Proposta recebida em March 2025 e aceite para publicação em August 2025.

PropBanks e representações semânticas: o que temos, o que queremos e o que podemos*

PropBanks and semantic representations: what we have, what we want and what we can do

Cláudia Freitas
Núcleo Interinstitucional de Linguística Computacional (NILC)
Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo

Thiago Alexandre Salgueiro Pardo
Núcleo Interinstitucional de Linguística Computacional (NILC)
Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo

Resumo

Neste artigo, detalhamos a metodologia utilizada na construção do Porttinari-base Propbank, um Propbank padrão-ouro com mais de 45 mil argumentos anotados sobre as dependências sintáticas do treebank Porttinari-base. Apresentamos os desafios subjacentes e as soluções adotadas na anotação de argumentos e predicados, e relatamos os resultados de um estudo sobre concordâncias na anotação. Nosso objetivo é duplo: apresentar o Porttinari-base Propbank enquanto recurso do PLN e, secundariamente, provocar reflexões acerca dos diálogos entre PLN e teorias linguísticas.

Palavras chave

papéis semânticos; anotação semântica; dependências sintáticas; dependências semânticas; dependências universais

Abstract

In this paper, we detail the methodology used to construct the Porttinari-base Propbank, a gold-standard Propbank with over 45,000 arguments annotated on the syntactic dependencies of the Porttinari-base treebank. We present the underlying challenges and the solutions adopted in the annotation of arguments and predicates, and report the results of an inter-annotator agreement study. Our goal is twofold: to present the Porttinari-base Propbank as a resource for NLP and, secondarily, to provoke reflections on the dialogues between NLP and linguistic theories.

Keywords

semantic roles; semantic annotation; syntactic dependencies; semantic dependencies; universal dependencies

1. Introdução

A anotação de papéis semânticos é uma das formas utilizadas pelo Processamento de Linguagem Natural (PLN) para representar computacionalmente informação semântica de frases. Papéis semânticos classificam e relacionam as entidades participantes de um evento – em geral um verbo. Ao indicarem *quem* fez *o quê* para *quem*, *onde*, *quando*, *como*, *por que*, *para que*, *com que*, *com quem* etc, estruturam a informação linguística de maneira explícita e interpretável.

PropBanks (Bancos de Proposições) são corpora que contêm a anotação de papéis semânticos. No PropBank, o conceito de proposição é tomado da Gramática de Casos de Fillmore (1968): uma proposição é um conjunto de relações entre nomes e verbos, sem informação relativa a tempo, modo, aspecto, negação ou modificadores modais. Do lado computacional, um PropBank é um recurso criado com a intenção de servir para o aprendizado automático da informação semântica que codifica.

Historicamente, papéis semânticos guardam uma relação estreita (ou, de dependência) com a sintaxe, sendo descritos pelos estudos linguísticos como elementos da interface sintaxe — semântica (Palmer et al., 2005; Levin, 1993; Levin & Rappaport Hovav, 2005). Daí que, tradicionalmente, anotação de papéis semânticos seja feita sobre frases já analisadas sintaticamente, e quanto maior a qualidade da análise subjacente, mais facilitado será o trabalho humano de anotação dos papéis semânticos.

Neste artigo, detalhamos o processo de criação do Porttinari-base Propbank (PBP), que corresponde à adição de uma camada de papéis semânticos ao Porttinari-base, subcorpus do treebank Porttinari (Pardo et al., 2021) que foi intensa e cuidadosamente revisto no que se

*Este artigo é uma versão significativamente ampliada de Freitas & Pardo (2024), incluindo diversos dados e análises novos.

DOI: 10.21814/lm.17.2.480

refere à anotação de dependências sintáticas. O Porttinari (cujo nome vem de *PORTuguese Treebank*) é um treebank multigênero anotado conforme a abordagem *Universal Dependencies* (UD) (de Marneffe et al., 2021), e um dos desdobramentos do projeto POeTiSA[1] (*POrtuguese processing – Towards Syntactic Analysis and parsing*). O POeTiSA tem como um de seus objetivos a ampliação de recursos linguístico-computacionais baseados na sintaxe, e o PBP é um dos recursos criados no âmbito do projeto: no ambiente dedicado à sintaxe, a anotação de papéis semânticos se apresenta como um caminho para o significado (Duran et al., 2023). Como ilustração, a Figura 1 traz a representação, conforme a anotação codificada no Porttinari-base Propbank, da frase "*Carlos andou 150 km a pé para chegar ao México após ser ameaçado por bandidos.*"

O primeiro propbank foi criado para a língua inglesa em 2005, tendo como motivação principal ir além das análises sintáticas produzidas por parsers sintáticos, que estariam "muito longe de representar o significado completo das frases analisadas." (Palmer et al., 2005, p. 71). No entanto, se em 2005 o aprendizado de máquina (majoritariamente supervisionado) era uma abordagem promissora para o PLN, duas décadas depois o cenário é outro, com a onipresença de arquiteturas neurais e de grandes modelos de língua, treinados normalmente de maneira auto-supervisionada com grandes quantidades de dados, sendo utilizados para os mais variados fins. E, apesar da ampla disseminação desses métodos em PLN/Inteligência Artificial, a falta de transparência e de interpretabilidade são um gargalo a ser resolvido, e uma tendência atual é investigar formas de combinar representações explícitas de conhecimento e arquiteturas neurais. A possibilidade de criar grafos de conhecimento a partir da anotação de papéis semânticos (Mohebbi et al., 2022) torna este tipo de representação linguística relevante para a investigação acerca de abordagens neuro-simbólicas, com a articulação entre fontes de conhecimento estruturado e grandes modelos de língua (Dong, 2023). Por outro lado, são igualmente conhecidos os altos custos envolvidos no treinamento dos grandes modelos de língua, fazendo com que abordagens computacionalmente menos custosas e semanticamente informativas, como os papéis semânticos, se mantenham como alternativas viáveis para a representação do significado no PLN.

Desde o surgimento do primeiro propbank, recursos similares para diferentes línguas têm sido produzidos, em maior ou menor alinhamento com o recurso original. Um dos mais utilizados datasets com anotação de papéis semânticos ao estilo PropBank é o Ontonotes (Hovy et al., 2006), que contém dados para as línguas inglesa, chinesa e árabe. Outros datasets de amplo uso que contêm anotação de papéis semânticos ao estilo PropBank são o material do CoNLL2005 (Carreras & Màrquez, 2005), do CoNLL2009 (Hajič et al., 2009), e do CoNLL2012 (Pradhan et al., 2012). Além disso, associados ao projeto PropBank original estão ainda treebanks para o hindi, chinês, árabe, finlandês, basco, turco e para a língua portuguesa (Duran & Aluísio, 2011). Desde 2017, o estilo PropBank também está alinhado ao projeto *Universal PropBank*[2].

1.1. Papéis semânticos e PropBanks de língua portuguesa

Em uma abordagem baseada em regras, Bick (Bick, 2007) propõe um anotador de papéis temáticos para o português, e relata que o sistema consegue um F-score de 88.6%. A língua portuguesa conta também com o CINTIL-PropBank (Branco et al., 2012) criado de maneira semiautomática e com um conjunto de papéis semânticos que é uma adaptação dos argumentos numerados de Palmer et al. (2005). Também desde 2012 está disponível à comunidade de língua portuguesa o PropBank-Br (Duran & Aluísio, 2011), que contém papéis semânticos ao estilo PropBank anotados sobre a porção brasileira do treebank Bosque, em sua versão de sintaxe de constituintes disponibilizada pela Linguateca. Pouco tempo depois, foi criado o Propbank.br v2, tendo em vista a criação de um material maior e mais lexicalmente diversificado, dedicado especificamente ao aprendizado de máquina (Duran et al., 2014; Hartmann et al., 2016). Diferentemente da versão anterior, este material foi construído sobre árvores sintáticas não revistas.

A criação do Porttinari-base Propbank se justifica por suas características: trata-se de um material construído sobre dependências sintáticas cuidadosamente revistas, com textos contemporâneos vindos do corpus Porttinari, e com mais instâncias anotadas. Até o momento, o maior propbank disponível para o português continha cerca de 7 mil classes anotadas; o PBP contém mais de 45 mil. No entanto, apesar de seu caráter de novidade, o material aqui apresentado herda

[1] https://sites.google.com/icmc.usp.br/poetisa

[2] https://universalpropositions.github.io/

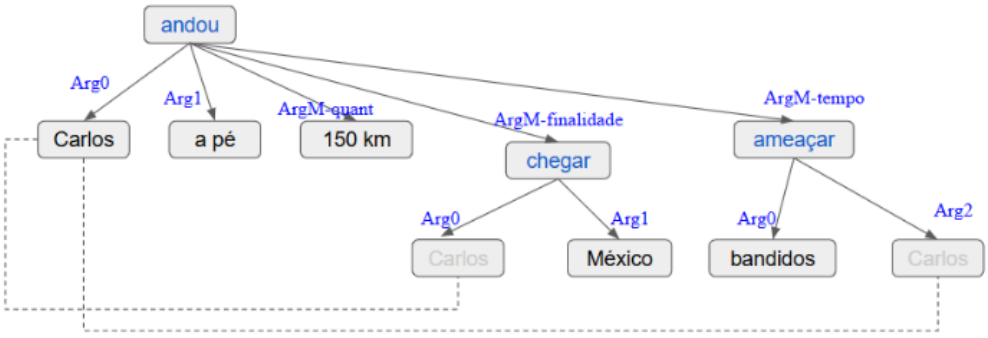

Figura 1: Representação da frase "Carlos andou 150 km a pé para chegar ao México após ser ameaçado por bandidos" conforme a anotação PropBank codificada no PBP

decisões e recursos criados ao longo da elaboração dos propbanks anteriores produzidos no Brasil, e sua construção teria sido muito mais complexa sem a existência de seus antecessores.

O restante do artigo se organiza da seguinte maneira: na Seção 2, nos aprofundamos no conceito de papéis semânticos e propbanks, destacando tanto os vínculos com teorias linguísticas quanto sua capacidade de generalização no PLN. Na Seção 3, apresentamos o Porttinari-base PropBank; e na Seção 4 detalhamos a metodologia utilizada na sua construção. Na Seção 5, apresentamos duas versões do corpus, e na Seção 6 relatamos os resultados de um estudo sobre concordâncias e discordâncias em anotações semânticas. Na Seção 7, apresentamos e analisamos os resultados da anotação de papéis semânticos. Na Seção 8, à luz dos resultados analisados, trazemos reflexões linguístico-computacionais motivadas pela construção deste tipo de recurso, e finalmente, na Seção 9, tecemos nossas considerações finais.

2. PropBanks e papéis semânticos

Papéis semânticos, também comumente chamados de papéis temáticos, papéis theta, casos semânticos e relações temáticas na literatura especializada, referem-se à maneira pela qual uma entidade está (semanticamente) envolvida em uma proposição. Associada à diversidade de termos, quantidades e tipos de propostas, a vasta literatura linguística sobre o tema traz algumas convergências, como a presença de elementos como "agente", "tema", "experienciador", "beneficiário", "instrumento", "localidade" e "objetivo" (Saeed, 2007).

O PropBank adota o termo "papéis semânticos", e contorna a diversidade mencionada com a utilização de i) argumentos numerados, que vão de Arg0 a Arg5 e consistem em um pequeno conjunto de etiquetas abstratas; e ii) argumentos modificadores (ArgM), um conjunto mais amplo de argumentos. A Figura 1 ilustra como os argumentos são anotados, onde se pode ver os dois tipos de argumentos.

A diferença entre argumentos numerados e argumentos modificadores está na natureza da relação sintática que o argumento mantém com o verbo: os argumentos numerados referem-se a elementos que, ou são exigidos pela valência (ou estrutura argumental[3]) de um predicado ou, se não exigidos, são aqueles que ocorrem com uma alta frequência no uso comum (Bonial et al., 2015). Já os argumentos modificadores (ArgM) são considerados opcionais. A distinção entre argumentos numerados e não numerados também é motivada pela assistematicidade semântica dos papéis com relação aos verbos, e por isso argumentos numerados são específicos de cada verbo (o Arg0 do verbo "abrir" em *abrir a porta* é "quem abre" e o Arg0 de "afastar" é "causador do afastamento"). Os ArgM, por outro lado, têm uma semântica específica e previsível, indicada pelo nome da etiqueta (ArgM-tmp para "tempo", ArgM-loc para "local", ArgM-prp para "finalidade" etc), e podem se associar a qualquer verbo.

Além de anotações ao estilo PropBank, que se centram no verbo e na análise sintática para seguir rumo à semântica, anotações conforme o projeto FrameNet (Baker et al. (1998) e, para a língua portuguesa (Salomão et al., 2013)[4] também são usadas no PLN para codificar simbolicamente a semântica das frases. Diferentemente das anotações ao estilo PropBank, que são genéricas com argumentos numerados e motiva-

[3]Podemos entender a estrutura argumental de um verbo como o número e o tipo de argumentos que o verbo exige para formar uma sentença gramaticalmente correta.

[4]Tanto a FrameNet quanto a FrameNet Brasil são projetos bastante ativos, conforme suas respectivas páginas: https://framenet.icsi.berkeley.edu/ e https://www2.ufjf.br/framenetbr/

das pela sintaxe, a anotação ao estilo FrameNet é centrada na noção *semântica* de *frame* (mais precisamente, na teoria da Semântica de Frames), sendo semanticamente mais detalhada. Ambos os recursos partem de ideias do linguista Charles Fillmore e tomaram rumos distintos ao longo do tempo, mas podem ser considerados membros de uma mesma família de recursos léxico-computacionais, como evidenciado pelo projeto SemLink (Stowe et al., 2021).

Na anotação PropBank, a semântica dos argumentos numerados é "revelada" com o alinhamento entre a anotação do corpus e o recurso lexical associado, que contém os chamados *Frames*[5] — em nosso caso, dispomos do Verbo-Brasil (Duran & Aluísio, 2015)[6]. Assim, a frase (1), segundo a anotação de um PropBank, recebe as etiquetas conforme os *frames* de *fechar.04* e *deixar.05* e é anotada como (2), levando a uma representação semântica como a indicada em (3) após a consulta ao Verbo-Brasil. Como a anotação é feita sobre dependências sintáticas e tem explicitação de sujeitos omitidos, temos um grafo de dependências semânticas, como na Figura 2.

1. Janot fechou 159 acordos de colaboração, deixando para Dodge uma série de inquéritos em curso.
2. Janot[Arg0] **fechou** 159 acordos[Arg1] de colaboração, **deixando** para Dodge[Arg2] uma série[Arg1] de inquéritos em curso.
3. Janot[agente] **fechou** 159 acordos [compromisso firmado] de colaboração, **deixando** para Dodge[beneficiário] uma série[bem legado] de inquéritos em curso.

Um PropBank não é apenas um corpus anotado, mas a associação entre um corpus anotado e um léxico, que indica como os argumentos devem ser anotados. Por isso, embora o corpus seja a parte visível do PropBank, sua existência depende do léxico, que indica como classificar os argumentos numerados, uma vez que estes têm sentido variável. A Tabela 1, adaptada de Bonial et al. (2018), indica os papéis *mais tipicamente* associados a cada argumento numerado.

É justamente a pouca regularidade entre o significado de um argumento numerado e o seu número que torna a anotação de papéis semânticos difícil de ser generalizada — por quem anota e pelos algoritmos. O papel de *instrumento*, por exemplo, pode ser tanto Arg2 do verbo *acalmar* na frase (1) como Arg3 do verbo *cortar* na frase (2). Uma ideia como *modo ou maneira* pode ser expressa como um argumento numerado do verbo *andar* da frase (3), ou como um argumento modificador do verbo *voar* na frase (4). A ideia de *paciente* corresponde, na maioria das vezes, a Arg1, mas em *Carlos foi ameaçado...* (Figura 1), *Carlos* é Arg2. Assim, cabe à anotação apenas codificar o que está especificado no léxico, que cristaliza as decisões relativas à numeração de cada argumento.

1. O presidente **acalmou** o ânimo dos correligionários com um discurso apaziguador [Arg2].
2. O vizinho que **cortou** a corda com facão [Arg3].
3. Carlos **andou** a pé [Arg1]
4. Eu **voei** de jatinho [ArgM-mnr]

Um dos motivos da variação semântica na atribuição de argumentos está relacionado às origens linguísticas do PropBank[7]. Ao partir da estrutura argumental dos verbos para atribuir os papéis semânticos — e não da semântica dos argumentos, por exemplo, como acontece na anotação de entidades mencionadas —, o resultado é uma classificação que, se por um lado é homogênea do ponto de vista do comportamento sintático, e permite investigar propriedades sintáticas e correlações entre sintaxe e semântica, por outro lado é heterogênea do ponto de vista da semântica dos argumentos, o que torna a anotação mais dependente do léxico que da interpretação da frase.

[5] Como informam Palmer et al. (2005, p. 8), não devemos confundir os "*frames* semânticos" da FrameNet de Fillmore com os "*frames* sintáticos" do PropBank, que se referem a realizações sintáticas.

[6] O Verbo-Brasil foi construído combinando dados do recurso lexical associado ao PropBank de língua inglesa, com o qual está alinhado, e dados do corpus PropBank-Br, de maneira a cobrir sentidos não presentes no recurso em inglês. No PropBank-Br, que não dispunha de recurso lexical associado, a anotação de sentidos verbais foi realizada manualmente e de maneira concomitante à anotação dos argumentos, com base nas diretivas do PropBank original. A estratégia de construção do Verbo-Brasil está descrita em (Duran & Aluísio, 2015)

[7] Originalmente, o PropBank também foi motivado por interesses linguísticos. "*Our objective with the Proposition Bank is not a theoretical account of how and why syntactic alternation takes place, but rather to provide a useful level of representation and a corpus of annotated data to enable empirical study of these issues.*" (Palmer et al., 2005, p. 74). E, algumas páginas adiante: "*The PropBank project and the FrameNet project (...) share the goal of documenting the syntactic realization of arguments of the predicates of the general English lexicon by annotating a corpus with semantic roles.*" (Palmer et al., 2005, p. 88). Do mesmo modo, Charles Fillmore, no texto que motiva o interesse moderno em listas de papéis semânticos, afirma: "O presente ensaio pretende ser uma contribuição ao estudo de universais sintáticos formais e materiais" (Fillmore, 1968, p. 278).

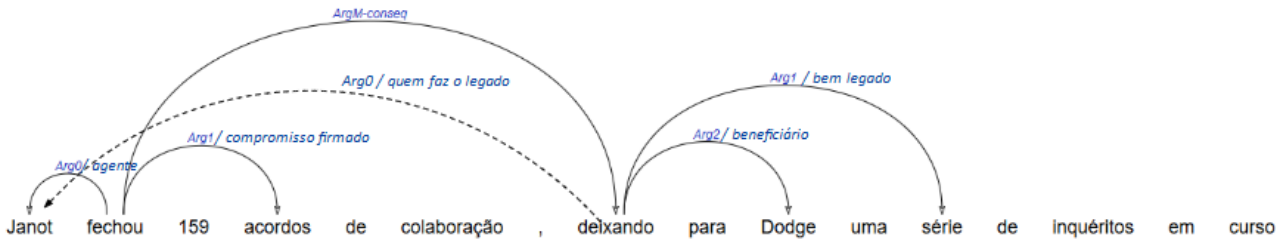

Figura 2: Representação do estilo PropBank no formato de dependências semânticas

Argumento numerado	**Papel Temático típico**
Arg0	Agente
Arg1	Paciente
Arg2	Beneficiário, Instrumento, Atributo, Estado Final
Arg3	Ponto de Partida, Beneficiário, Instrumento, Atributo
Arg4	Ponto Final
Arg5	Direção

Tabela 1: Mapeamento de argumentos gerais (fonte: Bonial et al. (2018))

No PropBank, a informação relativa aos argumentos dos verbos se baseia na classificação linguística proposta por Levin (1993), segundo a qual verbos que participam dos mesmos tipos de alternância sintática, isto é, verbos que têm o mesmo comportamento sintático, também compartilhariam aspectos do significado, e comporiam classes semanticamente coerentes. A classificação de Levin obedece primeiramente, portanto, critérios sintáticos, partindo da hipótese de que o comportamento sintático seria reflexo de aspectos do significado. Segundo Levin (1993), a opção de privilegiar a sintaxe se deve à dificuldade de identificação de significados com base apenas na intuição ou na definição de dicionários[8].

Verbos com sentidos parecidos ou simétricos — como os pares *alertar/avisar* (1), *causar/provocar* (2) e *assustar/ameaçar* (3) — porque têm comportamentos sintáticos diferentes, pertencem a grupos semânticos diferentes, e por isso podem atribuir papéis semânticos de maneira distinta. Em (1), se usamos o verbo *alertar*, o *aviso* (ou *alerta*) é codificado como Arg2; se usamos o verbo *avisar*, o mesmo *aviso* é codificado como Arg1. Em (2), se usamos o verbo *provocar*, o elemento *país* recebe a etiqueta Arg2 que, no caso de *provocar*, sinaliza a presença de um *beneficiário*. Entretanto, se a frase tiver sido construída com o verbo *causar*, não temos o que fazer com *país*, pois *causar* não prevê um espaço argumental para elementos do tipo *beneficiário*. Em consequência, precisamos escolher alguma das etiquetas dos argumentos modificadores (na anotação, foi escolhida *ArgM-adv*). Em (3), se usamos o verbo *ameaçar*, *Carlos* é Arg2, mas, se usamos o verbo *assustar*, *Carlos* é Arg1, uma vez que *assustar* exige 2 argumentos (*entidade que assusta* (Arg0) e *entidade assustada* (Arg1)), mas *ameaçar* exige 3 argumentos (*entidade que ameaça* (Arg0), *entidade ameaçada* (Arg2) e a *ameaça* propriamente (Arg1)).

Vale notar que esta "arbitrariedade" na atribuição de papéis já havia sido notada por Duran & Aluísio (2011, p. 1864), que apontam para a diferença na atribuição de papéis entre os verbos *renegociar* e *negociar*: em *renegociar*, o que é negociado é Arg1 e a "outra parte" envolvida no acordo é Arg2; já em *negociar*, os elementos se invertem: o que é negociado é Arg2, e a outra parte envolvida no acordo é Arg1.

1. (...) a CPPU enviara uma notificação à prefeitura (...) [**alertando/avisando**] que os infratores estavam sujeitos às penas previstas
2. (...) as concessões feitas a grupos que [**causaram/provocaram**] tanto sofrimento ao país.
3. Carlos foi [**ameaçado/assustado**] por bandidos

[8] "*The availability of this technique for investigating word meaning is important since it can be quite difficult to pin down the meanings of words using introspection alone. For instance, dictionaries provide rather different definitions of the sense of the verb whistle found in the context (...). They seem unsure whether to treat this sense as involving a verb of sound or a verb of motion.*" (Levin, 1993, p. 14-15)

Nos estudos linguísticos, o interesse na estrutura argumental dos verbos se relaciona ao interesse na caracterização do léxico mental. Um argumento nuclear/obrigatório de um item lexical é uma propriedade estável deste item no léxico mental[9]. No contexto de um propbank, o léxico corresponde à informação codificada nos *Frames* e, no PBP, às informações codificadas no Verbo-Brasil.

No PLN, PropBanks e VerbNets (Kipper et al., 2008) são as formas mais popularizadas de anotar o que estamos chamando de "papéis semânticos". Uma anotação ao estilo VerbNet parte da taxonomia da VerbNet, um léxico de verbos organizado hierarquicamente, independente de domínio e de ampla cobertura.

PropBanks e VerbNets compartilham uma série de pressupostos – como as classes de Levin (1993) como ponto de partida — , mas divergem quanto ao nível de granularidade. O PropBank é menos granular, com 6 etiquetas (considerando apenas os argumentos numerados). A VerbNet traz 23 etiquetas (considerando apenas os argumentos "instanciados", que seriam correspondentes aos numerados no PropBank). Além disso, a VerbNet não lista/descreve os elementos "opcionais" associados ao significado dos verbos (que seriam os argumentos modificadores do PropBank) devido ao seu caráter acidental e uso idiossincrático. O PropBank compartilha o mesmo pressuposto, mas o aplica de maneira moderada: distingue os tipos de argumentos, mas não deixa de anotá-los.

A motivação inicial para o conjunto restrito de papéis numerados no PropBank (em contraste ao conjunto da VerbNet) é facilitar a generalização no aprendizado de máquina. No entanto, trabalhos que comparam a capacidade de generalização no aprendizado de máquina do estilo PropBank e do estilo VerbNet de anotar papéis semânticos indicam que o segundo generaliza melhor quando se trata de classes/instâncias novas, e que o aprendizado, para argumentos numerados que não sejam Arg0 e Arg1, é limitado (Merlo & Van Der Plas, 2009; Yi et al., 2007; Bonial et al., 2018; Gung & Palmer, 2021). Talvez por isso, artigos mais recentes sobre o PropBank indiquem que a utilização de um conjunto de etiquetas numeradas tão enxuto, e não de etiquetas que tenham um rótulo mais descritivo, como propõe a VerbNet, é motivada pela dificuldade de "um conjunto universal de papéis semânticos ou temáticos capaz de cobrir todos os tipos de predicados" (Bonial et al., 2018, p.740).

Wang et al. (2022) fazem um levantamento dos resultados mais recentes para a anotação de papéis semânticos, e mesmo com a utilização das técnicas mais avançadas, os melhores resultados giram em torno de 85% no cenário mais realista, no qual o material é avaliado em um outro corpus, e não na partição de teste de um mesmo corpus (cenário *out of domain*).

3. O Porttinari-base PropBank

Nesta seção, trazemos uma visão geral do Porttinari-base PropBank, indicando características relevantes da anotação e do conjunto de etiquetas usadas e algumas especificidades decorrentes do alinhamento com a análise sintática. O material é um dos desdobramentos do *Porttinari-base treebank*, composto por 168 mil tokens distribuídos em 8.418 frases de textos jornalísticos.

3.1. Características gerais

A anotação da camada de papéis semânticos no Porttinari-base seguiu de perto as diretivas de anotação linguística do PropBank.br v2 (Duran, 2014), que foram atualizadas e ampliadas ao longo do projeto (Duran & Freitas, 2024). No entanto, há um ponto em que a anotação do PBP difere da anotação do PropBank.br: no PropBank.br, a anotação de papéis semânticos deveria seguir a análise sintática subjacente. Esta recomendação não foi levada em conta na anotação PBP, anotado sobre árvores analisadas sintaticamente conforme a teoria *Universal Dependencies* (UD) (de Marneffe et al., 2021). Justificamos esta opção com três argumentos:

1. A decisão torna a camada de papéis semânticos dependente das decisões de UD, o que tem como consequências i) a dificuldade de verificar a interferência de diferentes representações sintáticas no aprendizado de papéis semânticos; e ii) a impossibilidade de atribuir papéis semânticos caso a análise sintática, por princípio ou por lapso de anotação, esteja desalinhada com estrutura argumental do verbo descrita no Verbo-Brasil (Duran & Aluísio, 2015).
2. Não há impedimento formal para a anotação de papéis semânticos de maneira independente da camada sintática. Além disso, é possível recuperar as ocorrências em que um elemento foi tratado como argumento pela camada de

[9] Vale lembrar que a composição do léxico mental, bem como a própria existência de um léxico mental, são alvo de amplo debate dentro dos estudos linguísticos (veja-se Peeters (2000)).

papéis semânticos, mas não o foi pela camada sintática, para posterior investigação.

3. Este desalinhamento entre a anotação do treebank subjacente e a anotação de papéis semânticos já acontece no PropBank original.[10]

Assim, o primeiro compromisso do Porttinari-base PropBank é com seu léxico, o Verbo-Brasil (Duran & Aluísio, 2015). Ao longo do projeto, esta independência moderada entre as camadas de anotação se mostrou proveitosa: na camada sintática, o Porttinari anota como auxiliares apenas auxiliares de tempo e de voz. Recentemente, as diretivas da teoria UD para a língua portuguesa passaram a sinalizar que auxiliares de modo e de aspecto também deveriam ser incluídos na lista de verbos auxiliares[11]. Uma vez que, na camada propbank, verbos como *dever*, *poder*, *terminar* e *andar*, entre outros, já figuravam como auxiliares (porque assim estão codificados no Verbo-Brasil), nada precisou ser feito[12].

O Porttinari-base PropBank foi construído sobre a versão 1.0 do treebank Porttinari-base. Anotado a partir de um treebank de dependências, o PBP está no formato de dependências semânticas, no qual cada argumento é um único token, em um arquivo em formato *conllu*[13]. Seguindo esse formato, a anotação foi feita originalmente na coluna 9 ("deps") para os predicados verbais e na coluna 10 ("misc") para argumentos semânticos e seus núcleos. Foram anotados elementos implícitos (como sujeitos omitidos) e explícitos, como vimos nas Figuras 1 e 2. A Figura 3 traz a codificação da frase "*Janot fechou 159 acordos* (...)" no formato *conllu*. Para facilitar a leitura, omitimos os conteúdos das colunas 3 a 6 e renomeamos na figura as colunas 9 e 10 para maior clareza sobre o conteúdo que contêm.

3.2. Conjunto de etiquetas

O Porttinari-base Propbank utiliza 26 etiquetas, algumas delas específicas para a língua portuguesa — como as relacionadas aos verbos auxiliares e ao pronome "se", presentes desde o PropBank.br —, e outras que especificam as classes já existentes no PropBank original. Estas etiquetas, presentes apenas no PBP, foram criadas porque possuem alguma marca formal que torna sua identificação razoavelmente simples, e porque se referem a relações semânticas cujo significado pode ser relevante em tarefas subsequentes de PLN. Apresentamos a seguir cada uma das etiquetas específicas do PBP.

ArgM-cond: usadas para proposições que indicam condição. Originalmente ficariam sob o rótulo de ArgM-adv, mas, como estes casos são formalmente marcados em português com as conjunções "se" e "caso" e pelo tempo verbal, e são um grupo semanticamente relevante, decidimos separá-los de ArgM-adv.

- Na prioridade , caso cancelem[ArgM-cond] a corrida , eles são **colocados** no fim da fila.

ArgM-conseq: usadas para proposições que indicam consequências. Em termos tradicionais, são as orações adverbiais consecutivas. No PropBank original, também são anotadas como ArgM-adv, porque se referem a todo o conteúdo da oração, e não apenas ao verbo. No entanto, a ideia de consequência de uma ação parece interessante de ser codificada de maneira independente dos demais casos de ArgM-adv.

- Capítulos de narrativa fantástica **ligam** Evie a animais de a floresta , aproximando[ArgM-conseq] o personagem da Eva bíblica.

ArgM-comp: construções envolvendo comparações são anotadas no PropBank original como ArgM-ext, etiqueta aplicada a quantidades. Na anotação UD do Porttinari-base, construções comparativas são anotadas como orações adverbiais (advcl), assumindo que contêm um verbo implícito. Para lidar com esses casos, anotamos os advérbios (*mais*, *menos*, *melhor*, *pior*) como ArgM-ext e criamos a etiqueta ArgM-comp para as orações adverbiais. Apesar de ser uma etiqueta nova, a decisão facilita a identificação desses casos sem confundi-los com casos genuínos de quantidade — significado principal da etiqueta ArgM-ext — e não leva à falsa leitura de que o verbo conta com dois ArgM-ext.

- Nada **ajuda** mais[ArgM-ext] a divulgar a ciência do que um planetário[ArgM-comp].

ArgM-src: usada para indicar a fonte da informação (*source*) de uma proposição. De acordo com as diretrizes do Propbank original, este tipo de conteúdo é anotado no

[10] Conforme Palmer et al. (2005, p.82): "*Consider a sentence such as* (...). *The Penn Treebank's analysis assigns a single sentential (S) constituent to the entire string* (...), *making it a single syntactic argument (...). In the PropBank annotation, we split the sentential complement into two semantic roles* (...)."

[11] https://universaldependencies.org/pt/dep/aux_.htm

[12] Espera-se que a próxima versão do Porttinari-base já tenha esta atualização quanto aos auxiliares, diminuindo o desalinhamento entre as camadas sintática e semântica.

[13] http://universaldependencies.org/format.html

ID	FORM	LEMMA	UPOS	XPOS	FEAT	HEAD	DEPREL	FRAME	ARG:HEAD
1	Janot	_	_	_	_	2	nsubj	_	Arg0:2|Arg0:8
2	fechou	_	_	_	_	0	root	fechar.04	_
3	159	_	_	_	_	4	nummod	_	_
4	acordos	_	_	_	_	2	obj	_	Arg1:2
5	de	_	_	_	_	6	case	_	_
6	colaboração	_	_	_	_	4	nmod	_	_
7	,	_	_	_	_	8	punct	_	_
8	deixando	_	_	_	_	2	advcl	_	ArgM-conseq:2
9	para	_	_	_	_	10	case	_	_
10	Dodge	_	_	_	_	8	obl	_	Arg2:8
11	uma	_	_	_	_	12	det	_	_
12	série	_	_	_	_	8	obj	_	Arg1:8
13	de	_	_	_	_	14	case	_	_
14	inquéritos	_	_	_	_	12	nmod	_	_
15	em	_	_	_	_	16	case	_	_
16	curso	_	_	_	_	14	nmod	_	_

Figura 3: Anotação de papéis semânticos no arquivo conllu

Penn treebank como PRN (*parenthetical*), e não deve ser anotado na camada de papéis semânticos. Na análise sintática do Porttinari, estas estruturas são anotadas como *complemento oblíquo do verbo* (obl), e como tal devem receber algum papel semântico. Considerando a relevância argumentativa/retórica destas construções no que se refere à detecção da fonte da informação/proposição veiculada, decidimos anotá-las. Além dos casos prototípicos (1), a etiqueta também é usada em contextos menos óbvios, em que não se trata exatamente da fonte da informação, mas fonte de opinião relativa à proposição veiculada (2).

- (1) De acordo[ArgM-src] com a polícia, **trata-se** de uma "prisão significativa" para as investigações.
- (2) Pergunto-me por que tudo, para o articulista[ArgM-src], tem que se **resumir** a uma dicotomia PT-PSDB?

Auxiliares — ArgM-mod, ArgM-asp, Argm-tml, ArgM-pass: a camada sintática do Porttinari-base é bastante econômica no que se refere à classe dos verbos auxiliares[14], e apenas considera auxiliares de tempo composto e voz passiva. O PropBank original utiliza a etiqueta ArgM-mod para os auxiliares modais, e a versão anterior do PropBank.br, por sua vez, já utilizava, além de ArgM-mod, etiquetas ArgM-asp, Argm-tml e ArgM-pass para auxiliares de aspecto, tempo e voz. Na anotação do PBP, mantivemos a decisão do Propbank.br, e em Duran & Freitas (2024) indicamos os verbos e contextos em que essas etiquetas são utilizadas.

[14] A economia é motivada por diretrizes gerais que apenas recentemente foram reformuladas, como já comentamos.

3.3. Interferências da anotação sintática na anotação de papéis semânticos

Desalinhamentos quanto à segmentação de argumentos na anotação de papéis semânticos já foram referidos na literatura do PropBank original, ainda que com pouca ênfase (Palmer et al., 2005, p. 82). Na anotação do Porttinari-base PropBank, a divergência entre as segmentações sintático-semânticas (indicadas no Verbo-Brasil) e sintáticas derivam de quatro cenários:

- possibilidade de análises/segmentações sintáticas divergentes e legítimas, resultado da presença de substantivos predicadores/nominalizações;
- impossibilidade de cruzamento de arcos sintáticos, um direcionamento seguido pelo Porttinari-base, ainda que não seja uma exigência da teoria UD;
- economia do Porttinari-base no que se refere aos verbos auxiliares;
- tratamento especial dado aos verbos de cópula em UD.

Consideremos as frases (1) a (5) ao final dessa subseção. Um exemplo do primeiro cenário é a frase (1), na qual "sobre seu trabalho" foi considerado, na camada sintática, um modificador do substantivo *influência* (com leitura equivalente a "influenciar seu trabalho"), mas a estrutura argumental do verbo *exercer* no(s) léxico(s) indica a existência de três argumentos: alguém (Arg0)

exerce algo (Arg1) sobre algo ou alguém (Arg2). Dado o compromisso do PBP com o Verbo-Brasil, a análise anotada foi a indicada em (b).

As frases 2 e 3 ilustram o segundo cenário: a análise sintática do Porttinari-base não permite o cruzamento de arcos sintáticos, que resultariam em relações não-projetivas. Esta opção do Porttinari, se por um lado facilita a análise sintática automática e faz com que modelos treinados no Porttinari-base tenham um desempenho superior àqueles treinados em corpora em que esta não é uma restrição, por outro, leva à depreensão de argumentos que não correspondem ao que se esperaria da análise sintática convencional. Também em nome do não cruzamento de arcos, a segmentação feita pela análise sintática pode divergir da estrutura argumental do verbo indicada no Verbo-Brasil. Na frase 3, conforme a análise sintática, "Justiça" modifica o substantivo "processo", ou seja, não é um argumento do verbo "entrar", e "contra Grandes Irmãs" é modificador de "Justiça". No entanto, o *frame* de *entrar.04* (*submeter oficialmente (ação, pedido), apelar judicialmente*) prevê 4 argumentos: Arg0 (*iniciador de processo*); Arg1 (*coisa submetida (entrar com)*); Arg2 (*órgão que recebe a submissão*); Arg3 (*entidade contra a qual Arg1 é movido*). Assim, se "Justiça" e "Grandes Irmãs" são argumentos de *entrar.04*, deveriam estar vinculados ao verbo. Ao evitar o cruzamento de arcos envolvendo verbos, corremos o risco de codificar sintaticamente leituras com significados distantes daqueles convencionalizados ou de deixar de lado a estrutura argumental indicada no léxico, sobre o qual se sustenta o projeto de um Propbank. No PBP, uma vez que a anotação de papéis semânticos leva em conta o significado e está comprometida com o Verbo-Brasil, os "arcos semânticos" se cruzam.

A frase (4) exemplifica o terceiro ponto. Se verbos de significado lexical esvaziado (auxiliares) são tratados como verbos plenos, levarão a proposições indesejadas, uma vez que a relação entre núcleos e seus dependentes acontece de maneira diferente da esperada. Em (4), a opção (i) não é adequada para representar a frase.

A frase (5) exemplifica o último cenário. O fato de verbos de cópula não serem considerados verbos plenos em UD, mas auxiliares, significa que eles não atuam como núcleo de dependências sintáticas, e a relação sintática acontece entre dois nominais. Uma anotação propbank "ingênua" que decida anotar proposições de verbos de cópula seguindo as árvores sintáticas UD levará a uma representação como em 5(a), e não como em 5(b).

1. Segundo Thaler, Kahneman exerceu grande influência sobre seu trabalho

   (a) análise sintática: Segundo Thaler, Kahneman **exerceu** [grande influência sobre seu trabalho]

   (b) análise PBP: Segundo Thaler, Kahneman **exerceu** [grande influência] [sobre seu trabalho]

2. Mentir é *algo* que ela diz não **ensinar**

   (a) análise sintática: dizer algo[15]

   (b) análise PBP: ensinar algo[16]

3. Seus procuradores **entraram** com processos na Justiça contra cinco Grandes Irmãs: British Petroleum (...)

   (a) análise sintática: Seus procuradores **entraram** [com processos na Justiça] [contra cinco Grandes Irmãs]

   (b) análise PBP: Seus procuradores **entraram** [com processos] [na Justiça] [contra cinco Grandes Irmãs]

4. O time anda vencendo todos os jogos

   (a) análise sintática:

      i. time [Arg0] **anda** vencer[Arg1]

      ii. time [Arg0] **vencer** jogos [Arg1]

   (b) análise PBP: time [Arg0] anda[ArgM-aspecto] **vencer** jogos[Arg1]

5. Ela disse que o grupo está pronto

   (a) análise sintática: grupo [Arg?] **pronto** está [Arg?]

   (b) análise PBP: grupo [Arg1] **está** pronto [Arg2]

4. Metodologia

Todo o processo de anotação dos papéis semânticos foi feito com base em regras linguisticamente motivadas e de maneira não linear, seguindo o que foi feito por Freitas et al. (2023), para a anotação de entidades, e que por sua vez segue a metodologia de anotação semântica do projeto AC/DC (Santos & Mota, 2010). A ideia geral é aproveitar a anotação morfossintática e sintática já existente no corpus para criar regras que incluem outros tipos de anotação. Na criação de regras, utilizamos pistas lexicais e sintáticas

[15] Literalmente, "dizer *que*", que por sua vez tem o substantivo *algo* como antecedente.

[16] Literalmente, "ensinar *que*", que por sua vez tem o substantivo *algo* como antecedente.

derivadas de exemplos das diretivas preparadas para a anotação do PropBank.br v2, de exemplos de *frames* do Verbo-Brasil e de padrões observados no próprio corpus. Assim, humanos com conhecimento linguístico especializado analisam os dados, criam regras de anotação a partir deles, analisam os resultados da aplicação das regras (da anotação) e corrigem eventuais erros (caso a caso ou criando novas regras).

A anotação foi feita com a ferramenta ET (acrônimo de *Estação de Trabalho*) (de Souza & Freitas, 2021). Criada originalmente para permitir a busca, edição e avaliação de árvores sintáticas no formato *conllu*, a ferramenta foi aperfeiçoada ao longo do projeto para permitir a anotação, partindo do zero, de outros tipos de informação linguística.

A ferramenta traz uma típica interface de busca em corpus, com uma sintaxe de procura poderosa e visualização dos resultados como listas de distribuição ou como linhas de concordância — sendo ambas o ponto de partida para a construção das regras de anotação ou de correção. A ET permite anotar/corrigir por lote, criando regras que adicionam ou excluem etiquetas, e permite a edição caso a caso. Existe ainda o recurso de "filtro", que possibilita a seleção de frases/ocorrências (caso a caso ou com expressões regulares) e seu agrupamento em classes, que poderão ser tratadas separadamente, tanto por regras como caso a caso.

A anotação e revisão humanas com base em regras permitem avançar de maneira relativamente rápida (se comparada à anotação/revisão linear, caso a caso) e consistente: ganha-se em agilidade sem abrir mão da qualidade, uma vez que as condições sobre as quais as regras se aplicam, bem como o resultado da aplicação, passaram por inspeção humana.

As regras têm diferentes graus de generalização. Algumas, como regras para anotação de sujeitos de orações passivas, são de amplo alcance e anotam mais de 500 ocorrências. Por outro lado, há regras que se aplicam a menos de 10 casos. Uma vez que o objetivo do projeto era a construção de um propbank, e não a criação de um anotador com base em regras, casos com frequência até 4 ou 5 foram, em geral, tratados individualmente, uma vez que era mais rápido editá-los caso a caso do que criar uma regra para cada ocorrência.

A anotação dos *frames* dos verbos não foi o foco desta etapa do projeto, e privilegiou, de forma não exaustiva, verbos não monossêmicos. Os monossêmicos foram adicionados automaticamente, e atualmente 65% das ocorrências verbais têm informação de *frame*. Verbos ausentes do Verbo-Brasil, ou sentidos novos de formas verbais existentes, foram listados ao lado de exemplos, soluções provisórias e dúvidas, e foram encaminhadas para a equipe responsável pelo Verbo-Brasil. Este procedimento levou à documentação de cerca de 350 sentidos de verbos. No entanto, na busca por verbos similares para criação dos novos *frames*, nossa principal preocupação esteve na similaridade semântica, e não na verificação de contextos sintáticos, e, por isso, salientamos que são decisões provisórias.

A anotação dos argumentos foi feita em 3 fases, detalhadas a seguir:

- anotação de elementos explícitos;
- anotação de elementos implícitos (sobretudo identificação e classificação de sujeitos omitidos);
- aplicação de regras de validação para busca de erros formais e inconsistências.

4.1. Anotação de elementos explícitos: classificação

A anotação de elementos explícitos utilizou as seguintes estratégias:

1. identificação de padrões léxico-sintáticas;
2. identificação de padrões sintáticos simples;
3. anotação de argumentos numerados e regras complexas;
4. acabamentos;
5. validação inicial;

A anotação baseada em padrões léxico-sintáticos partiu de listas, descrições e padrões depreendidos dos exemplos contidos nas diretivas iniciais de anotação (Duran, 2014), que detalha os argumentos modificadores. A partir dos exemplos, criamos uma série de expressões de busca, que foram posteriormente transformadas em regras de anotação.

Padrões sintáticos simples são aqueles que envolvem argumentos cuja classificação é quase integralmente determinada pela sintaxe, como em frases na voz passiva, em que o sujeito paciente tende a ser Arg1 e o agente da passiva tende a ser Arg0. No entanto, esta anotação pode ser revista (e eventualmente corrigida — como fizemos), pois há casos em que o sujeito paciente não será Arg1, mas Arg2.

Em seguida, no terceiro momento, tratamos argumentos não cobertos pelas estratégias anteriores. De forma mais precisa, não estamos diante

de uma estratégia uniforme, mas de abordagens que recorrem ao Verbo-Brasil para dar conta dos casos não tratados nas etapas anteriores. Assim como nas diretivas, as frases de exemplo indicadas no léxico permitiram depreender padrões de anotação complexos, que poderiam estar associados a um *frame* verbal. Por exemplo, o *frame fazer.02* (reproduzido na Figura 4), que traz a ideia de fazer alguém fazer algo, se manifesta, na anotação UD, como uma estrutura que envolve a relação sintática xcomp. Analisamos o *frame* e os exemplos, "traduzimos" essas informações para a sintaxe UD – e para a sintaxe da ferramenta – e criamos uma regra de anotação (Figura 5), que anota em um único lance[17]:

- o token *xcomp* como *Arg2* na coluna *misc*;
- o token *obj* como *Arg1* na coluna *misc*;
- o token *nsubj* como *Arg0* na coluna *misc*;
- o token *fazer* como *fazer.02* na coluna *deps*.

Além disso, ao longo do projeto, foi construída uma interface alternativa para o Verbo-Brasil[18], que permite buscas não só pelos lemas, mas também pelo conteúdo dos seus vários campos do Verbo-Brasil. Por exemplo, podemos buscar por verbos que *não contêm* Arg0 em sua estrutura argumental, o que significa que, nesses verbos, o sujeito da voz ativa não será Arg0. No entanto, porque nem todas as formas verbais ou sentidos de verbos do Porttinari-base existem no Verbo-Brasil, esta estratégia, embora útil, tem alcance limitado.

O Verbo-Brasil também auxiliou a desambiguar padrões. Como já indicamos, argumentos com uma mesma forma e um mesmo sentido podem receber etiquetas diferentes conforme a estrutura argumental do verbo a que estão associados. Assim, enunciados com orações adverbiais (*advcl*) introduzidas pela preposição *para* e com verbo no infinitivo costumam indicar finalidade ou propósito, e são anotados como ArgM-prp (conforme exemplo 1 abaixo). No entanto, a ideia de propósito, caso seja considerada parte da estrutura argumental do verbo, como em *usar* (no exemplo 2), será um argumento numerado (em *usar*, Arg2 indica "propósito (*usar para*)").

1. Os rivais não **entravam** na área para ameaçar[ArgM-prp]
2. A ideia é **usar** a embalagem como diferencial para se apresentar[Arg2] aos potenciais clientes
3. (...) podem **criar** novos e mais fortes mecanismos para o país combater[ArgM-prp] a corrupção

O Porttinari-base tem 456 frases com a estrutura "para + infinitivo/advcl" — mais precisamente, 268 lemas verbais diferentes que ocorrem com esse padrão. Embora analisar cada um dos lemas seja uma alternativa quando se deseja um material de alta qualidade, uma consulta ao Verbo-Brasil otimiza o processo de criação de um material padrão-ouro. Pela interface, que admite expressões regulares, buscamos *frames* que contenham a palavra "finalidade" e variações, como "propósito", no atributo "Papéis", e indicamos que este papel deve estar associado a um argumento numerado. A busca retorna 23 *frames*/lemas diferentes. Com essa informação, fazemos uma busca, na ET, pelo padrão "para + infinitivo/advcl", especificando os lemas dos verbos. Ou seja, nesses casos teremos argumentos numerados, e em todos os demais casos (433 lemas verbais diferentes) estaremos diante de um ArgM-prp. Como é possível notar, este tipo de procedimento faz com que a anotação ganhe em qualidade/consistência e em facilidade/velocidade.

No entanto, nem sempre a distinção entre as classes de argumento é clara, como ilustra a frase (3), na qual *país* pode ser interpretado como beneficiário da criação (Arg3) ou como finalidade da criação: "para que o país combata a corrupção". A leitura codificada foi ArgM-prp, mas não nos parece que seja um caso óbvio. Este ponto será retomado na seção 6.2

Também experimentamos concatenar verbos e preposições por meio de padrões léxico-sintáticos. Por exemplo, buscamos argumentos de verbos associados a elementos *obl* introduzidos pela preposição *sobre* ainda não anotados, e analisamos a lista resultante. Neste caso, a busca devolveu 44 lemas diferentes distribuídos em 91 ocorrências. A análise da distribuição dos lemas associada à análise das linhas de concordância nos permite agrupar as ocorrências por tipo e criar regras para cada um dos casos:

- verbos em que obl é Arg1;
- verbos em obl é Arg2;
- verbos em obl é Arg3;
- verbos em obl é ArgM-loc;
- verbos em obl é outra classe.

[17] Este é um exemplo do que chamamos de regras complexas, que envolvem "mudança" de galho na árvore sintática.

[18] https://souelvis.dev/ui-verbobrasil/

Roleset id: fazer.02 , fazer alguém fazer uma ação, vcls: -, Mapeamento para o inglês: make.02

Roles:

Arg0: indutor (vnrole: -agent)

Arg1: agente e ação induzida (fazer alguém + infinitivo, fazer com que) (vnrole: -predicate)

Arg2: ação induzida, quando separada de arg1

Figura 4: *Frame* de *fazer.02* no Verbo-Brasil

```
if regex("VERB", token.head_token.upos) and regex("fazer", token.head_token.lemma) and regex("xcomp", token.deprel):
        for _nsubj in sentence.tokens:
                if _nsubj.dephead == token.head_token.id and _nsubj.deprel == "nsubj":
                        for _obj in sentence.tokens:
                                if _obj.dephead == token.head_token.id and _obj.deprel == "obj":
                                        token.misc = "Arg2"
                                        _nsubj.misc = "Arg0"
                                        _obj.misc = "Arg1"
                                        token.head_token.deps = "fazer.02"
```

Figura 5: Exemplo de regra de anotação. A regra sinaliza que SE o núcleo (head) do token em questão é um verbo, esse mesmo núcleo (head) tem o lemma *fazer*, e o token em questão é um xcomp e SE existe um token nsubj dependente deste mesmo head (lema *fazer*), e SE existe um token obj dependente deste mesmo head (lema *fazer*), ENTÃO o token em questão receberá a etiqueta *Arg2*, o nsubj receberá *Arg0*, o obj receberá *Arg1* e o head receberá *fazer.02*

Esta separação inicial dos casos é feita utilizando o recurso de "filtro" da ferramenta. Deste modo, podemos anotar cada grupo por lote, usando regras. No entanto, esta abordagem é pouco produtiva para preposições muito frequentes e com sentido esvaziado, como *em* ou *de*. Por outro lado, esse tipo de exploração permite detectar pistas que podem levar à depreensão de novos padrões.

Por fim, na fase de acabamentos, buscamos tokens dependentes de verbo que não haviam recebido papel semântico. Utilizando novamente o recurso de filtro, que viabiliza o agrupamento de frases, classificamos os tokens conforme o tipo de argumento que deveriam receber. Após a distribuição dos casos nos filtros/grupos adequados (por exemplo, grupo 1: *ArgM-tmp*; grupo 2: *ArgM-loc* etc), cada grupo é tratado individualmente, quer por regras, quer pela edição direta da frase. A Figura 6 ilustra a tela da ferramenta com a utilização dos filtros. A figura também mostra que o verbo (em vermelho) e o argumento alvo da busca (em azul) são destacados nas linhas de concordância, o que torna a leitura humana mais ágil.

Ao final do processo, na validação, buscamos candidatos a erros, como argumentos numerados idênticos (por exemplo, dois Arg0) associados a um mesmo verbo. Corrigimos todos os casos, e passamos à fase de anotação dos elementos implícitos.

4.2. Anotação de elementos implícitos: identificação e classificação

A anotação de elementos implícitos envolve, além da classificação dos argumentos, a sua identificação. Assim como na anotação de elementos explícitos, o trabalho também foi feito em etapas, caracterizadas desta vez não pelo tipo de abordagem, mas pelos fenômenos linguísticos tratados. A maior parte da anotação foi feita com regras de propagação de sujeitos associadas a certas estruturas sintáticas, sempre seguidas de uma avaliação do resultado das regras para revisar casos especiais. As regras de propagação de sujeitos lidavam com sujeitos em coordenação, sujeitos de estruturas que envolvem *xcomp*, orações adjetivas (*acl*), e orações adverbiais (*advcl*), tanto reduzidas quanto desenvolvidas. Aproveitamos padrões e regras de detecção de sujeitos de trabalhos anteriores (por exemplo, (Freitas & de Souza, 2024)), que foram aperfeiçoados e se beneficiaram da qualidade da anotação sintática subjacente ao Porttinari-base.

Quanto à identificação, os casos de sujeito *xcomp* são os mais simples de serem tratados, uma vez que a natureza da classe *xcomp*, quando aplicada a verbos, é justamente indicar que há um compartilhamento obrigatório de sujeito com o verbo do qual o *xcomp* depende. O desafio é o mesmo da fase anterior, a classificação quanto à natureza do argumento.

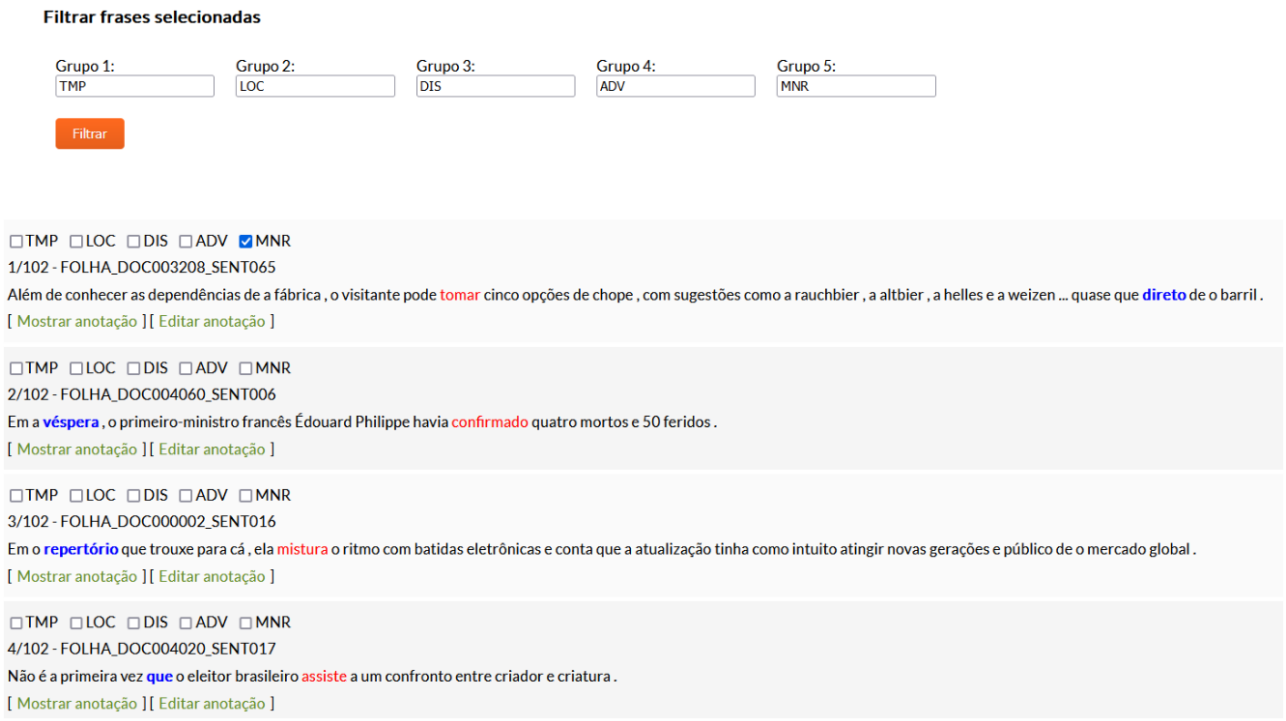

Figura 6: Tela da ferramenta com a utilização de filtros

Os casos de coordenação foram tratados por três tipos de regras, aplicadas nesta ordem:

- atribuição de Arg1 aos sujeitos de verbos na voz passiva, já tratando as exceções;
- atribuição de Arg1 aos sujeitos de verbos em que o sujeito não é Arg0 (regra que envolve lemas específicos, como *parecer*, *assemelhar*, *custar* etc). Nestes casos, cada frase foi lida e, com a utilização do filtros, foram excluídos os casos em que a regra de anotação não seria aplicada (casos em que o sujeito não seria Arg1);
- atribuição de Arg0 aos demais casos[19].

Antes da aplicação das regras, lemos as linhas de concordância dos casos que serão modificados. Na frase (1), por exemplo, embora *registrar* não exija um sujeito do tipo Arg1, *bebê*, apesar de sujeito, é *o que é registrado* (Arg1).

1. Hoje o bebê nasce e já **registra**

Para verbos na forma infinitiva, identificamos argumentos apenas se estes forem recuperáveis (caso do exemplo 1 abaixo). Em caso de dúvida ou em caso de argumentos não recuperáveis (exemplo 2), nada foi feito.

[19]Para uma ideia acerca da generalização das regras de coordenação de sujeitos, a primeira regra identificou e classificou cerca de 50 tokens, a segunda regra, 85 tokens e, a terceira regra, cerca de 650 tokens.

1. O presidente centrista optou por **garantir** pela a primeira vez em anos que (...). (*O presidente garantiu*)

2. A sua empresa o coloca em hotel em Guarulhos sempre que ele viaja a São Paulo, o que dificulta **pedir** o Uber (...) (*Não é possível determinar com certeza quem pedirá o Uber*)

Orações adverbiais gerundivas, quando equivalentes a "o que" (exemplo 1 abaixo), foram consideradas sem sujeito, e, portanto, não tiveram argumento associado à função sujeito. Nesses casos, o entendimento é de um agente oracional, que não pode ser localizado em um único sintagma/elemento nominal. Como sempre, há exceções, como em "*Janot fechou 159 acordos de colaboração, **deixando** para Dodge...*" em que *Janot* foi anotado como Arg0 de *deixar*.

1. Isso joga um pouco de água na propaganda petista, mas não significa que, sob Lula, o país não tenha enriquecido, **melhorando** também a situação dos pobres.

Os sujeitos de orações adjetivas (*acl*) e de orações adverbiais (*advcl*), ambas reduzidas de particípio, foram de resolução simples, uma vez que, na imensa maioria das vezes, se comportam como sujeitos de orações passivas, e recebem, em geral, a etiqueta Arg1. Regras de convergência morfológica foram especialmente úteis nesta fase, pois, mesmo quando a regra não se aplica, como

no caso (1), em que temos uma divergência morfológica que não é erro, separar os casos de convergência daqueles de divergência torna a análise das concordâncias mais rápida.

1. A valorização do indivíduo e a convivência exclusiva entre iguais têm acarretado o aumento da intolerância e dos preconceitos, amplamente **difundidos** nas redes sociais.

4.3. Validação final

Ao final do processo, passamos novamente o material por regras de validação, tendo em vista detectar e corrigir inconsistências ou erros formais de anotação introduzidas nesta etapa. Ao final das correções, as regras foram reaplicadas para garantir que não re-introduzimos erros. Diferentemente das regras de anotação, que em muitos casos são aplicáveis apenas ao corpus Porttinari-base (devido à utilização de filtros, por exemplo, que são aplicados a frases específicas), as regras de validação são gerais o suficiente para serem aplicadas a qualquer corpus anotado neste formato. Abaixo listamos algumas delas:

- Encontre um mesmo token que contenha dois ou mais argumentos diferentes que estejam associados a um mesmo núcleo (Esta regra garante que a um mesmo token não foram atribuídos dois papéis com relação ao mesmo verbo).
- Encontre 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).
- Nenhum token cuja relação sintática é *root* pode ter papel semântico.

5. Versões

Em (Freitas & Pardo, 2024), indicamos a presença de duas versões, "clássica" e "ud". Eliminamos essa divisão, uma vez que a principal diferença entre elas dizia respeito ao tratamento dado aos verbos auxiliares. Com a ciência de que UD ampliou a lista de auxiliares para a língua portuguesa, mantemos apenas o que em (Freitas & Pardo, 2024) chamamos de versão "clássica", isto é, uma versão na qual a anotação de argumentos segue a noção inicial de proposição, independentemente do que a abordagem sintática subjacente considere "verbo". Ainda assim, disponibilizamos duas versões:

1. *Versão completa*: contém todas as relações entre predicadores verbais e argumentos
2. *Versão apenas com relações explícitas*: contém apenas as relações entre predicadores verbais e argumentos que estão explícitos na frase.

Essas versões permitem decompor a anotação de papéis semânticos em duas tarefas – anotação de papéis explícitos e anotação de papéis implícitos – e permitem investigar se a incorporação de argumentos implícitos dificulta o aprendizado, apesar de adicionar mais dados.

6. Concordância interanotadores

Em (Freitas & Pardo, 2024), detalhamos o procedimento de concordância interanotadores — que contou com uma única anotadora e foi feito sobre uma amostra do corpus PropBank-Br v2, uma vez que tanto as diretivas de anotação e o Verbo-Brasil foram criados com a utilização desses recursos. A anotação original da amostra foi apagada, e as frases receberam uma nova anotação. A concordância (computada comparando-se a anotação nova com a existente) foi de 0,90 (*kappa*). Palmer et al. (2005), considerando o cenário de avaliação que leva em conta argumentos numerados e ArgMs (como fizemos), relatam uma concordância de 0,93, e Bonial et al. (2018) indicam uma média de 0,85, o que evidencia que os valores obtidos na nossa iniciativa estão condizentes com o que se observa na literatura.

6.1. Análise qualitativa

A análise qualitativa das discordâncias, que apresentamos aqui, foi feita com o *Julgamento*, ambiente da ferramenta ET voltado à avaliação de anotações. Embora o Julgamento tenha sido criado para facilitar a detecção de inconsistências na anotação morfossintática, ao longo deste projeto a ferramenta foi aperfeiçoada para permitir que outros tipos de anotação fossem comparados/julgados, incluindo o cálculo automático do Coeficiente Kappa de Cohen entre duas anotações.

A maioria das divergências aconteceu entre ArgM-adv (original) x ArgM-dis (anotação nova), ArgM-mnr (original) e ArgM-adv (anotação nova), ArgM-prd (anotação nova), Arg2 (anotação nova) e ArgM-ext (anotação nova), e entre Arg0, Arg1 e Arg2 (em ambas as direções). Todas as divergências foram analisadas individualmente, e percebemos que foram motivadas ou por falta de atenção de alguma das partes, ou por discordâncias na anotação de

classes genéricas que acabam servindo, em certas circunstâncias, como um guarda-chuva (como ArgM-mnr e ArgM-adv). Estas são exatamente as classes mencionadas por Palmer et al. (2005, p. 87) como sendo classes difíceis de distinguir.

Nesses casos, a discordância serviu como alerta para uma tentativa maior de uniformização com relação ao conteúdo das classes, com explicitação de informação e acréscimo de exemplos nas diretivas. As discordâncias referentes aos argumentos numerados, por sua vez, chamaram a atenção para um ponto relevante da anotação, ilustrado pelo exemplo abaixo de *cair no choro*, um caso de *expressão multipalavra*.

- Durante 78 minutos, ele **caiu** várias vezes no choro.

Este tipo de fenômeno ainda não passou por um tratamento sistemático no Verbo-Brasil, e apenas nas últimas atualizações do PropBank original tem recebido maior atenção (Pradhan et al., 2022). Em nosso caso, a divergência de análises aconteceu porque a anotação original seguiu o *frame* de *cair*, que indica o sujeito como Arg1. Já a anotação do PBP, por lapso, anotou o sujeito como Arg0 por analogia à estrutura argumental do verbo *chorar*, que tem sujeito Arg0, interpretando *cair no choro* como "chorar", ou "começar a chorar (muito)".[20]

Verbos que participam de expressões multipalavra estão entre os mais frequentes do corpus — por exemplo, *ter* (posição 1 no ranque de frequência), *fazer* (posição 3), *ficar* (posição 7), *ver* (posição 9), *dar* (posição 10), *levar* (posição 15), *tomar* (posição 47), *cair* (posição 53) — e respondem por 13% das ocorrências verbais, considerando apenas os verbos listados.

Apesar de alta, a concordância medida pelo índice *kappa* mascara a variedade de interpretações possíveis subjacentes a este tipo de anotação. Este mascaramento é consequência da maneira PropBank de anotar, que força artificialmente a escolha de uma interpretação quando sinaliza qual decisão tomar em caso de dúvida. Abaixo estão dois trechos das diretivas originais que indicam claramente como agir em caso de hesitação:

> *If an argument is both an agent and a patient, then Arg0 label should be selected* (Bonial et al., 2015, p. 8)

> *Thus, as a general rule, if the annotator cannot determine whether an argument is more appropriately purpose or cause, cause is the default choice.* (Bonial et al., 2015, p. 17)

Do mesmo modo, a existência de diretrizes ricas em exemplos (quer para argumentos numerados quer para modificadores) garante a uniformidade de análises, que podem ser feitas por analogia. Porém, essa homogeneidade, se por um lado assegura bons resultados na concordância interanotadores, por outro lado elimina a possibilidade de leituras alternativas e legítimas.

Ao longo do processo de anotação, a percepção de que um determinado elemento poderia ser interpretado de diferentes maneiras foi a motivação para um estudo sobre concordâncias e discordâncias na anotação de papéis semânticos, que relatamos a seguir.

6.2. Concordâncias e discordâncias na anotação de papéis semânticos

Tradicionalmente, a anotação de papéis semânticos não permite anotação múltipla, apostando que o valor semântico dos argumentos é centralmente monossêmico e que serão raros os casos de dúvida com relação ao papel semântico que deve ser atribuído em um dado contexto.

Do mesmo modo, a utilização do índice *kappa* pressupõe a existência de uma única alternativa consensual/correta, e não lida com alternativas distintas e não excludentes. Por outro lado, uma alta concordância é sem dúvida uma maneira válida (mas não infalível) de avaliar a confiança destas interpretações. Uma alta concordância valida uma maneira de analisar e classificar os dados linguísticos, e é indicativa do seu potencial de reprodutibilidade. No entanto, não há necessidade de restringir essa concordância a uma única alternativa; podemos concordar, simultaneamente, em várias análises – e reconhecemos que, nesses casos, será preciso lidar com dois gargalos relacionados: (i) como distinguir anotações divergentes legítimas daquelas que não encontram lastro no contexto ou que são resultado de lapso/falta de atenção, de diretivas mal feitas ou de esquemas de anotação mal desenhados, e (ii) como calcular a concordância entre anotadores, levando em conta (i).

Com o duplo objetivo de verificar se (a) a dificuldade de escolha entre duas (ou mais) etiquetas era idiossincrasia de uma única anotadora e se (b) estas anotações eram "confiáveis" (reforçando o resultado da medida *kappa*), fizemos um estudo

[20] Embora lapso, a anotação pela combinação "cair no choro" segue a diretriz de 2015 do Propbank original, que indica que anotadores devem ser generosos em sua definição de construções verbais leves" (Bonial et al., 2015, p. 48).

sobre concordâncias com algumas frases que suscitaram dúvidas durante a anotação do PBP. Utilizando o mesmo conjunto de frases, construímos dois cenários de anotação: cenário em que apenas uma anotação é possível (cenário PropBank, com anotação monossêmica) e cenário em que mais de uma alternativa de anotação é possível (anotação múltipla). Nossa hipótese era de que a discordância entre anotações seria alta em ambos os cenários e que, especificamente no cenário com anotação múltipla, a possibilidade de escolher mais de uma análise seria capaz de atenuar a (aparente) diversidade de classificações, funcionando como alternativa para a obtenção de consensos. Isto é, se dois anotadores estão cientes de que A e B são classificações possíveis para a frase 1, mas só podem escolher uma classe, é possível que um escolha A e outro escolha B, e ficamos com a falsa impressão de discordância. Por outro lado, se ambos podem escolher A e B, temos concordância.

O estudo verificou a convergência/divergência de análises entre os seguintes grupos de etiquetas do tipo ArgMs: *local*, *tempo*, *causa*, *maneira* e *finalidade*. A escolha por investigar ArgMs foi metodológica: por terem sentido genérico e poderem participar de qualquer frase, ArgMs tornam o exercício mais simples de ser feito, e não precisamos de pessoas especializadas na anotação de papéis semânticos.

O estudo foi feito utilizando Formulários Google, com respostas de múltipla escolha. A fim de manter a atenção e evitar um questionário longo, as frases foram distribuídas em 2 grupos distintos: um em que estavam em foco apenas as classes *local* e *tempo* (embora todas as alternativas estivessem disponíveis para escolha), e outro com todas as classes de interesse. Ambos os grupos continham 16 frases, algumas delas frases "controle": frases cuja classificação nos parecia claramente monossêmica. Dessa forma, foram criados 4 formulários (forms), com 2 conjuntos de frases:

1. Forms 1 e 2 continham as mesmas frases, na mesma ordem

   form 1: classificação única
   form 2: classificação múltipla

2. Forms 3 e 4 continham as mesmas frases, na mesma ordem

   form 3: classificação única
   form 4: classificação múltipla

As instruções de cada formulário estão na Figura 7, e a Figura 8 mostra a tela do formulário para uma frase.

Cinco pessoas responderam o form1, oito o form2, sete o form3 e sete o form4, com alguma sobreposição entre aquelas que responderam os forms 1 e 3 e aquelas que responderam os forms 2 e 4. Todas as pessoas que participaram tinham experiência na anotação de corpus, estavam comprometidas com a tarefa, e parte delas já tinha sido adjudicadora e/ou coordenadora de projetos de anotação.

Os resultados para os formulários 1 e 2 estão na Figura 9, e os resultados para os formulários 3 e 4 estão na Figura 10 (cada frase é identificada por "s" seguido de um número, e as frases usadas nos formulários estão no apêndice). Cada gráfico apresenta as análises para uma mesma frase. Como é possível observar, a divergência foi alta em ambos os cenários e para todos os formulários. Em uma análise global, vemos também que concordâncias integrais (gráficos de apenas uma cor) divergem entre os tipos de formulário (escolha única X múltipla), sinalizando que a possibilidade de selecionar mais de uma análise não necessariamente pulveriza as respostas.

As frases controle dos formulários 1 e 2 levaram a 100% de concordância e foram eliminadas da análise, que, como consequência, contou com 13 frases. Por outro lado, nos formulários 3 e 4, as frases controle só levaram a concordâncias integrais para as classes TEMPO e LOCAL. A frase 2 ("*Nós respeitamos a PGR e temos a soberania de decidir* ***por maioria***."), considerada por nós monossêmica da classe MANEIRA, foi interpretada por uma pessoa, dentre as 14 que responderam os forms 3 e 4 , como simultaneamente MANEIRA & CAUSA. É possível que a estrutura da frase, associada à leitura desatenta, tenha levado a essa análise ("*decidir por causa da maioria*" ou "*decidir porque a maioria quis*"), mas o contexto da frase não legitima essa interpretação. Por isso, continuamos a considerar esta frase monossêmica. As demais frases controle que se mostraram não-monossêmicas serão analisadas com as demais frases dos formulários 3 e 4. A seguir, analisamos mais detalhadamente cada grupo de formulários.

Formulários 1 e 2: tempo e local em foco. Analisando os resultados dos forms 1 (resposta única) e 2 (resposta múltipla), cujo foco estava em LOCAL e TEMPO, vemos que a discordância foi alta em ambos os cenários, como previsto, mas é difícil comparar os forms 1 e 2 porque, apesar de disporem das mesmas alternativas, respondentes do form1 só utilizaram as classes TEMPO e LOCAL. Mesmo assim, no form2, das únicas 5 frases em que notamos a existência de uma resposta majoritária (s3, s4,

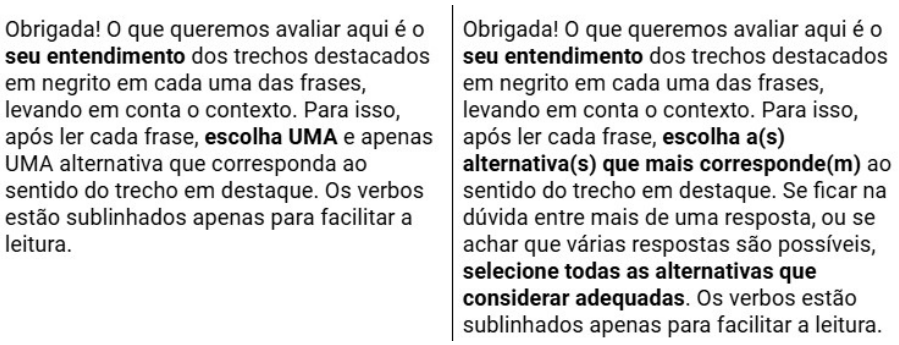

Figura 7: Instruções para o cenário de escolha única (esquerda) e de escolha múltipla (direita)

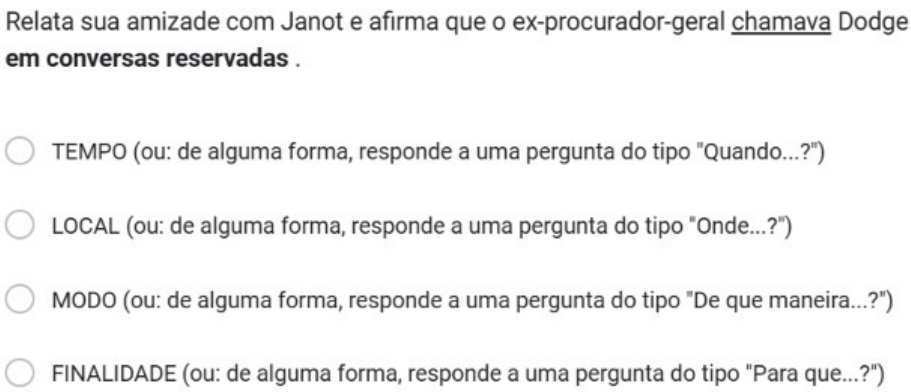

Figura 8: Apresentação das frases e opções de resposta no estudo de concordâncias

s5, s6, s15), 4 delas (s3, s4, s5, s6) referem-se à combinação TEMPO & LOCAL. Se considerarmos *concordâncias parciais* todas as classificações que selecionam LOCAL, TEMPO e LOCAL & TEMPO e julgamos aceitável uma concordância de 75% nas respostas para uma mesma frase, deixamos de ver discordâncias no form2 e passamos a ter concordâncias aceitáveis em 84% das frases (s1, s3, s4, s5, s6, s7, s8, s12, s13, s14 e s15). Deste ponto de vista, vemos a classificação múltipla como o caminho capaz de capturar a concordância.

Admitimos que apenas TEMPO e LOCAL são respostas válidas, e o form2 traz classificações não previstas inicialmente[21]. Em s1, por exemplo, não imaginávamos uma classificação MANEIRA ("*Relata sua amizade com Janot e afirma que o ex-procurador-geral chamava Dodge de "bruxa"* ***em conversas reservadas***"). O primeiro impulso foi descartar a análise, mas cientes de que os participantes eram leitores proficientes, insistimos nessa alternativa até perceber que "*... e afirma que,* ***reservadamente****, o ex-procurador-geral chamava Dodge de "bruxa""* é uma paráfrase possível de s1, legitimando a classificação MANEIRA.

Fizemos o mesmo com as outras frases e classificações, e raramente descartamos alguma análise, como será visto. Com esta ressalva, passamos a ter apenas concordâncias no form2[22].

Formulários 3 e 4: maneira, causa, finalidade (e tempo e local) em foco. Analisando s1 nos forms 3 e 4, vemos que as três classes presentes na divergência (MANEIRA, CAUSA, TEMPO) do form3 aparecem de forma individual e de forma combinada no form4, o que podemos ler como concordância, assim como fizemos na análise do form2. Situação parecida ocorre com s4, s7, s8; s9, s11, s13 e s14. Ou seja, desse outro ponto de vista, considerando ambos os forms e considerando ainda a presença de convergência

[21] O grupo de pessoas que preencheu o form2 trouxe mais possibilidades de leitura, e esse comportamento não era típico de certos anotadores, mas do grupo.

[22] Duas pessoas que responderam aos forms 1 e 2 manifestaram seu descontentamento com a ausência de uma classe do tipo EVENTO ou SITUAÇÃO, que eliminaria boa parte das dúvidas entre LOCAL e TEMPO. Transcrevemos o comentário de uma das participantes após o preenchimento do form2: *"(...) mas então [o PropBank] esquece/não considera as teorias semânticas que existem desde pelo menos os anos 80, como a DRT ou a davidsoniana. (...) Porque na maior parte das vezes em que pus TEMPO queria ter posto SITUAÇÃO ou EVENTO. (...) para mim, Acontecimento não é TEMPO E/OU LOCAL, mas sim outra categoria, mais básica.*". Concordamos com o descontentamento e também lamentamos a inexistência desta alternativa. No entanto, uma das ideias do estudo era lidar com as classes do PropBank, por mais discutíveis que possam ser.

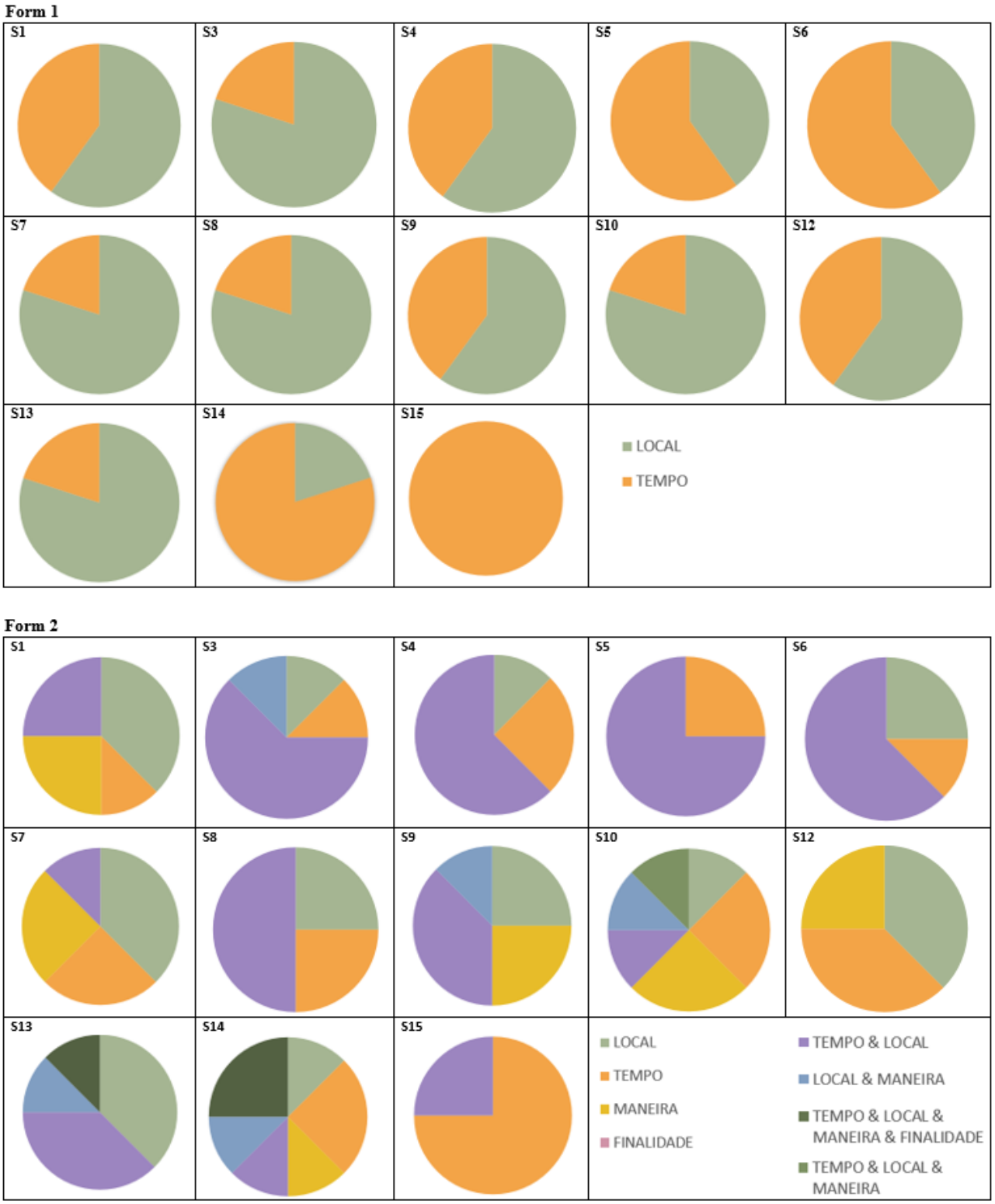

Figura 9: Respostas dos formulários 1 e 2

em s5, passamos a ter quase 70% das frases com concordância, em um cenário diferente do que teríamos com a medida kappa.

É interessante também analisar em mais detalhes alguns casos de frases controle. A frase 15, considerada monossêmica da classe FINALIDADE ("***Para se livrar do problema***, <u>dobrou</u> *a aposta e injetou mais R$ 2 bilhões na JBS*") foi interpretada como CAUSA por duas pessoas no form4, o que reconhecemos ser também possível[23]. A frase 12, considerada monossêmica

[23]Por exemplo, em *Por que dobrou a aposta e injetou... ? Para se livrar do problema — porque quis se livrar do problema.*

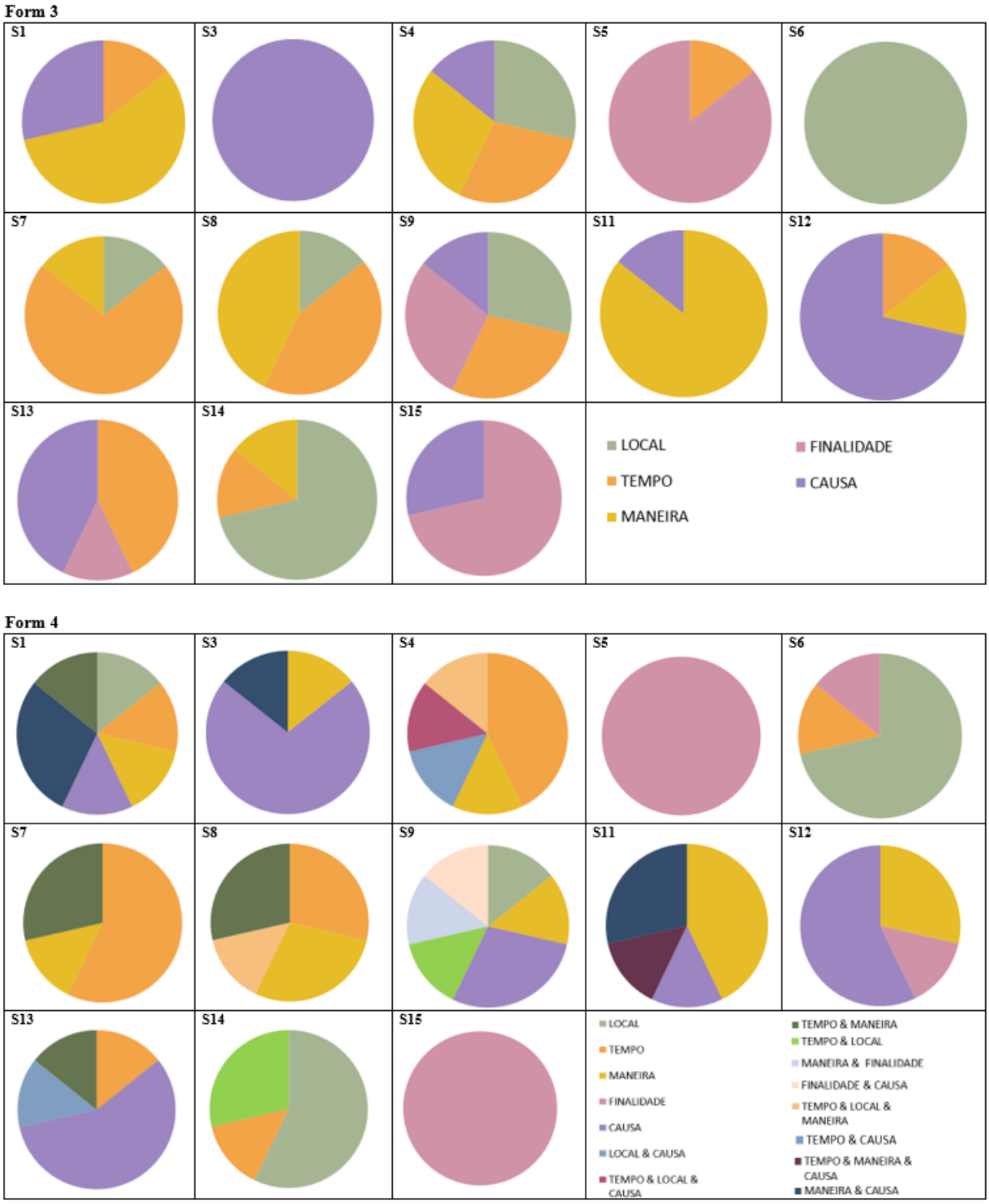

Figura 10: Respostas dos formulários 3 e 4

da classe CAUSA ("***Com a liderança caindo em seu colo***, *Lewis Hamilton ditou as ações da corrida.*") foi classificada como MANEIRA, FINALIDADE e TEMPO, e consideramos todas as análises legítimas.

Das 5 frases controle dos formulários 3 e 4, apenas 3 mantiveram esse status (para as classes TEMPO, LOCAL e MANEIRA). As demais entraram normalmente nas análises, que então contou com 13 frases. O fato de frases consideradas monossêmicas terem recebido classificações diferentes daquela prevista indica também que a polissemia é ainda mais frequente do que imaginamos e que, mesmo cientes de que pode haver

mais de uma leitura, nem sempre somos capazes de perceber todas as alternativas disponíveis.

Confiança na anotação. Nosso segundo objetivo com o estudo era verificar a confiança na anotação de papéis semânticos no Porttinari-base Propbank. Para tanto, comparamos as anotações codificadas no corpus com as análises dos forms 1 e 3 (escolha única) a fim de verificar se houve (i) ampliação ou construção de discordância onde não havia ou (ii) reforço de uma concordância já existente, quando a análise corroborava a decisão da maioria.

Os resultados estão na Tabela 2, e vemos tendências opostas conforme as classes em análise. No form1, com foco em LOCAL e TEMPO, em 61% dos casos a anotação indicada no corpus seguiu a classificação da minoria, o que resultou na acentuação de discordâncias e no desaparecimento de concordâncias. Já no form 3, os casos de ampliação ou criação de discordância foram minoria (35%), e a maioria das análises reforçou um consenso já existente. No entanto, em s5, s6 e s9, a classe/interpretação codificada no corpus foi única, isto é, tivemos um leitura completamente destoante das demais.

Em s6 ("*Em alguns roteiros, como na Croácia, é preciso pagar* ***na maioria dos trajetos***."), a anotação codificada no corpus é a única que indica TEMPO. A justificativa está na possibilidade de "na maioria dos trajetos" poder ser entendida como "sempre" ou "na maioria das vezes", e nesse caso a anotação seria TEMPO, conforme as diretivas de Duran (2014).

Em s9 ("***Em homenagem a Che Guevara,*** *Cuba reage às declarações de Trump.*"), a anotação do corpus é a única que indica MANEIRA. Curiosamente, exceto por MANEIRA, todas as demais classes estão presentes em pelo menos 1 resposta nesta frase. Nossa justificativa para a MANEIRA está na leitura "*Como/de que maneira Cuba reagiu? Homenageando Che Guevara*".

Por fim, o caso de s5 ("*Quem é que sai* ***para trabalhar*** *pensando em tomar um soco na cara?*") é digno de nota porque, além de a anotação do corpus ter sido a única a indicar LOCAL, tanto form3 quanto form4 apresentam uma rara situação de concordância quase total ou total entre eles, escolhendo FINALIDADE. De fato, esta foi uma frase que despertou idas e vindas no processo de anotação, com muita dificuldade entre FINALIDADE e LOCAL. Curiosamente, esta foi a única frase em que o trecho em destaque não era um ArgM, mas um argumento numerado, seguindo o *frame* do verbo *sair* — um Arg3 com o valor de "*lugar ou situação de destino (sair para)*", conforme o Verbo-Brasil (Duran & Aluísio, 2015). No entanto, e para simplificar o estudo das concordâncias, "*lugar ou situação de destino*" foi traduzido como LOCAL. Assim, podemos interpretar "*para trabalhar*" como a finalidade de sair de casa, mas também como "*situação de destino*": "*Saiu pra onde? Para trabalhar* ou "*Saiu para o trabalho*". Escolhemos LOCAL-Arg3 e reconhecemos que a motivação para a saída (FINALIDADE) também está presente. À luz das análises empreendidas, contudo, pensamos que nossa tradução de Arg3 foi equivocada, pois talvez Arg3, ao significar "*lugar* ou situação de *destino*", fosse melhor traduzida como LOCAL & FINALIDADE.

Estes três casos trazem desafios para a ideia de que a alternativa menos frequente tende a sinalizar erro de análise e, consequentemente, desafios para pensar a avaliação da concordância neste tipo de abordagem. O último passo da análise está relacionado a este aspecto: a distinção entre análises divergentes legítimas, como os casos apontados, daquelas não licenciadas pelo contexto. A leitura de todas as respostas identificou 10 análises, dentre as 432 repostas fornecidas, difíceis de serem legitimadas, o que corresponde a 2,3% das análises. As frases e análises envolvidas nestes casos estão abaixo:

- Form2-s9: "É uma grande honra receber uma proposta de um grande clube como esse , mas também é uma grande honra estar aqui no Liverpool, um grande clube, falou **em entrevista à ESPN**" — difícil justificar as leituras com MANEIRA (2 casos) e FINALIDADE (1);
- Form3-s7: "No momento em que a nova moeda for criada, se os preços forem convertidos **no pico**, não será necessário um aperto monetário, uma recessão temporária?" — difícil justificar a leitura com MANEIRA (3);
- Form3-s8: "Ou seja: o Emmy de 2017 foi politizado **do começo ao fim**." — difícil justificar a leitura com LOCAL (2);
- Form3-s13: "A equipe gaúcha perdeu a oportunidade de manter a distância para o líder **ao ser derrotada pela Chapecoense por 1 a 0**, em casa." — difícil justificar a leitura com FINALIDADE (1);
- Form4-s2: "Nós respeitamos a PGR e temos a soberania de decidir **por maioria**." — difícil justificar a leitura com CAUSA (1), ainda que tenha sido associada à MANEIRA.

Efeito	**Form 1**	**Form 3**
Ampliação ou construção de discordância	s1 s3 s4 s6 s7 s10 s12 s13	s5 s6 s9, s11, s14
Reforço de consenso	s5 s8 s9 s14 s15	s1 s2 s3 s4 s7 s8 s12 s13 s15

Tabela 2: Análises das anotações do corpus em comparação com outras anotações

A dificuldade de diferenciar interpretações legítimas e ilegítimas aumenta quando vemos que a frequência das análises não é pista confiável, já que interpretações únicas não necessariamente são ilegítimas e interpretações ilegítimas não necessariamente aparecem uma vez só. Por outro lado, o baixo índice de análises ilegítimas (2,3%) sinaliza que talvez o mais acertado seja confiar nas interpretações humanas, desde que sejam feitas por profissionais qualificados, com excelente capacidade de interpretação de texto (característica que não é exclusiva de linguistas), e, se for o caso, conhecimento do domínio. Do mesmo modo, os resultados nos levam a repensar o papel da adjudicação quanto ao seu caráter unificador de análises (Basile et al., 2021), bem como o papel das instruções de anotação. Nos habituamos à ideia de uma documentação detalhada como forma de garantir consistência. No entanto, esta indicação detalhada de como proceder nos "casos difíceis ou ambíguos", isto é, aqueles não prototípicos, pode ser uma armadilha. Ao garantir artificialmente a igualdade de análises, isto é, ao determinar como algo deve ser interpretado, podemos deixar de fora leituras legítimas, que talvez apenas acidentalmente não tenham sido percebidas por quem escreve as instruções de anotação. Por isso, talvez caiba uma distinção sutil entre *diretrizes* de anotação e *manuais* de anotação. Entendemos que uma diretriz aponta caminhos; um manual pode ter caráter mais determinístico — e, normalmente, determinístico com relação à unicidade na interpretação (e uma *documentação* detalha o que já foi feito, e como foi feito). Neste contexto, o desafio está em produzir diretivas claras o suficiente para apresentar o que significa escolher cada uma das classes de anotação sem eliminar o caráter interpretativo inerente à anotação.

Tomando um outro ponto de vista, vemos as análises das figuras não como discordâncias quanto ao significado da frase, mas como a presença de significados distribuídos. Por fim, o Porttinari-base é um corpus de frases, e portanto não dispomos dos contextos, que poderiam direcionar a interpretação e minimizar a pluralidade de leituras em alguns casos.

7. Resultados e análises da anotação de papéis semânticos

A Tabela 3 apresenta a distribuição dos argumentos no corpus, e a Figura 11 traz os mesmos dados de maneira mais detalhada, comparando versão completa e a versão que apenas contém argumentos de elementos explícitos. Já a Figura 12 traz a distribuição dos argumentos por função sintática, considerando a versão completa.

	PBP
Arg0	8.783 (19,2%)
Arg1	16.322 (35,6%)
Arg2	6.276 (13,7%)
Arg3	370 (0,8%)
Arg4	205 (0,4%)
Tmp	2.768 (6,0%)
Loc	1.465 (3,2%)
Mnr	1.428 (3,1%)
Adv	1.250 (2,7%)
ArgMs restantes	6.946 (15,0%)
Total	**45.813**

Tabela 3: Distribuição dos argumentos no PBP, considerando a versão completa

A partir da análise das Figuras 11 e 12, destacamos dois pontos: (i) a distribuição desigual de argumentos e (ii) a forte correlação entre certos argumentos e funções sintáticas. Quanto ao primeiro ponto, o desequilíbrio ocorre tanto do ponto de vista dos argumentos numerados (a frequência de Arg3 e Arg4 é bastante baixa quando comparada aos demais argumentos numerados), quanto dos argumentos modificadores, nos quais TEMPO, LOCAL e MANEIRA (além de *ArgM-Adv*, que tem função genérica, e *Neg*, usado para advérbios negativos) se destacam, mas mesmo assim têm uma frequência baixa quando comparados a Arg1 e Arg0.

Diferentemente do Propbank.br v2, o Porttinari-base foi pensado originalmente como conjunto de dados para treino de informação sintática, e não de papéis semânticos. Uma das consequências dessa escolha é que, do ponto de vista da anotação de papéis, certas construções

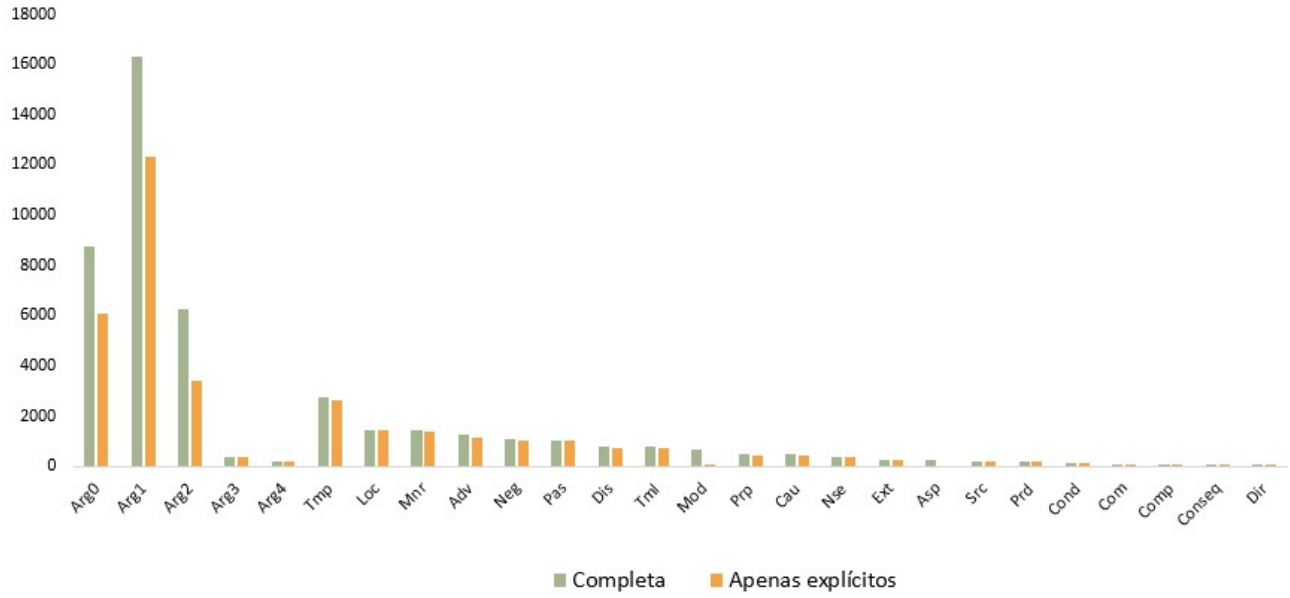

Figura 11: Distribuição detalhada dos argumentos, considerando ambas as versões

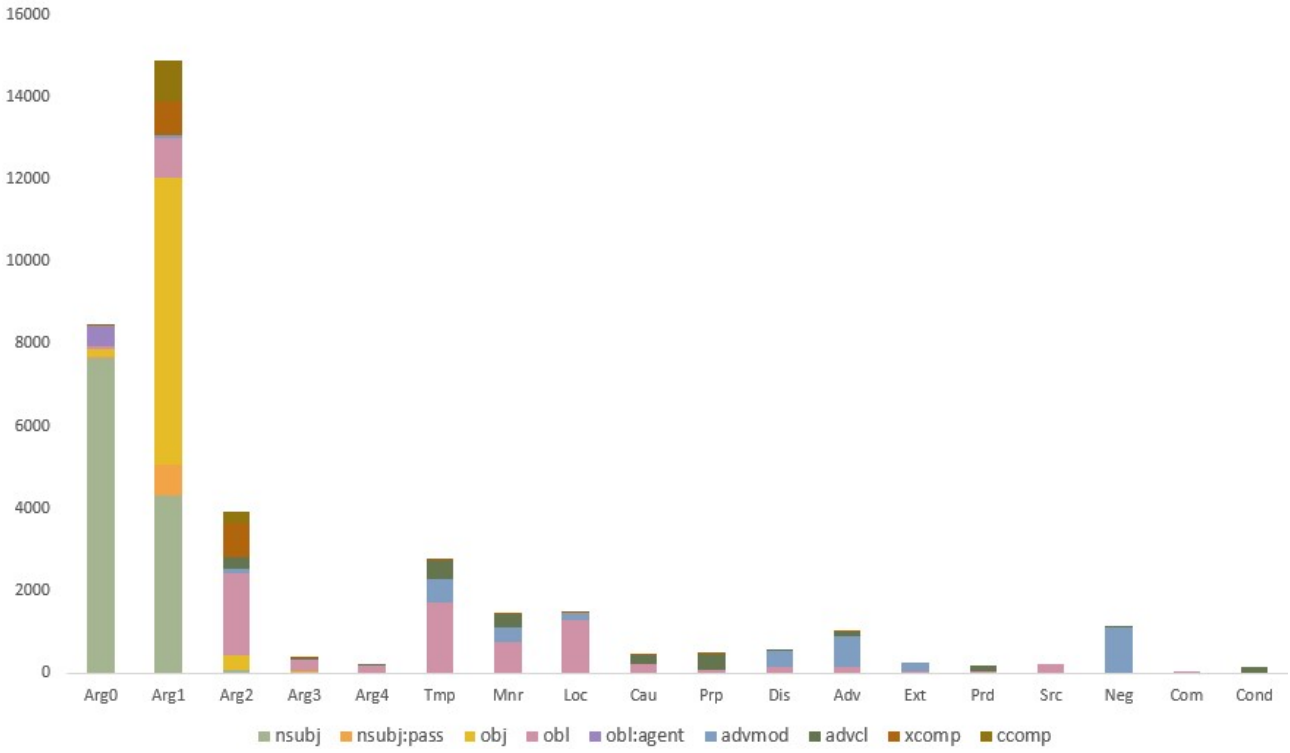

Figura 12: Distribuição de argumentos por função sintática (versão completa)

aparecem com frequência muito baixa. Uma estratégia para mitigar esta escassez é a utilização de dados/frases artificialmente criadas. Investigamos de maneira preliminar esta possibilidade em um grande modelo de língua, pedindo por meio de diferentes *prompts* a criação de frases com determinadas estruturas linguísticas — por exemplo, de frases com orações consecutivas, que levariam à inclusão de mais proposições com a relação ArgM-conseq. Os resultados parecem promissores[24], desde que as frases criadas passem pela revisão humana[25].

[24] O estudo foi apenas uma exploração inicial, não tendo sido possível analisar e quantificar os resultados.

[25] Apesar da revisão humana demandar algum trabalho, é mais rápido ler e eventualmente corrigir frases (quanto à naturalidade e coerência, e não quanto à correção gramatical, já que as frases geradas não apresentaram erros dessa natureza), do que encontrá-las em corpora reais, porque são frases em geral pouco frequentes.

Quanto ao segundo ponto, a associação mais frequente é entre obj e Arg1, com 92,03%, seguida da associação entre nsubj e Arg0, com 63,49%. Ispirados pelos dados de Palmer et al. (2005) para os elementos PP-in e PP-at (do inglês, *Prepositional Phrase* — sintagmas preposicionados), trazemos nas Tabelas 4, 5 e 6 a distribuição de argumentos pelas preposições *em*, *a* e *de*, mas o que vemos são tendências pouco marcadas. Para a preposição *em*, a distribuição entre ArgM-loc (36%) e ArgM-tmp (30%) é próxima, o que se alinha aos resultados do estudo das concordâncias, no que se refere às leituras de LOCAL e TEMPO.

Argumento ("em")	**Total**	**%**
Loc	1.108	36,28%
Tmp	990	30,75%
Arg2	440	13,66%
Arg1	222	6,89%
Mnr	152	4,72%
Dis	44	1,36%
Adv	37	1,14%
Arg3	30	0,93%
Prp	17	0,52%
Ext	16	0,49%
Cau	13	0,40%
Arg0	13	0,40%
Src	11	0,34%
Arg4	10	0,31%
Prd	8	0,24%

Tabela 4: Distribuição dos argumentos introduzidos pela preposição "em"

Argumento ("a")	**Total**	**%**
Arg2	443	46,83%
Arg1	145	15,33%
Tmp	107	11,31%
Arg4	80	8,46%
Mnr	47	4,97%
Arg3	23	2,43%
Loc	23	2,43%
Adv	17	1,80%
Cau	13	0,40%
Ext	6	0,63%
Dis	5	0,53%
Cau	3	0,32%
Arg0	2	0,21%

Tabela 5: Distribuição dos argumentos introduzidos pela preposição "a"

Argumento ("de")	**Total**	**%**
Arg2	363	40,11%
Arg1	230	25,41%
Mnr	103	11,38%
Arg3	64	7,07%
Tmp	33	3,65%
Src	27	2,98%
Adv	13	1,44%
Loc	10	1,10%
Dis	9	0,99%
Prd	7	0,77%
Arg4	5	0,55%
Ext	4	0,44%
Arg0	3	0,33%

Tabela 6: Distribuição dos argumentos introduzidos pela preposição "de"

Apesar da regularidade entre Arg0 e Arg1 e funções sintáticas de sujeito e objeto (e também predicativo do sujeito, para os verbos *ser* e *estar*), os demais argumentos numerados, com baixa frequência e sem correlações sintáticas explícitas, são de difícil generalização. Estas observações convergem com resultados para o inglês relacionados à generalização da anotação PropBank. Os trabalhos de Bonial et al. (2018), Merlo & Van Der Plas (2009) e Yi et al. (2007) indicam que, embora a anotação automática tenha um bom desempenho em Arg0 e Ag1, o desempenho nos argumentos numerados 2 a 6[26] é fraco. Se estamos diante de um sistema que leva em conta informação sintática, o bom desempenho em certas classes pode ser facilmente explicado. Como indicam Merlo & Van Der Plas (2009), um sistema de anotação PropBank ingênuo, que simplesmente atribuísse o rótulo mais frequente, estaria correto 52% das vezes atribuindo sempre um rótulo Arg1 (para o corpus do inglês).

Considerando os dados do PBP, um sistema de anotação ingênuo baseado em regras teria uma boa taxa de acerto se atribuísse Arg0 ao sujeito e Arg1 ao objeto, considerando que Arg0 e Arg1 respondem por 54,8% dos argumentos. Este desempenho poderia ser melhorado com a inclusão de um léxico enxuto, com a indicação de verbos cujo sujeito será sempre Arg1. Acrescentando à análise os dados das Tabelas 4, 5 e 6 no que se refere à anotação de ArgMs (que respondem por 30,2% dos argumentos no PBP), e alguma correlação entre certos verbos e argumentos numerados de baixa frequência (Arg2, Arg3

[26] Argumento atribuído aos substantivos, recentemente acrescentado ao projeto original.

e Arg4 respondem por 14,9% dos argumentos no PBP) um sistema baseado em regras e léxico é uma estratégia a ser levada em conta. Valores estado-da-arte para a anotação PropBank, conforme Wang et al. (2022), ficam entre 85% e 90% de F1 (números globais, e não temos acesso ao desempenho por classe). Em (Palmer et al., 2005), um sistema *baseline*, baseado em regras, consegue 83% de acertos — apenas 2% abaixo de valores estado-da-arte, considerando os cenários mais difíceis –, o que dá força ao argumento de que as generalizações facilitadas pela anotação estilo PropBank decorrem da sua forte correlação com a sintaxe.

Do ponto de vista da generalização e aprendizado de máquina, os argumentos numerados de baixa frequência, com sua semântica difusa e a pouca frequência, criam esparsidade e dificultam o aprendizado a partir de exemplos (Merlo & Van Der Plas, 2009). Ampliando a análise para incluir também argumentos modificadores, e trazendo os dados do PBP para ilustrar esta semântica pouco definida, observamos que a ideia de LOCAL se manifesta ora como Arg 1 (por exemplo, em *comparecer*), ora como Arg2 (em *plantar*, *instalar* e *acessar*), ora como Arg3 (em *apontar*), e na maioria das vezes como ArgM-loc. CAUSA pode ser Arg2 (em *culpar*, *elogiar* e *cansar*) e Arg5 (em *ampliar-se*), além de ArgM-cau. MANEIRA pode ser Arg1 (em *andar*), Arg2 (em *diminuir*, *tratar*, *atrapalhar* e *ampliar*) e ArgM-mnr. Nos dados de Yi et al. (2007), a ideia de *location*[27], por exemplo, aparece na totalidade de Arg5, mas também como Arg4 (em 5,6% dos Arg4), Arg3 (em 2% dos Arg3), Arg2 (6,5% dos Arg2) e, de forma marginal (menos de 1%) em Arg1 e Arg0. Porém, os dados de Yi et al. (2007) e Merlo & Van Der Plas (2009), assim como os nossos, levam em conta informações do léxico antes do processo de revisão, que levou a uma maior uniformização das classes (Bonial et al., 2018; Pradhan et al., 2022)[28].

Esta análise enseja alguns questionamentos. Quais as diferenças qualitativas entre os resultados de um sistema baseado em regras e um que leva em conta semântica distribuída? Podemos prever desempenhos diferentes conforme a natureza das classes e arquitetura dos sistemas, com argumentos numerados melhor classificados por um sistema baseado em regras, e argumentos modificadores melhor classificados por modelos que levam em conta semântica distribuída?

[27] Derivada de um alinhamento com as classes da VerbNet, e portanto restrita aos argumentos numerados.

[28] Porém, nos exemplos utilizados ao longo do artigo, sempre consultamos os *frames* atualizados da língua inglesa.

8. Papéis semânticos e o papel da sintaxe

A apresentação dos papéis semânticos — elementos da interface sintaxe/semântica — sugere o alinhamento entre esses dois níveis de análise linguística, o que nem sempre acontece. Por isso, se sintaticamente pode ser necessário colocar em grupos diferentes os pares *alertar* e *avisar*, ou *causar* e *provocar*, para as representações semânticas criadas para o PLN pode ser importante que os elementos de cada par tenham representações parecidas no que se refere à atribuição de argumentos — do mesmo modo que se espera que frases na voz passiva e na voz ativa tenham uma mesma representação.

Se, concordando com Saussure, admitimos que "o ponto de vista cria o objeto", devemos ter alguma cautela tanto com relação à naturalidade de classes de argumentos[29], quanto com a convergência (aparentemente também tida como natural) entre os interesses do PLN e de teorias linguísticas. No caso dos papéis semânticos — ou de encontrar *quem* faz *o que* etc –, estamos diante de pontos de vista diferentes para aspectos da língua que sem dúvida se cruzam e se sobrepõem — e que se informam mutuamente — mas que não são exatamente os mesmos, porque recortados de pontos de vista diferentes.

Ao surgir, o PropBank colocou lado a lado interesses de teorias linguísticas — e especificamente interesses mais voltados para a sintaxe –, e interesses aplicados do PLN. Ainda que tenham perdido espaço, esses objetivos iniciais, vinculados ao ponto de vista de teorias linguísticas, deixam vestígios no desenho do recurso.

Na criação de representações semânticas proposta pelo PropBank, a sintaxe tem um duplo papel de destaque: distinguindo argumentos numerados e não numerados, independentemente do papel semântico que exerçam na frase; e agrupando verbos em classes conforme seu comportamento sintático, em primeiro lugar, e apenas secundariamente conforme aspectos semânticos. Esta anterioridade da sintaxe sobre a semântica pode ser uma pista para explicar a "semântica difusa" e a dificuldade de generalização deste tipo de anotação quando se trata de argumentos que não têm correlação com uma posição sintática (todos aqueles que não são Arg0 e Arg1).

[29] Segundo Levin & Rappaport Hovav (2005), o compromisso dos papéis semânticos é com classes de argumentos "naturais": "*Each semantic role defines a natural class of arguments, with members of this natural class usually having a common semantic relation to their verbs and shared options for their morphosyntactic expression.*" (Levin & Rappaport Hovav, 2005, p. 36)

Por outro lado, a crítica que fazemos ao papel da sintaxe na construção de representações semânticas não nos leva a defender a eliminação da sintaxe da anotação de papéis semânticos (ou de demais tarefas do PLN). Pelo contrário, trata-se de redimensionar o seu lugar no fluxo linguístico do PLN. A sintaxe pode não ser uma etapa necessária ou inicial no desenvolvimento de sistemas de PLN em geral, ou de papéis semânticos, em particular, mas certamente é uma maneira de organizar os dados linguísticos fundamental para seres humanos criarem conjuntos de dados padrão-ouro com qualidade e agilidade. Classificações sintáticas do tipo "sujeito" e "objeto", como qualquer outra classificação[30], constroem igualdades que permitem generalizações, e por isso facilitam enormemente a tarefa de criação de datasets linguísticos, como mostram as regras de anotação utilizadas na preparação do Porttinari-base PropBank.

Por outro lado ainda, e retomando o interesse do PLN na construção de representações de papel semântico, podemos refletir sobre o objetivo do PropBank de servir como material de treino para o aprendizado de máquina. Apesar da ampla disseminação de modelos estatísticos/neurais, uma representação ao estilo PropBank não perde sua validade se tomada do ponto de vista de regras e analisadores simbólicos. Se, por um lado, a dificuldade de generalização para certas classes de argumentos parece incontornável, por outro, o tipo de representação semântica proposto, da maneira como foi formulado, continua sendo útil ao PLN e, conforme sugerido por alguns estudos e corroborado pelos resultados que apresentamos, talvez possa ser resolvido com regras e um léxico robusto — que ganha ainda mais robustez quando capaz de aproveitar *insights* derivados do uso, isto é, da anotação do corpus, como dados de frequência, que podem ser utilizados para informar analisadores baseados em regras.

9. Considerações finais

Neste artigo, apresentamos a anotação de papéis semânticos ao estilo PropBank incluída no corpus Porttinari-base, que deu origem ao Porttinari-base PropBank (PBP). Como indicam Pardo et al. (2021), mais do que um acrônimo, Porttinari é uma iniciativa que nos lembra dos imensos desafios e contribuições envolvidos na construção deste tipo de recurso, e o PBP acrescenta mais alguns elementos a este rico painel.

No que se refere à prática da concordância, mostramos que, apesar de ter sido feito por uma única pessoa, temos análises consistentes. Porém, a concordância na anotação de propbanks pode dizer pouco. Na concordância entre anotadores, algo que naturalizamos é o fato de a concordância ser calculada sobre *um* fenômeno: concordamos na unicidade. A partir dos resultados de nosso estudo sobre a concordância, defendemos que (i) é possível concordar sobre diferentes coisas ao mesmo tempo e (ii) divergências de análise podem não ser indicativas de erros. A alta divergência nas respostas aliada à experiência da maioria dos anotadores traz questionamentos também para a prática de adjudicação, que prevê uma solução única e correta para instâncias de anotação, e destaca a necessidade de se levar em conta de maneira mais enfática, ao menos no que se refere à anotação de papéis semânticos, a possibilidade de divergências legítimas de interpretação e de anotação (ver também (Basile et al., 2021)).

Como mais um elemento do Porttinari, o PBP se articula com as demais dimensões do projeto. A criação de uma nova interface de acesso ao Verbo-Brasil, que agora permite buscas complexas sobre o seu rico repositório lexical, contribuirá na fase de atualização do recurso. A criação de uma camada dedicada às *Enhanced Dependencies* (de Souza et al., 2024), por sua vez, permitirá a inferência complementar de relacionamentos semânticos, com possibilidades de mapeamentos para os papéis semânticos aqui investigados. Além disso, os papéis semânticos e o Verbo-Brasil devem ser a base para a anotação de mais dados segundo o modelo *Abstract Meaning Representation* (Banarescu et al., 2013), que tem iniciativas relacionadas para o português (por exemplo, o trabalho de Inácio et al. (2023)). A identificação e tipificação de entidades nomeadas (Zerbinati et al., 2024) também devem demonstrar alguma correspondência com os papéis semânticos específicos dos verbos. A construção de um NomBank (da Silva Barbosa, 2024) é diretamente complementar ao esforço feito aqui, focando no papel predicativo dos nomes. Todas essas iniciativas "semânticas" integram, direta ou indiretamente, o projeto maior ao qual o Porttinari se vincula.

Por fim, todos os recursos desenvolvidos ao longo do projeto e mencionados ao longo do artigo estão disponíveis na página do projeto[31] com lincença *Creative Commons* CC-BY.

[30] Para abordagens de viés universalista, no entanto, trata-se de classes especiais, "naturais" e "universais".

[31] https://sites.google.com/icmc.usp.br/poetisa/porttinari-base-propbank

Agradecimentos

Este trabalho foi realizado no âmbito do Centro de Inteligência Artificial da Universidade de São Paulo (C4AI[32]), 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 pelo apoio computacional e preparação dos arquivos para disponibilização, e a Magali Duran pelas discussões científicas e apoio na realização deste trabalho.

Referências

Baker, Collin F., Charles J. Fillmore & John B. Lowe. 1998. The Berkeley FrameNet project. Em *36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics*, 86–90. 10.3115/980845.980860

Banarescu, Laura, Claire Bonial, Shu Cai, Madalina Georgescu, Kira Griffitt, Ulf Hermjakob, Kevin Knight, Philipp Koehn, Martha Palmer & Nathan Schneider. 2013. Abstract Meaning Representation for sembanking. Em *7th Linguistic Annotation Workshop and Interoperability with Discourse*, 178–186.

Basile, Valerio, Michael Fell, Tommaso Fornaciari, Dirk Hovy, Silviu Paun, Barbara Plank, Massimo Poesio & Alexandra Uma. 2021. We need to consider disagreement in evaluation. Em *1st Workshop on Benchmarking: Past, Present and Future*, 15–21. 10.18653/v1/2021.bppf-1.3

Bick, Eckhard. 2007. Automatic semantic role annotation for Portuguese. Em *5th Workshop on Information and Human Language Technology*, 1713–1716.

Bonial, C., Kathryn Conger, Aous Hwang, Jena D.and Mansouri, Yahya Aseri, Julia Bonn, Timothy O'Gorman & Martha Palmer. 2018. Current directions in English and Arabic PropBank. Em Nancy Ide & James Pustejovsky (eds.), *The Handbook of Linguistic Annotations*, Springer. 10.1007/978-94-024-0881-2_27

Bonial, Claire, Julia Bonn, Kathryn conger, Jena Hwang, Martha Palmer & Nicholas Reese. 2015. English PropBank annotation guidelines. Relatório técnico. Institute of Cognitive Science, University of Colorado at Boulder.

Branco, António, Catarina Carvalheiro, Sílvia Pereira, Sara Silveira, João Silva, Sérgio Castro & João Graça. 2012. A PropBank for Portuguese: the CINTIL-PropBank. Em *8th International Conference on Language Resources and Evaluation (LREC)*, 1516–1521.

Carreras, Xavier & Lluís Màrquez. 2005. Introduction to the CoNLL-2005 shared task: Semantic role labeling. Em *9th Conference on Computational Natural Language Learning (CoNLL)*, 152–164.

de Souza, Elvis & Cláudia Freitas. 2021. ET: A workstation for querying, editing and evaluating annotated corpora. Em *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 35–41. 10.18653/v1/2021.emnlp-demo.5

Dong, Xin Luna. 2023. Generations of knowledge graphs: The crazy ideas and the business impact. Em *Very Large Data Bases Conference*, 4130–4137. 10.14778/3611540.3611636

Duran, Magali, Lucelene Lopes, Maria das Graças Nunes & Thiago Pardo. 2023. The dawn of the Porttinari Multigenre Treebank: Introducing its journalistic portion. Em *14th Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana*, 115–124. 10.5753/stil.2023.233975

Duran, Magali Sanches. 2014. Manual de anotação do PropBank-Br v2. Relatório técnico. Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo

Duran, Magali Sanches & Sandra Aluísio. 2015. Automatic generation of a lexical resource to support semantic role labeling in Portuguese. Em *4th Joint Conference on Lexical and Computational Semantics*, 216–221. 10.18653/v1/S15-1026

Duran, Magali Sanches & Sandra Maria Aluísio. 2011. Propbank-Br: a Brazilian Portuguese corpus annotated with semantic role labels. Em *8th Brazilian Symposium in Information and Human Language Technology (STIL)*,

Duran, Magali Sanches & Cláudia Freitas. 2024. Guia de anotação de papéis semânticos seguindo o modelo PropBank no corpus Porttinari-base. (no prelo). Relatório técnico. Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo

[32] http://c4ai.inova.usp.br/

Duran, Magali Sanches, Lianet Sepúlveda Torres, Marina Coimbra Viviani, Nathan Hartmann & Sandra Maria Aluísio. 2014. Seleção e preparação de sentenças do corpus PLN-BR para compor o corpus de anotação de papéis semânticos Propbank-Br.v2. Relatório técnico. Núcleo Interinstitucional de Linguística Computacional

Fillmore, Charles. 1968. Em favor do caso. Em Maria Pinheiro Lobato (ed.), *A semântica na lingüística moderna: o léxico*, Francisco Alves

Freitas, Cláudia & Thiago Pardo. 2024. PropBank e anotação de papéis semânticos para a língua portuguesa: O que há de novo? Em *15th Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana (STIL)*, 118–128. 10.5753/stil.2024.245377

Freitas, Cláudia & Elvis de Souza. 2024. Sujeito oculto às claras: uma abordagem descritivo-computacional. *Revista de Estudos da Linguagem* 29(2). 1033–1058. 10.17851/2237-2083.29.2.1033-1058

Freitas, Cláudia, Elvis Souza, Maria Clara Castro, Tatiana Cavalcanti, Patricia Ferreira da Silva & Fábio Corrêa Cordeiro. 2023. Recursos linguísticos para o PLN específico de domínio: o Petrolês. *Linguamática* 15(2). 51–68. 10.21814/lm.15.2.412

Gung, James & Martha Palmer. 2021. Predicate representations and polysemy in VerbNet semantic parsing. Em *14th International Conference on Computational Semantics (IWCS)*, 51–62.

Hajič, Jan, Massimiliano Ciaramita, Richard Johansson, Daisuke Kawahara, Maria Antònia Martí, Lluís Màrquez, Adam Meyers, Joakim Nivre, Sebastian Padó, Jan Štěpánek, Pavel Straňák, Mihai Surdeanu, Nianwen Xue & Yi Zhang. 2009. The CoNLL-2009 shared task: Syntactic and semantic dependencies in multiple languages. Em *13th Conference on Computational Natural Language Learning (CoNLL): Shared Task*, 1–18.

Hartmann, Nathan Siegle, Magali Sanches Duran & Sandra Maria Aluísio. 2016. Automatic semantic role labeling on non-revised syntactic trees of journalistic texts. Em *Computational Processing of the Portuguese Language (PROPOR)*, 202–212. 10.1007/978-3-319-41552-9_20

Hovy, Eduard, Mitchell Marcus, Martha Palmer, Lance Ramshaw & Ralph Weischedel. 2006. OntoNotes: The 90% solution. Em *Human Language Technology Conference of the NAACL*, 57–60.

Inácio, Márcio Lima, Marco Antonio Sobrevilla Cabezudo, Renata Ramisch, Ariani Di Felippo & Thiago Pardo. 2023. The AMR-PT corpus and the semantic annotation of challenging sentences from journalistic and opinion texts. *DELTA: Documentação e Estudos em Linguística Teórica e Aplicada* 39(3). 1–31. 10.1590/1678-460X202339355159

Kipper, Karin, Anna Korhonen, Neville Ryant & Martha Palmer. 2008. A large-scale classification of English verbs. *Language Resources and Evaluation* 42. 21–40. 10.1007/s10579-007-9048-2

Levin, Beth. 1993. *English verb classes and alternations: a preliminary investigation*. University of Chicago Press

Levin, Beth & Malka Rappaport Hovav. 2005. *Argument realization*. Cambridge University Press. 10.1017/CBO9780511610479

de Marneffe, Marie-Catherine, Christopher D Manning, Joakim Nivre & Daniel Zeman. 2021. Universal dependencies. *Computational Linguistics* 47(2). 255–308. 10.1162/coli_a_00402

Merlo, Paola & Lonneke Van Der Plas. 2009. Abstraction and generalisation in semantic role labels: PropBank, VerbNet or both? Em *Joint Conference of the 47thh Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing AFNLP*, 288–296.

Mohebbi, Majid, Seyed Naser Razavi & Mohammad Ali Balafar. 2022. Computing semantic similarity of texts based on deep graph learning with ability to use semantic role label information. *Scientific Reports* 12(1). 14777. 10.1038/s41598-022-19259-5

Palmer, Martha, Daniel Gildea & Paul Kingsbury. 2005. The proposition bank: An annotated corpus of semantic roles. *Computational Linguistics* 31(1). 71–106. 10.1162/0891201053630264

Pardo, Thiago, Magali Duran, Lucelene Lopes, Ariani Felippo, Norton Roman & Maria Nunes. 2021. Porttinari - a large multi-genre treebank for Brazilian Portuguese. Em *13th Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana (STIL)*, 1–10. 10.5753/stil.2021.17778

Peeters, Bert. 2000. Setting the scene: Some recent milestones in the lexicon-encyclopedia debate. Em *The Lexicon-Encyclopedia Interface*, Elsevier

Pradhan, Sameer, Julia Bonn, Skatje Myers, Kathryn Conger, Tim O'gorman, James Gung, Kristin Wright-bettner & Martha Palmer. 2022. PropBank comes of Age—Larger, smarter, and more diverse. Em *11th Joint Conference on Lexical and Computational Semantics (*SEM)*, 278–288. 10.18653/v1/2022.starsem-1.24

Pradhan, Sameer, Alessandro Moschitti, Nianwen Xue, Olga Uryupina & Yuchen Zhang. 2012. CoNLL-2012 shared task: Modeling multilingual unrestricted coreference in OntoNotes. Em *Joint Conference on EMNLP and CoNLL - Shared Task*, 1–40.

Saeed, John I. 2007. *Semantics*. Wiley-Blackwell

Salomão, Maria Margarida Martins, Tiago Timponi Torrent & Thais Fernandes Sampaio. 2013. A linguística cognitiva encontra a linguística computacional: notícias do projeto Framenet Brasil. *Cadernos de Estudos Linguísticos* 55(1). 7—-34. 10.20396/cel.v55i1.8636592

Santos, Diana & Cristina Mota. 2010. Experiments in human-computer cooperation for the semantic annotation of Portuguese corpora. Em *7th International Conference on Language Resources and Evaluation (LREC)*, 1437–1444.

da Silva Barbosa, Bryan Khelven. 2024. *Descrição sintático-semântica de nomes predicadores em tweets do mercado financeiro em Português*: Universidade Federal de São Carlos. Tese de Mestrado.

de Souza, Elvis, Magali Duran, Maria das Graças Nunes, Gustavo Sampaio, Giovanna Belasco & Thiago Pardo. 2024. Automatic annotation of enhanced universal dependencies for Brazilian Portuguese. Em *15th Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana (STIL)*, 217–226. 10.5753/stil.2024.245342

Stowe, Kevin, Jenette Preciado, Kathryn Conger, Susan Windisch Brown, Ghazaleh Kazeminejad, James Gung & Martha Palmer. 2021. SemLink 2.0: Chasing lexical resources. Em *14th International Conference on Computational Semantics (IWCS)*, 222–227.

Wang, Nan, Jiwei Li, Yuxian Meng, Xiaofei Sun, Han Qiu, Ziyao Wang, Guoyin Wang & Jun He. 2022. An MRC framework for semantic role labeling. Em *29th International Conference on Computational Linguistics (COLING)*, 2188–2198.

Yi, Szu-ting, Edward Loper & Martha Palmer. 2007. Can semantic roles generalize across genres? Em *Human Language Technologies: The Conference of the North American Chapter of the Association for Computational Linguistics*, 548–555.

Zerbinati, Michel Monteiro, Norton Trevisan Roman & Ariani Di Felippo. 2024. A corpus of stock market tweets annotated with named entities. Em *16th International Conference on Computational Processing of Portuguese (PROPOR)*, 276–284.

A. Apêndice: frases utilizadas no estudo de concordâncias

A.1. Frases dos formulários 1 e 2

s.1 Relata sua amizade com Janot e afirma que o ex-procurador-geral chamava Dodge de "bruxa" em **conversas reservadas**

s.2 A proposta **foi aprovada** em 2011, mas até hoje não decolou. (MONOSSÊMICA)

s.3 A tese é de que os insatisfeitos migram para os partidos populistas, como a Frente Nacional, que teve 34 % dos votos **nas eleições presidenciais francesas**.

s.4 **Nesse encontro** , encerra-se oficialmente o mandato do senador Aécio Neves

s.5 Os cartolas locais foram ajudados pela vitória **nas duas últimas edições da Copa América**.

s.6 **Em encontro**, Bono pergunta a Macri sobre argentino desaparecido

s.7 Não consigo deixar de amar essa magnífica e adorável figura paterna, mas está puxadíssimo ter prazer **em qualquer conversa**

s.8 **Nesses encontros**, o governador adota tom crítico à gestão de Temer em assuntos como privatizações, reformas econômicas e segurança.

s.9 É uma grande honra receber uma proposta de um grande clube como esse , mas também é uma grande honra estar aqui no Liverpool, um grande clube, falou **em entrevista à ESPN**

s.10 Livres que somos, nós estamos dispostos a defender nossa soberania e nossa democracia **em qualquer cenário** e modalidade

s.11 Ele disse que eu agora vivo de pijamas **em casa**. (MONOSSÊMICA)

s.12 **Em alguns casos**, manteve-se o caráter vitalício do poder. anotação PBP

s.13 Hernanes foi o primeiro jogador do São Paulo a falar **na discutida reunião com as torcidas uniformizadas da última quarta-feira**.

s.14 O sistema serve , por exemplo , para a empresa analisar **em quais situações** o produto é fotografado e usar esse tipo de informação para referenciar campanhas futuras.

s.15 **No final da reunião**, surpreendentemente, o Anselmo passa a adotar um comportamento mais duro, pressionando para delação.

s.16 Saí **derrotada** pouco depois das 14h. (MONOSSÊMICA)

A.2. Frases dos formulários 3 e 4

s.1 **Com o quarto mandato**, ela deve chegar aos 16 anos de poder. anotação

s.2 Nós respeitamos a PGR e temos a soberania de decidir **por maioria**. (MONONSSÊMICO)

s.3 **Na falta de notícias**, o pessoal acaba focando também em eleição

s.4 **Em um cenário extremo**, os bancos centrais temem que poderiam até perder o controle da base monetária

s.5 Quem é que sai **para trabalhar** pensando em tomar um soco na cara ?

s.6 Em alguns roteiros, como na Croácia, é preciso pagar **na maioria dos trajetos**.

s.7 No momento em que a nova moeda for criada, se os preços forem convertidos **no pico**, não será necessário um aperto monetário, uma recessão temporária?

s.8 Ou seja: o Emmy de 2017 foi politizado **do começo ao fim**.

s.9 **Em homenagem a Che Guevara**, Cuba reage às declarações de Trump.

s.10 Eles escreveram **na casa** a frase "SP não está à venda" (MONOSSÊMICO)

s.11 **Com mais capital privado**, a Sabesp poderá conquistar mercados hoje explorados pelas empresas privadas

s.12 **Com a liderança caindo em seu colo**, Lewis Hamilton ditou as ações da corrida.

s.13 A equipe gaúcha perdeu a oportunidade de manter a distância para o líder **ao ser derrotada pela Chapecoense por 1 a 0**, em casa.

s.14 Como avançar soluções de amplo espectro **nesse cenário**, com nosso sistema político-eleitoral?

s.15 **Para se livrar do problema**, dobrou a aposta e injetou mais R$ 2 bilhões na JBS

s.16 **Naquele ano**, o tucano tinha 30 % de aprovação (MONOSSÊMICO)