parsed corpora: From correction to evolution. In Abeillé, A., editor, *Treebanks: Building and Using Parsed Corpora*, pages 61–71. Springer Netherlands, Dordrecht.

Wang, N., Li, J., Meng, Y., Sun, X., Qiu, H., Wang, Z., Wang, G., and He, J. (2022). An MRC framework for semantic role labeling. In Calzolari, N., Huang, C.-R., Kim, H., Pustejovsky, J., Wanner, L., Choi, K.-S., Ryu, P.-M., Chen, H.-H., Donatelli, L., Ji, H., Kurohashi, S., Paggio, P., Xue, N., Kim, S., Hahm, Y., He, Z., Lee, T. K., Santus, E., Bond, F., and Na, S.-H., editors, *Proceedings of the 29th International Conference on Computational Linguistics*, pages 2188–2198, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Yi, S.-t., Loper, E., and Palmer, M. (2007). Can semantic roles generalize across genres? In Sidner, C., Schultz, T., Stone, M., and Zhai, C., editors, *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference*, pages 548–555, Rochester, New York. Association for Computational Linguistics.