UNIVERSIDADE DE SÃO PAULO

Instituto de Ciências Matemáticas e de Computação

ISSN 0103-2569

ANOTAÇÃO DE PAPÉIS SEMÂNTICOS NO CORPUS PORTTINARI-BASE: PROCEDIMENTOS, RESULTADOS E ANÁLISES

CLÁUDIA FREITAS

Nº 450

RELATÓRIOS TÉCNICOS

São Carlos – SP

Dez./2024

Natural Language Processing initiative (NLP2) of the Center for Artificial Intelligence (C4AI) of the University of São Paulo, sponsored by IBM and FAPESP

POeTiSA
POrtuguese processing – Towards Syntactic Analysis and parsing

Anotação de papéis semânticos no *corpus Porttinari-base*: Procedimentos, resultados e análises

Cláudia Freitas

Projeto: Anotação de Papéis Semânticos
Coordenação: Thiago Pardo

Dezembro 2024

Agradecimentos

Este trabalho foi realizado no âmbito do Centro de Inteligência Artificial da Universidade de São Paulo (C4AI), 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.

Agradeço também a Elvis de Souza pela preparação dos arquivos para disponibilização, e a Magali Duran pelos comentários a versões anteriores deste documento.

Sumário

1. Apresentação e enquadramento

Dentre os métodos utilizados para *representar* computacionalmente informação semântica em textos está a **anotação de papéis semânticos** ou SRL (*semantic role labeling*). Papéis semânticos são responsáveis por indicar *quem* fez *o quê* para *quem, onde*, *quando*, *como*, *por que, para que, com que, com quem* etc, e assim estruturam de maneira *explícita* e *interpretável* a informação contida em enunciados linguísticos. De maneira simplificada, papéis semânticos **classificam** semanticamente as entidades participantes de um evento – de forma mais precisa, as entidades que participam de uma *proposição* – e a ***relacionam*** a um predicado. Por isso, papéis semânticos também podem ser utilizados na construção de grafos de conhecimento, que são uma maneira flexível e intuitiva de modelar relações complexas entre dados.

A figura 1 apresenta uma representação da frase "Carlos andou 150 km a pé para chegar ao México após ser ameaçado por bandidos" conforme a anotação de papéis semânticos.

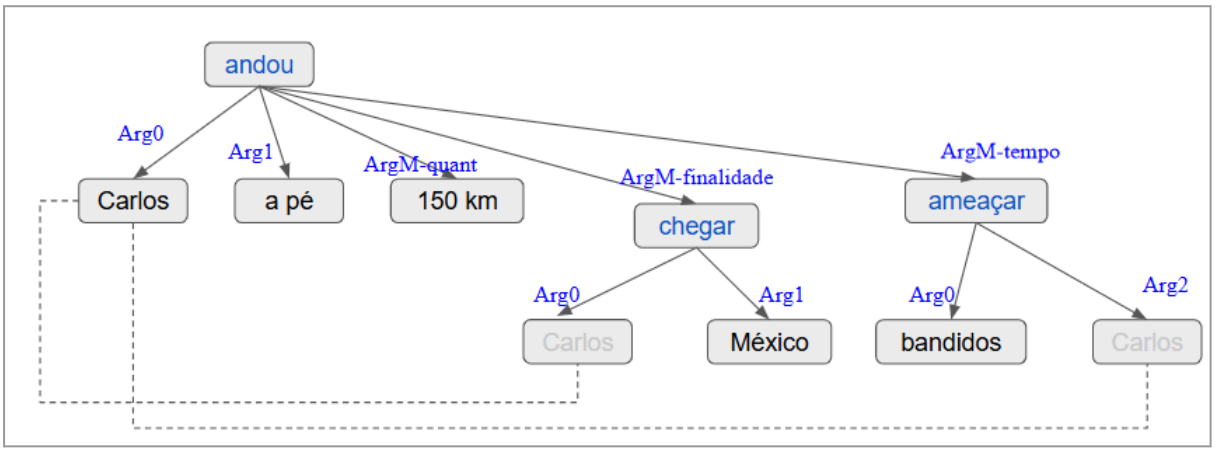


Figura 1: Representação semântica da frase "Carlos andou 150 km a pé para chegar ao México após ser ameaçado por bandidos"

A anotação de papéis semânticos é considerada um passo adiante na sintaxe rumo ao significado dos textos, e possibilita o aprendizado de *representações semânticas estáveis* ao longo de diferentes contextos.

Apesar da ampla disseminação dos métodos neurais no PLN/IA, a falta de transparência e de interpretabilidade, além das alucinações, são um gargalo a ser resolvido, e uma tendência atual é investigar formas de combinar representações explícitas de conhecimento e arquiteturas neurais. Nesse contexto, estruturas semânticas explícitas, como os papéis semânticos, oferecem maneiras interpretáveis de representar semanticamente enunciados linguísticos. Uma alternativa nesse caminho é conectar LLMs com fontes de conhecimento estruturado, como, por exemplo, grafos de conhecimento neurais (knowledge graph neural network) (Manning, 2022). Neste cenário, a possibilidade de criar grafos a partir da anotação SRL (por exemplo, Mohebbi et al., 2022) torna este tipo de representação semântica relevante para a investigação acerca da articulação entre LLMs e fontes de conhecimento estruturado, ampliando também as possibilidades de utilização de papéis semânticos.

Apesar de sua relevância, não há, para a língua portuguesa, conjuntos de dados com as características necessárias que permitam um pleno aprendizado deste tipo de classificação, o que justifica o presente projeto. Até o momento, o maior conjunto de dados disponível continha cerca de 7 mil classes anotadas; o material aqui apresentado contém mais de 43 mil.

Este relatório apresenta o processo e os resultados preliminares da anotação de papéis semânticos no *treebank* Porttinari-base. O *treebank* Porttinari (Pardo et al., 2021; Duran et al., 2023) é um dos recursos criados no âmbito do projeto **POeTiSA**[1] (*POrtuguese processing - Towards Syntactic Analysis and parsing*), uma das iniciativas do **C4AI**[2], *Centro de Inteligência Artificial da Universidade de São Paulo*, centro de pesquisa avançada em IA resultado da colaboração entre IBM e FAPESP.

O objetivo do projeto **POeTiSA** é a ampliação de recursos baseados na sintaxe e o desenvolvimento de ferramentas e aplicações para a língua portuguesa, procurando alcançar resultados estado-da-arte na área. A inclusão de papéis semânticos no corpus Porttinari é mais um desdobramento deste projeto: mais um recurso padrão-ouro que disponibilizamos à comunidade de PLN interessada no processamento computacional da língua portuguesa. Ao longo do projeto, dois outros recursos foram sendo desenvolvidos ou aperfeiçoados: (i) uma interface alternativa para o recurso Verbo-Brasil, que agora permite buscas complexas sobre o seu rico repositório lexical; (ii) uma ferramenta de anotação com base em regras capaz de otimizar a criação de datasets padrão ouro que utilizem ou partam de anotação sintática dependencial.

O corpus **Porttinari** (cujo nome vem de *PORTuguese Treebank*) é um *treebank* composto por textos de diferentes gêneros e domínios, como texto jornalístico, conteúdo gerado pelo usuário (tweets, resenhas e comentários), transcrições de fala e literatura. No entanto, como indicam Pardo et al. (2021), mais do que um acrônimo para um *treebank*, Porttinari é uma iniciativa que nos lembra dos imensos desafios e contribuições envolvidos na construção deste tipo de recurso no contexto do processamento computacional da língua portuguesa.

O Porttinari foi criado inicialmente como um *treebank* multigênero. O trabalho apresentado neste relatório acrescenta mais um elemento a este rico painel, relatando o processo de adição de uma camada de papéis semânticos ao material, o que faz dele, além de multigênero, um corpus multicamadas.

A porção jornalística do Porttinari divide-se em três subcorpora, cada um com suas características e finalidades (Duran et al., 2023):

> (i) *Porttinari-base*, intensa e cuidadosamente revisto, criado com o objetivo de ser material padrão ouro no que se refere à anotação de dependências sintáticas;

[1] https://sites.google.com/icmc.usp.br/poetisa
[2] https://c4ai.inova.usp.br/

(ii) *Porttinari-check*, um pequeno corpus com estrutura similar à do *Porttinari-base*, criado com o objetivo de ser um espaço de testes (testbed) e ilustrar o contraste entre anotação humana e automática;
(iii) *Porttinari-automatic*, um corpus amplo anotado automaticamente

Tradicionalmente, a anotação de papéis semânticas é feita sobre árvores sintáticas, e quanto maior a qualidade da análise sintática subjacente, mais facilitado o trabalho inicial humano de anotação dos papéis semânticos. Por isso, a adição de papéis semânticos se deu sobre a porção *Porttinari-base*, que passa a ser duplamente padrão ouro – nas dependências sintáticas e nos papéis semânticos – contribuindo para a disponibilização de recursos de alto nível para o PLN de língua portuguesa e, especificamente no que se refere aos papéis semânticos, dando seguimento ao trabalho de Duran & Aluisio (2012; 2013) e Duran et al. (2014). Assim, *Porttinari-base Propbank* – **PBP** – refere-se à anotação de papéis semânticos ao estilo PropBank (Palmer et al., 2005)[3] no *treebank Porttinari-base*, uma amostra do corpus Porttinari cuidadosamente revista por linguistas.

O material foi inicialmente anotado levando em conta o Manual de anotação do PropBank v2 (Duran, 2014) e os frames verbais elencados no recurso Verbo-Brasil (Duran & Aluísio, 2015). Ao longo do projeto, uma versão atualizada e ampliada das diretrizes linguísticas utilizadas na anotação do Porttinari foi produzida (Duran & Freitas, 2024).

A anotação de papéis semânticos pode assumir duas principais formas ou estilos: o estilo PropBank (Palmer et al., 2005) e o estilo VerbNet (Kipper et al., 2006). As diferentes maneiras de codificar papéis semânticos não são excludentes, pelo contrário, e trabalhos recentes (por exemplo, Li et al., 2023) têm investido em maneiras (automáticas) de tirar proveito de ambas as formas. A anotação do PBP seguiu o modelo PropBank, que será escrutinado nas próximas seções.

Do **ponto de vista da tarefa**, a anotação de papéis semânticos ao estilo Propbank se desdobra em duas tarefas ou etapas: (i) desambiguação dos sentidos dos verbos (por meio da consulta ao Verbo-Brasil) e (ii) anotação de argumentos. A etapa (i) determina o sentido específico de um predicador em uma frase, e a etapa (ii) identifica os argumentos do predicador e os anota com o papel semântico correspondente. Tradicionalmente, as duas etapas são tratadas separadamente, sendo comum a "eliminação" da primeira etapa por meio da incorporação de predicadores já padrão-ouro. Mas esta não é uma exigência, e trabalhos como He et al. (2018) tratam as duas etapas simultaneamente, com bons resultados.

1.1. PropBanks e papéis semânticos

Um Propbank (*Proposition Bank*) é, literalmente, um *Banco de Proposições*. Simplificadamente, podemos entender, por *proposição*, uma frase do ponto de vista do conteúdo que ela veicula. No PropBank, o conceito de proposição é tomado da proposta de

[3] Atualmente, o projeto PropBank está descrito e alocado em http://PropBank.github.io e https://github.com/PropBank.

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[4].

Papel semântico (também chamado de *papel temático* ou *papel theta*) refere-se à maneira pela qual uma entidade está envolvida em uma proposição, isto é, diz respeito ao papel (semântico) que a entidade exerce na frase com relação ao verbo.

Como mencionado, a anotação de papéis semânticos possibilita o aprendizado de *representações semânticas estáveis* ao longo de diferentes contextos. As frases 1 a 5 ilustram esta ideia, trazendo diferentes realizações do verbo "abrir". De um ponto de vista sintático, "porta" exerce diferentes funções (objeto direto em 1 e 2; sujeito de voz passiva em 3, sujeito em 4 e 5) e o mesmo acontece com "chave" (sujeito em 1 e adjunto adverbial em 2 e 3). Na atribuição de papéis semânticos, "porta" é a "coisa abrindo" em todas as 5 frases, e "chave" é o "instrumento de abertura" em todas as 5 frases, não importa a função sintática que exerçam. Na frase "O tempo abriu no feriado" teríamos uma outra estrutura argumental (e outra representação semântica), já que estamos diante de um outro sentido de "abrir".

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

Quanto aos papéis semânticos, o verbo "abrir" dos exemplos 1-5 pode ser associado aos papéis de *agente* (quem abre), *tema* (a coisa aberta), *instrumento* (com o que a coisa é aberta) e *beneficiário* (para quem a coisa é aberta), que podem ou não estar presentes na frase.

Apesar de ser um fenômeno linguístico amplamente estudado, **não há consenso relativo ao conjunto de papéis semânticos** da língua, com propostas que vão de 5 a mais de 30 tipos diferentes.

1.1.1. A taxonomia de argumentos no PropBank e o recurso lexical associado

No PropBank, a diversidade acerca dos papéis semânticos é contornada com a utilização de i) argumentos numerados, que vão de A0 a A5 e consistem em um pequeno conjunto de etiquetas genéricas, ou etiquetas abstratas; e ii) argumentos modificadores, um conjunto mais amplo de argumentos. Os papéis são atribuídos para cada par verbo-argumento individualmente. A motivação para o conjunto limitado de papéis numerados no PropBank é facilitar a generalização para o aprendizado de máquina. O quanto esta escolha tem sido bem sucedida é o tema da seção 1.4.

[4] No entanto, como indica Palmer et al. (2005), não devemos confundir os frames semânticos da FrameNet de Fillmore com os frames sintáticos do PropBank.

De maneira mais detalhada, argumentos numerados (argumentos *core*) refletem os argumentos que, ou são exigidos pela valência (ou estrutura de subcategorização) de um predicado ou, se não exigidos, são aqueles que ocorrem com uma alta frequência no uso comum[5]. Estes argumentos são considerados elementos especificados sintaticamente e exigidos pelo núcleo, enquanto os argumentos não numerados (tempo, lugar, etc.) podem modificar livremente um núcleo, estando sujeitos apenas a restrições de compatibilidade semântica.

A diferença entre argumentos numerados e argumentos modificadores está sobretudo na natureza da relação sintática que o argumento mantém com o verbo (*exigência*, nos argumentos numerados, vs *opcionalidade*, nos argumentos modificadores). A distinção entre os argumentos também é motivada pela assistematicidade semântica dos papéis com relação aos verbos, e por isso argumentos numerados são específicos de verbos (o Arg0 do verbo "abrir" é "quem abre", mas o Arg0 de "fechar" é quem fecha; o Arg0 de "gostar" é "aquele que gosta" e o Arg0 de "alagar" é "causador do alagamento") . Os ArgM, por outro lado, têm uma semântica específica e previsível (indicada pelo nome da etiqueta), e podem se associar a qualquer verbo.

Voltando aos exemplos 1-5 com o verbo "abrir": também podemos indicar "quando" a coisa é aberta, "com que finalidade" a coisa é aberta, "onde" a coisa foi aberta, ou "por que motivo" a coisa foi aberta, por exemplo. Essas outras informações ("quando", "para que", "onde", "por que" etc), porque não são consideradas exigências *desses* verbos (e portanto não participam da entrada lexical de tal verbo), recebem, na anotação PropBank, os argumentos modificadores ArgM-tmp (para "tempo"), ArgM-prp (para "finalidade"), ArgM-loc, (para "local"), ArgM-cau (para "causa"), respectivamente. No entanto, não é o tipo semântico do argumento que define seu status como argumento numerado ou modificador, mas o tipo de relação sintática entre o verbo e seu complemento.

A semântica subjacente aos argumentos numerados é "revelada" com o alinhamento entre a anotação do corpus e o recurso lexical associado ao PropBank, os chamados *Frame Files* – em nosso caso, dispomos do **Verbo-Brasil** (Duran & Aluisio, 2015). Assim, as frases (1), (2) e (3), conforme a anotação de um PropBank, recebem as seguintes etiquetas:

1. A chave[Arg2] **abriu** a porta [Arg1].
2. Ela [Arg0] **abriu** a porta [Arg1] com a chave[Arg2].
3. A porta [Arg1] **foi aberta** com a chave [Arg2].

e, do mesmo modo,

Ela[Arg0] **comprou** um carro[Arg1]
Ela[Arg0] **gosta** de sorvete [Arg1]
A chuva[Arg0] **alagou** os bairros [Arg1]

[5] "Numbered arguments reflect either the arguments that are required for the valency of a predicate (e.g., agent, patient, benefactive), or if not required, those that occur with high-frequency in actual usage" (Bonial et al., 2015:1).

Um PropBank, portanto, não é apenas um corpus anotado, mas a associação entre um corpus anotado e um léxico, que indica como os elementos devem ser anotados. Para a representação semântica completa da frase (2), isto é, para decodificar os valores semânticos de Arg0, Arg1 e Arg2, é necessário consultar a representação do verbo "abrir" no recurso lexical associado, e lá vemos que o Arg0 de abrir é o "agente da abertura"; que o Arg1 de abrir é a "coisa aberta", que o Arg2 é o "instrumento da abertura" e que o Arg3 (não expresso nos exemplos) é o "beneficiário". Por isso, embora o corpus anotado seja a parte visível do PropBank, não é exagero afirmar que o coração do projeto de anotação é o seu recurso léxico. O corpus anotado é "apenas" a exemplificação do que está codificado no léxico, e sem o léxico fazemos pouco, pois os argumentos numerados têm sentido variável.
Segundo Pradhan et al., (2022:279) "*The core of the PropBank paradigm consists of an annotation schema and a lexical inventory collectively referred to as the Frames"*.

1.1.2. Argumentos, estrutura argumental, subcategorização e valência

Argumento, neste contexto linguístico, refere-se ao elemento/expressão que ajuda a completar o significado de um predicado (verbo, nome ou adjetivo). Os argumentos, ainda nesse contexto, são elementos considerados necessários para completar o significado dos predicados/verbos; os adjuntos são considerados elementos opcionais para completar o significado do predicado/verbo. Já a **estrutura argumental** de um verbo corresponde à **quantidade** de argumentos que ele pode (opcionalidade/argumentos não numerados) ou deve (obrigatoriedade/argumentos numerados) tomar.

Para a linguística, o interesse na estrutura argumental dos verbos está associado ao interesse na caracterização do **léxico** mental[6] – especificamente, parte-se da ideia de que o léxico de uma língua contém informação sobre quais argumentos são indispensáveis e quais são facultativos em um enunciado, e a não realização da estrutura (argumental) inviabilizaria a interpretação do enunciado e levaria a frases mal formadas. Assim, um argumento nuclear/obrigatório/*core* de um item lexical é uma propriedade estável deste item no léxico mental.

Já o conceito de **subcategorização** diz respeito ao **ambiente sintático** no qual um verbo ocorre, sinalizando a exigência ou não de argumentos, bem como os tipos de argumentos sintáticos com os quais os núcleos co-ocorrem. Por exemplo, o verbo "pensar" exige um *sujeito* com características humanas, diferentemente do verbo "chover". Ainda segundo certas teorias linguísticas, informações de subcategorização fariam parte da entrada lexical de cada verbo (no léxico mental), uma vez que seriam informação considerada essencial[7].

No ambiente de um Propbank, o léxico corresponde à informação codificada nos *Frame Files,* e especificamente neste projeto, às informações codificadas no Verbo-Brasil.

[6]Lembramos que, desde seu estabelecimento como ciência autônoma, a Linguística sempre esteve preocupada, em primeiro lugar, com a linguagem humana enquanto fenômeno mental/cognitivo, o que não necessariamente se coaduna com as preocupações e interesses do PLN.
[7] 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).

O conceito de **subcategorização** se assemelha ao de **valência**, ainda que cada um venha de diferentes tradições nos estudos da linguagem. Frames de **subcategorização** têm origem em gramáticas de estrutura sintagmática (*phrase structure grammars*) dentro da tradição chomskyana (Chomsky, 1965) e integram abordagens como Head-Driven Phrase Structure Grammar, a Gramática Léxico-Funcional e o Minimalismo. A **valência** tem origem na abordagem de Lucien Tesnière, e integra Gramáticas de Dependências. A principal diferença entre *subcategorizaçao* e *valência* está relacionada ao status do sujeito: originalmente, a subcategorização não incluía o sujeito, ou seja, um verbo subcategorizava-se pelo seu complemento (objeto e argumentos oblíquos). Atualmente, é comum que o sujeito faça parte do frame de subcategorização de um verbo, como vemos na abordagem do Propbank. A noção de valência, por sua vez, desde sempre incluiu o sujeito.

1.1.3. Os datasets PropBank

O primeiro PropBank foi criado para a língua inglesa (Palmer et al., 2005). Desde então, PropBanks para diferentes línguas têm sido produzidos, em maior ou menor alinhamento com o inventário de papéis proposto pelo recurso original.

Um dos mais utilizados datasets com anotação de papéis semânticos ao estilo PropBank é o Ontonotes v5.0 (Weischedel et al., 2013), um corpus multicamadas anotado com informações sintáticas, de estrutura argumental (papéis semânticos), correferência e alinhamento de entre sentidos das palavras e uma ontologia (este último, não exaustivo), e 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 and Màrquez, 2005), do CoNLL2009 (Hajic 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 português. Para a nossa língua, temos o Propbank-br (Duran & Aluísio, 2012), o Propbank-br v2., utilizado em Hartmann et al., 2014 e, por fim, o corpus aqui apresentado, que herda decisões, diretivas e recursos lexicais criados ao longo da elaboração dos PropBanks anteriores.

1.1.3.1. Outras anotações de papéis semânticos: datasets VerbNet

Corpora com anotações de papéis semânticos mais genéricos, baseados no inventário da VerbNet (Kipper et al. 2006), também têm surgido para línguas como alemão (Mújdricza-Maydt et al. 2016), holandês (Monachesi et al. 2007), e Dinamarquês (Bick, 2017), por exemplo. Existem ainda corpora que contêm ambas as anotações, alinhadas, como o AnCora, criado para o Catalão e o Espanhol (Taulé et al., 2008) e corpora com representações aparentadas, como a camada de papéis “tectogramaticais” do Prague Czech-English Dependency Treebank 2.0 (Mikulová et al., 2005).

A anotação ao estilo VerbNet parte da taxonomia da VerbNet, um léxico de verbos organizado hierarquicamente, independente de domínio e de ampla cobertura[8]. As classes

[8] https://verbs.colorado.edu/verbnet/

da VerbNet ampliam as classes originais de Levin (1993) por meio do refinamento e da adição de subclasses para obter coerência sintática e semântica entre os membros de uma classe.

PropBank e VerbNet divergem quanto ao nível de granularidade dos papéis semânticos. O PropBank é menos granular, com 6 etiquetas (considerando apenas os argumentos numerados), enquanto a VerbNet traz 23 etiquetas (considerando apenas os "argumentos core" ou "argumentos numerados", conforme a terminologia PropBank). Além disso, a VerbNet, enquanto *léxico*, não lista/descreve os complementos opcionais associados aos verbos (os argumentos modificadores), devido ao seu caráter acidental e uso idiossincrático. O PropBank compartilha o mesmo pressuposto mas o faz de maneira moderada, apenas distinguindo os tipos de argumentos, mas sem deixar de anotá-los em uma frase.

1.1.4. PropBanks e anotação de papéis semânticos para a língua portuguesa

A língua portuguesa, como já mencionado, desde 2012 pode contar com o PropBank-Br (Duran & Aluísio, 2012), que contém papéis semânticos ao estilo PropBank sobre a parte brasileira do treebank Bosque, em sua versão de sintaxe de constituintes disponibilizada pela Linguateca. Este PropBank levou à formulação (e adaptação do inglês) de diretivas de anotação e permitiu a criação do recurso léxico que codifica os sentidos dos verbos e descreve seus frames sintáticos, o Verbo-Brasil (Duran & Aluísio, 2015). No entanto, devido à limitação de tamanho, o material, composto por 4213 frases e com 7107 argumentos anotados (quase mil associadas ao verbo ser), não permite um pleno aprendizado. A anotação seguiu os moldes do PropBank original: foi feita de maneira linear, frase a frase. Para além do recurso PropBank propriamente, a experiência de anotação possibilitou o estudo e posterior confecção de uma tabela de verbos auxiliares e seus padrões sintáticos de ocorrência, bem como de diretivas de anotação, que foram a base da anotação conduzida neste projeto.

Pouco tempo depois, a construção de um PropBank para a língua portuguesa foi retomada pela mesma equipe, tendo em vista a criação de um material maior e mais lexicalmente diversificado, dedicado especificamente para o aprendizado de máquina (Duran et al., 2014). Diferentemente da versão anterior, este material – chamado PropBank-BR v2 – foi construído sobre árvores sintáticas não revistas, e possui 7442 frases/7661 instâncias manualmente anotadas com rótulos de papéis semânticos (Hartman et al., 2016)[9].

A língua portuguesa conta ainda com o CINTIL-PropBank (Branco et al., 2012), um corpus de frases anotadas com estrutura de constituintes e papéis semânticos, composto de 10.039 frases e 110.166 tokens. O CINTIL-PropBank foi criado de maneira semiautomática, com etiquetas dos argumentos numerados anotadas automaticamente pelo módulo de análise gramatical do parser usado, e com um conjunto de papéis semânticos que é uma adaptação dos argumentos numerados de Palmer et al. (2005).

[9] Tanto o PropBankv1 como o PropBank v2 estão disponíveis em http://143.107.183.175:21380/portlex/index.php/en/downloadsingl

Em uma abordagem baseada em regras, Bick (2007) propõe um anotador SRL seguindo a abordagem da *Constraint Grammar*. Seu anotador contém cerca de 500 regras gerais e um pequeno conjunto de regras de desambiguação. Os papéis semânticos estão distribuídos em 35 papéis temáticos[10], e o artigo que descreve o sistema de anotação relata F-score de 88.6%. O trabalho também mediu o desempenho por classe de argumento, e as classes com F-score abaixo de 80% foram ADV (*dummy adverb*, que compreende orações adverbiais introduzidas por gerúndio - 72.2%) DES (*destino* 75.9%). BEN (*beneficiário* - 76.2%), RES (*resultado* -78.4%), CAU (*causa* - 78.8%), LOC-TMP (usada para "*temporal location*" - 79.3%).

1.2. Relevância da anotação de papéis semânticos

Passados mais de 10 anos da construção do primeiro PropBank para a língua portuguesa, quase 20 anos do primeiro PropBank, e levando em conta os direcionamentos recentes do PLN e a onipresença de arquiteturas neurais e grandes modelos de linguagem, nos perguntamos sobre a atual relevância, para o PLN, da anotação de papéis semânticos.

Dentre as críticas ao atual paradigma de IA está a falta de transparência e de explicabilidade/interpretabilidade, características relevantes não só no ambiente científico, mas também em aplicações em que a intolerância a erros é alta e a transparência relativa às decisões tomadas é crucial, como os campos jurídicos, da medicina, certos casos de pesquisa acadêmica, como nas Humanidades Digitais que mesclam análises/leituras de longe e de perto (*close* e *distant reading*) e ainda aplicações bastante específicas, como previsão de atrasos na partida de voos. Neste ambiente, papéis semânticos oferecem maneiras interpretáveis de representar semanticamente conteúdos verbais, e podendo servir de insumo para a criação de grafos de conhecimento (KG). Ao viabilizar o armazenamento de relações explícitas entre entidades, KGs são capazes de reduzir alucinações e de aumentar a precisão das respostas fornecidas por LLMs. Adicionalmente, contribuem para explicabilidade, uma vez que o conteúdo explicitado em um grafo é visível e rastreável.

Grafos de conhecimento neurais combinam a representação estruturada de KG com capacidades de aprendizado das redes neurais, e são parte de uma tendência mais ampla na IA que busca articular conhecimento simbólico e estatístico. Do lado simbólico, KG atuariam em pontos sensíveis de LLM, e a seguir estão alguns argumentos elencados por Dong (2023) para defender que, ao menos a curto prazo, KGs não serão superados por LLMs.

Uma vez que o treino de LLMs é caro, é difícil para LLMs incorporar rapidamente conhecimento atual. O GPT4, por exemplo, lançado em março de 2023, é treinado com conhecimento que vai até setembro de 2021. Além disso, um estudo conduzido pela equipe de Dong mostrou que para perguntas que podem ser respondidas usando dados da DBPedia, o ChatGPT tem uma taxa de alucinação de cerca de 20%, e não consegue responder cerca de 50% das perguntas (Sun et al., 2024). Por fim, LLMs só conseguem aprender se o conhecimento aparece com frequência nos dados de treinamento. O mesmo

[10] A descrição detalhada está em https://edu.visl.dk/~eckhard/pdf/TIL2007_roles.pdf

estudo de Sun et al.(2024) mostra que a acurácia na resposta para perguntas que envolvem fatos que estão na cauda longa (perguntas relativas a entidades que estão em 33% da popularidade) cai de cerca de 50% para cerca de 15%. Em todos esses casos, afirma Dong (2023), o conhecimento desejado poderia estar melhor alocado em tuplas e, em nosso caso, voltamos aos papéis semânticos.

De maneira complementar, mas tomando outra perspectiva, a anotação de SRL também pode ser usada para avaliação de modelos de linguagem quanto à capacidade de representar informação semântica estruturada e interpretável. Trabalhos como Tanney (2019a e 2019b) e He et al. (2020) investigam de que maneira modelos contextuais como BERT e ELMo codificam informação linguística, utilizando conjuntos de validação compostos por uma série de fenômenos linguísticos, utilizando para isso datasets com anotação de papéis semânticos e de dependências semânticas. Nessa vertente, notamos a escassez de conjuntos de validação criados para a língua portuguesa, que nos obriga a utilizar conjuntos de dados traduzidos do inglês, como é o caso do Albertina (Rodrigues et al., 2023), que utiliza o PLUE, uma tradução automática para o português do dataset GLUE, para o ajuste fino relativo a tarefas *downstream*.

Na articulação entre redes neurais e SRL, Mohebbi et al.(2022), propõem uma abordagem baseada em aprendizado de grafos profundos (deep graph learning) para computar *similaridade semântica* de documentos (STS), utilizando para isso a anotação de papéis semânticos. Segundo os autores, "one potential reason why SRL system output has not been used for computing STS is the lack of effective methods for applying and incorporating SRL information in neural networks."

A utilidade da anotação de papéis semânticos no PLN também pode ser indireta. Considerando o paradigma de avaliação extrínseca, Evans e Orasan (2019) utilizam SRL para verificar se a simplificação textual -- especificamente, um método de simplificação de frases -- é capaz de facilitar o desempenho na tarefa de SRL.

Por fim, temos visto que LLMs associados a técnicas como *ajuste fino* ou *few-shot prompting* avançaram o desempenho em tarefas como identificação e reconhecimento de entidades e a extração de relações. No entanto, para que o avanço se materialize e para que seja possível avaliar este avanço, medindo o grau de confiança dos resultados de LLMs, precisamos de materiais com as características desejáveis, o que torna relevante a existência de um recurso como o apresentado aqui.

1.3. Aprendizado automático de papéis semânticos

1.3.1. Resultados para a língua portuguesa

Mesmo considerado de tamanho restrito, Propbank-br continua sendo usado para o aprendizado com técnicas recentes. Falci et al (2019), usando a v1 PropBank-Br (Duran & Aluísio, 2012), conseguem F1 total de 68.18%. Exceto por Arg0 (84.28), Arg1 (71.71), Neg (92.31%) e ArgM-tmp (69, 90), o desempenho por classe é baixo, se tomarmos uma linha de corte de 70%. Alva-Manchego e Rosa (2012) conseguem F1 total de 79.6%, usando o

mesmo corpus. Ainda com o mesmo corpus, Oliveira (2020) utiliza uma abordagem cross-linguística para aprender SRL em português, e obtém F1 de 0.78%. Para o CINTIL-PropBank, a equipe responsável pela criação do corpus informa conseguir F-Score de 0.70%.

1.3.2. Resultados para outras línguas

O artigo de Wang et al. (2022) traz um levantamento dos resultados mais recentes para a anotação SRL, e mesmo com a utilização das técnicas mais avançadas os melhores resultados giram em torno de 85% (no cenário *out of domain/Brown corpus*), como vemos na figura 1.

	CoNLL05 WSJ			CoNLL05 Brown			CoNLL12 Test		
Model	P	R	F1	P	R	F1	P	R	F1
syntax-aware									
Zhou et al. (2019)+BERT	89.0	88.8	88.9	81.9	81.0	81.4	-	-	-
Mohammadshahi and Henderson (2021)+BERT	89.1	88.7	88.9	83.9	82.5	83.2	-	-	-
Xia et al. (2020)+RoBERTa	88.4	88.8	88.6	83.1	83.3	83.2	-	-	-
Marcheggiani and Titov (2020)+RoBERTa	87.7	88.1	88.0	80.5	80.7	80.6	86.5	87.1	86.8
syntax-agnostic									
Li et al. (2019b)	87.9	87.5	87.7	80.6	80.4	80.5	85.7	86.3	86.0
Conia and Navigli (2020)+BERT	-	-	-	-	-	-	86.9	87.7	87.3
Blloshmi et al. (2021)+BART	-	-	-	-	-	-	87.8	86.8	87.3
Shi and Lin (2019)+BERT	88.6	89.0	88.8	81.9	82.1	82.0	85.9	87.0	86.5
Jindal et al. (2020)+BERT	88.7	88.0	87.9	80.3	80.1	80.2	86.3	86.8	86.6
Paolini et al. (2021)+T5	-	-	89.3	-	-	82.0	-	-	87.7
Zhang et al. (2021)+RoBERTa	89.6	89.7	89.6	83.8	83.6	83.7	88.1	**88.6**	**88.3**
Ours+BERT	89.7	89.0	89.3	85.9	83.5	84.7	88.0	87.7	87.8
Ours+RoBERTa	**90.4**	**89.7**	**90.0**	**86.4**	**83.8**	**85.1**	**88.6**	87.9	88.3

Table 3: Argument labeling results on CoNLL2005 and CoNLL2012.

Figura 1.Resultados SOTA para SRL em inglês. Fonte: Wang et al.(2022)

Roth & Lapata (2016) argumentam que informações sintáticas são necessárias para se atingir resultados competitivos na anotação SRL baseada em dependências, o que não surpreende uma vez que tanto a teoria de papéis semânticos quanto à implementação da tarefa são vistas como um desdobramento da análise sintática. Por outro lado, com base em seus resultados, Shi & Lin (2019) defendem que "BERT can be adapted to relation extraction and semantic role labeling without syntactic features and human-designed constraints". A figura 1 também sinaliza que abrir mão de informação sintática, em modelos que incorporam representações distribuídas das palavras, não necessariamente piora os resultados, pelo contrário. Por outro lado, seria interessante verificar se há ganhos quando, usando modelos que levam ao melhor desempenho, informação sintática é incorporada (caso isto seja possível, e considerando que esta comparação já não foi feita nos trabalhos elencados).

1.4 Estudos sobre generalização de papéis semânticos no aprendizado de máquina

Desde o seu surgimento, a anotação de papéis semânticos em um PropBank está associada ao aprendizado de máquina (AM), no qual a generalização desempenha um papel fundamental. No entanto, são raros os estudos sobre a capacidade de *generalização das representações linguísticas* criadas pelos PropBanks.

Loper et al. (2007), Yi et al. (2007) e Merlo & Van der Plas (2009) – os dois primeiros trabalhos com a co-autoria de M. Palmer, uma das criadoras do PropBank e da VerbNet – investigaram as classes de papéis semânticos com relação à sua capacidade de generalização no AM, comparando as classes do PropBank com as classes da VerbNet. Para tanto, partiram de um mapeamento entre as classes, o SemLink, projeto iniciado em Loper et al.(2007) e atualmente na versão 2.0 (Stowe et al., 2021)[11]. (Importante notar que o mapeamento leva em consideração, do lado PropBank, apenas os argumentos numerados, uma vez que apenas esse tipo de argumento é classificado pela VerbNet.)

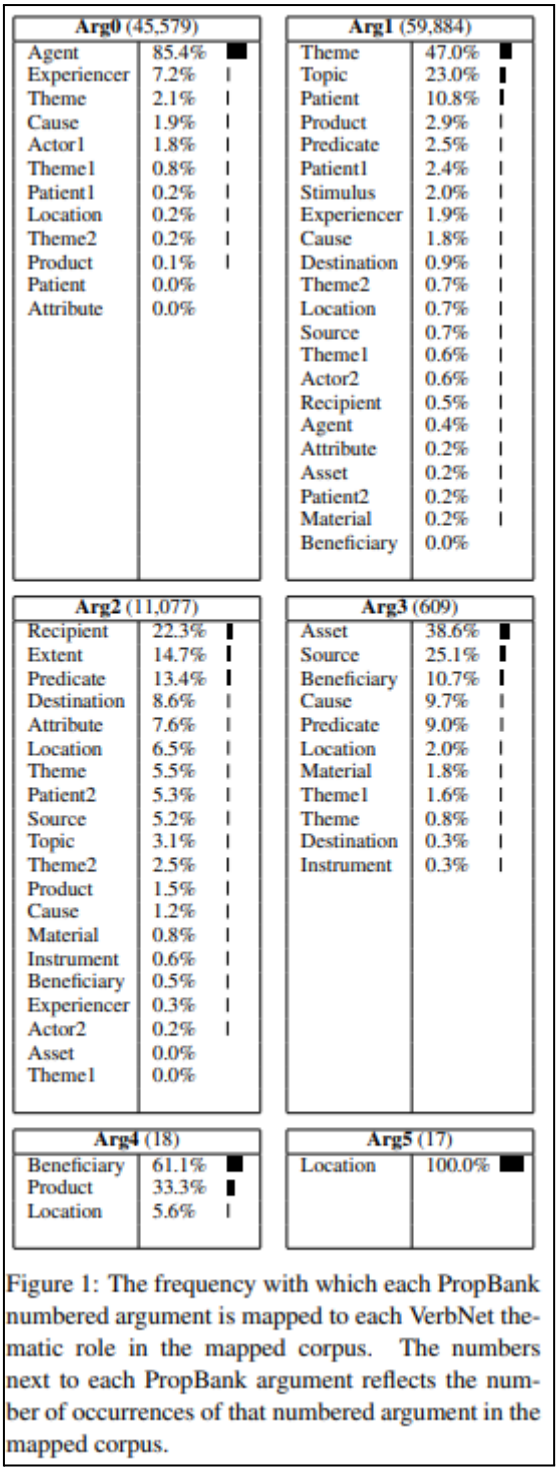
Segundo Loper et al. (2007), o mapeamento entre o PropBank e a VerbNet torna os dados mais consistentes, uma vez que as etiquetas do PropBank (específicas de cada verbo) são transformadas nos papéis temáticos – mais gerais – da VerbNet. Ainda conforme os autores, "diferentemente dos rótulos do PropBank, os rótulos da VerbNet são definidos de forma consistente entre verbos e, portanto, deve ser mais fácil para os sistemas estatísticos de SRL modelá-los. Além disso, uma vez que as etiquetas da VerbNet são significativamente menos dependentes do verbo do que etiquetas do PropBank, modelos de SRL devem generalizar melhor para verbos novos e para novos usos de verbos conhecidos."[12] Os autores treinaram duas vezes seu sistema de SRL: a primeira vez com a anotação PropBank, e a segunda vez com a anotação VerbNet. Inicialmente, os resultados foram pouco conclusivos devido à baixa frequência das classes PropBank Arg3-Arg5. Em um segundo experimento, limitaram o escopo do estudo apenas a instâncias mapeadas de Arg1 e Arg2. Os resultados confirmaram as expectativas: precisão e abrangência melhoraram significativamente com relação ao Arg2, "which demonstrates that the Arg2 label in the PropBank is quite overloaded. The Arg2 mapping improves the overall results (F1) on the WSJ by 6% and on the Brown corpus by almost 10%." Ou seja, considerando o material novo do Brown corpus (o sistema foi treinado no WSJ), a anotação com o tagset da VerbNet melhorou o desempenho do sistema de anotação em 10%. Os autores concluem ainda afirmando que "Our new SRL system handles these cases more robustly, demonstrating the consistency and usefulness of the thematic role categories."

---

[11] https://verbs.colorado.edu/semlink/ e https://github.com/cu-clear/semlink. Uma descrição detalhada do mapeamento entre VerbNet e PropBank está em https://verbs.colorado.edu/semlink/semlink1.1/vn-pb/README.TXT e ainda mais informações técnicas sobre o SemLink podem ser obtidas em https://www.nactem.ac.uk/meta-net/Narratives/SemLink.pdf .

[12] "In particular, we can use the mapping to transform the verb-specific PropBank role labels into the more general thematic role labels that are used by VerbNet. Unlike the PropBank labels, the VerbNet labels are defined consistently across verbs; and therefore it should be easier for statistical SRL systems to model them. Furthermore, since the VerbNet role labels are significantly less verb dependent than the PropBank roles, the SRL's models should generalize better to novel verbs, and to novel uses of known verbs." (Loper et al., 2007:8)

O trabalho de Yi et al. (2007), derivado de sua dissertação, traz uma apresentação mais detalhada dos mesmos resultados (os autores de ambos os artigos são os mesmos), na qual mostram a distribuição entre etiquetas de papéis semânticos do PropBank e etiquetas dos papéis temáticos da VerbNet (figura 2). Yi et al. (2007), assim como Merlo & Van der Plas (2009), mostram que 85% das ocorrências de Arg0 podem ser traduzidas para o papel de Agente, e que 45% das ocorrências de Arg1 podem ser alinhadas para papéis do tipo Paciente, considerando o esquema de anotação da VerbNet. Os autores concluem reafirmando que o mapeamento tornou Arg2 uma classe mais definida, sendo mais fácil distinguir entre Arg2 e demais argumentos. Uma análise da matriz de confusão apenas de Arg2, comparando o desempenho dos dois sistemas, mostrou que, de 1.539 instâncias que o sistema com a anotação "original" Arg2 não anotou corretamente, 233 não recebem argumento (confusão entre argumentos numerados (que devem anotados) e argumentos modificadores), e que os demais erros de Arg2 deveriam ser Arg1(716 ocorrências), ArgM (482 ocorrências), Arg0 (50 ocorrências) e Arg3 (1 ocorrência).



Arg0 (45,579)	
Agent	85.4%
Experiencer	7.2%
Theme	2.1%
Cause	1.9%
Actor1	1.8%
Theme1	0.8%
Patient1	0.2%
Location	0.2%
Theme2	0.2%
Product	0.1%
Patient	0.0%
Attribute	0.0%

Arg1 (59,884)	
Theme	47.0%
Topic	23.0%
Patient	10.8%
Product	2.9%
Predicate	2.5%
Patient1	2.4%
Stimulus	2.0%
Experiencer	1.9%
Cause	1.8%
Destination	0.9%
Theme2	0.7%
Location	0.7%
Source	0.7%
Theme1	0.6%
Actor2	0.6%
Recipient	0.5%
Agent	0.4%
Attribute	0.2%
Asset	0.2%
Patient2	0.2%
Material	0.2%
Beneficiary	0.0%

Arg2 (11,077)	
Recipient	22.3%
Extent	14.7%
Predicate	13.4%
Destination	8.6%
Attribute	7.6%
Location	6.5%
Theme	5.5%
Patient2	5.3%
Source	5.2%
Topic	3.1%
Theme2	2.5%
Product	1.5%
Cause	1.2%
Material	0.8%
Instrument	0.6%
Beneficiary	0.5%
Experiencer	0.3%
Actor2	0.2%
Asset	0.0%
Theme1	0.0%

Arg3 (609)	
Asset	38.6%
Source	25.1%
Beneficiary	10.7%
Cause	9.7%
Predicate	9.0%
Location	2.0%
Material	1.8%
Theme1	1.6%
Theme	0.8%
Destination	0.3%
Instrument	0.3%

Arg4 (18)	
Beneficiary	61.1%
Product	33.3%
Location	5.6%

Arg5 (17)	
Location	100.0%

Figure 1: The frequency with which each PropBank numbered argument is mapped to each VerbNet thematic role in the mapped corpus. The numbers next to each PropBank argument reflects the number of occurrences of that numbered argument in the mapped corpus.

Figura 2: Mapeamento entre argumentos numerados do PropBank e papéis temáticos da VerbNet feito por Yi et al.(2007). Fonte: Yi et al (2007)

Por fim, o trabalho de Merlo & Van der Plas (2009) busca aprofundar os achados dos trabalhos anteriores, também fazendo uma alinhamento entre as anotações PropBank e VerbNet e usando o mapeamento de Loper et al., (2007), o SemLink. A conclusão geral de Merlo & Van der Plas – que não se restringiram a Arg1 e Arg2 – reafirma que a anotação VerbNet leva a resultados melhores quando se trata de generalizar para novas instâncias de papéis semânticos, enquanto a anotação PropBank capta melhor algumas das restrições

estruturais (sintáticas) entre os papéis semânticos (e, justamente por isso, dificultando a generalização de papéis tanto entre verbos novos como entre verbos não vistos na fase de treino).

Para que algumas das comparações entre os dois tipos de anotação (PropBank e VerbNet) fossem válidas, os autores reduziram o inventário de etiquetas da VerbNet para o mesmo número de etiquetas do PropBank (considerando apenas os argumentos numerados). Assim como em Loper et al. (2007), Merlo & Van der Plas definiram classes de equivalência entre as etiquetas VerbNet. Os autores informam ainda que as análises foram feitas com base em um conjunto de 106.459 pares de papéis semânticos, uma vez que as anotações no SemLink 1.1 não estavam completas. Foram consideradas apenas ocorrências de papéis semânticos para os quais havia anotação tanto de etiquetas PropBank como de etiquetas VerbNet (conforme os autores, cerca de 45% dos papéis PropBank têm um papel semântico VerbNet). Como medidas de avaliação relativas à informatividade - um dos interesses do trabalho era verificar qual seria o tagset mais informativo – , os autores usaram "entropy, joint entropy, conditional entropy, mutual information, and a normalised form of mutual information which measures correlation between nominal attributes called symmetric uncertainty (Witten and Frank, 2005, 291)".

Merlo & Van der Plas (2009) mostram que o PropBank tem uma distribuição de classes mais enviesada que a VerbNet, com a imensa parte dos argumentos do PropBank se distribuindo entre Arg0 e Arg1, o que contribuiria para a dificuldade de generalização da anotação PropBank. Na VerbNet, os argumentos se distribuem de maneira um pouco mais equilibrada (figura 3). Um sistema de anotação PropBank ingênuo, que simplesmente atribuísse o rótulo mais frequente, estaria **correto 52% das vezes atribuindo sempre um rótulo A1**, mas estaria correto apenas 33% das vezes atribuindo o rótulo de Agente se a anotação fosse feita seguindo a VerbNet.

PropBank		VerbNet									
A0	38.8	Agent	32.8	Cause	1.9	Source	0.9	Asset	0.3	Goal	0.00
A1	51.7	Theme	26.3	Product	1.6	Actor1	0.8	Material	0.2	Agent1	0.00
A2	9.0	Topic	11.5	Extent	1.3	Theme2	0.8	Beneficiary	0.2		
A3	0.5	Patient	5.8	Destination	1.2	Theme1	0.8	Proposition	0.1		
A4	0.0	Experiencer	4.2	Patient1	1.2	Attribute	0.7	Value	0.1		
A5	0.0	Predicate	2.3	Location	1.0	Patient2	0.5	Instrument	0.1		
AA	0.0	Recipient	2.2	Stimulus	0.9	Actor2	0.3	Actor	0.0		

Table 1: Distribution of PropBank core labels and VerbNet labels.

Figura 3: Distribuição de etiquetas considerando anotação PropBank e VerbNet segundo Merlo & Van der Plas (2009). Fonte: Merlo & Van der Plas (2009)

Após uma série de comparações, os autores chegam às seguintes conclusões:

- A partir do cálculo da entropia da distribuição das classes conforme as etiquetas de cada um dos recursos (e já após a redução das classes VerbNet a 7, para fins de comparação), a distribuição dos rótulos do PropBank tem uma entropia de 1,37 bits e, a distribuição dos rótulos da VerbNet tem uma entropia de 2,06 bits. Portanto, a anotação VerbNet transmitiria mais informações. (cf. seção 3 do artigo).

- As etiquetas VerbNet fornecem uma especificidade maior. As etiquetas do PropBank estão concentradas em algumas poucas etiquetas da VerbNet de frequência mais alta. Isso não acontece com as etiquetas de baixa frequência, em que a VerbNet fornece desambiguações para as etiquetas "sobrecarregadas" do PropBank. Os autores concluem ainda que ambos os esquemas de anotação podem ser úteis em diferentes circunstâncias e em diferentes faixas de frequência (cf. seção 4 do artigo).
- As etiquetas do PropBank são abstrações de alto nível dos papéis da VerbNet e as etiquetas da VerbNet são mais específicas para verbos e classes (cf. seção 5 do artigo).
- A anotação PropBank captura melhor a hierarquia temática e está mais associada a funções gramaticais, o que a tornaria potencialmente mais útil para a anotação de papéis semânticos levando em conta modelos de AM baseados em árvores sintáticas (cf. seção 6 do artigo). Já as anotações da VerbNet são mais independentes da análise sintática.
- Reforçam, com novas evidências, que etiquetas da VerbNet são de aplicação mais gerais *entre verbos* (cf. seção 7 do artigo).

A **conclusão geral é que a anotação PropBank é mais fácil de ser aprendida** (se analisamos apenas os números globais do desempenho de AM, poderíamos acrescentar), mas **a anotação VerbNet é mais informativa como um todo** e **generaliza melhor para novas instâncias** de papéis.

Cerca de 10 anos depois, Hartmann et al. (2017) contrastam anotações conforme os estilos FrameNet, PropBank, e VerbNet após a construção de um dataset em alemão com anotações dos três tipos de representação linguística. O dataset continha pouco mais de 4 mil anotações de papéis para cada um dos três tipos de representação, o que fez os autores também investigarem os impactos da geração automática de dados de treino. Os resultados mostraram que FrameNet é a representação mais difícil de ser generalizada, que desempenho da generalização da representação PropBank é um pouco melhor que o da generalização VerbNet, e que esta última não se beneficiou da técnica de geração automática de dados de treino, que se aplica apenas a dados já vistos no treino. Os autores concluem afirmando que "Taking into account the semantic expressiveness of VerbNet, these results showcase the potential of VerbNet as an alternative to PropBank." (Hartmann et al., 2017:119).

Finalmente, nos referimos ao SemLink, projeto de alinhamento entre FrameNet, PropBank, VerbNet e WordNet, por meio dos grupamentos de sentidos do OntoNotes, atualmente na versão 2.0 (Stowe et al., 2021). Iniciado em 2007 (Loper et al., 2007), o SemLink contém mapeamentos entre os recursos e um corpus anotado com instâncias anotadas conforme PropBank, VerbNet e FrameNet. A versão 2.0 traz melhorias relativas à adição de novas instâncias e alinhamentos, correções manuais, vetores baseados em sentido ("sense-based vectors") e informação colocacional. Segundo os autores, "the mappings between the semantic role labels allow users to accurately convert annotations done with PB roles to VN roles and combine their respective data sets into a much larger corpus of training and test

data.” Mais recentemente, tem-se investido em formas automáticas de inferir a anotação VerbNet a partir da anotação PropBank, como em Li et al.(2023).

O trabalho de Li et al. (2023), em co-autoria com Palmer, também compara os resultados de AM para classes VerbNet e PropBank, e apesar de a técnica de conversão de papéis obter ótimos resultados, na comparação entre representações PropBank e VerbNet, PropBank leva a números mais baixos, corroborando resultados anteriores (figura 4).

	Joint	WSJ		Brown	
	SEMLINK	VN	PB	VB	PB
1	✗	91.42	89.12	85.90	82.56
2	✓	91.42	89.21	85.41	82.46

Table 5: Ablation of SEMLINK constraint during training using the *Joint* CRF. ✓ indicates SEMLINK constraint is applied; and ✗ indicates not.

Figura 4: Resultados para o AM considerando classes da VerbNet e do PropBank.
Fonte: Li et al.(2023)

Apesar da relevância da *generalização* dos papéis semânticos no contexto da tarefa de SRL, estudos comparativos sobre o papel de diferentes representações semânticas no AM são escassos. Uma pesquisa reversa no Google a partir do artigo de Merlo & Van der Plas (2009) – o mais citado deles – retorna apenas 53 citações/artigos, dos quais apenas 4 tratam diretamente do assunto.

A anotação ao estilo PropBank é anunciada como sendo mais fácil de generalizar para o AM. Em Palmer et al. (2014:14, grifo meu), lemos que “*PropBank has been shown to provide the **most effective** training data for supervised Machine Learning techniques*.“, e os trabalhos já citados de Merlo & der Plas (2009) e de Yi et al. (2007) são tomados como exemplo da bem sucedida generalização do modelo PropBank. Mas, como vimos, a afirmação sobre a superioridade dos dados PropBank para generalização é apenas parcialmente verdadeira. Para textos novos e novos verbos (cenário real de utilização dos modelos de AM), a representação VerbNet é a que generaliza melhor. De todos os trabalhos que encontramos sobre o tema – todos apresentados aqui –, três deles, que também mostram os limites da generalização PropBank para novos textos ou verbos, têm co-autoria de Martha Palmer.

Por fim, Gunger & Palmer (2021) afirmam que (grifos meus)

> “investigate how semantic role labeling performance on low-frequency predicates and out-of-domain data can be improved by using VerbNet(...). **We find that VerbNet classes provide an effective level of abstraction, improving generalization on low frequency predicates** (...). We also find that joint training of VerbNet role labeling and predicate disambiguation of VerbNet classes for polysemous verbs leads to improvements in both tasks, naturally

supporting the extraction of VerbNet's semantic representations." (Gunger & Palmer, 2021)

2. Preparação do Porttinari-base Propbank

A anotação da camada PBP seguiu estritamente as diretivas para a criação do PropBank.br (Duran, 2014), que foram sendo atualizadas, melhoradas e ampliadas ao longo do projeto, dando origem a novas diretivas (Duran & Freitas, 2024). No entanto, há um ponto crucial em que a anotação do PBP difere da anotação do PropBank, que diz respeito à relação entre a atribuição de papéis e a segmentação anterior fornecida pela análise sintática. No PropBank.br, a anotação de papéis semânticos deveria seguir estritamente a análise sintática fornecida pelo parser PALAVRAS. Na anotação PBP esta recomendação não foi levada em conta, e justificamos nossa escolha com três argumentos:

1. A decisão torna a camada SRL completamente dependente das decisões de UD, o que tem como consequências: a) se quisermos verificar o impacto de outro parser/gramática na atribuição de papéis semânticos, não podemos; b) se a sintaxe UD ou algum erro na anotação sintática tiver impedido a segmentação conforme descrita no frame do verbo, deixamos de recuperar a informação, mesmo ela estando disponível; c) ficamos reféns das diretivas de UD, dado que é um projeto em andamento. Caso as diretivas mudem em pontos sensíveis à anotação PBP, ficamos com um material datado ou precisaremos rever.

2. Não há impedimento formal para a anotação da camada PBP de maneira independente da camada sintática. Além disso, é possível recuperar de maneira automática os casos em que um elemento foi tratado como argumento pelo Verbo-Brasil e mas não o foi pela anotação sintática (busca por Args cuja deprel seja nmod).

3. Este desalinhamento entre a anotação do treebank subjacente e a anotação de papeis semânticos já acontece no PropBank original (Palmer et al., 2005):
   > *Consider a sentence such as Commonwealth Edison said the ruling could force it to slash its 1989 earnings by $1.55 a share. (wsj_0015). The Penn Treebank's analysis assigns a single sentential (S) constituent to the entire string it to slash . . . a share, making it a single syntactic argument to the verb force**. In the PropBank annotation, we split the sentential complement into two semantic roles for the verb** force, assigning roles to the noun phrase and verb phrase but not to the S node which subsumes them" (2005:82)*

Ou seja, embora seja um recurso que nasça da sintaxe, e embora esta decisão implique um pior desempenho em parsers que utilizam **exclusivamente** a camada UD como pista na detecção dos papéis semânticos, o **primeiro compromisso do PBP é com o Verbo-Brasil**. E, ao menos teoricamente, o que o Verbo-Brasil (e recursos similares nos quais se inspira) codifica são os argumentos dos verbos. Se a opção de segmentação na anotação sintática leva em conta outras variáveis, isso não deve interferir na atribuição dos papéis semânticos

(veja-se, por exemplo, o relato de Sampson (2000) sobre a ausência de consenso em um workshop de segmentação em constituintes na ACL de 1991)[13]. Este ponto será retomado na seção 3.1.

Ainda assim, isto não significa que decisões de UD não interfiram na anotação SRL, como será visto na seção 3.

2.1. Formato

A anotação de papéis semânticos ao estilo PropBank pode assumir duas configurações:

A. anotação SRL baseada em dependências, em que cada argumento é um único token (por exemplo, usada em Hajic et al. 2009)
B. anotação SRL baseada em intervalos (span-based), em que cada argumento é um intervalo constituído por um ou mais tokens (formato utilizado em Carreras & Marquez, 2005; Pradhan et al. 2012, por exemplo). Este formato é considerado mais difícil porque adiciona à tarefa a necessidade de detecção dos limites dos argumentos. A anotação baseada em span pode ainda ser codificada de duas formas: com base em grafos ou com base em sequências (anotação BIO). A anotação baseada em grafos consegue lidar com múltiplos predicados na mesma frase, enquanto a anotação BIO não consegue, como vemos na figura 5, retirada de He et al. (2018).

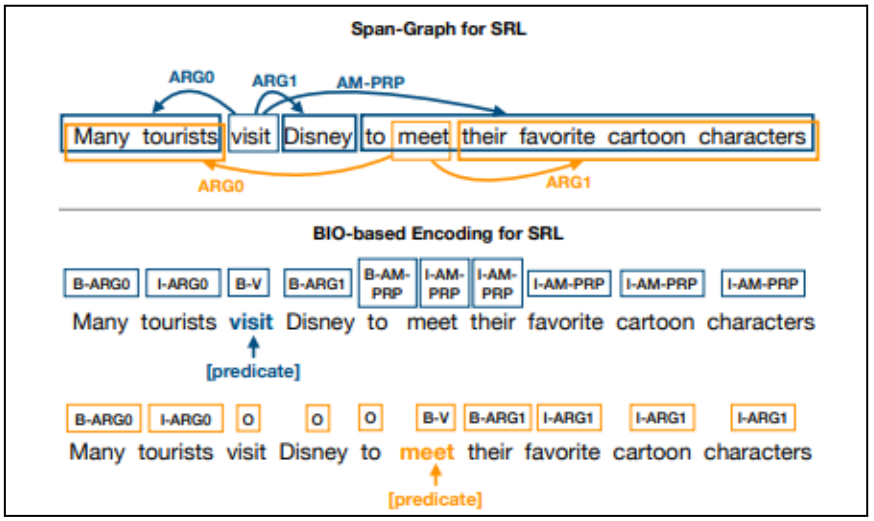


Figura 5. Dois formatos de anotação baseada em intervalo (span).
Fonte: He et al.(2018)

[13] Segundo Sampson, no workshop, pesquisadores de nove grupos de pesquisa diferentes receberam um mesmo conjunto de frases para analisar, seguindo as diretrizes de análise que seus grupos de pesquisa considerariam corretas. Não eram frases especialmente complicadas ou confusas, mas eram frases reais, extraídas de corpus. Para simplificar a tarefa, a comparação entre as análises não levaria em conta os rótulos dos constituintes, apenas a estrutura (a forma) das árvores, ou seja, a ideia era verificar apenas se os pesquisadores identificariam as mesmas subsequências de palavras como constituintes gramaticais, sem levar em conta a classificação desses constituintes. Para uma grande surpresa na época, o nível de concordância entre as análises foi extremamente baixo.

Cada um dos formatos baseados em intervalo (*span based*) codifica uma maneira de abordar o problema:
(i) A anotação BIO repete as frases conforme o número de verbos que contêm, e trata de um predicador de cada vez. Além disso, costuma partir de predicados padrão ouro. Com este tipo de formato, portanto, é preciso reutilizar o anotador para cada elemento predicador;
(ii) A anotação baseada em grafos realiza todo o processo em um único lance, sem a necessidade de replicação das frases por predicado e sem a necessidade de predicados padrão ouro já no input.

He et al. (2018) comparam diferentes formatos de anotação span-based no AM de SRL, e relatam resultados ligeiramente superiores usando a abordagem baseada em grafos.

O formato de anotação SRL descrito aqui refere-se à anotação (A), baseada em dependências (*dependency based SRL*), embora esteja prevista a incorporação de informação relativa aos *spans*, em formato ainda a ser decidido. O que apresentamos, portanto, são **dependências semânticas anotadas conforme o estilo PropBank**.

2.2. Conjunto de Etiquetas

A anotação de papéis semânticos do Porttinari utilizou **28 etiquetas**[14], incluindo algumas etiquetas provisórias, descritas a seguir. A versão anterior corpus Propbank contava com 25 etiquetas (argumentos numerados de Arg0 a Arg5 e 19 argumentos não numerados). O corpus PropBank original, por sua vez, conta com 20 etiquetas (6 para os argumentos numerados e 15 para os modificadores). A diferença numérica entre as duas versões brasileiras do PropBank resulta das seguintes decisões:

- Foi adicionada a etiqueta **ArgM-comp**, presente no PropBank original, para estruturas comparativas.

- Foram criadas as etiquetas **ArgM-src**, **ArgM-conseq**, e **ArgM-cond**.

- Não foi encontrada **nenhuma ocorrência de Arg5** no corpus. O Verbo-Brasil contém 5 frames com Arg5: ampliar.101; baixar.01; diluir.01; oscilar.01; pular.01

A seguir são apresentadas as novas etiquetas, bem como algumas especificações na anotação de etiquetas já existentes. A descrição completa de todas as etiquetas utilizadas na anotação PBP, bem como uma série de exemplos de uso, estão em Duran & Freitas (2024).

[14] Veja-se a página do projeto para detalhe das etiquetas usadas no PBP: https://sites.google.com/icmc.usp.br/poetisa/porttinari-base-propbank

2.2.1. Observações sobre etiquetas já existentes

ArgM-mnr

Inclui orações proporcionais, como aquelas introduzidas por “a medida que”, “a proporção que”

- A a medida que empresas **trabalham** para mostrar a os empregados que oferecem igualdade de oportunidades internas , o interesse feminino **começa a aumentar** , ainda que de forma discreta .
- Sob o efeito de a droga , **começou a enxergar** visualmente cada nota de a composição em o seu campo visual , a a medida que **ressoavam** .

ArgM-com

Indicam com quem uma ação foi realizada. Isso pode incluir pessoas ou organizações (entidades que possuem características de agentes prototípicos: animacidade, volição), mas exclui objetos, que seriam considerados modificadores instrumentais. Embora o termo formal para este modificador seja `comitativo', podemos pensar neste argumento como “companhia”, o que inclui também a ausência de companhia, como em “sozinho”.

- Hoje , ela fica **sozinha** com as **recreadoras** e se relaciona com outros cachorros .
- Quando o presidente Michel Temer se sentar à mesa com o **americano** Donald Trump para jantar , nesta segunda (18) , em Nova York , o cardápio pode ser indigesto .
- Então saiu de a casa de a mãe e foi morar **sozinho** , em outro bairro .

Porém, junto/juntos continua sendo **ArgM-mnr**:

- O trio teria , **junto** , cerca de 54 % de os votos .
- A dupla já havia trabalhado **junta** em o filme " Valsa com Bashir " (2008) , sobre a guerra em o Líbano .
- Escrevemos **juntos** , gravamos **juntos** .

ArgM-mod e ArgM-asp

Utilizadas para os verbos cujos frames no Verbo-Brasil ou na tabela de verbos auxiliares presente em Duran (2014) indicama presença de um *auxiliar modal* ou aspectual, respectivamente A atribuição de ArgM-Mod e ArgM-Asp está dissociada do que a anotação UD do Portinari considerou verbo auxiliar, uma vez que as diretivas UD são econômicas no uso de verbos auxiliares. Especificamente, foram considerados ArgM-mod os seguintes lemas/frames:

poder.201_aux-mod	393
dever.201_aux-mod	176
ter.203_aux-mod	70
ter.202_aux-mod	30

Foram considerados ArgM-asp os seguintes lemas/frames:

começar_aux-asp	62
passar.201_aux-asp	42
acabar.201_aux-asp	24
voltar.201_aux-asp	23
deixar.201_aux-asp	21
chegar.201_aux-asp	19
continuar.201_aux-asp	14
ficar_aux-asp	13
parar.201_aux-asp	10
acabar.202_aux-asp	8
continuar.202_aux-asp	8
seguir_aux-asp	4
acabar.203_aux-asp	3
andar.201_aux-asp	1
continuar_aux-asp	1
estar.301	1
ficar.202	1
sair_aux-asp	1
terminar.201_aux-asp	1
terminar.202_aux-asp	1
vir.201_aux-asp	1

Ao retomar as decisões de auxiliaridade descritas em Duran (2014), isto é, os verbos auxiliares listados na tabela 2 de Duran (2014), foram feitas pequenas mudanças nas classificações dos verbos[15], conforme documentado em Duran e Freitas (2024) e repetido abaixo:

- A combinação “vir” + gerúndio foi considerada ArgM-tml (era considerada “aspecto”)
- A combinação “estar” + gerúndio foi considerada ArgM-tml (era considerada “aspecto”)
- A combinação “estar” + particípio não foi mais considerada auxiliar de passiva (ArgM-pas), mas sim um verbo de cópula (“estar”) + adjetivo
- A combinação “ir” + gerúndio foi considerada ArgM-tml (não estava na lista)

Além disso, uma vez que nem todos os verbos listados como auxiliares possuem um *frame* específico no Verbo-Brasil, a informação de auxiliaridade também está indicada no campo “deps”, e variou da seguinte maneira:

- ter.201_aux-tml: existe o frame “ter.201” para indicar o “ter” como auxiliar temporal
- ir_aux-tml: o verbo ir está sendo usado como auxiliar temporal, mas não existe um frame para "ir" como auxiliar de tempo.

[15] As mudanças são consequência de decisões de anotação da camada sintática, conforme descrito em Duran (2022).

2.2.2. Etiquetas específicas do PBP (especificações da etiqueta genérica ArgM-adv)

ArgM-cond: condicionais

Indicam proposições condicionais. No PropBank original (2015), orações introduzidas por "se" devem ser anotadas com a etiqueta mais ampla ArgM-adv. Como estes casos são formalmente marcados em português com as conjunções se/caso e com o tempo verbal, e são um grupo semanticamente relevante, decidimos separá-los de ArgM-adv com uma etiqueta própria, ArgM-cond.

- Não podemos continuar se você não me **disser** a verdade .
- **Sofresse** ali o 0 a 1 , o intervalo seria um inferno e o segundo tempo um sofrimento .
- Caso seja **preciso** trocar los , o serviço custa a partir de R$ 100 .

ArgM-comp: comparações

Construções envolvendo comparações são anotadas no PropBank (Bonial et al., 2015) como ArgM-ext, etiqueta aplicada a quantidades. Na anotação UD, construções comparativas são anotadas como *advcl*, assumindo que contêm um verbo implícito. Para os exemplos abaixo

- Brasileiro diz pagar *menos* ***propina*** de o que **vizinhos** de a América Latina .
- O time tem dez *pontos a mais que* o segundo **colocado** em a tabela , o Grêmio .

, assume-se que as frases seriam versões reduzidas de

- Brasileiro diz pagar *menos* ***propina*** de o que pagam **vizinhos** de a América Latina.
- O time tem dez *pontos a mais que* tem o segundo **colocado** em a tabela , o Grêmio .

A principal dificuldade de lidar com esse tipo de construção é formal, uma vez que o "sintagma" que indica a comparação ("mais do que tem o segundo colocado"; "menos do que pagam os vizinhos") aparece segmentado, impedindo a associação entre os advérbios que puxam a comparação e os advcl da construção comparativa.
Foram pensadas três alternativas para lidar com esses casos:

a) anotar os advérbios como ArgM-ext (dedicada a quantidades) e deixar o elemento advcl como ArgM-mnr (maneira), o que "quebra" a ideia de comparação em duas: *Que quantidade de propina o brasileiro paga*? Menos. *De que maneira ele paga*? Da maneira como pagam os vizinhos da América Latina.

b) anotar os advérbios como ArgM-ext e criar uma etiqueta especial para esses casos de *advcl* (por ex, ArgM-comp)

c) anotar os advérbios como ArgM-ext e deixar o elemento *advcl* também como ArgM-ext (o que dá a entender a existência de dois ArgM-ext distintos ArgM-comp)

Opção foi pela alternativa (b), porque, **apesar** de criar uma etiqueta nova, é a opção que **possibilita a identificação posterior desses casos** sem confundi-los com casos genuínos *quantidade* e de *maneira*, como os exemplos abaixo, e porque não leva à falsa leitura de que o verbo conta com 2 ArgM-ext.

- As ações de a Petrobras **fecharam** *em alta (ArgM-ext)* , acompanhando (ArgM-mnr) a valorização de os preços de o petróleo em o exterior .
- Segundo o estudo , *sozinha (ArgM-mnr)* , **consumiu** , *em média (ArgM-ext)* , 46 % de o gasto público como um todo em a última década .
- " É muito interessante ver uma tecnologia que , sem substituir o ser humano(ArgM-mnr) , **aumentará** a segurança *substancialmente (ArgM-ext)* " , diz a gerente .

Assim, foi criada a etiqueta **ArgM-comp** para as construções *advcl* que participam de comparações**.**

- Causa tanta ou mais **aversão**(**Arg1**) que Temer(ArgM-comp) ou Eduardo Cunha , abjeções quase absolutas .
- Ficaria **melhor(Arg2)** que a bobeira(ArgM-comp) de noticiar picuinhas e abobrinhas.
- Os países que abraçaram sua herança colonial, de maneira geral, se saíram **melhor(Arg1)** do que aqueles(ArgM-comp) que o repudiaram.

ArgM-src: fonte da informação (source)

De acordo com as diretrizes do Propbank original (Bonial et al., 2015), sintagmas como o sublinhado abaixo – anotados como PRN (parenthetical) no treebank – não devem ser anotados.

> `It is the livestock sector, according to a new report by the United Nations Food and Agriculture Organization, that generates more greenhouse gases than any other industry.'

No entanto, considerando a relevância argumentativa/retórica de estruturas como a listada acima no que se refere à detecção da fonte da informação/proposição veiculada, decidimos anotá-las com a etiqueta **ArgM-src** (para indicar 'source', fonte da informação). A etiqueta inclui desde casos prototípicos, como construções introduzidas por *segundo Fulano*; *conforme Fulano*, *em sua opinião*, *Na visão de…*, *Nas palavras de*, *de acordo com…* , *na opinião de*, *do ponto de vista de…* , etc.

- Segundo o **balanço** de o banco em junho , o BNDES tinha R$ 173 bilhões em caixa .
- De **acordo** com a polícia , trata se de uma " prisão significativa " para as investigações .
- Em a **visão** feminista atual a música é machista , objetifica a mulher .

mas também em contextos menos óbvios, em que não se trata exatamente da fonte da informação veiculada na proposição, mas fonte de opinião relativa à proposição veiculada.

- Pergunto me por que tudo , para o **articulista** , tem que se resumir a uma dicotomia PT-PSDB ?
- **Para** os homens , vir o barbeiro e cortar o cabelo uma vez durante uma internação longa é suficiente .
- Além de isso , **para** os cidadãos cariocas , mais de a metade de os moradores de o Rio acredita que os usuários de droga são mais responsáveis por a violência em a cidade de o que os traficantes .

ArgM-conseq: consequência

Usada para indicar consequência das proposições. Em termos tradicionais, são as orações adverbiais consecutivas. Na anotação original do PropBank, ficam sob o guarda-chuva das ArgM-adv, decisão justificada pelo fato de se referirem a todo o conteúdo da oração, e não apenas ao verbo. No entanto, decidimos separá-las das ArgM-adv devido tanto à *relevância de seu conteúdo* (a ideia de "consequência" de uma ação parece interessante de ser codificada de maneira independente dos demais casos de ArgM-adv) e quanto à e*xistência de pistas formais*, em português, que auxiliam a sua detecção. Nas frases abaixo, o trecho sublinhado é a ArgM-conseq (e a etiqueta é atrelada ao token cujo deprel é advcl), e o verbo da proposição principal está em negrito.

> Mas a ciranda tecnológica vem **girando** **tão** rápido que voltar a atenção primordialmente a os estudantes *deixa* a já existente força de trabalho esquecida .
>
> A partir de o momento em que você **fica** **tão** rico que *precisa* contratar um segurança particular , seu nível de bem-estar cai substancialmente .

Também são anotadas como ArgM-conseq orações introduzidas por gerúndio, equivalentes a "o que":

> Deveria ter sido apoteótico , mas foi apenas tépido , **indicando** o clima geral de a apresentação . (= o que indica….)
> Quatro categorias respondem , em conjunto , por 90 % de a indústria , **sinalizando** de maneira inequívoca a preferência de o investidor . (= o que sinaliza…)

Etiquetas provisórias

Na versão atual do corpus (v1) existem ainda as seguintes etiquetas:

Etiqueta INC

INC vem de "incompleto", e a etiqueta foi utilizada para indicar (i) verbos sem frame no Verbo-Brasil e (ii) argumentos de um verbo não listado no Verbo-Brasil. Em ambos os casos, a utilização de INC não foi exaustiva, e, portanto, (i) nem todos os verbos ausentes do Verbo-Brasil têm INC no campo DEPS/coluna 9 e (ii) nem todos os argumentos de verbos ausentes do Verbo-Brasil têm INC no campo MISC/coluna 10, a maior parte deles têm a indicação de papel semântico, ainda que provisório.

Etiquetas Arg0_mod; Arg0_aux; Arg0_asp; Arg1_mod; Arg1_asp; Arg1_xcomp; Arg1_aux

Etiquetas que existem **apenas na versão base**, a partir da qual são criadas as versões clássica e ud (seção 4.8). Estas etiquetas foram criadas para destacar, provisoriamente, proposições envolvidas em construções com *verbos auxiliares*. Como mencionado na seção 3.2, a anotação sintática do Porttinari-base é bastante econômica no que se refere a classe de auxiliares, e na camada sintática do Porttinari apenas auxiliares de tempo composto e voz são considerados auxiliares. Em consequência, demais verbos, como *dever, poder*, etc, são classificados como VERB e, por isso, recebem papel semântico. Assim, em uma frase como abaixo, "fazer" é argumento do verbo *poder*, e não uma locução verbal "poder fazer". Do mesmo modo, "nós" é sujeito tanto de "poder" quanto de "fazer".

- "Nada que **nós possamos fazer** no Ministério pode ser contra isso", disse.

Para diferenciar esses casos daqueles em que temos de fato um verbo como argumento de outro, como nas frases abaixo, foram criadas etiquetas que sinalizam que, apesar de estarmos diante de um argumento de verbo, trata-se de um argumento "fraco", atribuído apenas devido à necessidade de classificar argumentos que, em outra abordagem gramatical, talvez não fossem classificados.

- Evair Evaristo Lima , técnico de a Acquazero , franquia de lavagem a seco , **ensina** a **fazer** a limpeza em casa .
- Foi por não vislumbrar um pré-candidato a presidente em quem ela mesma pudesse **votar** que Valéria **decidiu entrar** em o jogo .
- E se ele **resolver falar** tudo o que sabe pode , sim , prejudicar muita gente .

Assim, as etiquetas **Arg0_mod; Arg0_asp; Arg0_aux** indicam que estamos diante de um Arg0 (possível agente), mas Arg0 de um verbo que pode ser considerado ou auxiliar modal (Arg0_mod) ou um auxiliar de aspecto (Arg0_asp) outro tipo (Arg0_aux). Assim, na frase (1), "eu" é Arg0_mod do verbo "poder" (e Arg1 do verbo "sair"). Na frase (2), "Mercedez" é Arg0_asp do verbo "começar".

- **Eu posso sair** de esse país quando e como quero .
- **Mercedez começou** a **ficar** deprimida quando percebeu que praticamente todos os seus amigos haviam partido .

Do mesmo modo, as etiquetas **Arg1_mod; Arg1_asp; Arg1_xcomp; Arg1_aux** foram atribuídas aos verbos que seriam Arg1 de verbos listados como auxiliares em Duran (2014) ou considerados auxiliares nos *frames* do Verbo-Brasil.

Uma vez que essas etiquetas são atribuídas de maneira direta em função do lema do verbo, e levando em conta também a alta frequência na língua de construções com verbos auxiliares, a utilização das etiquetas provisórias associadas à auxiliaridade ajuda a diminuir a quantidade de Arg0 e Arg1 de atribuídos de forma 100% regular, mascarando o aprendizado/generalização dessas classes onde realmente importam.

2.3. Elementos não anotados (provisoriamente)

2.3.1. Sujeitos e discurso relatado

Não foi anotada a relação entre sujeitos de verbos que introduzem discurso relatado (ccomp:speech) e o conteúdo do discurso relatado. Assim em

- "Fizemos progressos consideráveis e continuaremos fazendo o nosso melhor para reduzir as ameaças em esse país ", **disse** **Cressida** também a a BBC .

Indicamos que Cressida é Arg0/sujeito de “disse”, mas não anotamos que Cressida é Arg0/sujeito de “fizemos”.

Esta é uma decisão que pode - e talvez deva - ser deixada de lado. Como UD indica a presença de discurso relatado com a etiqueta ccomp:speech, a identificação desses casos para posterior anotação não deverá ser complexa. Abaixo, mais alguns exemplos, nos quais há situações de compartilhamento de sujeito (frase 1), e nos quais não há (frases 2 e 3).

- " Treino muito e procuro ser frio a o cobrar " , **disse** **o atacante** que anotou seis gols , todos com paradinha , mas deslocando o goleiro , em o Brasileiro .
- " Tão lindo , e de perto parece um mendigo " , **disse** **uma fã** a o ver lo .
- " Sempre trabalhou muito , mas não deixava de viver para a gente " , **lembra** a **filha** Claudia .

2.3.2. Explicitação de objetos

Casos como o abaixo, não foram (ainda) tratados de forma sistemática, ou seja, não foi explicitada a relação entre entre “sapos” e “ferver”

- Pura lenda urbana (na verdade , rural - dificilmente você encontrará sapos pra **ferver** em a cidade) .

2.3.3. Correferência

Não foram estabelecidas relações entre os elementos em destaque nas frases abaixo:

- Para isso , o **usuário** responde a quatro perguntas : qual o objetivo financeiro , em quanto tempo pretende usar o dinheiro , qual seu perfil como investidor e como **ele** prefere receber os rendimentos .
- Há muitos casos de **venezuelanas** em estágio avançado de gravidez **que** tentam chegar a os hospitais de Cartagena antes de o parto para garantir melhor atendimento a seus recém-nascidos.
- Talvez a **liga** a unir em as tramoias políticos poderosos a empresários próximos de o Estado fosse tão resistente que apenas órgãos de controle dotados de superpoderes seriam capazes de romper **la** .

3. Especificidades (ou interferências) de UD na anotação SRL

Tradicionalmente, a anotação de papéis semânticos parte de treebanks, já segmentados quanto aos constituintes. Subjacente a esta prática, a ideia de que a identificação de constituintes de um enunciado é tarefa objetiva e inequívoca. Como ilustrado desde Sampson (2002), trata-se de uma crença discutível, e alguns exemplos de diferentes possibilidades de segmentação estão nesta seção. O PropBank original, por exemplo, apesar de não discutir diferenças de segmentação, informa a existência de desalinhamentos entre a segmentação do *Penn TreeBank* e as escolhas de segmentação da anotação de papéis semânticos, como citado na seção 2.

A anotação do PropBank é feita sobre o treebank Porttinari-base, anotado conforme a teoria *Universal Dependencies* (de Marneffe et al., 2021). As chamadas Dependências Universais (UD) não são apenas um aparato formal, mas uma teoria linguística com seus princípios próprios. Nesta seção, elencamos os pontos em que as opções teóricas da anotação sintática em geral, e de UD em especial, influenciaram, ou precisaram ser deixadas de lado, na anotação de papéis semânticos.

3.1. Segmentação em constituintes

Desalinhamentos referentes à segmentação já foram referidos na literatura do PropBank original, ainda que com pouca ênfase. Na anotação do PBP, em alguns casos a segmentação em argumentos derivada do Verbo-Brasil não encontrava respaldo na anotação sintática subjacente. Nos exemplos da quadro 1, abaixo, os elementos em azul correspondem a argumentos de verbos/frames que estão indicados nos quadros 2 a 5, a seguir, mas foram considerados modificadores dos nomes (sublinhados) na anotação UD, porque ou se trata de um argumento de um verbo considerado verbo-suporte com nome predicativo (frase 1); ou porque a anotação sintática não pode violar o princípio do não cruzamento de arcos (frases 2 e 4), que será abordado na seção 3.2; ou porque a anotação sintática escolheu uma análise alternativa, na qual o elemento nominal foi considerado o argumento de um nome predicativo. Assim, por exemplo, na frase (3), "sobre seu trabalho" é um argumento previsto pelo substantivo "influência" (e equivalente a "influenciar seu trabalho"), mas a análise do Verbo-Brasil referente ao verbo *exercer* (alguém (Arg0) exerce algo (força, influência, Arg1) sobre algo ou alguém (Arg2)), indicada no quadro 4, é igualmente legítima no contexto da frase.

1. (...) isto pode **ter** efeito positivo em o **grupo** .
2. Seus procuradores **entraram** com processos em a **Justiça**
3. Segundo Thaler , Kahneman **exerceu** grande influência sobre seu **trabalho** .
4. São Paulo **dispõe** de um plano B para diversos **gostos** e endereços: as piscinas públicas .

Quadro 1: Divergências de segmentação entre a camada sintática e a camada SRL. Em negrito, a anotação SRL, com tokens (em azul) associados aos verbos. Em sublinhado, anotação da camada sintática, com os tokens em azul associados aos substantivos sublinhados.

Roleset id: **ter.09 ,** *exercer*, vncls: , Mapeamento para o inglês:
Arg0: entidade ou coisa que exerce (vnrole: -agent)
Arg1: poder exercido (vnrole: -theme)
Arg2: sobre o que o poder é exercido (vnrole: -patient)

Quadro 2: Frame de ter.09 conforme Verbo-Brasil

Roleset id: **entrar.04**, submeter oficialmente (ação, pedido), apelar judicialmente, vncls: 9.10, 31.4-3, Mapeamento para o inglês: file.01, appeal.02:
Arg0: iniciador de processo (vnrole: 9.10-agent)
Arg1: coisa submetida (entrar com) (vnrole: 9.10-theme)
Arg2: órgão que recebe a submissão
Arg3: entidade contra a qual arg1 é movido

Quadro 3: Frame de entrar.04 conforme Verbo-Brasil

Roleset id: **exercer.01**, chegar exercer.01 , provocar; acarretar, vncls: , Mapeamento para o inglês: exert.01:
Arg0: causa ou causador (vnrole: -agent)
Arg1: coisa exercida (pressão, influência, poder) (vnrole: -theme)
Arg2: afetado (exercer sobre) (vnrole: -patient)

Quadro 4: Frame de exercer.01 conforme Verbo-Brasil

Roleset id: **dispor.01**,ter, possuir, vncls: -, Mapeamento para o inglês: possess.01:
Arg0: possuidor
Arg1: coisa possuída (dispor de)
Arg2: finalidade a que se presta a coisa possuída (dispor para)

Quadro 5: Frame de dispor.01 conforme Verbo-Brasil

3.1.1.As mwes verbais

De forma associada está o tratamento das mwes verbais, e a relevância, para qualquer PropBank, de um tratamento sistemático de construções verbais com os chamados verbos leves ou suporte. Nestes casos, a diferença na anotação é decorrente de decisões de anotação da camada UD posteriores à criação do Verbo-Brasil. Na camada sintática, em construções com verbo suporte, a opção foi anotar os argumentos do nome predicativo, associado ao verbo leve/suporte, como como argumentos dos nomes (nmod) e não como obl associados ao verbo. Em consequência, frases como 5, 6 e 7, divergem na identificação de constituintes, como indicado no quadro 6.

O exemplo da frase 5 é curioso porque envolve leituras distintas, cuja divergência de análise talvez fosse capturada em um estudo de concordância. Na camada sintática, “de todas as

formas” foi considerado um modificador nominal de “pedradas”. Ou seja, *levamos pedradas* “grandes”, “arredondadas”, “marrons”, etc. Na camada SRL, “de todas as formas” foi lido como modificador de “levar” (ou melhor, de “levar pedrada”, sinônimo de “apanhar/ser criticado”).

5. E existem companheiros que até hoje dizem assim, ' poxa , nós buscamos fazer o melhor e **levamos** pedradas de todas as **formas** '
6. Além , é claro de ainda **levar** encoxadas de os **homens**
7. Em a época , o episódio também **deu** início a uma **série** de protestos .

Quadro 6: Divergências de segmentação entre a camada sintática e a camada SRL. Em negrito, a anotação SRL, com tokens (em azul) associados aos verbos. Em sublinhado, anotação da camada sintática, com os tokens em azul associados aos substantivos sublinhados.

Já as frases 6 e 7 não envolvem diferentes leituras possíveis, mas os verbos “levar” e “dar” como verbo suporte. Mais uma vez, o frame do Verbo-Brasil propõe uma análise distinta (quadros 7 e 8) , que foi seguida na anotação SRL.

Roleset id: **dar.401**, *iniciar*, vncls: , Mapeamento para o inglês: initiate:
Arg0: iniciador (vnrole: -agent)
Arg1: início (fixo)
Arg2: *evento iniciado (dar início a)* (vnrole: -theme)
Argm: modo ou instrumento (dar início com, dar início + gerúndio) (vnrole: -instrument)

Quadro 7: Frame de dar.401 (dar início) conforme Verbo-Brasil

Roleset id: **levar.05.**, *receber (oposto de "dar"), vncls: , Mapeamento para o inglês:*
Arg1: paciente (vnrole: -patient)
Arg2: ação sofrida (surra, tapa, coice, pontapé, safanão, golpe, coice, susto, calote, etc.)
Arg3: agente da ação descrita pelo arg2 (levar x de) (vnrole: -source))

Quadro 8: Frame de levar.05 conforme Verbo-Brasil

Por fim, cabe notar que a anotação da camada SRL foi concomitante à revisão final da camada sintática, e é possível que alguns casos considerados “desalinhamento” desapareçam. Por outro lado, os casos remanescentes se oferecem como um campo rico de estudo.

3.2. Cruzamento de arcos sintáticos, locuções verbais e verbos auxiliares

Conforme explicam Pagano et al. (2023), existe uma orientação geral na sintaxe de dependências para se evitar cruzamento de arcos sintáticos. No entanto, relações sintáticas de longa distância, bem como línguas de ordem livre, acabam levando a árvores com arcos cruzados, chamadas relações não-projetivas. E uma das alegadas vantagens da gramática de dependências é justamente lidar melhor com esses fenômenos – longas distâncias e ordem livre.

Assim, e voltando ao que afirmam Pagano et al., há casos pontuais em que o cruzamento de arcos é necessário para estabelecer a relação correta entre duas palavras, como na frase exemplo *A menina postou uma foto no Instagram com filtro* (figura 6). No exemplo, para evitar o cruzamento de arcos a sentença precisaria ter uma outra ordem: "*A menina postou no Instagram uma foto com filtro.*"[16] Por isso, se não quisermos violar a restrição do cruzamento de arcos, ficamos com o sentido de que o sintagma "com filtro" está associado a "Instagram" e não ao substantivo "foto":

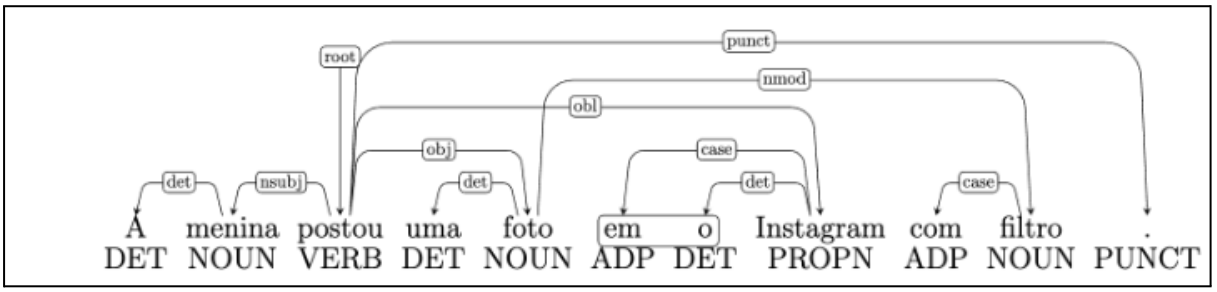


Figura 6: Cruzamento de arcos do sintagma "com filtro" para a frase "A menina postou uma foto no Instagram com filtro". Fonte: Pagano et al. (2023).

Embora não seja uma restrição de UD, na anotação sintática do Porttinari-base os arcos das relações sintáticas não podem se cruzar: "a atribuição de relações de dependência deve observar o princípio da projetividade, ou seja, os arcos das relações não devem se cruzar"(Duran, 2022:9). Duran (2022) também informa que o cruzamento "deve ser evitado a todo custo, **mesmo que isso implique uma interpretação menos fiel** de alguma relação sintática" (Duran, 2022:9, grifo nosso). Para a frase exemplo, a leitura decorrente da anotação seria, portanto, a do sintagma "Instagram com filtro", e não "foto com filtro".

Assim, ao evitar o cruzamento de arcos envolvendo especificamente verbos (o que se relaciona com a anotação SRL), corremos o risco de codificar sintaticamente leituras com significados distantes daqueles convencionalizados. Para a frase 1 (que tem a locução verbal "demorar a reconhecer", exemplificando o cenário descrito por Duran 2022:129), ficamos com a leitura 1a quando evitamos o cruzamento de arcos na anotação sintática, enquanto a leitura adequada seria 1b.

[16] No entanto, e como sabemos, a língua não acontece como deseja a teoria, é precisamente o contrário: a teoria vai sendo elaborada a partir (dos dados) da língua, como maneira de organizá-los e explicá-los.

1. O defeito, **que** a Takata demorou a **reconhecer** , foi revelado em 2014 e, desde então , várias montadoras foram obrigadas a anunciar uma série de recalls , que afetaram quase 100 milhões de veículos .
    a. Leitura sem cruzamento: Takata demorou o defeito
    b. Leitura com cruzamento: Takata reconheceu o defeito

Associado ao cruzamento de arcos com locuções verbais está o tratamento dado por UD e pela anotação sintática do Porttinari-base (Duran, 2022) aos **verbos auxiliares**. UD é bastante econômica na adoção de auxiliares[17], o que traz impactos para uma análise do significado adequada. Na frase 2, o verbo "acabar" é considerado um verbo pleno camada sintática, mas é possível considerá-lo um auxiliar de aspecto[18] (inclusive uma tradução possível é "He quoted poems he had ***just*** read" (e porque a leitura de "citava poemas que havia lido até o final" é menos plausível)[19]. Com isso, ficamos com um verbo pleno e, dada a restrição do cruzamento de arcos, torna-se impossível indicar que "poemas" é o que ele *acabou de ler* (leitura 2b), e somos obrigados a vincular "poema" a "acabar" (leitura 2a), o que não tem nenhuma interpretação possível no contexto da frase[20].

2. Citava poemas **que** acabara de **ler** ou, em jovem, aprendera com os tios.
    a. Leitura sem cruzamento: Ele acabou os poemas
    b. Leitura com cruzamento: Ele leu os poemas

Considerar o verbo "acabar" em contextos como esse um verbo auxiliar (AUX) - **análise codificada no Verbo-Brasil** - permite uma análise sintática sem cruzamento de árvore, uma vez que permite associar "poemas" a "ler", resolvendo o impasse e permitindo a análise adequada da frase. Apesar dessa observação, nenhuma alteração nas camadas de anotação sintática foi feita na anotação SRL.

Outro exemplo/tipo de cruzamento de arcos com consequências relevantes para a anotação SRL está na frase 3. Uma leitura convencional indicaria que o sintagma "a mãe cozinhar" é o Arg1 de "observando", ou seja, "a mãe cozinhar" é a *coisa observada*. No entanto, a análise UD impede que a atribuição de papéis se dê dessa forma, porque isto implicaria em "mãe" ser sujeito de "cozinhar", e xcomp - etiqueta atribuída a "cozinhar" na anotação sintática -, por definição, não tem sujeito, embora, na frase, saibamos que quem cozinha é a mãe. A anotação sintática leva a uma leitura como 3a, e a anotação SRL leva a uma leitura como

[17] Verificamos que, recentemente, as diretivas para auxiliares em português foram atualizadas em UD https://universaldependencies.org/pt/dep/aux_.html, mas a anotação do Porttinari-base ainda não passou por esta atualização.
[18] A análise de "acabar" como auxiliar aspectual está codificada no Verbo-Brasil sob os frames acabar.201; acabar.202 e acabar.203.
[19] Reconhecemos a dificuldade - e talvez impossibilidade - de delimitar com clareza os limites entre auxiliares e verbos plenos, sobretudo porque se trata de um *processo* de lexicalização e, enquanto *processo*, existem elementos mais claramente lexicalizados (como o "just read") e outros nem tanto (como em "acabou a leitura do poema"). E embora em alguns casos a leitura permita mais de uma interpretação, em outros é inequívoca, como em "ele acabou de chegar", em que a leitura de "acabar" como verbo pleno não é possível.
[20] Esta leitura até seria possível se a frase fosse "Citava poemas que acabou de escrever/criar" → ele acabou os poemas.

3b. No entanto, além da semântica pouco fiel, a **anotação da camada sintática traz outro complicador para a anotação SRL**: cria-se um terceiro espaço argumental para "observar" (espaço preenchido com "cozinhar"), que precisará receber alguma etiqueta Arg. Ou seja, se a anotação SRL fosse feita diretamente sobre a anotação UD, ficaríamos com um xcomp associado a um verbo sem papel argumental (ou que receberá um papel argumental genérico e cujo sentido não corresponde ao sentido do enunciado no contexto, como ArgM-adv)[21].

3. Na pequena Itajubá, no sul de Minas, Regina **cresceu observando a mãe cozinhar.**
    a. Leitura sem cruzamento: observando [a mãe] [cozinhar]
    b. Leitura com cruzamento: observando [a mãe cozinhar]

A figura 7 mostra como fica a anotação na camada SRL em situações como essa:

12	Regina	Regina	PROPN	_	_	13	nsubj	_	Arg1
13	cresceu	crescer	VERB	_	Mood=Ind|Number=Sing|Person=3|Tense=Past|VerbForm=Fin	0	root	_	_
14	observando	observar	VERB	_	VerbForm=Ger	13	advcl	_	ArgM-mnr
15	a	o	DET	_	Definite=Def|Gender=Fem|Number=Sing|PronType=Art	16	det	_	_
16	mãe	mãe	NOUN	_	Gender=Fem|Number=Sing	14	obj	_	Arg0:17
17	cozinhar	cozinhar	VERB	_	Number=Sing|Person=3|VerbForm=Inf	14	xcomp	_	Arg1:14
18	.	.	PUNCT	_	_	13	punct	_	SpaceAfter=No

Figura 7: Exemplo de frase com cruzamento de arcos (em rosa) no PBP

Em todos esses casos, a decisão de anotação da camada SRL foi ignorar a camada sintática e atribuir papéis semânticos de acordo com o sentido das frases, conforme indicado em 1 e 2, abaixo.

1. O defeito, **que** a Takata demorou a **reconhecer** , foi revelado (...) → "que" (defeito) é **Arg1** do verbo "reconhecer"
2. Citava poemas **que** acabara de **ler** ou, em jovem, aprendera com os tios. → "que" (poemas) é **Arg1** do verbo "ler"

Ou seja, na anotação da camada SRL, o arco "cruza": considerando a frase 2, "poemas" é Arg1 (documento/tema) de "ler", e não foi anotada qualquer relação entre o obl "que" "acabar" (figura 8).

[21] Futuramente, esta situação poderá ser resolvida pela equipe do Verbo Brasil, com a alteração do frame do verbo "observar" e de demais verbos que admitem sujeito acusativo, como "mandar", "ver", "obrigar", "sentir", "ouvir", "esperar", prevendo mais um espaço argumental.

sent_id = FOLHA_DOC004025_SENT012
text = Citava poemas que acabara de ler ou, em jovem, aprendera com os tios.

1	Citava	citar	VERB	_	Mood=Ind|Number=Sing|Person=1|Tense=Imp|VerbForm=Fin	0	root	_	_
2	poemas	poema	NOUN	_	Gender=Masc|Number=Plur	1	obj	_	Arg1
3	que	que	PRON	_	PronType=Rel	4	obl	_	Arg1:6
4	acabara	acabar	VERB	_	Mood=Ind|Number=Sing|Person=1|Tense=Pqp|VerbForm=Fin	2	acl:relcl	acabar.202_aux-asp	_
5	de	de	ADP	_	_	6	mark	_	_
6	ler	ler	VERB	_	VerbForm=Inf	4	xcomp	_	Arg1_asp_xcomp

Figura 8: Exemplo de frase com cruzamento de arcos no PBP

Além do já mencionado desalinhamento entre as anotações sintáticas e anotações SRL, outra consequência do não cruzamento de arcos para a anotação SRL é uma **explosão da classe ArgM-loc**, dedicada a locais, ainda que virtuais e, consequentemente, a perda de informação, sobretudo vinculada aos **nomes predicativos**. Na frase 4, a informação veiculada é "limitações na capacidade de ensinar", e não *sentir [limitações] [em algum lugar]*, que é a segmentação/anotação possível decorrente da análise sintática que evita o cruzamento de arcos. Na frase 5, e levando em conta o contexto da frase, temos que a *vaga no torneio está assegurada*, e não que a vaga foi assegurada ao jogar o torneio.

4. Certamente, deveríamos ouvir os professores quando expressam as *limitações* **que** sentem em sua **capacidade** de ensinar o currículo .
    a. Leitura sem cruzamento: sentir [o que: limitações] [onde: na capacidade]
    b. Leitura com cruzamento: sentir [limitações na capacidade]
5. Ou leve e solto exatamente porque com a **vaga** assegurada em o **torneio** continental ?
    a. Leitura sem cruzamento: assegurar [o que: vaga] [onde:no torneio]
    b. Leitura com cruzamento: assegurar [o que: vaga no torneio]

De modo inverso, o não cruzamento de arcos pode ter o efeito oposto: não realizar a segmentação esperada, e com isso dificultar a atribuição de papéis semânticos, uma vez que ficarão mais difíceis de serem encontrados por deixarem e estar associados ao verbo. Este é o caso da frase 6, na qual para evitar o cruzamento de arcos é necessário concatenar os sintagmas "em Miami" e "na segunda", e assim deixamos de especificar a informação relativa a *quando* o pouso foi feito.

6. Pousaram em **Miami em a segunda** (4) , onde ficariam até embarcarem em um cruzeiro por o Caribe em o dia 16 .
    a. Leitura sem cruzamento: pousaram [onde: em Miami na segunda]
    b. Leitura com cruzamento: pousaram [onde:em Miami] [quando:na segunda]

No entanto, diferentemente dos demais casos, em que forçamos o desalinhamento, **nestes casos** (do tipo 4, 5 e 6) **a análise SRL acompanhou a anotação sintática**, como indica a figura 9 abaixo, referente à frase 4, e fica como ponto a ser discutido o tipo de análise desejável, nesses casos, para a camada SRL.

10	limitações	limitação	NOUN	_	Gender=Fem|Number=Plur	8	obj	_	Arg1
11	que	que	PRON	_	PronType=Rel	12	obj	_	Arg1
12	sentem	sentir	VERB	_	Mood=Ind|Number=Plur|Person=3|Tense=Pres|VerbForm=Fin	10	acl:relcl	_	FEITO
13	em	em	ADP	_	_	15	case	_	_
14	sua	seu	DET	_	Gender=Fem|Number=Sing|Person=3|Poss=Yes|PronType=Prs	15	det	_	_
15	capacidade	capacidade	NOUN	_	Gender=Fem|Number=Sing	12	obl	_	ArgM-loc
16	de	de	ADP	_	_	17	mark	_	_
17	ensinar	ensinar	VERB	_	VerbForm=Inf	15	acl	_	FEITO
18	o	o	DET	_	Definite=Def|Gender=Masc|Number=Sing|PronType=Art	19	det	_	_
19	currículo	currículo	NOUN	_	Gender=Masc|Number=Sing	17	obj	_	Arg1

Figura 9: Exemplo de anotação do PBP

Como vemos, a restrição de cruzamento de arcos é um problema para a anotação de papéis semânticos, uma vez que, se formos guiadas pela sintaxe, chegaremos a frases cuja representação semântica – objetivo principal de uma anotação de papéis semânticos – é inadequada. O quadro 6 abaixo traz mais alguns exemplos de diferenças nas análises.

Frase	leitura **sem cruzamento** de arcos (camada sintática)	leitura **com cruzamento** de arcos (camada PBP)
Teve até congresso em Londres para discutir o que fazer com o *cheiro* de borracha queimada **que** a uva costuma **produzir** (além dos vinhos pouco convincentes).	costumar cheiro	produzir cheiro
As medidas replicam normas **que** o **mercado** financeiro é obrigado a **seguir**.	obrigar normas	seguir normas
Em sua luta para se livrar da zona de rebaixamento do Brasileiro , São Paulo tem fracassado justamente contra os adversários **que** mais **precisava vence**r .	precisar adversários	vencer adversários
E os dois com um tom bem diferente daquele **que** ele **costuma exibir** na TV	costumar um tom	exibir um tom
Ao fazer uma busca em o local, os agentes acharam, **armazenados** em um depósito, 213 **tabletes** de cocaína .	acharam armazenados	tabletes armazenados
Vítima de uma doença genética , essas são as ações **que** ele consegue **fazer** .	conseguir ações	fazer ações
No ano passado , **viu** os **negócios crescerem** 130 % ,	viu [os negócios] [crescerem]	viu [os negócios crescerem]

mas nem sempre a trajetória de suas empresas foi positiva assim .[22]		

Quadro 6: Diferença entre as análises conforme a anotação sintática e anotação SRL

3.3. Tratamento especial dos verbos ser e estar (cópula)

O fato de verbos de cópula não serem considerados "verbos", mas "auxiliares" – e terem seus argumentos sintáticos descolados – torna a procura por estas estruturas sintáticas mais complexa. A frase (1), por exemplo, precisa de uma regra distinta da frase (2), inventada, apesar de estarmos diante de estruturas muito parecidas.

1. Em a parte inicial de o show , cantando sozinha , Maria Rita **foi** muito bem , em peças variadas de jazz .
2. Em a parte inicial de o show , cantando sozinha , Maria Rita **performou** muito bem, em peças variadas de jazz .

Além disso, devido à anotação sintática do Porttinari-base não distinguir entre argumentos preposicionados do tipo *core* e *não-core* (entre obl e obl:arg), precisamos distinguir entre verdadeiros obl de uma oração com verbo "ser" (3 e 4) dos obl que modificam os ADJ que, devido à abordagem UD, passaram a root (5, 6 e 7):

3. Em o barco , o espaço é limitado , e a convivência , intensa.
4. Em os últimos anos , Zé Mineiro estava distante de o processamento de carnes.
5. Não é benéfico para as seleções jogar amistosos .
6. Entretanto, ainda estamos distantes de o ajuste de as contas públicas .
7. Esse comportamento não é exclusivo de esse caso .

3.4. Preferência pela literalidade e pela composicionalidade

A preferência pela literalidade e composicionalidade em UD se manifesta, por exemplo, nos critérios de identificação das *mwes fixed* e no tratamento dado às combinações verbais. No primeiro caso, combinações como "devido a", "a seguir" e "a partir de" são analisadas literalmente, e "dever", "seguir" e "partir" são considerados verbos, o que os "obriga" a ter um frame e argumentos associados ("devido ao anúncio"; "a partir de sábado").

3.5 Auxiliares

Como já indicado, a sintaxe UD, à época da construção do Porttinari-base, era bastante econômica quanto ao que deve ser considerado auxiliar, e no Porttinari estão anotados

[22]Como indicado na nota 21, esta situação poderá ser resolvida com a alteração dos frames de verbos que admitem sujeito acusativo.

como AUX apenas auxiliares de tempo e voz. Na atribuição de papéis semânticos, porém, esta economia tem como resultado a necessidade de atribuir papéis a elementos que sabemos, em português, não estarem atuando como verbos plenos, e que portanto não recebem papéis semânticos. Por isso, como indicado na seção 2.2.1 (ArgM-mod e ArgM-asp) e em 2.2.2 (Etiquetas provisórias), implementamos algumas estratégias para que a anotação final dos papéis semânticos corresponda ao que se espera de uma *proposição*. As frases 1 e 2 exemplificam as diferenças de anotação.

1. Algumas empresas, como a Vale, oferecem formulários que podem ser preenchidos…

a) *Anotação se seguíssemos a sintaxe UD:* "que" (formulários) é Arg0 de "poder"
b) *Anotação intermediária:* "que" (formulários) é Arg0_ de "poder"
c) *Anotação final (pós-processada):* "que" (formulários) é Arg1 de "preencher"

2. O projeto do Senado ***prevê*** que <u>as deduções só possam ocorrer a partir de 2021</u>

a) *Anotação se seguíssemos a sintaxe UD*: "deduções" é Arg0 de poder; "poder" é Arg1 de "prever"; "ocorrer" é Arg1 de "poder"
b) *Anotação intermediária:* "deduções" é Arg0_mod de "poder"; "poder" é Arg1 de "prever" & MOD; "ocorrer" é Arg1_aux_xcomp de "poder"
c) *Anotação final (pós-processada):* "deduções" é Arg1 de "ocorrer"; "poder" é ArgM-mod; "ocorrer" é Arg1 de "prever"

Adicionalmente, a anotação do PropBank (Bonial et al., 2015) só admite a indicação de auxiliares de modo (ArgM-mod), o que faz com que, para a língua portuguesa, frases como "o mercado **acaba ficando** mais preocupado" levem à identificação de "mercado_ACABAR_ficar" como proposição. Por isso, seguimos **integralmente** a decisão adotada em Duran (2014) na anotação do PropBank-br versões 1 e 2, que consiste em aplicar etiquetas para todos os tipos de verbos auxiliares do português. Deste modo, além da etiqueta *ArgM-mod*, prevista no PropBank original, o PBP conta também com a etiqueta *ArgM-asp* para verbos auxiliares de *aspecto*, uma vez que estes estão anotados como verbos plenos. Do mesmo modo, também utilizamos as etiquetas *ArgM-tml* (para auxiliares de tempo) e *ArgM-pas* (para auxiliares de voz), apesar de os respectivos verbos auxiliares já estarem anotados como AUX na camada sintática, e portanto não têm espaços argumentais vinculados a eles. Embora redundante, esta decisão garante a independência entre as camadas de anotação sintática e de anotação de papéis semânticos (outros modelos de gramática podem variar na lista de verbos que consideram auxiliares de tempo, por exemplo). Do ponto de vista da anotação codificada no corpus, esses casos "redundantes" (*ArgM-tml* e *ArgM-pas*) não levam à anotação de Arg0_d e Arg1_d na versão UD, uma vez que estes elementos não têm dependentes sintáticos.

4. Metodologia

A anotação da camada PBP foi realizada a partir de uma versão do Porttinari-base anotado sintaticamente conforme a gramática UD. A anotação dos papéis semânticos foi feita, originalmente, na colunas 10 (coluna misc) para os argumentos dos verbos, e na coluna 9 (deps) para os frames dos verbos do arquivo conllu.

Em termos gerais, a anotação de papéis semânticos no PropBank seguiu as seguintes etapas:

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

A anotação dos **frames (sentidos) dos verbos** não foi prioridade desta fase, e foi concomitante à anotação dos papéis semânticos.

Todo o processo de anotação foi feito com base em **regras linguisticamente motivadas**, e de maneira **não linear**, seguindo o que já fizemos em Freitas et al. (2023), que por sua vez se inspira em Santos e Mota (2010). Fazemos uma anotação /revisão transversal, que permite ver os fenômenos em questão de forma ampliada e garante uma revisão consistente (Wallis, 2003).

No PropBank original, a anotação partiu de um sistema de anotação baseado em regras e léxico, cuja saída foi manualmente corrigida. O processo de anotação aconteceu em duas fases: uma primeira rodada de anotação dupla, seguida de uma fase de adjudicação das discordâncias detectadas na fase anterior. A anotação/revisão foi feita por lema verbal, e não seguindo a sequência do texto. O tratamento de cada lema, por sua vez, foi feito frase a frase. Metodologicamente, diferimos parcialmente do processo de anotação do PropBank original, e do PropBank-br versões 1 e 2, que seguiu a metodologia original, exceto pela fase inicial de anotação automática.

A anotação humana com base em regras permite avançar de maneira de maneira relativamente rápida (se comparada à anotação 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**.

A inspeção humana é parte fundamental de qualquer projeto de criação de datasets padrão ouro, mesmo quando baseado em regras linguisticamente motivadas. A importância da revisão reside em uma característica das línguas humanas, mais precisamente, em uma propriedade distribucional das línguas: a presença dos casos raros e não previstos (veja-se Freitas 2017 para uma discussão relativa ao caráter simultaneamente regular e irregular das línguas, com implicações para o PLN, e Sampson (2001, capit 10) para uma crítica à visão de linguagem como fenômeno centralmente regular). De fato, a insuficiência de abordagens

inteiramente guiadas por regras no PLN, quando se trata de processar “textos gerais”, ajuda a entender os limites de métodos simbólicos do mesmo modo que a insuficiência de abordagens inteiramente guiadas por frequência ajuda a entender os limites de métodos estatísticos (Freitas, 2017).

Os exemplos 1-4 demonstram a importância da análise linguística cuidadosa na preparação de material padrão ouro, uma vez que mesmo fenômenos linguísticos aparentemente de grande regularidade podem fugir às regras.

1. Ou sei lá qual é , hoje em dia , o termo "politicamente correto" para "cucaracha".
2. Se não colocam em risco a saúde humana ali , por que o fariam no resto do país?
3. Uruguai (70%) e Costa Rica (66 %) aparecem depois do Brasil.
4. A investigação da Zelotes corre em Brasília e não inclui até aqui o ex-prefeito Luiz Marinho.

No primeiro caso, temos o advérbio *lá*, típico de “local”, funcionando como uma negativa; no segundo, uma oração introduzida pela conjunção *se*, elemento tipicamente condicional, indica “causa”, e não “condição”; no terceiro, temos *depois*, tipicamente temporal, indicando “local” e no quarto exemplo, *aqui*, típico de “local”, indicando “tempo”. Com os exemplos indicamos, também, porque um bom desempenho em tarefas SRL pode ser difícil quando estamos diante de um material bem preparado e cuja anotação corresponde ao entendimento humano.

A anotação dos papéis semânticos no PBP foi feita em 3 fases:

- anotação de elementos explícitos;
- anotação de elementos implícitos;
- verificação final para busca de erros ou inconsistências.

A seguir, detalhamos cada uma das fases da anotação de papéis semânticos, bem como a anotação relativa aos frames dos verbos (alinhamento com o Verbo-Brasil)

4.1 Anotação de elementos explícitos

A anotação de elementos explícitos foi feita em 5 etapas:

1. Padrões léxico sintáticos derivados de exemplos
2. Pistas sintáticas simples
3. Argumentos numerados e regras complexas
4. Acabamentos
5. Validação

4.1.1. Etapa 1: Padrões léxico-sintáticos derivados das diretivas

A anotação tomou como ponto de partida os exemplos de frases indicados nas diretivas (associados aos ArgM), que serviram para a busca de padrões linguísticos, que por sua vez

foram sendo refinados e ampliados. Especificamente, esta etapa foi baseada em léxicos e padrões léxico-sintáticos, e aproveitou as listas, descrições e padrões presentes nas diretivas de anotação (Duran, 2014).

A partir de pistas indicadas nas diretivas, criamos uma série de padrões de busca e anotação. Considerando o trecho abaixo, referente à anotação de ArgM-tmp:

> Eles podem ser expressos por uma única palavra (*ontem, futuramente, frequentemente*), por um PP (*em novembro, de 5 a 9 de julho, desde 1933, durante a exibição do filme, na semana passada, cinco vezes por semana*) ou por uma oração completa (*ao acordar, quando veio ao Brasil, enquanto dormia, sempre que tem vontade, nunca ao se deitar, até não querer mais, assim que cheguei, tão logo cheguei, uma vez acordado*). (DURAN, 2014)

Podemos derivar uma série de padrões, que precisarão ser verificados quanto à adequação da etiqueta ArgM-tmp:

1. advérbios como o*ntem, futuramente, frequentemente* (e demais advérbios já listados nas diretivas)
2. PPs com função sintática obl e dependentes do verbo, cujo núcleo seja um mês do ano, uma data, um dia da semana, e demais palavras indicativas de tempo como "semana, mês, década, ano, minuto, vez")
3. sintagmas com função obl ou advcl introduzidos por elementos como *quando, enquanto, durante, assim que, sempre que, nunca, ao + verbo no infinitivo* etc.

A seguir, exemplificamos cada um três dos padrões listados acima.

Para buscas como as do item 1, criamos regras de anotação simples do tipo abaixo, que indica que se alguns dos lemas dos advérbios listados for encontrado, e for dependente de um verbo (e não de um adjetivo, por exemplo), ele receberá a etiqueta ArgM-temp. Alguns advérbios da lista podem ter sua natureza temporal questionada, como "recorrentemente", que também indica *maneira*. No entanto, uma vez que as diretivas especificam que "frequência" deve ser considerada ArgM-tmp, optamos por essa classe, mais específica que ArgM-mnr. A regra listada anotou 52 advérbios como ArgM-tmp;

```
if
regex("atualmente|recentemente|diariamente|regularmente|anter
iormente|imediatamente|inicialmente|posteriormente|rapidament
e|antecipadamente|anualmente|brevemente|eventualmente|raramen
te|recorrentemente|novamente|repetidamente|tradicionalmente",
token.lemma) and regex("VERB", token.head_token.upos):
      token.misc = "ArgM-tmp"
```

A regra diz que

se existe um token cujo lema (*token.lemma*) é “atualmente” ou “recentemente” ou “diariamente” (e assim por diante) e o head desse token tem o atributo upos (*token.head_token.upos*) do tipo “VERB”, então o token terá o campo misc (*token.misc*) anotado como “ArgM-tmp”

O item 2 indica que elementos temporais como meses do ano, dias da semana etc costumam se referir a argumentos temporais, e usamos a regra abaixo para anotar este tipo de ArgM-tmp.

```
if regex("[Ee]m", token.previous_token.word) and
regex("janeiro|fevereiro|março|abril|maio|junho|julho|agosto|
setembro|outubro|novembro|dezembro", token.lemma) and
regex("obl", token.deprel) and regex("VERB",
token.head_token.upos):
    token.misc = "ArgM-tmp"
```

A regra diz que

se existe um token cuja forma (word) é “em” ou “Em” e após esse token (que está indicado por “token.previous_token.word”) há um token cujo lema corresponde aos meses do ano (que está indicado por “token.lemma”), e este mesmo token tem deprel “obl” e o head desse token é um verbo (que está indicado por ("VERB", token.head_token.upos), então o token receberá o token.misc “ArgM-tmp”

A regra abaixo exemplifica o item 3. Fazemos uma busca por elementos cuja deprel é obl ou advcl, introduzidos por “após” e cujo dephead é um VERBO. O último elemento da busca (token – mais precisamente, token.head_token, ou seja, o obl ou advcl) garante que o obl/advcl não contém nenhuma anotação de papel semântico.

```
if regex("após", token.lemma) and regex("obl|advcl",
token.head_token.deprel) and regex("VERB",
token.head_token.head_token.upos) and not regex("A.*",
token.head_token.misc):
    token.head_token.misc = "ArgM-tmp"
```

Especificamente, a regra busca por:

- um token cujo lemma é “após”
- esse token “após” depende de um token cujo deprel é obl ou advcl (ou: o pai desse token “após” é token cujo deprel é obl ou advcl)
- O pai do pai do obl ou advcl é um verbo
- O token cujo deprel é obl ou advcl não tem nenhum elemento iniciado por “A” no campo misc.

Quando encontrar essa condição, o token cujo deprel é obl ou advcl (que é o token.head_token) terá o seu campo misc anotado com “ArgM-tmp”.

O procedimento de criar padrões a partir de pistas linguísticas presentes nas diretivas foi feito para todos os casos de ArgM em que era possível depreender padrões. Está em preparação um documento com instruções relativas à utilização de regras com a ET, mas ao longo do relatório ilustramos/explicamos a utilização de algumas regras.

4.1.2. Etapa 2: Pistas sintáticas simples

Na etapa 2, foram anotados todos os argumentos que poderiam ser anotados diretamente pela sintaxe: esses são os casos de voz passiva, em que o sujeito paciente tende a ser Arg1 e o obl/agent tende a Arg0. No entanto, a anotação não pode ser feita de maneira direta pois há casos em que o sujeito paciente não será Arg1, mas Arg2. Esses casos foram cobertos por regras como as abaixo, para sujeitos simples com verbos de voz passiva e sujeitos oracionais de voz passiva, respectivamente. As duas regras poderiam estar juntas em uma regra única.

A regra abaixo procura por

- um token cujo deprel seja “nsubj”
- este mesmo token deve ter o campo misc vazio (o elemento vazio, pela anotação do corpus, está indicado com “_” ou com “SpaceAfter..”, e por isso a expressão "_|S.*", token.misc)
- o pai deste token (token.head_token) deve ter o atributo feats preenchido com “Voice=Pass”.

```
if regex("nsubj", token.deprel) and regex("_|S.*", token.misc) and
regex(".*Voice=Pass.*", token.head_token.feats):
   token.misc = "Arg1"
```

A regra abaixo faz a mesma coisa, mas busca por csubj, e não nsubj:

```
if regex("csubj", token.deprel) and regex("_|S.*", token.misc) and
regex(".*Voice=Pass.*", token.head_token.feats):
   token.misc = "Arg1"
```

Se fizermos analogia a montar um quebra-cabeças, as etapas 1 e 2 seriam equivalentes a começar pelas peças das bordas - aquelas que, pela forma, temos certeza de onde se encaixam no quebra cabeça. Já a etapa 3, descrita a seguir, seria o preenchimento do “miolo”, o coração da anotação.

4.1.3. Etapa 3: Argumentos numerados e casos não identificados nas etapas anteriores

A terceira e última etapa trata da anotação dos argumentos numerados e dos casos não cobertos – que não são poucos – pelas etapas anteriores.

Nesta etapa, a estratégia consiste em tirar proveito de todos os recursos associados à anotação de SRL em portugues já produzidos, nomeadamente o corpus PropBank 2.0 e o recurso léxico Verbo-Brasil.

Um exemplo deste tipo de estratégia está relacionado à atribuição de argumentos à posição sintática de sujeito. Sabemos que, em geral, o sujeito tende a ser agente (Arg0), mas isto não é regra. O desafio justamente está em encontrar os casos em que o sujeito sintático **não** é agente (ou causador), e portanto **não é Arg0**. Para isso, exploramos os recursos existentes de duas maneiras:

A. No *Propbank-BR 2.0*, buscamos pelos casos de sujeitos (função sintática) cujo papel semântico não fosse Arg1.
B. No *Verbo-Brasil*, buscamos por verbos cujo "roleset ID" não contivesse Arg0, *utilizando a interface alternativa* adaptada exatamente para este fim, e que permite buscas sobre o conteúdo do Verbo-Brasil.

No entanto, e devido (i) ao caráter idiossincrático da anotação de qualquer PropBank (esta idiossincrasia será detalhada na seção *Resultados*), e (ii) à existência de verbos (ou sentidos de formas verbais) do Porttinari inexistentes no corpus PropBank-Br 2.0, tais estratégias, embora úteis, foram de alcance limitado.

Uma segunda estratégia utilizada consiste em concatenar verbos e preposições:

A regra abaixo anota como Arg1 o elemento obl introduzido por "em" associado aos verbos "pensar" e "circular"

```
if regex("em", token.lemma) and regex("obl",
token.head_token.deprel) and regex("_|S.*", token.head_token.misc)
and regex("VERB", token.head_token.head_token.upos) and
regex("pensar|circular", token.head_token.head_token.lemma):
   token.head_token.misc = "Arg1"
```

A anotação, entretanto, é feita em duas fases. Na primeira, buscamos frases que contêm elemento obl introduzido por "em" associado aos verbos "pensar" e "circular". Todas elas são lidas, e os casos que não se enquadram são eliminados, utilizando para isto o recurso Filtro. A regra é então aplicada aos casos em que de fato o elemento obl introduzido por "em" se refere ao Arg1 dos verbos "pensar" e "circular".

Uma maneira de avançar com esta etapa é buscando padrões entre verbos e preposições, como indicado na próxima seção.

Exemplos de padrões léxico-sintáticos

Uma busca como a abaixo retorna elementos com deprel "obl" dependentes de verbo e introduzidos pela preposição "sobre". Em outras palavras, queremos saber *quais verbos* tem elementos "obl" introduzidos pela preposição "sobre", e ainda não receberam nenhuma etiqueta:

```
token.lemma == "sobre" and token.head_token.deprel == "obl" and
@token.head_token.head_token.upos == "VERB"
```

O passo seguinte é analisar a lista de distribuição dos lemas verbos (indicados na busca pelo sinal @), que, nesse caso, devolve 44 lemas diferentes distribuídos em 91 ocorrências. A análise da distribuição dos verbos (figura 9) associada à análise das linhas de concordância (figura 10), nos permite agrupar as ocorrências por tipo:

- 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 coisa

e tratar cada um dos casos separadamente. Se cada caso será tratado individualmente ou por regras depende da quantidade de lemas. A regra abaixo foi utilizada para os casos de Arg2, e indica que elementos obl introduzidos pela preposição sobre que estiverem associados aos lemas `incidir|orientar|jogar|ouvir|recair|exercer` serão anotados como Arg2:

```
if regex("sobre", token.lemma) and regex("obl",
token.head_token.deprel) and
regex("incidir|orientar|jogar|ouvir|recair|exercer",
token.head_token.head_token.lemma):
      token.head_token.misc = "Arg2"
```

Distribuição de lemma

Relatório gerado dia 20 de mai. 2024 16:34

Busca: token.lemma == "sobre" and token.head_token.deprel == "obl" and @token.head_token.head_token.upos == "VERB"
Corpus: Porttinari-b-SRL_FINALIZADO.conllu

Quantidade de ocorrências: 91
Quantidade de **lemma** diferentes: 44

lemma	frequência	em arquivos
falar	16	1
dizer	12	1
escrever	5	1
questionar	5	1
contar	3	1
conversar	3	1
pronunciar	3	1
depor	2	1
incidir	2	1
indagar	2	1
manifestar	2	1
perguntar	2	1
posicionar	2	1
refletir	2	1
abordar	1	1
acumular	1	1
afirmar	1	1
aplicar	1	1
cobrar	1	1
comentar	1	1
comunicar	1	1
consultar	1	1
debater	1	1
deliberar	1	1
descartar	1	1
enrolar	1	1
erguer	1	1
especular	1	1
estudar	1	1
exercer	1	1
explicar	1	1
fazer	1	1
haver	1	1
jogar	1	1
lembrar	1	1
marcar	1	1
montar	1	1
navegar	1	1
orientar	1	1
ouvir	1	1
recair	1	1
reinar	1	1
saber	1	1
tratar	1	1

Figura 9: Distribuição dos lemas verbos com “obl” introduzidos pela preposição “sobre”

Figura 10: Linhas de concordâncias para as buscas de verbos com “obl” introduzidos pela preposição “sobre”.

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.

Regras complexas: elementos do léxico + sintaxe complexa + conhecimento linguístico

Pronomes relativos e antecedentes

Tomando como exemplo a anotação de ArgM-tmp, o padrão a seguir lida com a anotação dos papéis semânticos em pronomes relativos, levando em conta seu antecedente.
Na regra abaixo – em (1) como sintaxe de busca e em (2) como regra para anotação – , pedimos por um pronome relativo (ainda não anotado, por isso o campo MISC “vazio”) cujo head tem upos Verbo e cujo head desse verbo é uma palavra como *tempo, dia, noite* ou *manhã*. Traduzindo para a linguagem da gramática tradicional, queremos **pronomes relativos que tenham como antecedente palavras *tempo, dia, noite ou manhã***, pois serão candidatos a ArgM-tmp.

(1)

```
token.feats == ".*Rel.*" and token.misc == "_|S.*" and
token.head_token.upos == "VERB" and
token.head_token.head_token.lemma == "tempo|dia|noite|manhã"
```

(2)

```
if regex(".*Rel.*", token.feats) and regex("_|S.*", token.misc) and regex("VERB", token.head_token.upos) and
regex("tempo|dia|noite|manhã", token.head_token.head_token.lemma):
        token.misc = "ArgM-tmp"
```

A expressão retorna frases como

- Foi se o *tempo* em **que** as empresas se **adequaram** a os conceitos de os eventos .
- Em o *dia* em **que** foi **anunciada** a retirada de as tropas federais , dois jovens foram torturados por bandidos em o morro .

Podemos **aproveitar esse modelo de regra** para buscar candidatos a ArgM-loc, apenas trocando as palavras de *tempo* por palavras de *lugar*, por exemplo.

Outras regras complexas: mudança de galho na árvore

À medida que a anotação e a familiarização com a ferramenta avançam, é possível aplicar regras mais complexas que envolvem mudança de galho na árvore, como abaixo:

Se um verbo tem Arg2, mas objeto direto e nsubj estão sem argumentos anotados, então nsubj será Arg0 e obj será Arg1. A busca é feita usando essa expressão, que não envolve anotação (ainda), mas apenas busca pelos casos indicados a fim de verificar se o nsubj será Arg0 e o obj será Arg1[23]:

```
for token in sentence.tokens:
        if token.misc == "Arg2" and regex(".*", token.deprel) and token.head_token.upos == "VERB":
                for _nsubj in sentence.tokens:
                        if _nsubj.dephead == token.head_token.id and _nsubj.deprel == "nsubj" and regex("_|S.*", _nsubj.misc):
                                for _obj in sentence.tokens:
                                        if regex("_|S.*", _obj.misc) and _obj.dephead == token.head_token.id and _obj.deprel == "obj":
                                                bold_tokens.append(_nsubj)
                                                bold_tokens.append(_obj)
                                                bold_tokens.append(token.head_token)
```

Assim, a análise dos resultados nos mostra se, nos casos em destaque, o nsubj será realmente Arg0 e o obj será Arg1. Esta verificação antes da anotação é sempre essencial, pois é possível que alguns casos, apesar de seguirem o padrão, escapem à regra. Por isso é importante analisar os casos da busca, para identificar os casos excepcionais que precisarão ser tratados).

A regra encontra casos como

Capítulos de narrativa fantástica **ligam** Evie a animais de a floresta , aproximando o personagem de a Eva bíblica .

O acidente em a largada **colocou** a liderança em o colo de Lewis Hamilton

[23] Regras do tipo “sobe e desce”. Neste caso, trata-se da regra “SUBJ e OBJ vazios com Arg2 VARIOS”

Tudo correndo bem, em um segundo movimento criamos a regra de anotação correspondente à regra de busca, na qual apenas as duas últimas linhas foram alteradas

```
for token in sentence.tokens:
        if token.misc == "Arg2" and regex(".*", token.deprel) and token.head_token.upos == "VERB":
                for _nsubj in sentence.tokens:
                        if _nsubj.dephead == token.head_token.id and _nsubj.deprel == "nsubj" and regex("_|S.*", _nsubj.misc):
                                for _obj in sentence.tokens:
                                        if regex("_|S.*", _obj.misc) and _obj.dephead == token.head_token.id and _obj.deprel == "obj":
                                                _nsubj.misc = "Arg0"
                                                _obj.misc = "Arg1"
```

E com isso temos a anotação. Esta regra alterou/anotou **258 tokens**, e a figura 10 ilustra um pedaço do "relatório", arquivo produzido pela ferramenta ao final de cada regra aplicada que contém tudo o que foi modificado pela regra.

```
FOLHA_DOC000170_SENT019
*Capítulos* de narrativa fantástica ligam Evie a animais de a floresta , aproximando o personagem de a Eva bíblica .
ANTES: 1        Capítulos       capítulo        NOUN    _       Gender=Masc|Number=Plur 5       nsubj   _       *_*
DEPOIS: 1       Capítulos       capítulo        NOUN    _       Gender=Masc|Number=Plur 5       nsubj   _       *Arg0* (head: ligam)

FOLHA_DOC000170_SENT019
Capítulos de narrativa fantástica ligam *Evie* a animais de a floresta , aproximando o personagem de a Eva bíblica .
ANTES: 6        Evie    Evie    PROPN   _       _       5       obj     _       *_*
DEPOIS: 6       Evie    Evie    PROPN   _       _       5       obj     _       *Arg1* (head: ligam)

FOLHA_DOC003105_SENT043
O *acidente* em a largada colocou a liderança em o colo de Lewis Hamilton .
ANTES: 2        acidente        acidente        NOUN    _       Gender=Masc|Number=Sing 6       nsubj   _       *_*
DEPOIS: 2       acidente        acidente        NOUN    _       Gender=Masc|Number=Sing 6       nsubj   _       *Arg0* (head: colocou)

FOLHA_DOC003105_SENT043
O acidente em a largada colocou a *liderança* em o colo de Lewis Hamilton .
ANTES: 8        liderança       liderança       NOUN    _       Gender=Fem|Number=Sing  6       obj     _       *_*
DEPOIS: 8       liderança       liderança       NOUN    _       Gender=Fem|Number=Sing  6       obj     _       *Arg1* (head: colocou)
```

Figura 10: Parte de um "relatório", arquivo que contém um registro das alterações feitas por uma regra

Este tipo de regra foi utilizado na anotação dos argumentos de diversos verbos, e foi usada com algumas variações.

Abaixo, a regra anota **nsubj como Arg0 e obl introduzido pela preposição "a" como Arg1 dos verbos** `obedecer|atender|referir|andar|resistir`.
Diferentemente do exemplo anterior, a anotação agora acontece em dois passos, pois é possível que uma frase só tenha nsubj, e não tenha obl introduzido pela preposição "a", ou o contrário.

```
for token in sentence.tokens:
        if regex("obedecer|atender|referir|andar|resistir", token.head_token.lemma):
                for _nsubj in sentence.tokens:
                        if _nsubj.dephead == token.head_token.id and _nsubj.deprel == "nsubj" and regex("_|S.*", _nsubj.misc):
                                _nsubj.misc = "Arg0"
for token in sentence.tokens:
        if regex("obedecer|atender|referir|andar|resistir", token.head_token.lemma):
                for _obl in sentence.tokens:
                        if _obl.dephead == token.head_token.id and _obl.deprel == "obl" and regex("_|S.*", _obl.misc):
                                for _a in sentence.tokens:
                                        if _a.lemma == "a" and _a.dephead == _obl.id:
                                                _obl.misc = "Arg1"
```

A regra encontrou e anotou casos como abaixo, e foram tratados/anotados **40 tokens**.

> Foi socorrido por policiais de o seu batalhão , mas não **resistiu** a os ferimentos e já chegou morto a o hospital .
>
> Naza é o modo como muitos moradores de Belém se **referem** a a Nossa Senhora de Nazaré .

```
FOLHA_DOC004070_SENT009
Foi socorrido por policiais de o seu batalhão , mas não resistiu a os *ferimentos* e já chegou morto a o hospital .
ANTES: 15       ferimentos      ferimento       NOUN    _       Gender=Masc|Number=Plur 12      obl     _       *_*
DEPOIS: 15      ferimentos      ferimento       NOUN    _       Gender=Masc|Number=Plur 12      obl     _       *Arg1* (head: resistiu)

FOLHA_DOC000062_SENT019
Naza é o modo como muitos *moradores* de Belém se referem a a Nossa Senhora de Nazaré .
ANTES: 7        moradores       morador NOUN    _       Gender=Masc|Number=Plur 11      nsubj   _       *_*
DEPOIS: 7       moradores       morador NOUN    _       Gender=Masc|Number=Plur 11      nsubj   _       *Arg0* (head: referem)

FOLHA_DOC000062_SENT019
Naza é o modo como muitos moradores de Belém se referem a a *Nossa* Senhora de Nazaré .
ANTES: 14       Nossa   Nossa   PROPN   _       _       11      obl     _       *_*
DEPOIS: 14      Nossa   Nossa   PROPN   _       _       11      obl     _       *Arg1* (head: referem)
```

Na variação abaixo, temos uma regra que anota todos os argumentos de um verbo (no exemplo, o verbo "causar"). A regra anota **sujeitos como Arg0 e obj como Arg1 e obl introduzido pela preposição "a" como Arg2 do verbo `causar`**.

```
for token in sentence.tokens:
        if regex("causar", token.head_token.lemma):
                for _nsubj in sentence.tokens:
                        if _nsubj.dephead == token.head_token.id and _nsubj.deprel == "nsubj" and regex("_|S.*", _nsubj.misc):
                                _nsubj.misc = "Arg0"
for token in sentence.tokens:
        if regex("causar", token.head_token.lemma):
                for _obj in sentence.tokens:
                        if _obj.dephead == token.head_token.id and _obj.deprel == "obj" and regex("_|S.*", _obj.misc):
                                _obj.misc = "Arg1"
for token in sentence.tokens:
        if regex("causar", token.head_token.lemma):
                for _obl in sentence.tokens:
                        if _obl.dephead == token.head_token.id and _obl.deprel == "obl" and regex("_|S.*", _obl.misc):
                                for _a in sentence.tokens:
                                        if _a.lemma == "a" and _a.dephead == _obl.id:
                                                _obl.misc = "Arg2"
```

Esta regra encontra casos como abaixo

> Seu caso , ocorrido em o fim de o ano passado , **causou** comoção em a rede estadual .
> Essas tempestades não vão mudar o que eles dizem em público , mas podem vir a **causar** uma mudança de atitude , mesmo que seja apenas em os bastidores .

que são então anotados, como indica a imagem abaixo do relatório. A regra não encontrou casos com Arg2, porque já haviam sido anotados. Foram encontrados (e tratados) **29 tokens** com essa regra.

```
FOLHA_DOC003215_SENT089
Seu *caso* , ocorrido em o fim de o ano passado , causou comoção em a rede estadual .
ANTES: 2        caso    caso    NOUN    _       Gender=Masc|Number=Sing 13      nsubj   _       *SpaceAfter=No*
DEPOIS: 2       caso    caso    NOUN    _       Gender=Masc|Number=Sing 13      nsubj   _       *Arg0* (head: causou)

FOLHA_DOC003215_SENT089
Seu caso , ocorrido em o fim de o ano passado , causou *comoção* em a rede estadual .
ANTES: 14       comoção comoção NOUN    _       Gender=Fem|Number=Sing  13      obj     _       *_*
DEPOIS: 14      comoção comoção NOUN    _       Gender=Fem|Number=Sing  13      obj     _       *Arg1* (head: causou)
```

4.1.4. Etapa 4: Verbo-Brasil como fonte de padrões e pistas linguísticas

Esta etapa envolve a depreensão de padrões linguísticos e de regras complexas, e foi realizada em uma fase já avançada da anotação. No entanto, **nada impede que seja**

utilizada no início do processo. Pelo contrário, é possível que, utilizada no início do processo de anotação, otimize ainda mais a anotação.

O Verbo-Brasil, ao listar uma série de exemplos do corpus para ilustrar um *frame*, também permite a depreensão de padrões léxico sintáticos (como na etapa 1), não apenas otimizando a anotação, mas também **permitindo avançar com a anotação dos frames,** realizadas na coluna 9/"deps"**,** que aparece como mais uma linha da regra.

Três exemplos ilustram o procedimento. Os dois primeiros estão associados aos verbos "fazer", e "ter", formas verbais polissêmicas na medida em que frequentemente esvaziadas de sentido (verbo leve/verbo suporte). O terceiro exemplo mostra a utilização do Verbo-Brasil como um "banco de dados", em uma busca que auxilia a **desambiguação** de padrões com verbos no infinitivo.

Na análise dos 16 frames associados ao "fazer" (sem contar mwes como "fazer parte"; "fazer força" etc), percebemos um padrão associado ao frame fazer.02 (figura 11):

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

Exemplo 1:
bosC. s3473: A uma pergunta sobre se a volta dos investimentos externos não dependeria do selo de aprovação do Plano Real através de um acordo formal entre o FMI e o Brasil, Camdessus disse que o que vai fazer a taxa de juros cair não é um acordo com o FMI, mas a credibilidade interna e externa do Plano Real.

Arg0: o que
Rel: fazer
Arg1: a taxa de juros cair

Figura 11. Frame de fazer.02 no Verbo-Brasil

A ideia de *fazer alguém fazer algo* se manifesta, em UD, como uma estrutura que envolve a deprel *xcomp*, como em

> A estatal reajustou para cima , e **isso faz a receita de ela crescer**.
> O **episódio fez lembrar** alguns casos recentes de ataques terroristas em cidades europeias .

Passos até a correção:

1. Busca no Interrogatório pela estrutura em questão, e verificação se a intuição da correlação entre estrutura sintática e frame se mantém (verificada). Como é mais fácil compreender a busca pelo nó mais baixo da árvore (elemento token), e ir

subindo, vamos ler a regra abaixo pela seguinte ordem das linhas: linha 3, linha 2, linha 1, linha 4.

Quero um token que não tenha campo misc (token.misc) anotado - linha 3

Quero que esse mesmo token tenha campo deprel (token.deprel) anotado com "xcomp" - linha 2

Quero que o head desse mesmo token (token.head_token) tenha o campo upos (token.head_token.upos) anotado com VERB - linha 1

Quero que o head desse mesmo token (token.head_token) tenha o campo lemma (token.head_token.lemma) anotado com fazer - linha 4

```
token.head_token.upos == "VERB" and
token.deprel == "xcomp" and
token.misc == "_|S.*" and
token.head_token.lemma == "fazer"
```

Como podemos perceber, a linha 1 é redundante quando temos a linha 4, pois se o lema é "fazer", estamos diante de um verbo. Nesse caso, a linha 1 poderia ser eliminada da regra sem qualquer prejuízo. Por outro lado, nem sempre há redundância: o lema "poder", por exemplo, pode ser substantivo ou verbo.

2. Análise das frases devolvidas (análise das linhas de concordância) para eventual eliminação/correções de estruturas que não sejam as desejadas
3. Criação e aplicação de regra para anotação[24]. No exemplo abaixo, anotamos em um único lance:
    a. o token xcomp como Arg2 na coluna misc
    b. o token obj como Arg1 na coluna misc
    c. o token nsubj como Arg0 na coluna misc
    d. o token "fazer" como fazer.02 - na coluna deps

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

Do mesmo modo, o frame **ter.10 , alimentar no pensamento,** traz a seguinte informação:

Arg0: *pensador* (vnrole: -agent)

Arg1: *nome eventivo (certeza, dúvida, convicção, ilusão, esperança, etc)* (vnrole: -theme)

Arg2: *conteúdo do pensamento, complemento nominal do arg1 (de, de que, sobre, se)* (vnrole: -topic)

que por sua vez leva à criação da seguinte regra de busca

```
@token.lemma   ==   "razão|certeza|dúvida|convicção|esperança"
and token.deprel == "obj" and
```

[24] Ver regra " FAZER Arg1 Arg2 DEPS", em regras do tipo "sobe e desce".

```
token.head_token.lemma == "ter"
```

e à correspondente regra de anotação

```
if          regex("razão|certeza|dúvida|convicção|esperança",
token.lemma) and regex("obj", token.deprel) and regex("ter",
token.head_token.lemma):
      token.deprel = "Arg1"
      token.head_token.deps = "ter.10”
```

Para anotar também o sujeito como Arg0 a regra é mais complexa e envolve “mudança de galho” da árvore. Como os sujeitos do verbo **ter** já haviam sido anotados, a regra de sujeito não foi necessária.

As pistas linguísticas do Verbo-Brasil também auxiliam a **desambiguar** padrões, e aqui usamos o Verbo-Brasil como um *banco de dados* por meio da **interface alternativa**, como será indicado na seção *Verbo-Brasil como banco de dados*.
Enunciados com uma mesma forma e um mesmo sentido podem receber etiquetas diferentes conforme os frames dos verbos - conforme a sua subcategorização. Assim, enunciados com orações adverbiais (*advcl*) introduzidas pela preposição “para” e com *verbo no infinitivo*, por exemplo, costumam indicar *finalidade* ou *propósito*, e são anotados como **ArgM-prp** (exemplos 1-3).

1. O foco acabou mudando, **ficamos** no Brasil para **fazer** faculdade, mas retomamos os planos.
2. Os rivais não **entravam** na área para **ameaçar** .
3. Vale **investir** tempo para **conhecer** as alternativas no mercado local .

No entanto, essa mesma ideia de *propósito*, caso faça parte da estrutura argumental do verbo, como nos exemplos 4 e 5, será um argumento numerado. O Porttinari tem 456 frases com a estrutura para + verbo no infinitivo/advcl - ou, mais precisamente, 268 lemas verbais diferentes que ocorrem com esse padrão[25], e embora analisar cada um dos lemas seja uma alternativa quando se deseja um material padrão-ouro, uma consulta ao Verbo-Brasil pode otimizar o processo.

Especificamente, pedimos, na interface desenvolvida, frames *que contenham a palavra “finalidade” no atributo “Papéis” e que este papel não esteja associado* `(busca negativa)` *a um ArgM* (em outras palavras, que este papel “finalidade” esteja associado a um argumento numerado) (fig. 11). A busca retorna 23 frames de verbos. Com essa informação, fazemos uma busca refinada do padrão para + advcl infinitivo, especificando os lemas dos verbos, como abaixo.

```
token.lemma == "para" and @token.head_token.upos
== "VERB" and token.head_token.deprel == "advcl"
and  token.head_token.feats  ==  ".*Inf.*"  and
token.head_token.lemma == "lemas_dos_23_verbos"
```

[25] A expressão de busca é token.lemma == "para" and @token.head_token.upos == "VERB" and token.head_token.deprel == "advcl" and token.head_token.feats == ".*Inf.*"

Lemos os resultados para confirmar que está tudo bem, anotamos, e lemos as restantes para confirmar que são ArgM-prp, e anotamos. Esta leitura, já pré-organizada quanto aos sentidos/classes, faz com que a anotação ganhe em qualidade/consistência e em facilidade/velocidade.

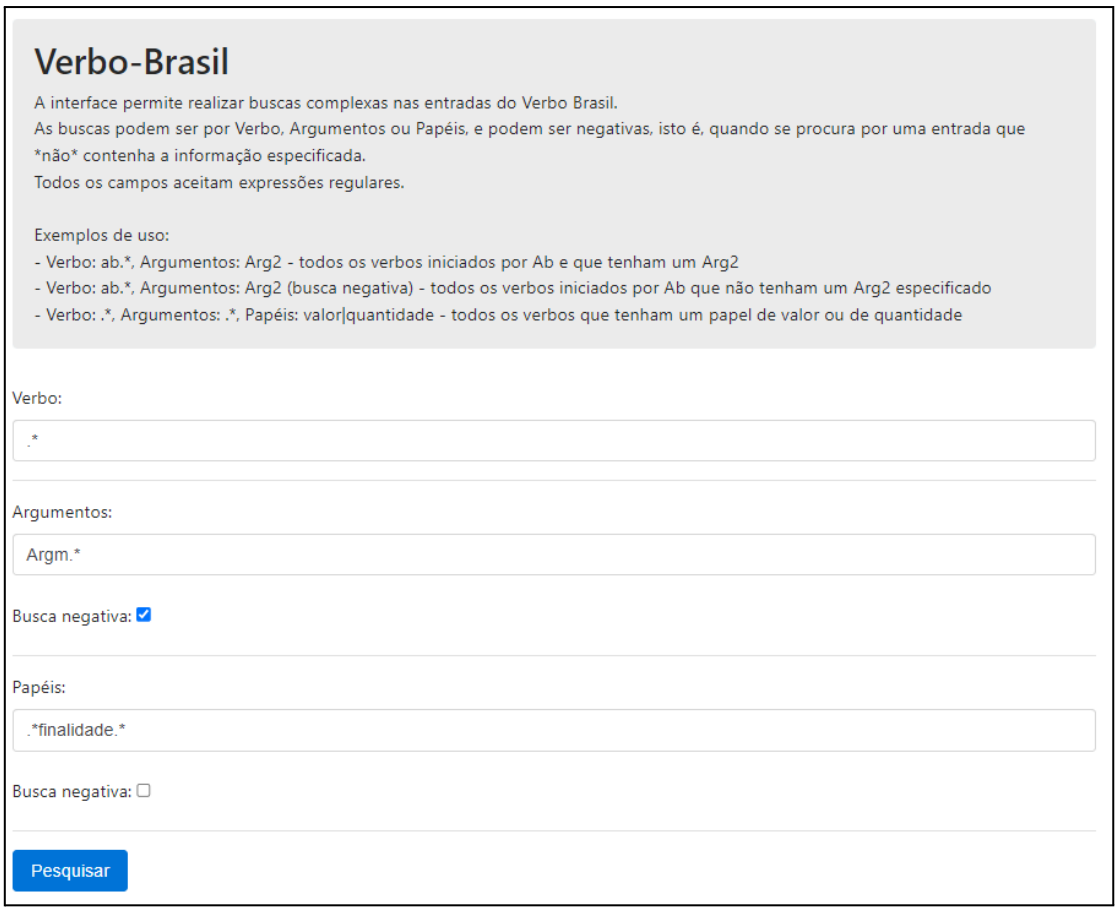


Figura 11: Interface do Verbo-Brasil para a busca de frames que não tenham qualquer tipo de ArgM com papel associado à finalidade.

4.1.5. Etapa 5: Acabamentos

Nesta etapa, buscamos elementos que deveriam ter recebido alguma anotação, porque dependentes de verbos, mas que não receberam.

Para tanto, utilizamos o recurso **filtro** associado à leitura das linhas de concordância. Os filtros aceleram o processo porque não é preciso editar cada frase, basta (i) a leitura, (ii) a classificação adequada, e (iii) a correção final, feita por lote (por regra), como nos exemplos anteriores.
Abaixo, temos uma lista com os últimos OBLs sem papel semântico atribuído. Ou seja, fazemos uma busca que pede elementos obl associados a verbos e com o campo "misc" vazio:

```
token.misc == "_|S.*" and token.deprel == "obl" and
token.head_token.upos == "VERB"
```

A regra encontrou 103 casos , como mostra a figura 12, que traz também a indicação dos filtros:

Figura 12: Tela do Interrogatório com a aplicação de filtros às concordâncias

Após a distribuição dos casos nos filtros adequados (grupo 1: ArgM-tmp; grupo 2: ArgM-loc; grupo 3: ArgM-loc; grupo 4: ArgM-dis; grupo 5: ArgM-mnr), cada grupo é tratado individualmente, quer por regras, quer pela edição direta da frase.

4.1.6. Limitações das regras (nem tudo pode ser tratado por regras)

Apesar de otimizarem imensamente a tarefa de anotação, nem tudo que pode ser buscado por meio de uma regra pode ser resolvido, de maneira simples, com outra regra. Por isso, a utilização dos filtros é de grande ajuda.

Os padrões como “brincar com” e “trabalhar com” exemplificam o ponto. Nas frase 3 (padrão “brincar com”), o sintagma introduzido por “com” é ArgM-mnr (= brincando “*sob a supervisão de recreadoras*”), e na frase 4, é Arg1 (*brincar com alguém ou com algo*).

1. O proprietário paga por o tempo desejado (de 30 minutos a cinco horas) para deixar seu amigo sem coleiras , **brincando** com a supervisão de recreadores.
2. Em " A Galinha " , o arquiteto **brinca** com as letras de o alfabeto para criar bichos .

Considerando o padrão “trabalhar com”, trabalhar *com crianças* e *com design* é Arg1 (frases 5 e 6), mas trabalhar *com alguém* (co-trabalhador, frase 7) é Arg3.

3. A cada hora chega uma criancinha ... eu **trabalho** com criança , conta .
4. Não há dados precisos de quantas pessoas **trabalham** com design em o país , de acordo com a Associação Brasileira de as Empresas de Design .
5. Vinda de o country e **trabalhando** com seu então marido , Mutt Lange , produtor de bandas como AC/DC e Def Leppard , ela fazia música grandiosa , eclética .

E ainda assim há casos vagos, como a frase 8, em que o jornal pode ser considerado o trabalho ou pode fazer referência, por metonímia, aos colegas de trabalho, e portanto Arg3. Foi anotado como Arg3.

6. O suíço também **trabalhou** com o " Corriere del Ticino " .

Mas, em geral, a uma regra de identificação corresponde uma regra de correção.

4.1.7. Etapa 6: Validação

Ao fim do processo de anotação, buscamos por casos que seriam indicativos de erros, como argumentos numerados idênticos associados a um mesmo verbo (figura 12). Foram detectados **121 erros** nesta etapa, considerando a anotação explícita.

```
# A regra abaixo verifica se há dois tokens numerados identicos apontando para o mesmo verbo, o que é
if regex("Arg1|Arg2|Arg3|Arg4", token.misc): # Qual a etiqueta de MISC vc quer?
    for _token in sentence.tokens: # Começo a procurar outro _token que tenha o mesmo MISC e DEPHEAD
        if _token.id != token.id: # Preciso garantir que esse outro _token não é o próprio token, sen
            if _token.misc == token.misc and _token.dephead == token.dephead: # Aqui de fato verifico
                bold_tokens.append(token)
                bold_tokens.append(_token)
```

Figura 12: Exemplo de regra de validação

Na figura 13, vemos um erro: os tokens 2 e 8 foram anotados como Arg2 do verbo **“dar”**, e o token 8 precisa ser corrigido para “obl” - o erro foi motivado pela estrutura “**dar X para Y**”, mas nesse caso o “para Y” é ArgM-prp:

3/118 - FOLHA_DOC000041_SENT069

Não **me** deu um centavo para o **tratamento** .

[Esconder anotação] [Editar anotação]

```
# text = Não me deu um centavo para o tratamento.
# sent_id = FOLHA_DOC000041_SENT069
1       Não     não     ADV     _       _       3       advmod  _       ArgM-neg
2       me      me      PRON    _       Case=Dat|Number=Sing|Person=1|PronType=Prs      3       iobj    _       Arg2
3       deu     dar     VERB    _       Mood=Ind|Number=Sing|Person=3|Tense=Past|VerbForm=Fin   0       root    dar.01  _
4       um      um      NUM     _       Gender=Masc|NumType=Card        5       nummod  _       _
5       centavo centavo NOUN    _       Gender=Masc|Number=Sing 3       obj     _       Arg1
6       para    para    ADP     _       _       8       case    _       _
7       o       o       DET     _       Definite=Def|Gender=Masc|Number=Sing|PronType=Art       8       det     _       _
8       tratamento      tratamento      NOUN    _       Gender=Masc|Number=Sing 3       obl     _       Arg2
9       .       .       PUNCT   _       _       3       punct   _       SpaceAfter=No
```

Figura 13: Exemplo de erro detectado por regra de validação

Na figura 14, temos dois tokens Arg2 (tokens 5 e 15) associados a "**ficar**". Apesar da análise fazer sentido, por padrão, conforme as instruções do PropBank, o primeiro elemento fica como Arg2 (Argumento numerado) e o segundo como ArgM-loc (argumento modificador)

2/118 - FOLHA_DOC003013_SENT015

Janot ficou em seu **gabinete** de subprocurador-geral durante a posse , em o mesmo **prédio** onde Dodge assumia a sua cadeira .

[Esconder anotação] [Editar anotação]

```
# text = Janot ficou em seu gabinete de subprocurador-geral durante a posse, no mesmo prédio onde Dodge assumia a sua cadeir
# sent_id = FOLHA_DOC003013_SENT015
1       Janot   Janot   PROPN   _       _       2       nsubj   _       Arg1
2       ficou   ficar   VERB    _       Mood=Ind|Number=Sing|Person=3|Tense=Past|VerbForm=Fin   0       root    _       _
3       em      em      ADP     _       _       5       case    _       _
4       seu     seu     DET     _       Gender=Masc|Number=Sing|Person=3|Poss=Yes|PronType=Prs   5       det     _       _
5       gabinete        gabinete        NOUN    _       Gender=Masc|Number=Sing 2       obl     _       Arg2
6       de      de      ADP     _       _       7       case    _       _
7       subprocurador-geral     subprocurador-geral     NOUN    _       Gender=Masc|Number=Sing 5       nmod    _       _
8       durante durante ADP     _       _       10      case    _       _
9       a       o       DET     _       Definite=Def|Gender=Fem|Number=Sing|PronType=Art       10      det     _       _
10      posse   posse   NOUN    _       Gender=Fem|Number=Sing  2       obl     _       ArgM-tmp
11      ,       ,       PUNCT   _       _       15      punct   _       _
12-13   no      _       _       _       _       _       _       _       _
12      em      em      ADP     _       _       15      case    _       _
13      o       o       DET     _       Definite=Def|Gender=Masc|Number=Sing|PronType=Art       15      det     _       _
14      mesmo   mesmo   DET     _       Gender=Masc|Number=Sing|PronType=Dem    15      det     _       _
15      prédio  prédio  NOUN    _       Gender=Masc|Number=Sing 2       obl     _       Arg2
16      onde    onde    ADV     _       _       18      advmod  _       ArgM-loc
17      Dodge   Dodge   PROPN   _       _       18      nsubj   _       Arg0
18      assumia assumir VERB    _       Mood=Ind|Number=Sing|Person=3|Tense=Imp|VerbForm=Fin   15      acl:relcl       _
19      a       o       DET     _       Definite=Def|Gender=Fem|Number=Sing|PronType=Art       21      det     _       _
20      sua     seu     DET     _       Gender=Fem|Number=Sing|Person=3|Poss=Yes|PronType=Prs   21      det     _       _
21      cadeira cadeira NOUN    _       Gender=Fem|Number=Sing  18      obj     _       Arg1
22      .       .       PUNCT   _       _       2       punct   _       SpaceAfter=No
```

Figura 14: Exemplo de erro detectado por regra de validação

Na figura 15, temos dois Arg1 (tokens 2 e 8) associados ao verbo "**questionar**". O token 2 deve ser transformado em Arg2, e sua anotação como token 1 se refere à regra de sujeitos de voz passiva (que *tendem* a ser Arg1 mas que, como podemos ver, também podem exercer outros papéis)

79/118 - FOLHA_DOC003091_SENT006

O **técnico** Unai Emery foi questionado sobre **isso** durante a semana e disse que a situação estava resolvida .

[Esconder anotação] [Editar anotação]

```
# text = O técnico Unai Emery foi questionado sobre isso durante a semana e disse que a situação estava resolvida.
# sent_id = FOLHA_DOC003091_SENT006
1       O       o       DET     _       Definite=Def|Gender=Masc|Number=Sing|PronType=Art       2       det     _       _
2       técnico técnico NOUN    _       Gender=Masc|Number=Sing 6       nsubj:pass      _       Arg1
3       Unai    Unai    PROPN   _       _       2       nmod    _       _
4       Emery   Emery   PROPN   _       _       3       flat:name       _       _
5       foi     ser     AUX     _       Mood=Ind|Number=Sing|Person=3|Tense=Past|VerbForm=Fin   6       aux:pass        _       ArgM-pas
6       questionado     questionar      VERB    _       Gender=Masc|Number=Sing|VerbForm=Part|Voice=Pass        0       root    _       _
7       sobre   sobre   ADP     _       _       8       case    _       _
8       isso    isso    PRON    _       Gender=Masc|Number=Sing|PronType=Dem    6       obl     _       Arg1
9       durante durante ADP     _       _       11      case    _       _
10      a       o       DET     _       Definite=Def|Gender=Fem|Number=Sing|PronType=Art       11      det     _       _
11      semana  semana  NOUN    _       Gender=Fem|Number=Sing  6       obl     _       ArgM-tmp
12      e       e       CCONJ   _       _       13      cc      _       _
13      disse   dizer   VERB    _       Mood=Ind|Number=Sing|Person=3|Tense=Past|VerbForm=Fin   6       conj    dizer.01        _
14      que     que     SCONJ   _       _       18      mark    _       _
15      a       o       DET     _       Definite=Def|Gender=Fem|Number=Sing|PronType=Art       16      det     _       _
16      situação        situação        NOUN    _       Gender=Fem|Number=Sing  18      nsubj   _       NSUBJ_COP
17      estava  estar   AUX     _       Mood=Ind|Number=Sing|Person=1|Tense=Imp|VerbForm=Fin   18      cop     _       _
18      resolvida       resolvido       ADJ     _       Gender=Fem|Number=Sing|VerbForm=Part    13      ccomp   _       Arg1
19      .       .       PUNCT   _       _       6       punct   _       SpaceAfter=No
```

Figura 15: Exemplo de erro detectado por regra de validação

4.2. Anotação de elementos implícitos

Após a anotação dos elementos explícitos, foi feita a anotação de elementos implícitos, especificamente sujeitos omitidos e argumentos – na maioria, sujeitos – de orações reduzidas. Boa parte desse processo foi feita de forma automática, por regras associadas de busca e de anotação, sempre levando em conta a revisão humana.

Nesta etapa, foi introduzida a etiqueta provisória "FEITO" para cada verbo alvo de análise, de maneira a garantir que, ao final do processo, todos os verbos haviam sido analisados. Na ET, a atribuição do argumento é feita usando o comando *append*, uma vez que, com frequência, o campo MISC já possui alguma etiqueta de papel semântico, atribuída na fase anterior.

A explicitação privilegiou a busca por argumentos numerados dos verbos, em especial sujeitos, que são os "elementos nucleares" mais frequentemente omitidos em português. Assim, **a anotação dos elementos implícitos se concentrou em identificar todos verbos** do corpus e atribuir a eles, sempre que possível, papéis semânticos numerados.

Para verbos na forma infinitiva, identificamos (explicitamos) argumentos apenas se estes forem **recuperáveis** (frases 1 e 2). Em caso de dúvida ou em caso de argumentos não recuperáveis (frases 3 e 4), nada foi feito.

1. O presidente centrista *optou* por **garantir** por a primeira vez em anos que o deficit orçamentário estará abaixo de o limite europeu de 3 % de o PIB (produto interno bruto) , como insiste a Alemanha . (O presidente garantiu)
2. O delegado responsável por o inquérito , César Matoso , *informou* que não irá se **manifestar** até que seja encerrada a investigação . (O delegado (não) se manifestará)

3. São mecânicas que *pressionam* a **entender** que isso tem custo . (NADA - não é possível determinar quem entenderá)

4. A sua empresa o coloca em hotel em Guarulhos sempre que ele viaja a São Paulo , o que *dificulta* **pedir** o Uber , já que a distância é curta . NADA - não é possível determinar quem pedirá o Uber)

Não foram anotados argumentos implícitos do tipo *objetos ou argumentos não numerados*. Na frase abaixo, por exemplo, apenas "você" foi anotado como Arg0 das duas ocorrências do verbo "montar", e "time" é Arg1 apenas da primeira ocorrência (em itálico). Planejamos explicitar os objetos na 2a versão do corpus.

> Quando você ***monta*** um time , tem que **montar** com dois critérios .

Do mesmo modo, em frases como

> Em a obra de 2015 , o **artista** usa um ralador para **desmanchar** uma imagem de Nossa Senhora Aparecida .

, apenas "artista" foi anotado como Arg0 de **desmanchar**, mas "ralador" também deveria ser um argumento (instrumento) de **desmanchar** (verbo sem frame), e uma segunda rodada de revisão do corpus, principalmente após a inclusão dos verbos faltantes no Verbo-Brasi, será benéfica. Aparentemente, orações que envolvem verbos como "**servir**" ("servir para") e "**usar**", bem como sintagmas do tipo ArgM-prp, têm boas chances de conterem implícitos não numerados. No caso abaixo (frase 1), "vasilhames" seria ArgM-loc de **armazenar**. Na frase 2, "volumes" seria argumento (instrumento?) de **incriminar** (verbo sem frame)

1. Os **vasilhames** , feitos de plástico , papel e alumínio , seriam usados para **armazenar** leite , molho de tomate , sucos e iogurtes .
2. Estantes eram devassadas em busca de volumes **que** servissem para **incriminar** os respectivos leitores .

Argumentos que não podiam ser localizados em um único sintagma também **não foram anotados**. Na frase abaixo, não sabemos bem o que vinha sendo solicitado (a *contratação*, a *consultoria*, uma *preparação para a sucessão*, tudo isso?), e, por isso, o Arg1 de *solicitar* não foi preenchido.

> A família sugeriu então o nome de Zé Mineiro e o presidente de o conselho , Tarek Farahat , disse que a JBS contrataria uma consultoria para preparar a sucessão , conforme vinha sendo **solicitado** por o BNDES .

Não foram explicitadas formas pronominais recuperáveis apenas pela flexão verbal, como em

> Cozinhar, **detesto** e não **sei**.

Na frase abaixo, apesar da correferência lógica (a gente = "nós", que está implícito pela forma verbal), **não indicamos "gente" como Arg0 de "acostumar",** mas talvez esta decisão deva ser revista.

" A **gente** olha muito para a indústria local , nos **acostumamos** a ter o CDI [juros médios de empréstimos entre instituições financeiras] como parâmetro para tudo " , critica o especialista .

Por fim, também **não foram anotadas** relações entre epítetos predicativos e verbos, como no trecho abaixo (decisão que pode ser revista)

Leitor voraz desde garoto , aos 20 anos começou a **vender** contos para revistas .

4.2.1. Desalinhamento entre sintaxe e semântica na anotação de elementos implícitos

A anotação de elementos implícitos levou a um outro tipo de **desalinhamento entre sintaxe e semântica**, uma vez que nem sempre há coincidência entre os elementos/sintagmas identificados na árvore original, na qual se baseia a anotação de elementos explícitos, e os argumentos envolvidos nos implícitos.

Na frase (1), o sintagma "contato com o general Mourão" é Arg1 do verbo "pedir", mas apenas "general Mourão" é sujeito/Arg0 de "comentar".

A **Folha** pediu [contato com o general Mourão] , para que **comentasse** suas declarações, mas o centro de comunicação social de o Exército informou que as respostas serão dadas por meio de o órgão .

A anotação de elementos implícitos também foi feita em etapas, caracterizadas não pelo tipo de abordagem (como na anotação de elementos explícitos), mas pelos fenômenos linguísticos tratados.

4.2.2. Etapa 1: Preparação

Esta etapa é decorrente do tratamento dado por UD a fenômenos como cópula e verbos auxiliares.

Identificação de sujeitos nsubj e de verbos xcomp envolvidos em cópulas

Inclusão da etiqueta provisória COP apenas para sinalizar que se trata de um sujeito "especial", e que não está associado ao seu head. Abaixo, o token "objetivo", cuja deprel é nsubj (de "elevar"), recebe a etiqueta NSUBJ_COP.

O objetivo em a época era **elevar** a oferta de crédito e evitar a retração de a economia .

Posteriormente, uma regra associou esses sujeitos como Arg1 do verbo "ser".

Xcomp e nsubj associados a auxiliares

xcomp verbais dependentes de verbos considerados auxiliares no Verbo-Brasil receberam a etiqueta Arg1_[tipodeuxiliar]_xcomp. A etiqueta indica que esses casos sempre precisarão ter o sujeito explicitado, e que o sujeito será o sujeito do "verbo auxiliar". Na frase abaixo, "ajudar" foi anotado como "Arg1_aux_xcomp"

Argentino , **Pizzi** **pode** **ajudar** a eliminar a seleção de o seu país .

Inclusão da etiqueta FEITO

Todos os verbos que têm um *nsubj* apontando para si e não se encaixam nos casos anteriores receberam uma etiqueta provisória FEITO, atribuída no campo MISC. A função da etiqueta é sinalizar, de forma simples, verbos que têm omissão de sujeito (verbos que, por exclusão, não têm o campo MISC anotado com FEITO).

Embora descrita na fase de *Preparação*, esta etapa aconteceu após o início da anotação dos elementos implícitos. A motivação para a inclusão desta etiqueta é evitar regras excessivamente complexas de busca e de anotação que envolvem ausência (verbos que *não* têm sujeito).
Além disso, assim que um elemento era explicitado, o verbo a que estava associado recebia a etiqueta FEITO.

4.2.3. Etapa 2: Identificação e anotação

Sujeitos omitidos - verbos com deprel conj

Compartilhamento de um mesmo sujeito por diferentes verbos.
A figura 16 ilustra uma relação de coordenação (conj) entre os verbos **desprezar** (token19) e **legislar** (token 24).

17	o	o	DET	_	Definite=Def|Gender=Masc|Number=Sing|PronType=Art	18	det	_	_
18	movimento	movimento	NOUN	_	Gender=Masc|Number=Sing	19	nsubj	_	Arg0
19	despreza	desprezar	VERB	_	Mood=Ind|Number=Sing|Person=3|Tense=Pres|VerbForm=Fin	0	root	_	_
20	a	o	DET	_	Definite=Def|Gender=Fem|Number=Sing|PronType=Art	21	det	_	_
21	complexidade	complexidade	NOUN	_	Gender=Fem|Number=Sing	19	obj	_	Arg1
22	humana	humano	ADJ	_	Gender=Fem|Number=Sing	21	amod	_	_
23	e	e	CCONJ	_	_	24	cc	_	_
24	legisla	legislar	VERB	_	Mood=Ind|Number=Sing|Person=3|Tense=Pres|VerbForm=Fin	19	conj	_	_
25	com	com	ADP	_	_	26	case	_	_
26	dogmas	dogma	NOUN	_	Gender=Masc|Number=Plur	24	obl	_	ArgM-mnr
27	.	.	PUNCT	_	_	19	punct	_	SpaceAfter=No

Figura 16: Exemplo de anotação de coordenação

Para encontrar casos desse tipo - um mesmo sujeito é compartilhado por diferentes verbos - foi construída um uma regra[26] que

- encontra um verbo com deprel *conj* que não tem sujeito associado.

[26] A regra apresentada foi formulada/utilizada antes da aplicação da etiqueta FEITO, conforme descrita na seção Preparação.

- busca o verbo que é head do verbo que está em coordenação
- encontra o *nsubj* desse verbo
- indica esse *nsubj* como provável sujeito do verbo coordenado.

```
for token in sentence.tokens:
        if token.deprel == "conj" and token.upos == "VERB" and token.head_token.upos == "VERB":
                for _nsubj in sentence.tokens:
                        if _nsubj.deprel == "nsubj" and _nsubj.dephead == token.head_token.id:
                                verbo = token
                                tem_sujeito = False
                                for _token in sentence.tokens:
                                        if _token.dephead == verbo.id and regex("nsubj|nsubj:pass|csubj", _token.deprel):
                                                tem_sujeito = True
                                                break
                                if not tem_sujeito:
                                        bold_tokens.append(token)
                                        bold_tokens.append(_nsubj)
```

A regra encontra casos como os ilustrados na figura 17.

51/487 - FOLHA_DOC002007_SENT054

O **anfitrião** ainda declamou poemas e **fez** um discurso emocionado em a hora de apagar as velas .

[Mostrar anotação] [Editar anotação]

52/487 - FOLHA_DOC000041_SENT021

As **mulheres** se calavam e depois de algumas semanas , **desistiam** de o tratamento .

[Mostrar anotação] [Editar anotação]

53/487 - FOLHA_DOC003152_SENT003

A **história** digna de heroína de fábula vendeu mais de 2,7 milhões de cópias , foi **traduzida** em 30 idiomas

[Mostrar anotação] [Editar anotação]

54/487 - FOLHA_DOC000100_SENT062

A **Folha** enviou e-mail sobre isso a a Fifa , mas não **obteve** resposta .

[Mostrar anotação] [Editar anotação]

Figura 17: Exemplos de frases devolvidas com a regra de coordenação de sujeitos

No entanto, a **frase 53** é um subtipo desta regra, já que "**traduzir**" está na voz passiva (sujeitos da passiva recebem Arg1 ou 2 ou 3, a depender do verbo, mas nunca recebem Arg0). Após a análise das linhas de concordância, a regra geral de coordenação foi desmembrada em 3 regras:

- regra para coordenação de sujeitos com **voz passiva** (23 sujeitos explicitados e anotados)
- regra para coordenação de sujeitos com **sujeitos do tipo Arg1** (42 sujeitos explicitados e anotados)
- regra para os demais casos (153 sujeitos explicitados e anotados)

A regra a seguir trata os casos em que o verbo coordenado está na voz passiva.

```
for token in sentence.tokens:
        if token.deprel == "conj" and token.upos == "VERB" and token.head_token.upos == "VERB" and regex(".*Pass.*",
        token.feats):
                for _nsubj in sentence.tokens:
                        if _nsubj.deprel == "nsubj" and _nsubj.dephead == token.head_token.id:
                                verbo = token
                                tem_sujeito = False
                                for _token in sentence.tokens:
                                        if _token.dephead == verbo.id and regex("nsubj|nsubj:pass|csubj",
                                        _token.deprel):
                                                tem_sujeito = True
                                                break
                                if not tem_sujeito:
                                        bold_tokens.append(token)
                                        bold_tokens.append(_nsubj)
```

e devolve casos como os da figura 18

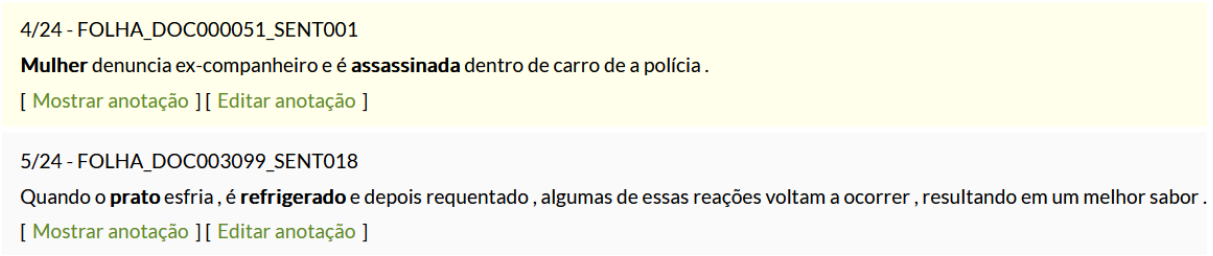


Figura 18: Exemplos de frases devolvidas com a regra de coordenação de sujeitos e passiva

A regra a seguir trata os casos que têm sujeitos do tipo Arg1 (verbos do tipo “parecer” “assemelhar”, “custar” etc etc). Cada frase é lida e, com a utilização do filtro, são excluídos os casos em que a regra de anotação não seria aplicada (casos em que o sujeito não seria Arg1)

```
for token in sentence.tokens:
     if token.deprel == "conj" and token.upos == "VERB" and
token.head_token.upos == "VERB" and not regex(".*Pass.*",
token.feats) and not regex(".*FEITO.*", token.misc) and
regex("aplicar|apoiar|aumentar|bater|cair|chamar|classificar|coloc
ar|começar|comer|concentrar|contar|conter|continuar|correr|crescer
|custar|dar|definir|desaparecer|durar|estancar|estrear|faltar|fech
ar|ficar|figurar|haver|inchar|integrar|ir|manter|morrer|mostrar|mu
dar|parar|parecer|passar|pegar|perder|permanecer|precisar|preparar
|recuperar|revelar|sair|sangrar|seguir|servir|significar|sumir|ten
der|terminar|tornar|transmitir|vender|vir|virar|voltar",
token.lemma):
          for _nsubj in sentence.tokens:
               if _nsubj.deprel == "nsubj" and _nsubj.dephead ==
token.head_token.id:
                    verbo = token
                    tem_sujeito = False
                    for _token in sentence.tokens:
                         if _token.dephead == verbo.id and
regex("nsubj|nsubj:pass|csubj", _token.deprel):
                              tem_sujeito = True
                              break
                    if not tem_sujeito:
                         bold_tokens.append(token)
```

```
bold_tokens.append(_nsubj)
```

Após o tratamento desses casos, a regra geral (apresentada no início da seção) é aplicada aos casos verbo/conj que ainda não foram anotados (os que não têm a etiqueta "FEITO"), com atribuição de Arg0 aos nsubj do tipo. As linhas de concordância são analisadas, e, eventualmente, corrigidas.

Exemplos de ocorrências que precisaram ser corrigidas manualmente

No exemplo abaixo, embora **registrar** não exija um subj Arg1, o que temos é "bebê" como objeto de **registrar** – ou, alternativamente, como sujeito de registrar em uma leitura incoativa. De qualquer maneira, "bebê" será Arg1 em qualquer uma das análises.

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

Já nos exemplos abaixo, apesar da coordenação, não temos um mesmo sujeito compartilhado entre os verbos "**encantar**" (sujeito é samba) e "**achar**" (o sujeito – não explicitado -- é "nós")[27]:

Encantava nos o **samba** e o choro e **achávamos** nos livres de preconceito .

A **Formiga** ainda tem condições de jogar por a seleção brasileira e **pretendo** também conversar com ela , completou .

Além de - coordenação de verbos

Construções com "além de" foram tratadas separadamente porque, na anotação sintática, não envolvem a deprel *conj*. Para lidar com casos como os apresentados na figura 19, em que a coordenação de sujeitos envolve "além de", usamos a expressão de busca descrita a seguir (que depois é adaptada para uma regra de anotação)

1/5 - FOLHA_DOC003118_SENT012

Além de **custear** a expansão de a rede , **ela** também subsidia os consumidores de baixa renda .

[Mostrar anotação] [Editar anotação]

2/5 - FOLHA_DOC003169_SENT005

" Hoje , além de **disputar** atenção com os concorrentes , os **produtos** precisam encontrar espaço entre as distrações de o mundo moderno " , diz Leonardo Massarelli , sócio de a consultoria Questtonó .

[Mostrar anotação] [Editar anotação]

3/5 - FOLHA_DOC003208_SENT065

Além de **conhecer** as dependências de a fábrica , o **visitante** pode tomar cinco opções de chope , com sugestões como a rauchbier , a altbier , a helles e a weizen ... quase que direto de o barril .

[Mostrar anotação] [Editar anotação]

Figura 19: Exemplos de frases com coordenação do tipo "além de"

Expressão de busca:

[27] Porque a preocupação não era fazer um anotador, não foi feita regra de convergência morfológica.

```
for token in sentence.tokens:
        if token.upos == "VERB" and not regex(".*FEITO.*", token.misc) and token.head_token.lemma == "além" and
        token.head_token.misc == "ArgM-adv" and token.head_token.head_token.upos == "VERB":
                for _nsubj in sentence.tokens:
                        if _nsubj.deprel == "nsubj" and _nsubj.dephead == token.head_token.head_token.id and not
                        regex("NSUBJ_COP", _nsubj.misc):
                                verbo = token
                                tem_sujeito = False
                                for _token in sentence.tokens:
                                        if _token.dephead == verbo.id and regex("nsubj|nsubj:pass|csubj",
                                        _token.deprel):
                                                tem_sujeito = True
                                                break
                                if not tem_sujeito:
                                        bold_tokens.append(token)
                                        bold_tokens.append(_nsubj)
```

Regra de anotação correspondente:

```
for token in sentence.tokens:
        if token.upos == "VERB" and not regex(".*FEITO.*", token.misc) and token.head_token.lemma == "além" and
        token.head_token.misc == "ArgM-adv" and token.head_token.head_token.upos == "VERB":
                for _nsubj in sentence.tokens:
                        if _nsubj.deprel == "nsubj" and _nsubj.dephead == token.head_token.head_token.id and not
                        regex("NSUBJ_COP", _nsubj.misc):
                                verbo = token
                                tem_sujeito = False
                                for _token in sentence.tokens:
                                        if _token.dephead == verbo.id and regex("nsubj|nsubj:pass|csubj",
                                        _token.deprel):
                                                tem_sujeito = True
                                                break
                                if not tem_sujeito:
                                        _nsubj.misc = append_to(_nsubj.misc, "Arg0:" + token.id)
                                        token.misc = append_to(token.misc, "FEITO")
```

Acl com formas participiais

As estruturas de *acl* com verbos nas **formas participiais** foram encontradas com a seguinte expressão de busca

```
token.upos == "VERB" and token.deprel == "acl" and token.feats == ".*Part.*"
and token.head_token.upos == "NOUN"
```

A análise das linhas de concordâncias mostrou que o papel semântico do sujeito associado ao verbo *acl* será, em geral, Arg1 – como nas orações passivas – , mas nem sempre. Por isso é fundamental a análise humana.

- Arg1: A **cartilha** de o rebaixamento , **seguida** a a risca em o Morumbi , foi repetida por grandes clubes de o planeta .

- Arg2:**General Motors** lidera (18,2 %) , **seguida** por a Fiat (13 %) .

Não foram tratados

Verbo *dizer* na combinação “propriamente dito/dita/ditos/ditas”

- Bem mais durante os debates matinais de o que durante os filmes **propriamente** ditos .

Acl e regras com convergência de traços morfológicos

Para facilitar a detecção dos sujeitos, especificamos na regra de busca (e de anotação) a necessidade de convergência entre traços morfológicos de gênero e número entre a forma participial e o token candidato a sujeito paciente/Arg1[28]:

> **Criado** em 2016 , o **bacharelado** de a PUC deve abrir matrículas em 2019 .

```
for token in sentence.tokens:
        if regex("VERB", token.upos) and token.deprel == "acl" and not regex(".*FEITO.*", token.misc) and
        regex(".*Part.*", token.feats) and regex("NOUN|PRON|PROPN", token.head_token.upos) and token.gender ==
        token.head_token.gender and token.number == token.head_token.number:
                bold_tokens.append(token)
                bold_tokens.append(token.head_token)
```

Do mesmo modo, foi levada em conta a possibilidade de *coordenação* (conj) dos acl participiais (exemplo abaixo), levando à seguinte regra[29] (abaixo, a regra de anotação)

> O garimpo em o armário revelou valiosos e úteis **acessórios** , escondidos ou **esquecidos** .

```
if regex("conj", token.deprel) and not regex(".*FEITO.*", token.misc) and regex("acl", token.head_token.deprel) and
regex("VERB", token.head_token.upos) and regex("FEITO", token.head_token.misc) and regex(".*Part.*", token.head_token.feats)
and regex("NOUN", token.head_token.head_token.upos):
        token.head_token.head_token.misc = append_to(token.head_token.head_token.misc, "Arg1:" + token.id)
        token.misc = append_to(token.misc, "FEITO")
```

No entanto, mesmo a regra de convergência morfológica não é absoluta, pois há casos que morfologicamente divergem e que não são erros, como abaixo. No entanto, separar os casos de convergência daqueles de divergência torna o processo de análise das concordâncias muito mais rápido.

> A valorização de o indivíduo e a convivência exclusiva entre iguais têm acarretado o aumento de a **intolerância** e de os preconceitos , amplamente **difundidos** em as redes sociais .
> Por outro lado , um **molho** a a bolonhesa ou um frango a o curry **requentados** ... hummm !

A mesma estratégia de convergência de traços morfológicos também foi utilizada para detectar casos de erro na anotação - neste caso, buscando por divergências morfológicas (as frases foram enviadas para a equipe responsável pela camada de anotação sintática).

> O aplicativo de mensagens instantâneas WhatsApp disponibilizou para download sua **versão** para **negócios** anunciada em setembro . (a dependência é entre versão e anunciada, mas está como sendo entre “negócios” e “anunciada”)

[28] Regra “ACL ARg1 identidade feats BUSCA”
[29] Regra “ACL ARg1 identidade feats com CONJ”.

Demais casos de Acl

Diferentemente da previsibilidade dos *acl* com formas participiais, em que o nsubj tenderá a Arg1, nos acl com formas no gerúndio e no infinitivo a atribuição de papel semântico ao head NOUN precisa levar em conta cada verbo. Abaixo, na frase 1, bate-estaca é Arg1 de **vir**. Na frase 2, homem é Arg0 de **querer**. Na frase 3, agrados é Arg2 de **incluir**.

1. Enquanto Twain evitava cuidadosamente ouvir música atual , ela não conseguia evitar ouvir o bate-estaca **vindo** de o quarto de o filho .
2. Mais um homem **querendo** minimizar a cultura de o estupro .
3. O presidente já fez vários agrados a o senador , **incluindo** um telefonema em que diz ser " pró-etanol " .

Ainda assim, nem sempre o substantivo head será o sujeito, como nos exemplos 4 e 5

4. É para ter a sensação de que ninguém , absolutamente ninguém está **ouvindo** nossa profunda desafinação .
5. Carlos Ambrósio , vice-presidente de a Anbima , atribui o resultado histórico a os primeiros sinais de que a economia brasileira está se **recuperando** .

Para tratar desses casos, mais algumas regras foram formuladas.

Acl com formas infinitivas

A regra abaixo encontra frases como 1 e 2, nas quais o sujeito negritado será Arg0 (a certeza de Arg0 se deve ao fato de a regra já exigir a presença de um Arg1 associado ao verbo acl)

1. O **grupo** HMB e os rappers Virgul e Carlão estão acostumados a **dividir** shows em a tímida cena portuguesa de o gênero .
2. Livres que somos , **nós** estamos dispostos a **defender** nossa soberania e nossa democracia em qualquer cenário e modalidade .

```
for token in sentence.tokens:
        if regex("acl", token.deprel) and token.upos == "VERB" and regex(".*Inf.*", token.feats) and not regex(".*FEITO.*", token.misc) and
        regex("ADJ", token.head_token.upos):
                for _nsubj in sentence.tokens:
                        if regex("nsubj.*", _nsubj.deprel) and _nsubj.dephead == token.head_token.id:
                                for _obj in sentence.tokens:
                                        if regex("Arg1", _obj.misc) and _obj.dephead == token.id:
                                                bold_tokens.append(token)
                                                bold_tokens.append(_nsubj)
```

A regra para anotação é

```
for token in sentence.tokens:
        if regex("acl", token.deprel) and token.upos == "VERB" and regex(".*Inf.*", token.feats) and not regex(".*FEITO.*", token.misc) and
        regex("ADJ", token.head_token.upos):
                for _nsubj in sentence.tokens:
                        if regex("nsubj.*", _nsubj.deprel) and _nsubj.dephead == token.head_token.id:
                                for _obj in sentence.tokens:
                                        if regex("Arg1", _obj.misc) and _obj.dephead == token.id:
                                                _nsubj.misc = append_to(_nsubj.misc, "Arg0:" + token.id)
                                                token.misc = append_to(token.misc, "FEITO")
```

Frases do tipo 3 e 4, abaixo, que envolvem verbo ser/cópula e nos quais o sujeito é Arg1 do verbo negritado, são tratadas com a regra de anotação a seguir, após análise:

3. A maior " **culpa** " paterna em esse bolo não é difícil de **entender** .
4. " Eles só preferem negar , mas as **evidências** são difíceis de **contestar** " , ele diz .

```
for token in sentence.tokens:
        if regex("acl", token.deprel) and token.upos == "VERB" and regex(".*Inf.*", token.feats) and not regex(".*FEITO.*",
        token.misc) and regex("ADJ", token.head_token.upos):
                for _nsubj in sentence.tokens:
                        if regex("nsubj.*", _nsubj.deprel) and _nsubj.dephead == token.head_token.id:
                                _nsubj.misc = append_to(_nsubj.misc, "Arg1:" + token.id)
                                token.misc = append_to(token.misc, "FEITO")
```

Já a regra abaixo encontra casos como 1-3 abaixo, nos quais temos um acl no infinitivo dependente de um ADJ, que por sua vez depende de um NOUN (candidato a argumento do verbo acl)

```
for token in sentence.tokens:
        if regex("acl", token.deprel) and token.upos == "VERB" and regex(".*Inf.*", token.feats) and not regex(".*FEITO.*",
        token.misc) and regex("ADJ", token.head_token.upos) and regex("amod", token.head_token.deprel) and regex("NOUN|PROPN",
        token.head_token.head_token.upos):
                bold_tokens.append(token)
                bold_tokens.append(token.head_token.head_token)
```

1. A tropa deve ter orgulho e confiança em uma **liderança** capaz de **enfrentar** desafios de maneira austera , racional e ponderada .
2. Muitos casos de automutilação estariam associados a a necessidade de chamar a atenção de os pais ou a a dificuldade de lidar com uma **dor** difícil de **explicar** .
3. Mas não demorou a receber treinamento de os EUA , sendo beneficiado também por a chegada de agentes de a **CIA** (Agência Central de Inteligência) , ansiosa por **ver** Che Guevara morto .

Como é possível observar, **a mesma estrutura traz casos com soluções diferentes**: às vezes o nome negritado é sujeito (Arg0) do verbo (*liderança; CIA*), e às vezes é objeto (Arg1) do verbo (*dor*). Caso não fosse percebida uma prevalência de nenhum dos casos, poderíamos usar o recurso de filtro da ET, e com isso separamos os casos Arg0 dos casos Arg1 para anotação posterior (figura 20).

Figura 20: Exemplo de utilização do recurso filtro no ambiente Interrogatório da ET

Na frase (1) abaixo, o Arg0 é o objeto da oração principal; na frase 2, o Arg0 é o sujeito da oração principal.

(1) Em abril , a polícia prendeu dois **homens** " radicalizados " em Marselha , suspeitos de **preparar** um ataque " iminente " em a França , cinco dias antes de as eleições

(2) **Pedro Botelho** , 10 , tirava uma foto antes de **entrar** em a fila para o lounge de a NBA , o mais lotado em a tarde de domingo .

A análise atenta permite criar uma regra para cada caso.
Para frases como (1), temos a regra abaixo

```
for token in sentence.tokens:
        if regex("acl", token.deprel) and token.upos == "VERB" and regex(".*Inf.*", token.feats) and not regex(".*FEITO.*",
        token.misc) and regex("ADJ", token.head_token.upos) and regex("NOUN|PROPN|PRON", token.head_token.head_token.upos) and
        regex("obj", token.head_token.head_token.deprel) and regex("VERB", token.head_token.head_token.head_token.upos):
                for _nsubj in sentence.tokens:
                        if regex("nsubj.*", _nsubj.deprel) and _nsubj.dephead == token.head_token.head_token.head_token.id:
                                bold_tokens.append(token)
                                bold_tokens.append(token.head_token.head_token)
```

que identifica frases como

A. Entre as novidades , Cambuhy destaca um **trator** capaz de **fazer** duas funções a o mesmo tempo .
B. Mas muitas pessoas simplesmente não possuem os **recursos** necessários para **explorar** seus interesses e preocupações .

No entanto, a ambiguidade persiste. Na frase (a), não temos dúvida quanto ao *trator* ser Arg0 de **fazer**. Já na frase (b), *recursos* são o instrumento da exploração (ArgM-mnr), e o Arg0 é *pessoas*. Cada caso é então tratado individualmente, e pode estar associado a uma regra, como a regra abaixo, que anota frases do tipo (a):

```
for token in sentence.tokens:
        if regex("acl", token.deprel) and token.upos == "VERB" and regex(".*Inf.*", token.feats) and not regex(".*FEITO.*",
        token.misc) and regex("ADJ", token.head_token.upos) and regex("NOUN|PROPN|PRON", token.head_token.head_token.upos) and
        regex("obj", token.head_token.head_token.deprel) and regex("VERB", token.head_token.head_token.head_token.upos):
                for _nsubj in sentence.tokens:
                        if regex("nsubj.*", _nsubj.deprel) and _nsubj.dephead == token.head_token.head_token.head_token.id:
                                token.head_token.head_token.misc = append_to(token.head_token.head_token.misc, "Arg0:" + token.id)
                                token.misc = append_to(token.misc, "FEITO")
```

Para anotar frases como a 2, a regra é

```
for token in sentence.tokens:
        if regex("acl", token.deprel) and token.upos == "VERB" and regex(".*Inf.*", token.feats) and not regex(".*FEITO.*", token.misc)
        and regex("ADV", token.head_token.upos) and regex("NOUN|PROPN|PRON", token.head_token.head_token.upos) and regex("obj",
        token.head_token.head_token.deprel) and regex("VERB", token.head_token.head_token.head_token.upos):
                for _nsubj in sentence.tokens:
                        if regex("nsubj.*", _nsubj.deprel) and _nsubj.dephead == token.head_token.head_token.head_token.id:
                                bold_tokens.append(token)
                                bold_tokens.append(_nsubj)
```

para a qual existe a seguinte regra de anotação (e que só detectou um único caso)

```
for token in sentence.tokens:
        if regex("acl", token.deprel) and token.upos == "VERB" and regex(".*Inf.*", token.feats) and not regex(".*FEITO.*", token.misc)
        and regex("ADV", token.head_token.upos) and regex("NOUN|PROPN|PRON", token.head_token.head_token.upos) and regex("obj",
        token.head_token.head_token.deprel) and regex("VERB", token.head_token.head_token.head_token.upos):
                for _nsubj in sentence.tokens:
                        if regex("nsubj.*", _nsubj.deprel) and _nsubj.dephead == token.head_token.head_token.head_token.id:
                                _nsubj.misc = append_to(_nsubj.misc, "Arg0:" + token.id)
                                token.misc = append_to(token.misc, "FEITO")
```

Ainda quanto aos acl infinitivos, as coisas podem ficar ainda mais complexas:

Sujeitos e objetos do acl com verbo ser

O fato de UD dar um tratamento sintático especial ao verbo *ser* faz com que sejam necessárias regras especiais para esses casos. A regra a seguir identifica casos como o abaixo. No entanto, apenas 1 caso correspondeu à estrutura buscada.

> Vamos ver quanto **dinheiro** o seu **dinheiro** é capaz de **colocar** em o seu bolso em o longo prazo ?

```
for token in sentence.tokens:
        if regex("acl", token.deprel) and token.upos == "VERB" and not regex(".*FEITO.*", token.misc) and
        regex("ADJ", token.head_token.upos) and regex("obj", token.head_token.head_token.deprel):
                for _nsubj in sentence.tokens:
                        if regex("nsubj.*", _nsubj.deprel) and _nsubj.dephead == token.head_token.id:
                                bold_tokens.append(token)
                                bold_tokens.append(token.head_token)
                                bold_tokens.append(token.head_token.head_token)
```

Acl:relcl e explicitação de sujeitos

Estes casos também foram tratados por meio de regras, como abaixo, que busca pelos heads NOUN do verbo anotado como acl:relcl e pelos heads VERB desse mesmo NOUN, indicando que o sujeito desse head VERB (em vermelho) é candidato ao sujeito do acl:relcl (em negrito)

```
for token in sentence.tokens:
        if regex(".*relcl", token.deprel) and token.upos == "VERB" and not regex(".*FEITO.*", token.misc) and regex("NOUN",
        token.head_token.upos) and regex("VERB", token.head_token.head_token.upos):
                for _nsubj in sentence.tokens:
                        if regex("nsubj.*", _nsubj.deprel) and _nsubj.dephead == token.head_token.head_token.id:
                                bold_tokens.append(token)
                                bold_tokens.append(_nsubj)
```

A regra de busca acima identifica frases como

> Os **alunos** apenas reproduzem em a escola a violência que **experimentam** desde a mais tenra idade em suas famílias .
>
> A **banda** de rock U2 e o cantor Ed Sheeran cancelaram seus shows que **fariam** em este sábado e domingo , respectivamente , em Saint Louis , citando questões de segurança .
>
> Em " The Crown " , **Claire** Foy , concorrendo como melhor atriz de série dramática , destaca se por a docilidade com que **interpreta** Elizabeth 2ª .

Mas também traz “erros”, como abaixo:

> Mas as **entradas** de Denílson e Jucilei tiraram de o São Paulo a agressividade que **tinha** até a metade de o segundo tempo.

ou abaixo, em que não há sujeito a ser buscado/anotado:

> As **eficiências** que vamos **trazer** agora vão ser compensadas por a retomada de o mercado .

Após a análise e tratamento dos casos errados ou inadequados (por ex, quando o sujeito era Arg1), foi aplicada a regra geral abaixo:

```
for token in sentence.tokens:
        if regex(".*relcl", token.deprel) and token.upos == "VERB" and not regex(".*FEITO.*", token.misc) and regex("NOUN",
        token.head_token.upos) and regex("VERB", token.head_token.head_token.upos):
                for _nsubj in sentence.tokens:
                        if regex("nsubj.*", _nsubj.deprel) and _nsubj.dephead == token.head_token.head_token.id:
                                _nsubj.misc = append_to(_nsubj.misc, "Arg0:" + token.id)
                                token.misc = append_to(token.misc, "FEITO")
```

Xcomp no padrão “ter + OBJ + xcomp_pcp”

Devido à previsibilidade, os primeiros casos tratados de xcomp foram do tipo “teve seu sucesso baseado...”, “tiveram o acesso negado...”; “ter sua meta vazada”, com a estrutura verbo ter + obj_Arg1 + xcomp_Arg2_verbo_pcp

Nesses casos, o token obj_Arg1 será Arg1 do xcomp. Abaixo estão a regra de busca e a correspondente regra de anotação[30]. Esta regra anotou 19 ocorrências.

```
token.upos == "VERB" and token.deprel == "xcomp" and
token.feats == ".*Pass.*" and token.head_token.lemma == "ter"
```

```
for token in sentence.tokens:
        if token.deprel == "xcomp" and not regex(".*FEITO.*", token.misc) and regex(".*Pass.*", token.feats) and token.head_token.lemma == "ter":
                for _obj in sentence.tokens:
                        if _obj.deprel == "obj" and _obj.dephead == token.head_token.id:
                                _obj.misc = append_to(_obj.misc, "Arg1:" + token.id)
                                token.misc = append_to(token.misc, "FEITO")
```

[30] Regra “SUBJ de TER obj xcomp PCP Arg1”

Xcomp no particípio

Neste caso, a regra busca xcomp na forma participial, cujo head é VERBO, e associado a este verbo está um obj. O token obj será sujeito paciente (Arg1 na maioria das vezes) do xcomp[31]. Esta regra anotou 3 ocorrências.

Dodge defende manter **Joesley** e Saud , de a JBS , **presos** preventivamente .

```
for token in sentence.tokens:
        if token.deprel == "xcomp" and not regex(".*FEITO.*", token.misc) and regex(".*Part.*", token.feats) and
        regex("VERB", token.upos) and token.head_token.upos == "VERB":
                for _obj in sentence.tokens:
                        if _obj.deprel == "obj" and _obj.dephead == token.head_token.id:
                                _obj.misc = append_to(_obj.misc, "Arg1:" + token.id)
                                token.misc = append_to(token.misc, "FEITO")
```

Xcomp no infinitivo

Em seguida, foram tratados xcomp na forma infinitiva que complementam ADJ, como em

"**Os** que furiosamente bradam em nome de (...) não são capazes de se **encarar**_xcomp no espelho".

ou que complementam verbos, como em

"o sistema consegue **distinguir_xcomp** plásticos..."

Em construções como essas a heurística foi:

buscar o head do verbo/xcomp (capazes|consegue)
buscar o sujeito desse head (os|sistema)
indicar esse nsubj como provável sujeito do verbo no infinitivo.

Há ainda casos com obj_arg1, em que precisaremos de outra(s) regra(s):

"Janot acusa **Temer** de **obstruir** a Justiça"

Foram criadas regras em que o obj_Arg1 é Arg1 do xcomp infinitivo[32](15 ocorrências anotadas); regras em que o obj_Arg1 é Arg0 do xcomp infinitivo[33] (52 ocorrências anotadas)

Isso só pode ser um sintoma de que o governo Temer tá fazendo o **Brasil**(Arg1) **decolar** . (Brasil = Arg1)
"Janot acusa **Temer**(Arg0) de **obstruir** a Justiça"

Ao final, foram elaboradas "regras de finalização" para os casos não detectados pelas regras anteriores.

Advcl e gerúndio: não anotado

Orações adverbiais gerundivas, quando equivalentes a "o que" (abaixo), foram consideradas sem sujeito, e **não tiveram Arg0 anotado**. Nesses casos, o entendimento é de um agente oracional, que não pode ser localizado em um único sintagma/elemento nominal.

---

[31] Regra "SUBJ de varios obj xcomp PCP Arg1"
[32] Regra "XCOMP Verbo ARg2 xcomp é Arg1 ANOTA"
[33] Regra "XCOMP Verbo ARg2 xcomp é Arg0 ANOTA"

1. Isso joga um pouco de água em a propaganda petista , mas não significa que , sob Lula , o país não tenha enriquecido , **melhorando** também a situação de os pobres.
2. Em o ano seguinte , o grupo expulsou de Gaza as forças de Abbas , **deixando** o presidente em o controle apenas de áreas autônomas em a Cisjordânia.
3. Quatro categorias respondem , em conjunto , por 90 % de a indústria , **sinalizando** de maneira inequívoca a preferência de o investidor .
4. Quando o prato esfria , é refrigerado e depois requentado , algumas de essas reações voltam a ocorrer , **resultando** em um melhor sabor .

Como sempre, há exceções, e nas frases abaixo, "fibras" e "Janot" foram anotados como Arg0 de "liberar" e "deixar", respectivamente. Neste caso, podemos afirmar que as fibras liberam o material gelatinoso. Na frase anterior, não são as reações que resultam em um melhor sabor, mas a oração completa.

5. Quando esfriam e são requentados , tornam se mais viscosos porque as **fibras** de a proteína se decompõem , **liberando** o material gelatinoso que fica entre as células .
6. **Janot** fechou 159 acordos de colaboração , **deixando** para Dodge uma série de inquéritos em curso .

E, também como sempre, temos os casos vagos/ambíguos (mas cuja vagueza não impede a compreensão.) Na frase abaixo, podemos dizer que *Rajoy seria o responsável por desmontar o aparato separatista* (e, portanto, Arg0) , ou que *o fato de ele forçar a Catalunha a antecipar as eleições é o responsável pelo desmonte* (e, portanto, não há um Arg0 delimitado)?

> Com essa ferramenta , Rajoy poderia forçar a Catalunha a antecipar suas eleições , **desmontando** o aparato separatista .

Advcl e formas infinitivas : anotados às vezes

Para as formas infinitivas, se não fosse possível reconhecer o sujeito, nada era feito. Ou seja, na frase 1 explicitamos o sujeito na anotação, mas na frase 2 não fizemos nada.

1. Autoridades de a área econômica planejavam **vender** os 56 terminais de a empresa e extinguir la até 2018 . (*quem venderá? Autoridades da área econômica*)
2. Evair Evaristo Lima , técnico de a Acquazero , franquia de lavagem a seco , ensina a **fazer a limpeza** em casa . (*quem fará a limpeza? Como não está no texto, nada foi feito*)

Do mesmo modo, **foi anotada** a relação entre argumento agente e verbo (eliminar) em frases como (3). Para uma pergunta do tipo *Pizi eliminou a seleção?* A resposta "sim" é

aceitável, ainda que talvez imprecisa, mas a resposta “não” é mais imprecisa ainda. E mais um argumento para fazer de SRL verificador de compreensão (o que quer que seja isso)

3. Argentino , **Pizzi** ajudou a eliminar a seleção de o seu país .

Advcl anotados

Devido à previsibilidade, foram tratados inicialmente os sujeitos de verbos **ArgM-prd**. Naqueles com **verbo no gerúndio**, a tendência era de os sujeitos serem **Arg0**, como na frase abaixo.

> **Vestindo** uma camiseta com a frase " somos escravos de a saúde " , **ele** explicou a os indígenas que " queria trabalhar em harmonia " .

Após a análise dos resultados da regra de busca, e eventual tratamento dos casos em que o sujeito não seria Arg0, foi aplicada a seguinte regra de anotação[34] (que anotou 3 casos):

```
for token in sentence.tokens:
        if token.misc == "ArgM-prd" and token.feats == "VerbForm=Ger" and token.head_token.upos == "VERB":
                for _nsubj in sentence.tokens:
                        if _nsubj.deprel == "nsubj" and _nsubj.dephead == token.head_token.id:
                                _nsubj.misc = append_to(_nsubj.misc, "Arg0:" + token.id)
                                token.misc = append_to(token.misc, "FEITO")
```

O mesmo procedimento foi feito para os sujeitos de verbos **ArgM-prd** com **verbo no particípio**. Neste caso, porém, a tendência era de os sujeitos serem **Arg1** (a regra anotou 33 ocorrências)[35].

> **Formado** por quatro músicos , **ele** leva a os espectadores composições feitas principalmente em os séculos 20 e 21 .

```
for token in sentence.tokens:
        if regex("ArgM\-prd", token.misc) and regex(".*Part.*", token.feats) and token.upos == "VERB" and
        token.head_token.upos == "VERB":
                for _nsubj in sentence.tokens:
                        if _nsubj.deprel == "nsubj" and _nsubj.dephead == token.head_token.id:
                                _nsubj.misc = append_to(_nsubj.misc, "Arg1:" + token.id)
                                token.misc = append_to(token.misc, "FEITO")
```

Em seguida, foram tratados (isto é, identificados os sujeitos e seus papéis semânticos) dos verbos advcl do tipo **ArgM-cau**. Para aqueles ArgM-cau com verbo no Infinitivo, se houver sujeito na oração principal, ele é o sujeito da advcl, mas a leitura dos casos é importante para confirmar. O mesmo vale para **ArgM-tmp**.

As seguintes regras foram usadas para anotar **ArgM-cau**, e a diferença entre elas está no “grau de profundidade” do nó da árvore do candidato a sujeito e seu verbo associado. A primeira regra – nó mais baixo da árvore – dá conta de frases do tipo (1 e 2), e anotou 20 tokens[36]. A segunda regra - com mais um nível da árvore - dá conta de frases do tipo (3 e 4), e anotou 5 tokens[37].

[34] Regra “ArgM-prd GER ANOTA”
[35] Regra “ArgM-prd ANOTA”
[36] Regra “ArgM-cau ANOTA 2”
[37] Regra “ArgM-cau ANOTA 3”

1. **Eu** não posso demonstrar raiva , porque **tenho** o lado profissional .
2. O **assunto** provocou comoção porque **colocou** em xeque a qualidade de a carne brasileira .

3. **Gilberto Uzmán** , 35 , fugiu de Sesuntepeque , El Salvador , após ser surrado por uma gangue por **soltar** fogos .
4. A **líder** de fato de o Mianmar , Aung San Suu Kyi , afirmou por meio de sua porta-voz que não viajará por **ter** problemas mais graves para resolver .

```
for token in sentence.tokens:
        if regex("ArgM\-cau", token.misc) and not regex(".*FEITO.*", token.misc) and token.upos == "VERB" and
        token.head_token.upos == "VERB":
                for _nsubj in sentence.tokens:
                        if regex("nsubj.*|csubj", _nsubj.deprel) and _nsubj.dephead == token.head_token.id:
                                _nsubj.misc = append_to(_nsubj.misc, "Arg0:" + token.id)
                                token.misc = append_to(token.misc, "FEITO")
```

```
for token in sentence.tokens:
        if regex("ArgM\-cau", token.misc) and not regex(".*FEITO.*", token.misc) and token.upos == "VERB" and
        token.head_token.upos == "VERB" and token.head_token.head_token.upos == "VERB":
                for _nsubj in sentence.tokens:
                        if regex("nsubj.*|csubj", _nsubj.deprel) and _nsubj.dephead == token.head_token.head_token.id:
                                _nsubj.misc = append_to(_nsubj.misc, "Arg0:" + token.id)
                                token.misc = append_to(token.misc, "FEITO")
```

Como sempre, após a expressão de busca – e anterior à aplicação da regra de anotação – é feita uma análise e os casos minoritários (especificamente aqui, aqueles em que o candidato a sujeito não era Arg0) foram tratados individualmente, e a regra final foi aplicada aos demais casos.

Para **ArgM-tmp, ArgM-mnr** e **demais casos de advcl** foi feito o mesmo procedimento.
Exemplificando com **ArgM-tmp**, foram criadas duas regras, que diferem apenas quanto ao papel semântico atribuído ao sujeito: Arg0 ou Arg1.
A regra 1 (frase 1) - que atribui Arg0 ao sujeito – anotou 50 tokens[38] e a regra 2 (frase 2) - que atribui Arg1 ao sujeito – anotou 16 tokens[39].

1. Não consigo enxergar responsabilidade de o MAM , **que** cumpriu sua obrigação a o **orientar** os visitantes sobre a existência de um nu .

```
for token in sentence.tokens:
        if regex("ArgM\-tmp", token.misc) and token.upos == "VERB" and token.head_token.upos == "VERB":
                for _nsubj in sentence.tokens:
                        if regex("nsubj.*|csubj", _nsubj.deprel) and _nsubj.dephead == token.head_token.id:
                                _nsubj.misc = append_to(_nsubj.misc, "Arg0:" + token.id)
                                token.misc = append_to(token.misc, "FEITO")
```

2. Sabe se até onde pode ir o **STF** quando **exposto** a insinuações tão constrangedoras

[38] Regra “ArgM-tmp ANOTA Arg0”
[39] Regra “ArgM-tmp ANOTA Arg1”

```
for token in sentence.tokens:
        if regex("ArgM\-tmp", token.misc) and token.upos == "VERB" and regex(".*Part.*", token.feats) and
        token.head_token.upos == "VERB":
                for _nsubj in sentence.tokens:
                        if regex("nsubj.*|csubj", _nsubj.deprel) and _nsubj.dephead == token.head_token.id:
                                _nsubj.misc = append_to(_nsubj.misc, "Arg1:" + token.id)
                                token.misc = append_to(token.misc, "FEITO")
```

ArgM-prp

Na anotação e criação de regras, partimos da ideia de que, preferencialmente, orações subordinadas ArgM-prp sem sujeito herdarão o sujeito da principal. A frase 1 corresponde à regra 1, que anotou 88 ocorrências[40]

1. **Evo** Morales reúne líderes de esquerda para **homenagear** Che Guevara .

```
for token in sentence.tokens:
        if regex("ArgM\-prp", token.misc) and token.upos == "VERB" and not regex(".*FEITO.*", token.misc) and
        token.head_token.upos == "VERB" and not regex(".*FEITO.*", token.head_token.misc):
                for _nsubj in sentence.tokens:
                        if regex("nsubj.*|csubj", _nsubj.deprel) and _nsubj.dephead == token.head_token.id:
                                _nsubj.misc = append_to(_nsubj.misc, "Arg0:" + token.id)
                                token.misc = append_to(token.misc, "FEITO")
```

O sujeito "herdado", no entanto, pode estar mais acima na árvore sintática, como na frase 2, tratada pela regra 2, que anotou 25 ocorrências[41].

2. Saí de aqui um comunistinha **que** tinha que roubar para **comer** .

```
for token in sentence.tokens:
        if regex("ArgM\-prp", token.misc) and token.upos == "VERB" and not regex(".*FEITO.*", token.misc) and
        token.head_token.upos == "VERB" and token.head_token.head_token.upos == "VERB":
                for _nsubj in sentence.tokens:
                        if regex("nsubj.*", _nsubj.deprel) and _nsubj.dephead == token.head_token.head_token.id:
                                _nsubj.misc = append_to(_nsubj.misc, "Arg0:" + token.id)
                                token.misc = append_to(token.misc, "FEITO")
```

No entanto, trata-se de uma preferência, e há casos que fogem ao padrão:

3. Ela conta que o **desejo** de fazer a mãe feliz **a** incentivou , inclusive , a abandonar os estudos para se **dedicar** a a música .
4. Eu cobiço o que a Alemanha tem , **que** consegue deixar tanto tempo um **treinador** para **trabalhar** , afirmou Tite .

Para facilitar a identificação desses casos, após o esgotamento de regras que buscavam o sujeito em orações adverbiais (não necessariamente ArgM-prp), foi feita outra regra apenas para detectar casos em que o objeto seria um candidato a argumento do verbo:

```
for token in sentence.tokens:
        if regex("advcl", token.deprel) and token.upos == "VERB" and not regex(".*FEITO.*", token.misc) and
        token.head_token.upos == "VERB":
                for _obj in sentence.tokens:
                        if regex("obj", _obj.deprel) and _obj.dephead == token.head_token.id:
                                bold_tokens.append(token)
                                bold_tokens.append(_obj)
```

E, ainda assim, existem os casos vagos:

[40] Regra "ArgM-prp ANOTA"
[41] Regra "ArgM-prp BUSCA niveis ANOTA"

A **estudante** Ana Braga , 22 , começou a mobilizar amigos de a comunidade LGBT para **levar** a informação a a montadora japonesa .

Nos demais casos de advcl, as regras – várias – foram criadas do mesmo modo, levando em consideração o papel semântico atribuído ao sujeito, a profundidade da árvore sintática e a presença de verbo *ser*, que obriga o head da oração a ser um ADJ ou NOUN.

- A **frase 1** foi tratada pela regra geral 1, que anotou 61 ocorrências[42].
- A **frase 2**, na qual o sujeito está um nível acima na árvore (profundidade), foi tratada pela regra 2, que anotou 11 ocorrências[43].
- A **frase 3,** que exemplifica frase com verbo ser, foi tratada pela regra 3, que anotou 6 ocorrências[44]).

1. Todas essas frentes consumiram investimentos **que** levam mais tempo para **dar** retorno .

```
for token in sentence.tokens:
        if regex("advcl", token.deprel) and token.upos == "VERB" and not regex(".*FEITO.*", token.misc) and
        token.head_token.upos == "VERB":
                for _nsubj in sentence.tokens:
                        if regex("nsubj.*", _nsubj.deprel) and _nsubj.dephead == token.head_token.id:
                                _nsubj.misc = append_to(_nsubj.misc, "Arg0:" + token.id)
                                token.misc = append_to(token.misc, "FEITO")
```

2. **Dodge** diz que o chamou por e-mail , assim como **fez** com todos os procuradores .

```
for token in sentence.tokens:
        if regex("advcl", token.deprel) and token.upos == "VERB" and not regex(".*FEITO.*", token.misc) and
        token.head_token.upos == "VERB" and token.head_token.head_token.upos == "VERB":
                for _nsubj in sentence.tokens:
                        if regex("nsubj.*", _nsubj.deprel) and _nsubj.dephead == token.head_token.head_token.id:
                                _nsubj.misc = append_to(_nsubj.misc, "Arg0:" + token.id)
                                token.misc = append_to(token.misc, "FEITO")
```

3. A **Tesla** (EUA) será a segunda maior fabricante a o **concluir** fábrica de US$ 5 bilhões em Nevada .

```
for token in sentence.tokens:
        if regex("advcl", token.deprel) and token.upos == "VERB" and not regex(".*FEITO.*", token.misc) and regex("NOUN|
        PROPN|ADJ", token.head_token.upos):
                for _nsubj in sentence.tokens:
                        if regex("nsubj.*", _nsubj.deprel) and _nsubj.dephead == token.head_token.id:
                                _nsubj.misc = append_to(_nsubj.misc, "Arg0:" + token.id)
                                token.misc = append_to(token.misc, "FEITO")
```

De fato, a identificação de sujeitos implícitos de orações advcl foi a que levou ao maior número de regras de atribuição de papéis semânticos: 13.

4.3 Validação final

Ao final do processo, passamos o material por um segundo e último processo de validação, tendo em vista detectar inconsistências ou erros de anotação. Para tanto, criamos 4 regras que detectam os seguintes casos[45]:

[42] Regra “ADVCL geral ANOTA”
[43] Regra “ADVCL geral ANOTA níveis”
[44] Regra “ADVCL head é N ou ADJ ANOTA”
[45] Agradeço ao Elvis de Souza pela implementação das regras na etapa de validação do corpus.

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

O quadro 7 indica a quantidade de ocorrências encontradas por cada regra, bem como exemplifica cada um dos fenômenos. Todos os erros detectados foram corrigidos manualmente, caso a caso. Ao final das correções, as regras foram reaplicadas para evitar a introdução de novos erros, e eventuais problemas detectados foram corrigidos. As regras de validação e a correção foram aplicadas no corpus antes de qualquer pós-processamento, ou seja, antes da criação das versões UD e clássica.

Regra	erros	Exemplo	Exemplo corrigido
1	10	interesse Arg1|Arg1:3|**Arg2:3**	Arg1|Arg1:3
2	59	afirmação 10 obj Arg1|**Arg1:13**|Arg2:13 pergunta 13 obl Arg1	Arg1|Arg2:13 Arg1
3	30	trata **root** **ArgM-tmp**	_
4	2	ambos eram falsos positivos.	

Quadro 7: Candidatos a erros encontrados pelas regras de validação

4.4. Anotação de frames

Como já mencionado, a adição da camada SRL no Porttinari só foi possível porque já existia o recurso Verbo-Brasil. Para além da utilização na atribuição dos papéis e sentidos dos verbos, o Verbo-Brasil foi crucial de duas outras maneiras: como fonte de pistas para encontrar padrões linguísticos (etapa 4, já descrita) e como um banco de dados. É deste último aspecto que trataremos aqui.

O Verbo-Brasil está disponível para a comunidade para consultas online, no endereço http://143.107.183.175:21380/verbobrasil/

No entanto, esta interface, assim como a interface do recurso original em inglês, só permite buscar pelo lema dos verbos, e não admite buscas sobre os atributos dos frames (como papel, tipo do argumento, etc) ou sobre o texto utilizado na descrição de cada atributo. Por isso, ao longo do projeto de anotação foi criada uma outra interface de acesso que permite buscas mais específicas no contexto da anotação. Estas buscas foram importantes para a

verificação e confirmação final dos papéis atribuídos, tendo em vista certas "idiossincrasias semânticas" codificadas em um PropBank (ver seção Resultados e análises).

Especificamente, aproveitando a informação presente no atributo "Papéis", podemos, por exemplo, buscar pelos frames em que a ideia de *finalidade* está indicada em um argumento numerado (e não como ArgM-prp, criado justamente para anotar a ideia de finalidade).

A figura 21 mostra a interface de acesso[46], e nela vemos o pedido de uma busca por *frames que contenham a palavra "finalidade" no atributo "Papéis" e que este papel não esteja associado* `(busca negativa)` *a um ArgM* (em outras palavras, que este papel "finalidade" esteja associado a um argumento numerado). A busca retorna 23 frames de verbos, que foram revistos no corpus com relação aos ArgM-prp que (erradamente) pudessem ter.

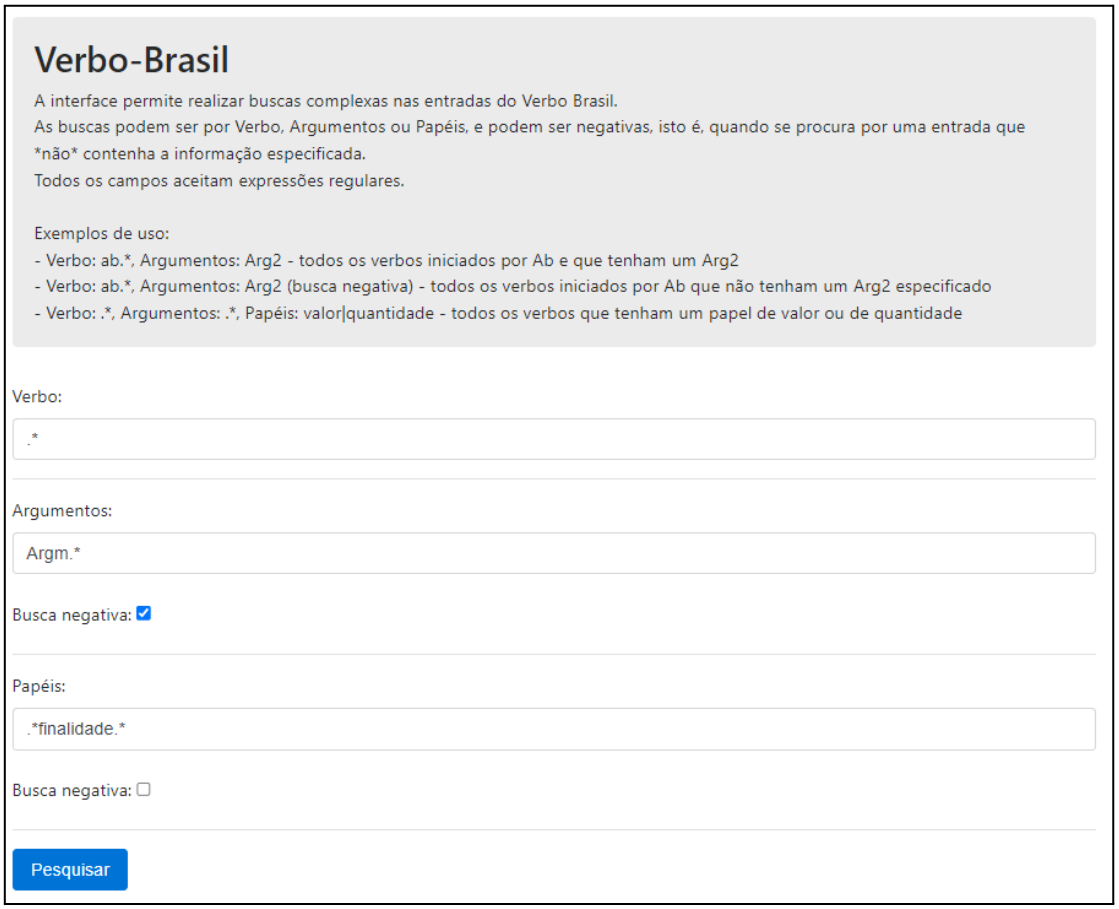


Figura 21: Interface do Verbo-Brasil para a busca de frames que não tenham papel de *finalidade* que não esteja associado a um argumento numerado (e não a um arg modificador).

Do mesmo modo, podemos pedir verbos que não contenham Arg0, o que retorna uma lista de 333 frames distintos (figura 22). Por fim, a interface também pode servir para verificação

[46] Disponível em https://souelvis.dev/ui-verbobrasil/

e melhoria do próprio Verbo-Brasil, quando pedimos, por exemplo, pelos verbos que não têm Arg1, Arg2 ou Arg3, o que pode sinalizar a presença de um frame incompleto (figura 2).

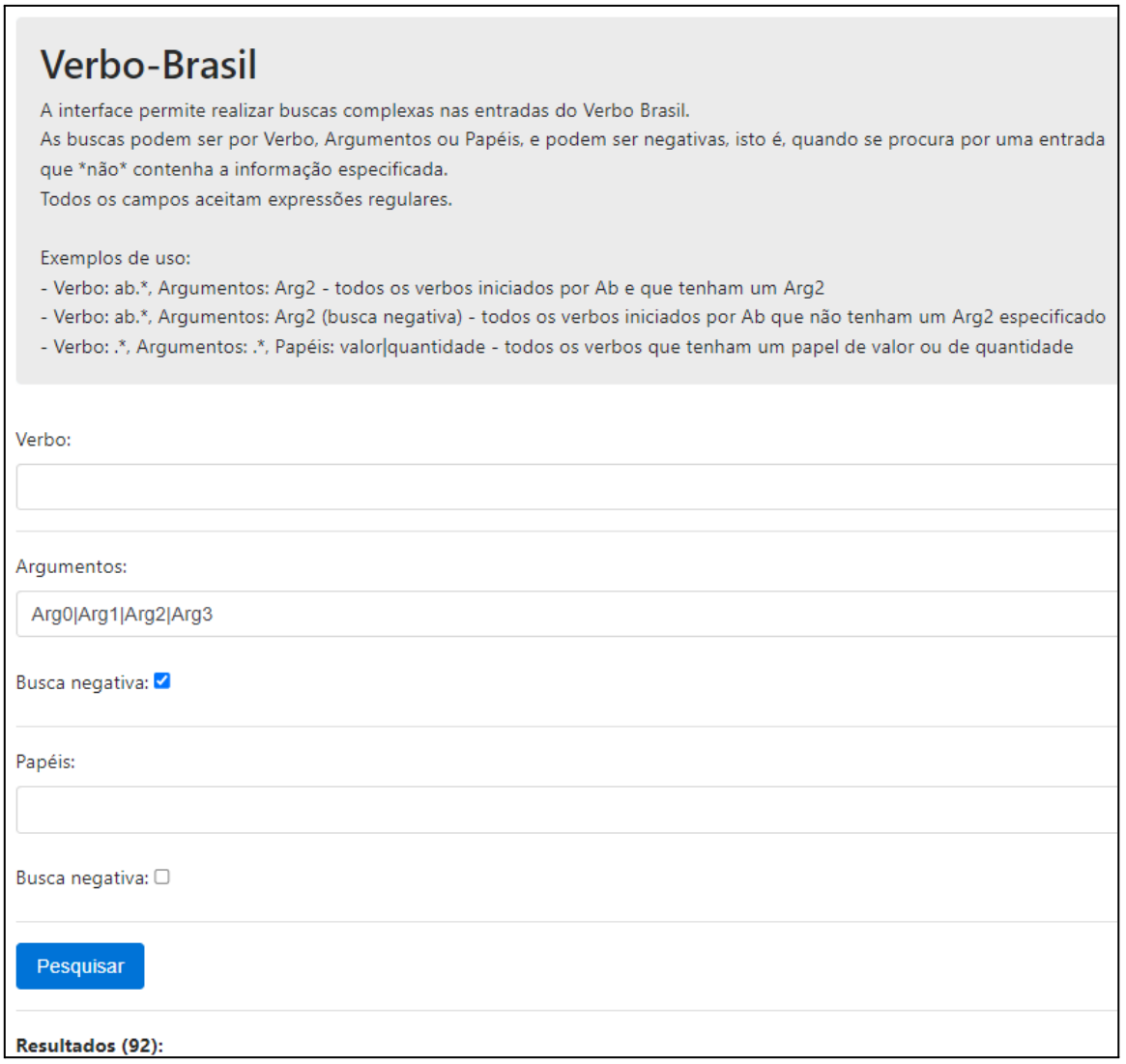


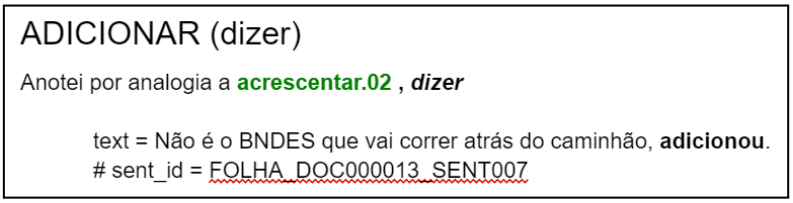
Figura 22: interface alternativa para o Verbo-Brasil

4.5. Anotação de verbos sem frames: frames e anotações provisórios

Como já indicado, o processo de anotação dos frames dos verbos foi concomitante à anotação dos papéis semânticos. Uma vez que o foco da anotação da camada SRL do PropBank esteve na atribuição dos papéis semânticos, o processo de atribuição de sentidos aos verbos não foi exaustivo. E, além de não exaustiva, a anotação privilegiou o alinhamento com verbos não monossêmicos, uma vez que a anotação de verbos monossêmicos pode ser feita de forma automática.

No entanto, apesar de já dispor de um recurso como o Verbo-Brasil (Duran e Aluísio, 2015), um corpus novo sempre traz novas formas verbais e novos sentidos para formas verbais já descritas. Para os sentidos (provisoriamente) sem frames, foram adotadas as seguintes soluções:

Busca por um verbo similar/análogo no Verbo-Brasil, e indicação da solução em um documento para posterior verificação pela pesquisadora responsável (estratégia similar à utilizada no PropBank original). Por exemplo, para o verbo “adicionar”, sem frame, a indicação no documento está assim (figura 23):


ADICIONAR (dizer)

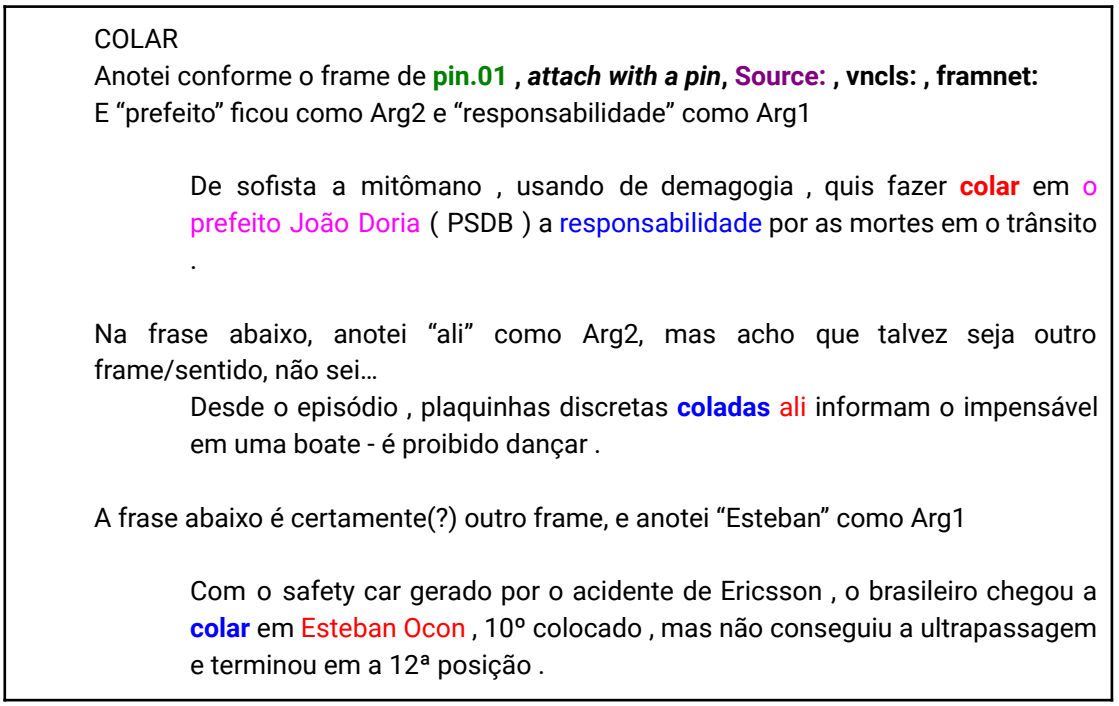
Anotei por analogia a **acrescentar.02** , ***dizer***

text = Não é o BNDES que vai correr atrás do caminhão, **adicionou**.
# sent_id = FOLHA_DOC000013_SENT007

Figura 23: Trecho do documento que contém verbos sem frames e dúvidas na anotação de frames

Quando isto não era possível, ou quando nenhum sinônimo imediato vinha a mente ou estava disponível no Verbo-Brasil:

- criação de um frame com base em outros, não necessariamente similares
- tradução automática (usando Google Translator) da frase do corpus, e busca do verbo em inglês no recurso original. A figura 24 ilustra o caso do verbo COLAR, conforme está no documento que contém dúvidas sobre a anotação de frames:


COLAR
Anotei conforme o frame de **pin.01** , ***attach with a pin***, **Source: , vncls: , framnet:**
E “prefeito” ficou como Arg2 e “responsabilidade” como Arg1

De sofista a mitômano , usando de demagogia , quis fazer **colar** em o prefeito João Doria (PSDB) a responsabilidade por as mortes em o trânsito .

Na frase abaixo, anotei “ali” como Arg2, mas acho que talvez seja outro frame/sentido, não sei...

Desde o episódio , plaquinhas discretas **coladas** ali informam o impensável em uma boate - é proibido dançar .

A frase abaixo é certamente(?) outro frame, e anotei “Esteban” como Arg1

Com o safety car gerado por o acidente de Ericsson , o brasileiro chegou a **colar** em Esteban Ocon , 10º colocado , mas não conseguiu a ultrapassagem e terminou em a 12ª posição .

Figura 24: Trecho do documento que contém verbos sem frames e dúvidas na anotação de frames

Ao final do processo, foram documentados **cerca de 350 sentidos de verbos** sem frames no Verbo-Brasil, com exemplos do corpus e soluções provisórias, em boa parte deles. Existem ainda mais algumas dezenas de frames com alguma necessidade de ajuste (por incompletude ou atualização do recurso em inglês).

4.6 Anotação automática de frames monossêmicos

A anotação de frames verbais conforme o Verbo-Brasil foi feita de maneira automática para os casos de verbos monossêmicos (ou seja, que só dispunham de um frame), desde que a referida forma verbal não contivesse a etiqueta INC na coluna 9[47]. Este procedimento levou à inclusão de 580 frames verbais.

4.6.1.Ausência de contexto

Como o Portinari é um corpus de frases, a ausência de contexto pode impedir a correta atribuição do frames, como na frase abaixo, em que não sabemos se o sentido de **bater** é o de *golpear* (por exemplo em uma luta de boxe), ou o de *chutar*:

> Mas ele treina demais aquela forma de **bater** e sabe muito bem fazer aquilo , afirma o ex-jogador .

4.7. Concordância Interanotadores

Todo o processo de anotação foi conduzido por uma pessoa, a partir das informações contidas no *Manual de anotação do PropBank v2* (Duran, 2014) e nos frames dos verbos conforme listados no recurso *Verbo-Brasil* (Duran et al., 2013; Duran e Aluisio, 2015). A fim de avaliar a qualidade das análises, foi feita uma concordância inter-anotadores *a posteriori,* tomando como base **uma amostra do corpus PropBank-Br v2**, sobre o qual se baseiam as diretivas de anotação e a construção do Verbo-Brasil. Essa amostra foi então reanotada pela anotadora da camada PBP, e os resultados foram comparados utilizando o índice kappa.

Especificamente

1. Selecionamos as 100 frases com a maior quantidade de argumentos anotados no corpus Propbank-Br v2
2. Eliminamos a anotação dos papéis semânticos, bem como a anotação sintática anterior (a anotação do Propbank-Br v.2 foi feita sobre a anotação sintática automática feita pelo parser PALAVRAS)
3. Anotamos o material com o melhor modelo UD disponível
4. Inserimos a amostra PropBank-BR v2 na ferramenta ET, para iniciar o processo de anotação

O processo de anotação da amostra aconteceu na metade do processo de anotação do PBP. A escolha, apesar de arriscada, visava medir exatamente o grau de convergência de duas anotadoras já experientes ao longo do processo de anotação.
Importante notar também que, apesar de termos escolhido as 100 frases com a maior quantidade de argumentos anotados, nem todos os verbos do PropBank-Br v2 têm seus argumentos anotados, mas apenas aqueles mais frequentes no corpus. A comparação, medida com o índice kappa, foi sobre 443 tokens anotados por ambas as anotadoras, e **resultou em uma convergência de 0.907.**

[47] Agradeço ao Elvis de Souza pela implementação desta etapa.

A análise qualitativa dos dados foi feita com a interface Julgamento, ambiente da ferramenta ET para avaliação de anotações. Geramos então uma matriz de confusão comparando ambas as anotações, e analisamos caso a caso cada uma das divergências.

Pela matriz de confusão (figura 25), vemos que a maioria das divergências aconteceu entre ArgM-adv (original) x ArgM-dis, ArgM-mnr (original) e ArgM-adv, ArgM-prd, Arg2 e ArgM-ext, e entre Arg0, Arg1 e Arg2. Todas foram analisadas individualmente, e 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). 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 que, como já indicadas, foram sendo enriquecidas ao longo do projeto.

secundário	Arg0	Arg1	Arg2	Arg3	Arg4	ArgM-adv	ArgM-asp	ArgM-cau	ArgM-dis	ArgM-ext	ArgM-loc	ArgM-mnr	ArgM-mod	ArgM-neg	ArgM-nse	ArgM-pas	ArgM-prp	ArgM-tmp	_	All
principal																				
Arg0	75	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	76
Arg1	1	135	4	0	0	0	0	0	0	1	1	0	0	0	1	0	0	0	0	143
Arg2	2	1	36	0	0	1	0	0	0	1	1	1	0	0	0	0	0	0	0	43
Arg3	0	0	0	4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	4
Arg4	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2
ArgM-adv	0	0	0	0	0	10	0	0	2	0	0	3	0	0	0	0	0	0	0	15
ArgM-asp	0	0	0	0	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0	3
ArgM-cau	0	0	0	0	0	0	0	2	1	0	0	0	0	0	0	0	0	0	0	3
ArgM-dis	0	0	0	0	0	8	0	0	1	0	0	0	0	0	0	0	0	0	0	9
ArgM-ext	0	0	0	0	0	0	0	0	0	5	0	1	0	0	0	0	0	0	0	6
ArgM-loc	0	0	0	1	0	0	0	0	0	0	17	0	0	0	0	0	0	0	0	18
ArgM-mnr	0	0	0	0	0	0	0	0	0	0	0	13	0	0	0	0	0	0	0	13
ArgM-mod	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	1	0	0	0	3
ArgM-neg	0	0	0	0	0	0	0	0	0	0	0	0	0	16	0	0	0	0	0	16
ArgM-nse	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1
ArgM-pas	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	12	0	0	0	12
ArgM-pin	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	2
ArgM-prd	0	0	0	0	0	0	0	0	0	0	1	3	0	0	0	0	0	0	0	4
ArgM-prp	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	13	0	0	14
ArgM-tmp	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	56	0	56
_	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2355	2355
All	78	137	40	5	2	19	3	2	4	7	20	22	2	16	4	13	13	56	2355	2798

Figura 25: Matriz de confusão relativa às divergências na concordância inter-anotadores.

4.7.1.Estudo sobre concordâncias e discordâncias

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. Na versão mais recente das diretivas originais (Bonial et al., 2015), temos apenas duas passagens que mencionam como proceder em caso de dúvida: especificamente para dúvida entre *agente* e *paciente*, e entre "causa" (ArgM-cau) e "finalidade" (ArgM-prp):

> If an argument is both an agent and a patient, then Arg0 label should be selected (2015:8)

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

No entanto – e apesar do alto grau de concordância verificado na anotação da amostra PropBank –, a prática da anotação mostrou que em diversos casos, e envolvendo diferentes classes de ArgM, foi difícil escolher entre duas (ou mais) etiquetas, sendo todas as alternativas disponíveis possíveis no contexto da frase.

Com o objetivo de verificar se este fenômeno era idiossincrasia e falsa impressão de uma única anotadora, foi realizado um pequeno estudo sobre concordâncias com algumas das frases que suscitaram dúvidas.

O estudo verificou a convergência/divergência de análises entre os seguintes grupos de etiquetas do tipo ArgMs:

- local e tempo
- causa e maneira
- causa e tempo
- modo e tempo
- finalidade e local
- tempo, local e maneira

O estudo foi feito por meio de Formulários Google. A fim de manter a atenção dos respondentes e evitar um "questionário" longo, as frases foram distribuídas em 2 formulários distintos: um apenas para as classes *local* e *tempo*, e outro com as demais classes. Ambos continham 16 frases, 3 delas cuja classificação era claramente monossêmica.

Além da variável tipo de classe, o estudo também levou em conta a forma de classificação/anotação, comparando respostas quando apenas uma única alternativa/classe deveria ser selecionada (modelo de anotação do PropBank) e a possibilidade de respostas múltiplas. Foram criados 4 formulários:

Formulário 1. tempo e local com escolha única: 5 respondentes
Formulário 2: tempo e local com escolha múltipla: 8 respondentes
Formulário 3: todos os demais escolha única: 7 respondentes
Formulário 4: todos os demais escolha múltipla: 7 respondentes

Buscamos responder duas perguntas com o estudo:

1) Em que medida a polissemia na anotação de papéis semânticos modificadores (descritivos) é um fenômeno periférico? (análise global dos formulários)
2) Em que medida a obrigatoriedade de classificação única é um fator que contribui para a divergência na anotação? (comparação da convergência e divergência nas respostas entre formulários que permitiam escolha única (formulários 1 e 3) e que permitiam escolha múltipla (formulários 2 e 4)

Diferentemente de uma concordância interanotadores clássica, que conta com um número pequeno de anotadores e muitas frases, neste estudo optamos por um número relativamente pequeno de frases e um maior número de anotadores, justamente para perceber se seria possível falar de tendências quanto ao modo de anotar. Além disso, todos os respondentes eram linguistas com vasta experiência na anotação de corpus e estavam comprometidos com a tarefa. Na anotação do PropBank original, os anotadores recrutados tinham formações variadas, com alunos de graduação a doutorandos. Ao final, mais de 30 pessoas contribuíram com a anotação, algumas por poucas semanas e outras por mais de 3 anos.

O Apêndice 2 traz as frases e respostas para todos os formulários. De modo geral, vemos uma grande variação nas respostas, e uma avaliação aprofundada do material está em Freitas e Pardo (submetido).

A fim de avaliar o grau de confiança nas análises codificadas no PBP, as anotações do corpus foram analisadas levando em conta a ampliação ou construção de discordância onde não havia ou, por outro lado, o reforço de concordância já existente, com a análise de acordo com a da maioria. Como a anotação foi feita seguindo o modelo PropBank, que não admite classificação múltipla, as anotações sempre foram comparadas com os forms 1 e 3, de escolha única. Os resultados estão no quadro 8 abaixo, onde * sinaliza a construção de uma divergência em frases de aparente consenso (4x1) e ** sinaliza o reforço de um consenso em frases consideradas monossêmicas. Os resultados mostram tendências contrárias conforme o tipo de classificação que estava em jogo. No Form1, cujo foco estava nas classes LOCAL e TEMPO, na maioria dos casos a anotação indicada no corpus seguiu a classificação da minoria, o que resultou na acentuação de discordâncias (casos 3x2 passaram a 3x3) e no desaparecimento de concordâncias (4x1 passaram a 4x2). Já no Form 3, com outras classes em foco, os casos de ampliação ou criação de discordância foram minoria.

Efeito	Form 1	Form 3
Ampliação ou construção de discordância	s1 s3 s4* s6 s7* s10* s12 s13 (**61%**)	s5* s6* s9*, s11, s14 (**35%**)
Reforço de consenso	s5 s8 s9 s14 s15 (**38%**)	s1 s2** s3 s4 s7 s8 s12 s13 s15** (**64%**)

Quadro 8: Análise da anotação do PBP em contraste com demais anotações

A alta divergência nas respostas aliada ao fato de todos os respondentes serem anotadores_pesquisadores_linguistas experientes (e todos poderiam ser ou já foram adjudicadores de projetos de anotação) 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. De fato, questionamentos linguísticos acerca da possibilidade de uma única entidade preencher dois ou mais papéis simultaneamente não são novos. Um ponto central da teoria de Princípios e Parâmetros, por exemplo, é o *Critério Theta*, que estabelece que só pode haver correspondência um-para-um entre sintagmas nominais e papéis temáticos. Jackendoff (1972; 1990, apud Saeed, 1997), por outro lado, sugere que

uma mesma entidade pode preencher mais que um papel (Saeed, 1997:141). Levin & Rappaport Hovav (2005:42) reconhecem que "*Dual semantic role assignment is especially useful in distinguishing between certain verbs of transfer of possession*".

Do ponto de vista da tarefa, especificamente da avaliação, a opção pela classificação única também tem consequências. Por exemplo, se um sistema/modelo atribuir um ArgM de *finalidade* (ArgM-prp) a alguma das frases em que arbitrariamente (por default) foi atribuída a etiqueta de *causa* (ArgM-cau), ele será penalizado, mesmo que a leitura/análise não codificada no corpus seja legítima. Além disso, não é raro que diferentes leituras, não excludentes, estejam presentes no mesmo enunciado e levem a dúvidas legítimas na anotação. O sintagma "para trabalhar" na frase abaixo ilustra o ponto.

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

Podemos interpretar "para trabalhar" como a finalidade (ArgM-prp) de sair de casa: *Sair de casa para que? Para trabalhar*. Mas essa mesma ideia também pode ser considerada a "situação de destino", codificada como Arg3: *saiu pra onde? Para trabalhar* (ou: *Saiu pra onde? Para o trabalho*) e, nesse caso, estaríamos diante de um Arg3. A classe de anotação escolhida foi Arg3, mas a decisão não deixa de ser arbitrária, pois a *motivação* para a saída (ArgM-prp) está presente.

4.8. Disponibilização do PBP: diferentes versões

O material que disponibilizamos com anotação de papéis semânticos está no formato de dependências, e cada relação entre predicadores e argumentos está codificada no nível do token.

A relação estreita entre anotação sintática e anotação de papéis semânticos, por um lado, e as interferências da anotação sintática UD (cf. seção 3), por outro, levaram à criação de diferentes versões do corpus, caracterizadas por um maior ou menor alinhamento à sintaxe codificada no Porttinari-base ou aos princípios e etiquetas do PropBank. Especificamente, as versões diferem quanto ao tipo de anotação atribuída em verbos cópula e em verbos que, no Verbo-Brasil (e em consonância com gramáticas de língua portuguesa), são considerados auxiliares.

O material está disponibilizado em dois formatos – que chamamos provisoriamente de versão UD e versão clássica – detalhados a seguir[48].

4.8.1. PBP na versão UD

Esta versão se caracteriza pela atribuição de papéis semânticos apenas aos elementos considerados verbos plenos (upos = VERB) na camada de anotação UD. Em consequência:

- Não foram anotados os papéis semânticos de argumentos do verbo "ser".

[48] Agradeço ao Elvis de Souza a preparação dos arquivos para disponibilização.

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

4.8.2. PBP na versão clássica

Esta versão se caracteriza pela atribuição de papéis semânticos conforme o modelo PropBank, priorizando o conceito de *proposição*, independentemente do que foi considerado verbo pleno na camada de anotação UD. Em consequência:

- Foram anotados os papéis semânticos de argumentos do verbo "ser"
- Não foram anotados os papéis semânticos de argumentos de verbos que, apesar de considerados plenos pela anotação UD, são considerados auxiliares no Verbo-Brasil e nas diretivas Duran (2014) – especificamente, auxiliares de modo e aspecto.
- Foram anotados como auxiliares os verbos que, apesar de considerados plenos pela anotação UD, são considerados auxiliares no PBP (ArgM-mod, ArgM-asp), além daqueles considerados sempre auxiliares (ArgM-tml; ArgM-pas).

Para a frase

Mas os outros são piores

temos a seguinte representação em cada uma das versões:

- *versão UD:* Não há anotação de papéis argumentais
- *versão clássica:* "outros" é Arg1 do verbo "ser"; "piores" é Arg2 do verbo ser

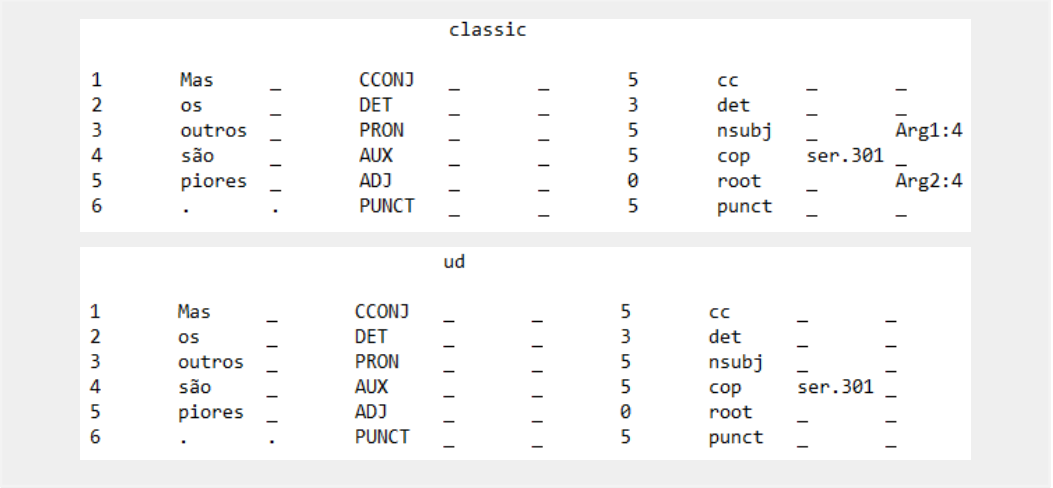
Para a frase

Algumas empresas, como a Vale, oferecem formulários que podem ser preenchidos…

temos a seguinte representação em cada uma das versões:

- *versão UD:* "que" (formulários) é Arg0_d de "poder" (leia-se: "formulários podem"); "preenchidos" é Arg1_d de "poder" (leia-se: "poder preencher")
- *versão clássica:* "que" (formulários) é Arg1 de "preencher" (leia-se: "preencher formulários"), e não há argumento associado a "poder", tomado como verbo auxiliar nesse contexto.

Cada uma das representações está indicada nas figuras 26 e 27 abaixo:



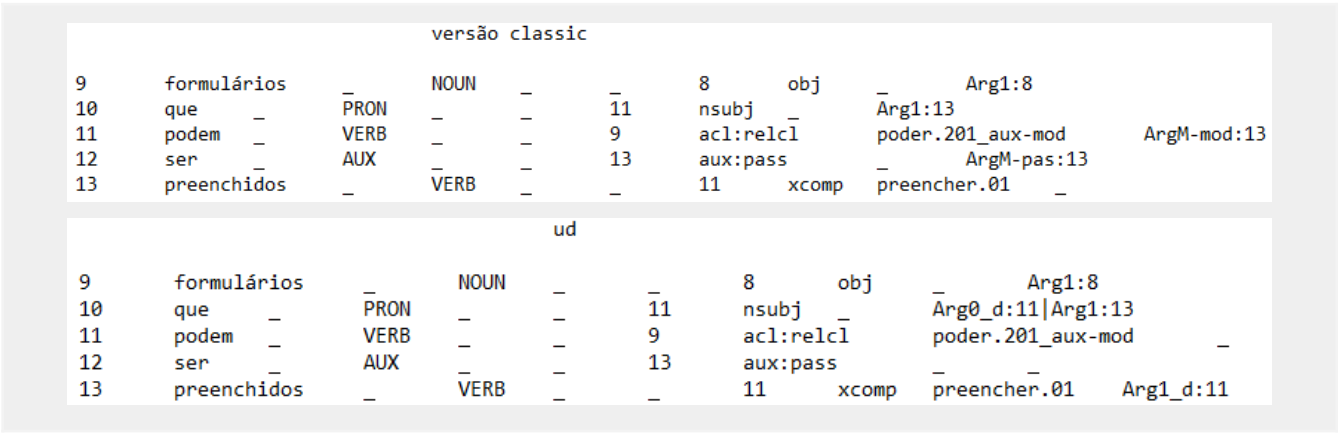
classic

1	Mas	_	CCONJ	_	_	5	cc	_	_
2	os	_	DET	_	_	3	det	_	_
3	outros	_	PRON	_	_	5	nsubj	_	Arg1:4
4	são	_	AUX	_	_	5	cop	ser.301	_
5	piores	_	ADJ	_	_	0	root	_	Arg2:4
6	.	.	PUNCT	_	_	5	punct	_	_

ud

1	Mas	_	CCONJ	_	_	5	cc	_	_
2	os	_	DET	_	_	3	det	_	_
3	outros	_	PRON	_	_	5	nsubj	_	_
4	são	_	AUX	_	_	5	cop	ser.301	_
5	piores	_	ADJ	_	_	0	root	_	_
6	.	.	PUNCT	_	_	5	punct	_	_

Figura 26: Diferença na anotação das versões clássica e ud no que se refere a frases com cópula



versão classic

9	formulários	_	NOUN	_	_	8	obj	_	Arg1:8
10	que	_	PRON	_	_	11	nsubj	_	Arg1:13
11	podem	_	VERB	_	_	9	acl:relcl	poder.201_aux-mod	ArgM-mod:13
12	ser	_	AUX	_	_	13	aux:pass	_	ArgM-pas:13
13	preenchidos	_	VERB	_	_	11	xcomp	preencher.01	_

ud

9	formulários	_	NOUN	_	_	8	obj	_	Arg1:8
10	que	_	PRON	_	_	11	nsubj	_	Arg0_d:11|Arg1:13
11	podem	_	VERB	_	_	9	acl:relcl	poder.201_aux-mod	_
12	ser	_	AUX	_	_	13	aux:pass	_	_
13	preenchidos	_	VERB	_	_	11	xcomp	preencher.01	Arg1_d:11

Figura 27: Diferença na anotação das versões clássica e ud no que se refere a frases com auxiliares

Consequentemente, a (i) **versão clássica contém muito mais instâncias do tipo Arg1** (para o sujeito) **e Arg2** (para o predicativo do sujeito) que a versão UD**,** dado que o verbo "ser" só admite esses dois tipos de argumentos; (ii) **a versão UD mais instâncias anotadas com Arg1 e Arg0** (mais precisamente, Arg1_d e Arg0_d, assim indicados para que possam ser diferenciados) **que a versão clássica**, dada a necessidade de atribuir papéis semânticos a predicadores verbais considerados auxiliares.

Cada uma das versões é gerada automaticamente a partir da versão base, que contém as etiquetas provisórias elencadas em 2.2.2.

SUBSTITUIÇÃO DE ETIQUETAS para o aprendizado de máquina

- As etiquetas ArgM:src, ArgM:conseq, ArgM:cond; ArgM:comp são especificações da etiqueta ArgM:adv e, porque são pouco frequentes no corpus, podem ser substituídas por ArgM:adv.

- Arg1_d e Arg0_d, presentes apenas na versão UD, podem ser substituídos por Arg1 e Arg0, respectivamente.

4.8.3. PBP versão completa e versão apenas com relações explícitas

Além das versões clássica e UD, disponibilizamos mais duas versões, pensando também na facilidade de consulta do material em pesquisas linguísticas por meio da ferramenta ET.

- *Versão completa*: contém todas as relações entre predicadores verbais e argumentos
- *Versão apenas com relações explícitas*: contém apenas as relações entre predicadores verbais e argumentos que estão explícitos na frase, e corresponde ao resultado da anotação da fase explícita. Apesar de não conter todas as relações entre papéis semânticos e predicadores, resolvemos disponibilizá-la por dois motivos: i) é uma versão mais amigável para buscas linguísticas na ET se o interesse está em pesquisas linguísticas leves; ii) permite facilmente decompor a tarefa de SRL em duas tarefas: anotação de SRL explícitos e anotação de SRL implícitos. Diferentemente da versão completa, esta versão não contém o id do head predicador na coluna 10, apenas o tipo de argumento. Esta versão não contém relações de predicação entre verbos de cópula, mesmo que todos os argumentos estejam explícitos na frase.

Ambas as versões estão disponíveis na versão clássica e na versão UD.

Dispondo desses materiais, criamos condições para que algumas questões sejam investigadas futuramente:

Considerando apenas a versão UD:

- As etiquetas Arg1_d e Arg0_d, utilizadas para Arg1 e Arg0 de verbos auxiliares, contribuem ou não para generalização e melhor desempenho de modelos de AM, em contraste com a utilização apenas de Arg1 e Arg0 em todos os casos?

Considerando a versão UD e a versão clássica:

- Qual versão leva ao melhor desempenho no AM?
- Qual versão vai melhor com modelos que não levam em conta a sintaxe?

Considerando a explicitação de argumentos:

- A incorporação de argumentos implícitos dificulta o aprendizado, apesar de adicionar mais dados?

Em todos os cenários, a análise qualitativa dos resultados poderá ser feita com a utilização do ambiente Julgamento da ferramenta UD, que passou por melhorias ao longo do projeto para que viabilizasse a comparação qualitativa e quantitativa entre dois datasets linguísticos com anotações diferentes.

Por fim, a criação de duas versões (UD e Clássica) permite comparar a anotação PBP com outros PropBanks que tenham seguido estritamente a anotação UD.

5. Resultados e análise

A Figura 28 lista as etiquetas utilizadas na anotação de cada versão, associadas à quantidade de ocorrências.

Versão clássica

Classe	Ocorrências	Explícitos	Implícitos
Arg0	8783	6105	2678
Arg1	16322	12351	3971
Arg2	6276	3409	2867
Arg3	370	367	3
Arg4	205	203	2
ArgM-tmp	2768	2629	139
ArgM-loc	1465	1423	42
ArgM-mnr	1428	1372	56
ArgM-adv	1250	1158	92
ArgM-neg	1109	1024	85
ArgM-pas	1007	1007	0
ArgM-dis	820	738	82
ArgM-tml	789	747	42
ArgM-mod	684	4	680
ArgM-prp	471	441	30
ArgM-cau	466	413	53
ArgM-nse	380	380	0
ArgM-ext	269	264	5
ArgM-asp	258	0	258
ArgM-src	213	170	43
ArgM-prd	179	169	10
ArgM-cond	163	128	35
ArgM-com	50	48	2
ArgM-comp	45	43	2
ArgM-conseq	37	33	4
ArgM-dir	6	6	0
Total	45813	34632	11181

Versão ud

Classe	Ocorrências	Explícitos	Implícitos
Arg0	8859	6122	2737
Arg0_d	697	653	44
Arg1	14001	12503	1498
Arg1_d	975	971	4
Arg2	3473	3416	57
Arg3	370	367	3
Arg4	205	203	2
ArgM-tmp	2735	2734	1
ArgM-loc	1451	1448	3
ArgM-mnr	1423	1416	7
ArgM-adv	1240	1239	1
ArgM-neg	1109	1109	0
ArgM-dis	819	819	0
ArgM-prp	468	467	1
ArgM-cau	460	460	0
ArgM-nse	380	380	0
ArgM-ext	269	268	1
ArgM-src	191	191	0
ArgM-prd	179	179	0
ArgM-cond	163	163	0
ArgM-com	50	49	1
ArgM-comp	43	43	0
ArgM-conseq	37	37	0
ArgM-dir	6	6	0
Total	39603	35243	4360

Figura 28: Distribuição da frequência dos argumentos em cada uma das versões do PBP

5.1. Distribuições

Na figura 29, temos os dados agrupados, apresentados maneira comparativa e, na figura 30, a distribuição das funções sintáticas por anotação de argumento, isto é, as funções sintáticas com mais anotações na camada PBP.

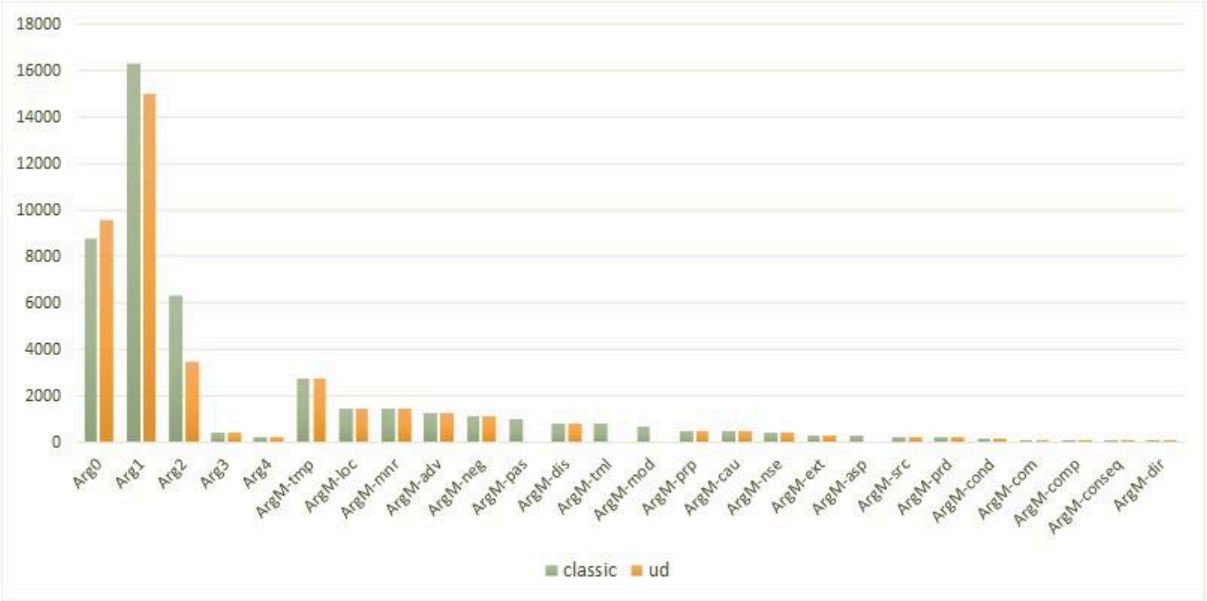


Figura 29: Distribuição dos argumentos, por frequência, em ambas as versões do corpus

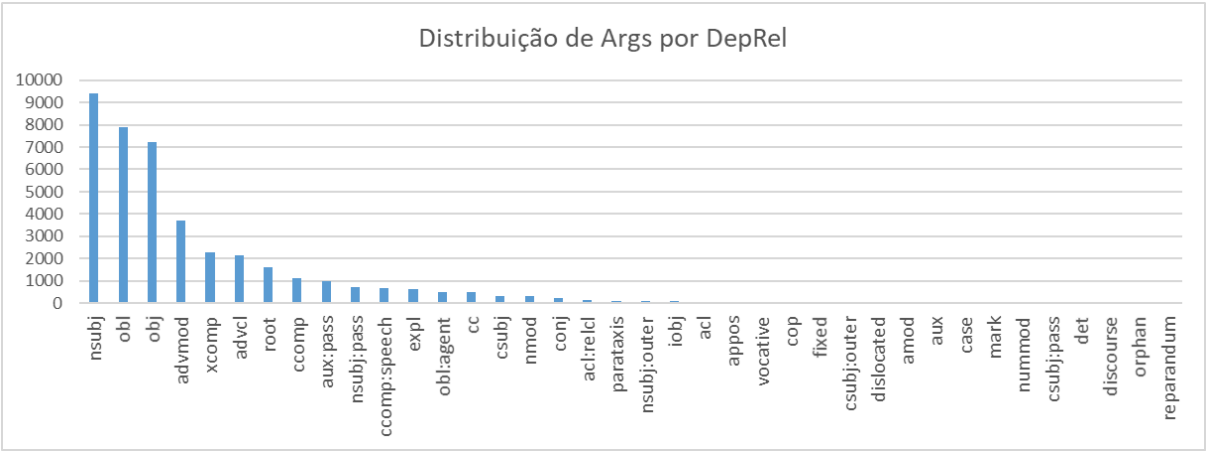


Figura 30: Distribuição das relações sintáticas (deprel) com anotação na camada PBP[49]

5.1.1. Distribuição de argumentos por tipo de relação sintática

Na figura 31, vemos a distribuição dos papéis semânticos tendo em vista a relação sintática mais frequente com que aparece associado, e na figura 32 vemos o mesmo tipo de cálculo feito para o PropBank original de Palmer et al.(2005). Em seguida, na figura 33, temos a mesma informação da figura 31, mas com mais detalhes numéricos, e na figura 11 o mesmo tipo de cálculo feito para o PropBank original de Palmer et al.(2005).

[49] Expressão de busca: token.misc == ".*Arg.*" , pedindo distribuição de deprel

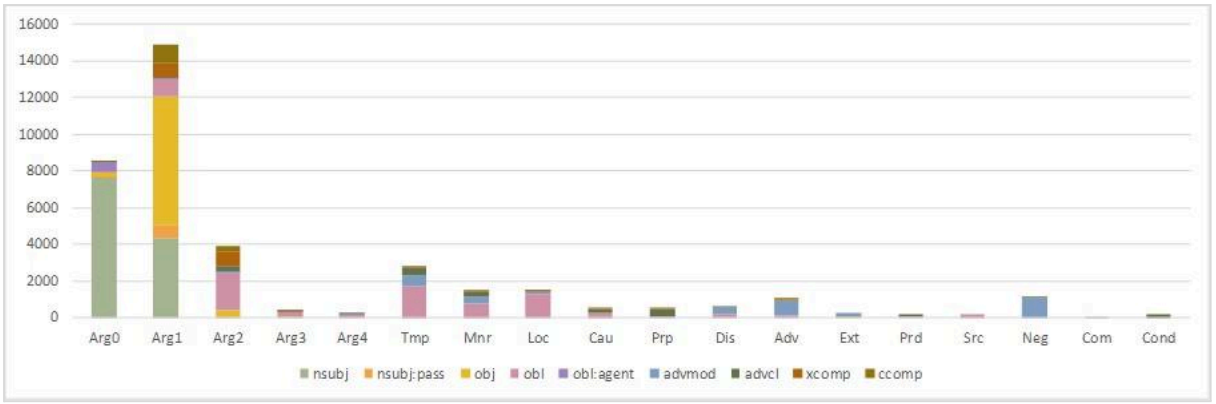


Figura 31: Distribuição de papéis semânticos por deprel

Table 7
Most frequent syntactic positions for each semantic role.

Roles	Total	Four most common syntactic positions (%)								Other positions (%)
Arg1	35,112	Obj	51.7	S	21.9	Subj	17.9	NP	5.2	3.4
Arg0	30,459	Subj	96.9	NP	2.4	S	0.2	Obj	0.2	0.2
Arg2	7,433	NP	35.7	Obj	28.6	Subj	12.1	S	10.2	13.4
TMP	6,846	ADVP	26.2	PP-in	16.2	Obj	14.6	SBAR	11.9	31.1
MOD	4,102	MD	98.9	ADVP	0.8	NN	0.1	RB	0.0	0.1
ADV	3,137	SBAR	30.6	S	27.4	ADVP	19.4	PP-in	3.1	19.5
LOC	2,469	PP-in	59.1	PP-on	10.0	PP-at	9.2	ADVP	6.4	15.4
MNR	2,429	ADVP	54.2	PP-by	9.6	PP-with	7.8	PP-in	5.9	22.5
Arg3	1,762	NP	56.7	Obj	9.7	Subj	8.9	ADJP	7.8	16.9
DIS	1,689	ADVP	69.3	CC	10.6	PP-in	6.2	PP-for	5.4	8.5

Figura 32: Distribuição de papéis semânticos por relação sintática no PropBank original. Fonte: Palmer et al.(2005).

A figura 33 traz os papéis semânticos mais frequentes para cada relação sintática (mesma informação da figura 31) considerando relações que poderiam receber papéis (isto é, não são dependentes de adj, por exemplo), e com mais algum detalhe[50]. Vemos que a generalização mais frequente é entre obj e Arg1, com 92,03%, seguida da generalização entre nsubj e Arg0, com, 63, 49%.

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

Figura 33: Distribuição de papéis semânticos por relação sintática

[50] Para saber a quantidade de *sujeitos* de que são *Arg0*, a expressão de busca é token.deprel == "nsubj" and token.misc == ".*Arg0.*" . Para a quantidade de *sujeitos* que são *Arg1*, a expressão de busca é token.deprel == "nsubj" and token.misc == ".*Arg1.*" ; Para a quantidade de objetos que são Arg0, a expressão de busca é token.deprel == "obj" and token.misc == ".*Arg0.*"

A figura 34 traz os mesmos números para o PropBank original (Palmer et al., 2005:90). No entanto, uma vez que cada gramática recorta e define os elementos que lhes parecem relevantes, é necessário algum cuidado antes de compararmos os números. O elemento sentencial/oracional S, presente no PropBank original (derivado da anotação do Penn Treebank), não existe em UD, e está distribuído entre alguns casos de xcomp e alguns casos de ccomp, por exemplo. Os elementos PP-in e *PP-at* precisam ser comparados com elementos obl introduzidos por certos tipos de preposição (e, em português, teríamos a preposição "em" na maioria dos casos, o que traz uma outra variável complicadora para a comparação).

Table 6
Most frequent semantic roles for each syntactic position.

Position	Total	Four most common roles (%)								Other roles (%)
Sub	37,364	Arg0	79.0	Arg1	16.8	Arg2	2.4	TMP	1.2	0.6
Obj	21,610	Arg1	84.0	Arg2	9.8	TMP	4.6	Arg3	0.8	0.8
S	10,110	Arg1	76.0	ADV	8.5	Arg2	7.5	PRP	2.4	5.5
NP	7,755	Arg2	34.3	Arg1	23.6	Arg4	18.9	Arg3	12.9	10.4
ADVP	5,920	TMP	30.3	MNR	22.2	DIS	19.8	ADV	10.3	17.4
MD	4,167	MOD	97.4	ArgM	2.3	Arg1	0.2	MNR	0.0	0.0
PP-in	3,134	LOC	46.6	TMP	35.3	MNR	4.6	DIS	3.4	10.1
SBAR	2,671	ADV	36.0	TMP	30.4	Arg1	16.8	PRP	7.6	9.2
RB	1,320	NEG	91.4	ArgM	3.3	DIS	1.6	DIR	1.4	2.3
PP-at	824	EXT	34.7	LOC	27.4	TMP	23.2	MNR	6.1	8.6

Fig 34: Distribuição de papéis semânticos por relação sintática no PropBank original. Fonte: Palmer et al.(2005)

5.1.2. Correlações entre sintaxe e anotação SRL: busca de regularidades nos dados anotados

Associando os dados das figuras 31 e 33, temos as seguintes informações:

A. **Arg1** se distribui principalmente entre as deprel *obj*, *nsubj*, *obl*, *xcomp*, *ccomp* e *nsubj:pass*
B. **Arg0** se concentra em *nsubj*, mas também ocorre em *obj* e *obl:agent*
C. **Arg2** se distribui principalmente entre obl, xcomp e obj
D. **ArgM-tmp** se distribui entre as deprel *obl*, *advmod* e *advcl*
E. **ArgM-mnr** se distribui entre as deprel *obl*, *advmod* e *advcl*
F. **ArgM-loc** se concentra em *obl*

Além disso, e tomando a perspectiva da análise sintática, que é o nosso ponto de partida:

- A deprel obj participa de Arg1 e tudo indica ser aquela que leva à generalização mais fácil: se é um obj, então é Arg1 em 92% das vezes
- A deprel xcomp participa igualmente de Arg1 e Arg2

- A deprel obl participa de Arg2, Arg1, tempo, modo, local
- A deprel advmod está bem distribuída entre neg, adv, tempo, dis e modo
- A deprel advcl participa de tempo, finalidade, modo e Arg2

A partir disso, e analisando mais atentamente os dados da figura 33 em busca de correlações entre a anotação sintática e a anotação SRL, partindo sempre da sintaxe e considerando apenas a versão clássica, alinhada à versão 1.0 do Porttinari-base, temos o seguinte cenário:

- A deprel **nsubj** está bem distribuída entre Arg1(35,79%) e Arg0(63,49%), sinalizando que, aparentemente, é difícil de generalizar. No entanto, se analisamos a distribuição dos heads de nsubj do tipo Arg1[51] (quadro 9, lado esquerdo), vemos que em cerca de 60% dos casos o head do nsubj Arg1 é um verbo pleno, e em cerca dos 35% restantes o nsubj Arg1 está em uma construção com verbo cópula (ser ou estar), com heads nominais. Os heads de nsubj do tipo Arg0[52] (quadro 9, lado direito), são verbos plenos em 98% dos casos. Ou seja, se levamos em conta apenas o fato de o nsubj aparecer com um head nominal já conseguimos alguma diferenciação entre Arg1 e Arg0. Para os demais casos em que o head é um verbo, e assumindo que a preferência por nsubj é ser Arg0, olhamos apenas para os verbos em que nsubj é Arg1[53] (tabela 1). Os lemas mais frequentes são aqueles que não têm Arg0 em sua estrutura argumental, como “ficar”, “acontecer”, “vir”, “ocorrer”, “existir”, “surgir”. Os demais podem ou não ter Arg0, conforme o sentido/frame. Um caso certamente difícil de generalizar é o do verbo “ter” com sentido de “existir” (ter.03). No entanto, e tomados como um todo, estes resultados sugerem que regras simples associadas a um léxico são capazes de um bom desempenho na atribuição de Arg1 e Arg0 associados a nsubj.

Distribuição do head do nsubj Arg1

UPOS	Total	(%)
VERB	2323	58,28
NOUN	798	20,02
ADJ	577	14,47
PRON	107	2,68
PROPN	64	1,61
ADV	50	1,25
NUM	43	1,08
SYM	18	0,45
AUX	3	0,08
X	3	0,08

Distribuição do head do nsubj Arg0

UPOS	Total	(%)
VERB	5976	98,37
ADJ	59	0,97
NOUN	27	0,44
ADV	5	0,08
PRON	4	0,07
PROPN	3	0,05
NUM	1	0,02

[51] A expressão de busca é token.deprel == "nsubj" and token.misc == ".*Arg1.*" and @token.head_token.upos == ".*"
[52] A expressão de busca é token.deprel == "nsubj" and token.misc == ".*Arg0.*" and @token.head_token.upos == ".*"
[53] A expressão de busca é token.deprel == "nsubj" and token.misc == "Arg1" and token.head_token.upos == ".*" and @token.head_token.upos == "VERB", usada apenas na versão base.

Quadro 9: Distribuição dos heads de nsubj Arg1 e Arg0 por classe gramatical

Distribuição dos 10 lemas mais frequentes de verbos com nsubj Arg1		
LEMA	**Total**	**(%)**
ficar	92	6,95
acontecer	71	5,37
vir	61	4,61
ocorrer	51	3,85
passar	45	3,40
ter	41	3,10
existir	37	2,80
começar	30	2,27
surgir	30	2,27
tornar	30	2,27

Tabela 1: Lemas mais frequentes em que nsubj é Arg1, considerando todo os verbos

- A correlação simples entre **obj** e Arg1, como mostra a Figura 33, leva a cerca de 92% de acertos. Se consideramos que o Arg1 é composto principalmente por nsubj e obj (e minoritariamente, xcomp) (figura 31), vemos que abordagem baseada em regras é capaz de ter um desempenho excelente nesses casos.
- A deprel xcomp está igualmente distribuída entre Arg1 e Arg2, e uma correlação simples entre **xcomp** e Arg2 leva a quase 46% de acertos (Figura 33). Tentamos buscar alguma outra pista linguística que desse mais confiança à distinção entre Arg1 e Arg2 nos casos de xcomp. A análise de upos mostra que, tanto nos casos de xcomp Arg1 quanto em xcomp Arg2, a deprel xcomp tem grandes chances de ser um verbo. No entanto, se xcomp for um elemento nominal (ADJ, NOUN, PRON ou PROPN), ele será Arg 2 em quase 60% dos casos. Aqui, também, uma abordagem baseada em regras simples pode ser promissora.

A deprel obl é aquela mais distribuída entre os papéis semânticos, comparecendo principalmente em Arg2 (31.43%), Tempo (26.70%) e Local (20.07%) e Arg1 (15.29%). Buscamos correlações entre o tipo de preposição e o papel semântico atribuído. A tabela 2 traz as preposições mais frequentes associadas a obl. Analisando individualmente as três preposições mais frequentes associadas a obl (“em”[54], “a”, “de”[55]) em busca de correlações entre a sintaxe e a anotação SRL, vemos a figura 34, analisada a seguir:

- *padrão obl + em:* formalmente, Tempo e Local são as classes mais difíceis de distinguir, e prevemos confusão na generalização (e na concordância humana). Se verificarmos os léxicos mais associados a cada um deles, temos uma situação um pouco mais definida: o léxico do *tempo* é ligeiramente mais marcado que o de *local*. Para *local*, o lema mais frequente, “Brasil”, responde por apenas 3% das ocorrências de ArgM-loc com “em”, sendo a maioria dos casos ocorrências únicas. Já para

[54] A expressão de busca é token.upos == "ADP" and token.lemma == "em" and @token.head_token.deprel == "obl" and token.head_token.head_token.upos == "VERB"
[55] A expressão de busca é token.upos == "ADP" and token.lemma == "de" and @token.head_token.deprel == "obl" and token.head_token.head_token.upos == "VERB"

tempo, o lema mais frequente, “ano”, responde por 8.7% das ocorrências, o segundo mais frequente, “dia”, responde por 6,5% e o terceiro mais frequente, “domingo”, por 3%, sinalizando que uma estratégia simples de anotação automática que combine léxico e regras pode ser um pouco mais bem sucedida em tempo que em local. A pouca especificidade de *local* pode estar relacionada, também, à explosão de “lugares virtuais”, caso mencionado nos arcos cruzados (quando, por exemplo, no trecho *"expressam as limitações que sentem em sua capacidade de ensinar o currículo*", a informação veiculada é *sentir* [*limitações na capacidade de ensinar]*, e não *sentir [limitações] [em algum lugar]*, que é a segmentação/anotação possível decorrente da análise sintática que evita o cruzamento de arcos.

- *padrão obl + “de”*: vemos uma preferência clara por Arg2, e um léxico dos *verbos* usados em cada papel (Arg2 e Arg1) já garantiria um acerto de aproximadamente 75% se utilizamos uma regra simples de anotação automática que combina léxico e regras.
- *padrão obl + “a”*: a mesma estratégia do padrão *“obl” + de* pode ser usada, levando, a princípio, aos mesmos resultados.

Preposições mais frequentes associadas a obl

Preposição	Total	(%)
em	3219	41,29
a	947	12,15
de	905	11,61
com	793	10,17
por	594	7,62
para	456	5,85
como	204	2,62

Tabela 2: Preposições mais frequentes em “obl”

Tipos de argumentos introduzidos pela preposição "em"

Argumento	Total	%
ArgM-loc	1168	36,28%
ArgM-tmp	990	30,75%
Arg2	440	13,66%
Arg1	222	6,89%
ArgM-mnr	152	4,72%
ArgM-dis	44	1,36%
ArgM-adv	37	1,14%
Arg3	30	0,93%
ArgM-prp	17	0,52%
ArgM-ext	16	0,49%
ArgM-cau	13	0,40%
Arg0	13	0,40%
Argm-src	11	0,34%
Arg4	10	0,31%
Argm-prd	8	0,24%

Tipos de argumentos introduzidos pela preposição "a"

Argumento	Total	(%)
Arg2	443	46,83
Arg1	145	15,33
ArgM-tmp	107	11,31
Arg4	80	8,46
ArgM-mnr	47	4,97
Arg3	23	2,43
ArgM-loc	23	2,43
ArgM-adv	17	1,80
ArgM-ext	6	0,63
ArgM-dis	5	0,53
ArgM-cau	3	0,32
Arg0	2	0,21

Tipos de argumentos introduzidos pela preposição "de"

Argumento	Total	(%)
Arg2	363	40,11
Arg1	230	25,41
ArgM-mnr	103	11,38
Arg3	64	7,07
ArgM-tmp	33	3,65
ArgM-src	27	2,98
ArgM-adv	13	1,44
ArgM-loc	10	1,10
ArgM-dis	9	0,99
ArgM-prd	7	0,77
Arg4	5	0,55
ArgM-ext	4	0,44
Arg0	3	0,33

Figura 34: Argumentos mais frequentes introduzidos por "em" "a" e "de"

Algumas perguntas e explorações possíveis

Já dispondo de um corpus anotado com papéis semânticos ao estilo PropBank, podemos explorar alguns pontos:

- Dada a regularidade sintática dos fenômenos anotados, podemos esperar um bom desempenho de um anotador simples baseado em regras?
- Qual a frequência de verbos com estrutura argumental "completa"? Quantos com apenas A0, A1, e A2?. Ou seja, podemos explorar a distribuição da frequência dos verbos por "completude" da sua estrutura argumental. De fato, quais verbos têm A0,

A1, A2, A3, por exemplo, já sabemos porque estão listados no Verbo-Brasil. Por isso o dado “novo” é a frequência com que a completude ocorre no corpus, uma vez que argumentos numerados são aqueles considerados essenciais ao sentido do verbo ou muito frequentes.

- De modo complementar, nos verbos com menos argumentos numerados, qual a distribuição dos argumentos modificadores? Existem verbos que sempre (ou quase sempre) aparecem com certos tipos de ArgM? A partir desses resultados, é possível falar em “exigência”?

- Sobre anotação SRL auxiliando a anotação sintática: todos os obl que são args numerados são obl:arg? A tabela 3 lista os 20 verbos mais frequentes cujo complemento obl é um argumento numerado.[56]

Lema / Freq		Lema/ Freq	
chegar	83	**falar**	49
fazer	74	**partir**	49
ir	73	**dizer**	48
ficar	67	**tratar**	44
entrar	66	**colocar**	43
dar	63	**vir**	37
levar	62	**voltar**	36
sair	56	**trabalhar**	31
ter	54	**participar**	30
passar	52	**ver**	**27**

Tabela 3: 20 lemas mais frequentes cujo obl é um arg numerado.

5.2. Cruzamento de arcos sintáticos e anotação SRL

Como indicado na seção 3.2, o princípio sintático de não cruzamento de arcos traz uma série de consequências para a representação do significado das frases. A seguir, uma lista não exaustiva de outras frases em que foi detectado o cruzamento de arcos sintáticos[57]. A figura 35 traz um trecho da frase 1 com a representação dos arcos sintáticos (que não se cruzam) e a figura 36 traz o mesmo trecho com a representação dos arcos semânticos (que se cruzam)

[56] A expressão usada em ambas as buscas foi token.deprel == "obl" and token.misc == ".*Arg.*" and token.misc != ".*ArgM.*" and @token.head_token.upos == "VERB"

[57] Lembramos que o cruzamento de arcos não está no corpus, já que a anotação do Porttinari-base não permite. As frases foram detectadas devido ao estranhamento que causam do ponto de vista do significado ao longo da anotação SRL.

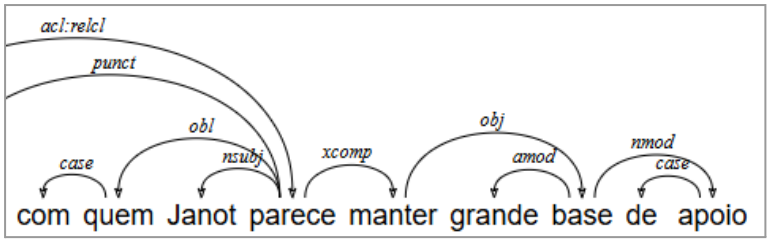


Figura 35: Trecho da análise sintática de “As críticas não ecoam entre os procuradores , com quem Janot parece manter grande base de apoio” , sem cruzamento de arcos

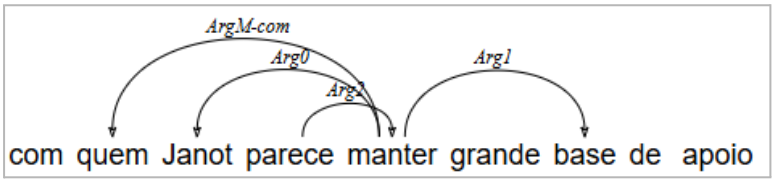


Figura 36: Trecho da análise semântica de “As críticas não ecoam entre os procuradores , com quem Janot parece manter grande base de apoio” , com cruzamento de arcos

1. As críticas não ecoam entre os procuradores , com **quem** Janot parece **manter** grande base de apoio .
2. Em os últimos 12 meses , a equipe francesa seguiu cartilha de tudo **o** que o romeno acredita **dever** ser evitado .
3. Que empresa não quer uma demanda que paga o valor **que** você quiser **cobrar** ?
4. Guimarães cobra a partir de R$ 1.000 para criar um rótulo , serviço **que** consegue **entregar** em cerca de 20 dias .
5. É um lugar livre para que eu diga tudo **o** que eu quiser **dizer** .
6. Não são poucos os críticos que apontam o programa como uma metáfora de o **que** os conservadores planejam **fazer** com os EUA .
7. Mentir é algo **que** ela diz não **ensinar** , até porque fazer isso deixa o corpo fraco , flácido .
8. Somos nós , e não os deuses , os agentes de as mudanças **que** queremos **ver** em o mundo.
9. A lista poderia ser montada para manter em o poder justamente os políticos **que** queremos **trocar** .
10. Nem sempre a reação de os mexicanos a os vizinhos é solidária , embora eles sejam os principais prejudicados com as restrições migratórias e o muro **que** Trump pretende **construir** em a fronteira .
11. É preciso algum choque para desencadear uma mudança de paradigma, algo que fará as pessoas se **darem conta** de que não podem lidar com esses problemas individualmente .
12. Muitas de as questões **que** nós fazemos hoje sobre **polícia** , autoridade , direito a o porte de armas e violência têm a ver com aquilo que escolhemos ser enquanto nação " , afirma .
13. Um a um , farão o **que** querem **fazer**.

5.3. Anotação de frames

A figura 37 apresenta a distribuição da anotação de frames no PBP[58]. Como podemos ver, cerca de 35% das ocorrências verbais estão sem frame (o número corresponde à soma de verbos sem anotação e de casos INC).

[58] Expressão de busca: token.upos == "VERB|AUX" , com Distribuição de deps.

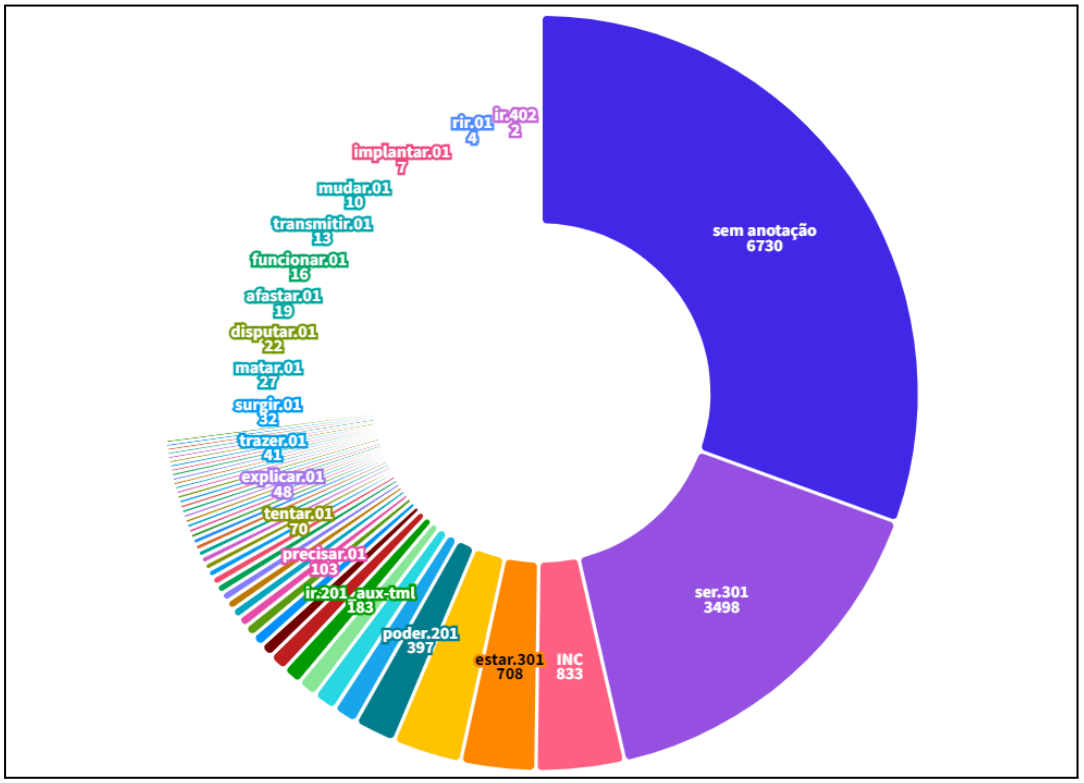


Figura 37: Distribuição das etiquetas de frames[59]

A análise dos dados do ponto de vista dos lemas, e não das ocorrências, informa que 1009 **(56%) formas verbais diferentes – todas polissêmicas – ainda precisam de anotação** quanto aos frames (tabela 4). A Figura 38 traz a distribuição dos lemas conforme seu status de anotação, levando em conta a frequência de cada lema.

Situação da anotação de frames	Quantidade de lemas
lemas diferentes **com** anotação de frame	780
lemas diferentes **sem** anotação de frame	1009
Total de lemas verbais diferentes	1789

Tabela 4: Distribuição da anotação de frames por lemas

[59] A figura pode ser explorada de maneira interativa em https://public.flourish.studio/visualisation/18907295/

Figura 38: Distribuição dos lemas por status de anotação[60]

6. Análise ampla: Propbanks e representações semânticas

Concluído o processo de construção do PBP, apresentamos nesta seção algumas considerações tendo em vista o objetivo mais amplo de criar representações semânticas estruturadas para o PLN.

6.1.Sobre FrameFiles e a criação de Frames

Indicamos no início deste relatório que o recurso léxico (*Frames*) é o coração do projeto PropBank. Uma outra maneira de dizer o mesmo é tomar o corpus PropBank como a instanciação do recurso léxico: tudo o que está na anotação do corpus é o resultado de decisões relativas à criação e ao design do recurso. Apesar disso, a construção de *Frames*, diferentemente da anotação dos papéis semânticos, não conta com diretivas detalhadas, o que pode sugerir que se trata de uma atividade com poucos desafios. Esta posição é indiretamente reforçada na apresentação do recurso, quando se afirma que a classificação dos argumentos numerados (que está especificada nos *Frames*) é "teoricamente neutra":

> *Although the semantic role labels are purposely chosen to be quite generic and theory neutral, Arg0, Arg1, etc., they are still intended to consistently annotate the same semantic role across syntactic variations.* (Bonial et al., 2018:738)

[60] A figura pode ser explorada de maneira interativa em https://public.flourish.studio/visualisation/18907923/

Nesta seção, tomamos como ponto de partida da análise o elemento "responsável" pelas decisões de anotação, os *Frames*. Especificamente, partimos das instruções relativas a como proceder caso a pessoa que anota se depare com um verbo para o qual não há um frame sintático/estrutura argumental prevista:

> *If a necessary Frame File is not found in the PropBank inventory, linguists or "framers" are given that lemma and all of the instances in which it occurs. They identify its senses or "Rolesets," assign roles for each Roleset, and create a Frame File listing all Rolesets* (Bonial et al., 2018: 739).

> "*If this is not the case* [se não há frame para o verbo], *the annotator will have to do some research to determine what an appropriate argument structure would be. Brainstorm other verbs that have* ***similar syntax and semantics****, and see if any of those verbs already have frame files. Try to model your annotation after that frame file. Again, fill out a problem report for this instance, taking note of the task and instance number along with an outline of the numbered arguments used and their semantic role correspondences. Also, include the roleset that served as a basis of comparison if one was used.(...)*
> *the annotators can consult the Unified Verb Index: https://verbs.colorado.edu/verb-index/. This has links to existing VerbNet (VN) and FrameNet (FN) information on a particular predicate, which can be used to understand what are commonly thought of as a verb's core arguments. Where possible, PropBank is mapped to VerbNet thematic roles. If a mapping is appropriate, a roleset's VN class will be listed at the top of the roleset on the frames website.*
> *VerbNet uses more canonical thematic roles instead of numbered arguments that are unique to the predicate. Thus, the annotator can use this as a resource for brainstorming argument structures for new verbs as well as coming to a better understanding of existing frame files by examining another analysis of that verb's thematic roles.*" (Bonial et al., 2015:8)

E, ainda, na documentação de 2015:

> *A polysemous verb may have more than one frameset when the differences in meaning are distinct enough to require a different set of roles, one for each frameset* (Bonial et al., 2015: 76).

Como vemos, a criação de frames parte de dois pressupostos:

a) **a identificação de diferentes sentidos de um verbo é uma tarefa razoavelmente objetiva**, uma vez que não necessita de diretivas da parte de quem cria os frames, bastando ter acesso aos diferentes contextos de ocorrência;
b) **a distinção entre argumentos *core* e *não-core*** (ou entre argumentos numerados e não numerados) **é igualmente uma tarefa objetiva**, e pode ser estabelecida sem muita discussão, cabendo aos "criadores dos frames sintáticos" – aos criadores dos frames – apenas codificar essa informação, distinguindo argumentos numerados,

que integrarão os frames, dos argumentos não numerados. Este pressuposto é assumido também pela VerbNet.

6.1.2. Desambiguação de sentidos (primeiro pressuposto)

Kilgarriff (1997), no contexto da tarefa de desambiguação automática de sentidos, aborda a desambiguização em uma perspectiva lexicográfica. O artigo é rico em exemplos e explicações de por que a visão que assume a "naturalidade" da identificação de diferentes sentidos de uma palavra está bem distante da realidade lexicográfica – e a criação de frames não deixa de ser uma atividade lexicográfica. (Vale notar que a tarefa de desambiguação de sentidos era considerada promissora no PLN da época, como atesta a criação, em 2000, da avaliação conjunta *Senseval* (*International Workshop on Evaluating Word Sense Disambiguation Systems*), precursora do *SemEval*[61], ativo até hoje, e da qual Adam Kilgarriff, ao lado de Martha Palmer, foi o primeiro organizador (Kilgarriff & Palmer, 2000)). Segundo Kilgarriff (1997), embora, para a comunidade PLN que trabalha com a desambiguação de sentidos, os significados das palavras sejam, de certo modo, o que o dicionário diz, teóricos oferecem diversos motivos para acreditar que não há tais conjuntos de significados. Significados só existem com relação a um objetivo, que pode ser o de escrever dicionários – ou, em nosso caso, criar *frames* para construir representações semânticas de proposições. Por isso, não há motivo para esperar que o mesmo conjunto de sentidos de uma palavra seja relevante para diferentes aplicações de PLN. *Corpora* diferentes e diferentes objetivos irão levar a diferentes significados (1997:92).

> "*Word senses are simply undefined unless there is some underlying rationale for clustering, some context which classifies some distinctions as worth making and others as not worth making*" (Kilgarriff 1997: 107)

Conforme Kilgarriff (1997), a divisão do significado das palavras em **significados discretos**, que não se sobrepõem, é forçada ao lexicógrafo pelo ambiente econômico e cultural no qual ele trabalha (a própria estrutura dos dicionários).

Uma visão do significado que também se opõe à ideia de significados como entidades discretas está em Manning (2022):

> "**Meaning is not all or nothing**; in many circumstances, we partially appreciate the meaning of a linguistic form. I suggest that meaning arises from understanding the network of connections between a linguistic form and other things, whether they be objects in the world or other linguistic forms. If we possess a dense network of connections, then we have a good sense of the meaning of the linguistic form."

Nesse sentido, significados também são instáveis, já que novas conexões sempre podem ser feitas, e conexões antigas podem perder a força (o que podemos ler como "ressignificar"). Quanto aos papéis semânticos, podem ser vistos como mais uma camada de conexão, viabilizando uma representação do significado mais complexa.

[61] https://semeval.github.io/

Kilgarriff (1997) salienta ainda a escassa literatura lexicográfica sobre a **utilização de critérios** na tarefa de divisão dos significados das palavras, o que não apenas dificulta o trabalho de lexicógrafos como contribui para a visão estabelecida no senso comum de que tais significados existem organizados de maneira discreta e independentemente de objetivos. O artigo segue apresentando críticas aos testes de detecção de ambiguidade frequentemente utilizados, que não são confiáveis e frequentemente levam a resultados ambíguos:

> *The tests are generally presented with the aid of an unproblematical example of ambiguity and an unproblematical example of vagueness. This is done in order to demonstrate what the test is and what the two contrasting outcomes are. However, this is not to use the tests in anger. What we want of a test is that it is consistent with our intuitions, where our intuitions are clear, and that it resolves the question, where our intuitions are unclear. (...) But do the tests help where intuitions are unclear? There is little if any evidence that they do. Here I discuss three classes of problems* (Kilgarriff 1997:97)

O autor sugere, então, abordar a questão por meio da lexicografia de corpus, sinalizando porém que não há motivo para esperarmos encontrar distinções claras entre os casos ambíguos e os casos vagos.

Na criação e anotação de frames, dada à natureza da tarefa, somos abrigadas lidar com o sentido como unidade discreta, mas para que os resultados (os próprios frames) tenham consistência, precisamos de critérios que nos permitam:

- **distinguir** um sentido/frame de outro sentido/frame, com segurança e de maneira motivada;
- **detectar** a insuficiência de um frame para um dado uso;
- **escolher** com segurança entre um frame e outro

Por exemplo, em *bater na porta*, temos o sentido de *golpear (bater.01*), de *chocar contra* (*bater.03*), ou precisaremos de um novo sentido, considerando a granularidade dos demais frames de *bater*?

1. Depois, eles **batem em a porta** de o Anselmo Lopes [procurador que atua em a Greenfield] para uma reunião com a PGR

Abaixo, considerando a frase com o verbo “plantar” e os frames disponíveis para anotação (figura 39), como saber se precisamos criar um novo frame? Como avaliar se *“differences in meaning are distinct enough”*, para retomar as instruções do PropBank? Novamente, podemos concordar com a solução proposta, mas precisamos de critérios que nos ajudem a decidir até que ponto uma similaridade se sustenta.

> 2. O partido **plantou** as bases da reforma no Senado (ou *The policeman **planted** a camera in the suspect's home*)

plantar.01 cultivar; introduzir semente ou muda de vegetal na terra **Arg0**: putter, agent **Arg1**: entity planted **Arg2**: location	**plant.01 , to put or place, sewing in the ground** **Arg0-PAG:** putter, agent (vnrole: 9.7-2-agent, 9.1-1-agent, 55.5-1-agent) **Arg1-PPT:** entity planted (vnrole: 9.7-2-theme, 9.1-1-theme, 55.5-1-theme) **Arg2-GOL:** location (vnrole: 9.7-2-destination, 9.1-1-destination)

Figura 39: Frames do Verbo-Brasil e dos Frames para *plantar.01*

Os exemplos 3 e 4 a seguir mostram outra situação: frames diferentes existem, mas a partir de qual critério escolher o frame adequado? Considerando o verbo "dividir" (frase 3): estamos diante do sentido de *partir* (dividir.01) ou do sentido de *compartilhar* (dividir.02) (figura 40)? Na atribuição de argumentos, a escolha é indiferente, uma vez que para ambos os sentidos a "coisa dividida" é Arg1.

3. Com a prática já estruturada, o próximo passo é **dividir** mais as responsabilidades.

dividir.01 , *repartir* **Arg0**: *repartidor* (vnrole: 23.1-agent, 45.4-agent) **Arg1**: *coisa sendo dividida* (vnrole: 23.1-patient1, 45.4-patient) **Arg2**: *partes em que arg1 está sendo dividido (dividir em, dividir por, dividir entre)* (vnrole: 23.1-patient2)	dividir.02 , *partilhar; compartilhar* **Arg0**: entidade ou coisa que compartilha (vnrole: -actor1) **Arg1**: coisa compartilhada (vnrole: -theme) **Arg2**: compartilhar com (vnrole: -actor2)

Figura 40: Frames do Verbo-Brasil para o verbo *dividir*

O mesmo vale para "resolver", nas frases 4a e 4b. Como decidir entre os sentidos de *solucionar* (resolver.01) ou de *decidir* (resolver.02) (figura 41)?

4a. Após a partida, Emery comentou sobre o ocorrido: vamos **resolver** no grupo.
4b. Outro efeito colateral de esse processo é a impaciência : os jovens estão com dificuldade de lidar com quaisquer aspectos da vida que não se **resolvam** em brevíssimo tempo .

resolver.01, decidir **Arg0**: *entidade que toma a decisão* **Arg1**: *decisão*	resolver.02, solucionar **Arg0**: pessoa que provê solução **Arg1:** problema **Arg2**: maneira ou instrumento usado para solucionar (resolver com)

Figura 41: Frames do Verbo-Brasil para o verbo *resolver*

6.1.3. Distinção entre argumentos core e não-core (segundo pressuposto)

Tanto PropBanks quanto VerbNets assumem uma distinção clara (categórica) entre verbos e seus argumentos (em oposição a verbos e seus modificadores/adjuntos), atribuindo papéis semânticos numerados (ou *papéis theta/papéis temáticos*) exclusivamente aos elementos do primeiro tipo.

Conforme Manning (2003), a distinção entre argumentos e adjuntos que se combinam com um núcleo é uma noção central de todas as teorias formais de gramática (incluindo a Regência e Ligação (Chomsky 1981), o Programa Minimalista (Chomsky 1995), a Gramática Léxico-Funcional (Bresnan 2001), Head-driven Phrase Structure Grammar (Pollard & Sag 1994), Categorial Grammar (Morrill 1994), Tree-Adjoining Grammar (Joshi & Schabes 1997), e outras abordagens).

No entanto, a **necessidade de distinção** entre esses dois tipos de complemento, dificilmente é colocada em xeque, apesar da ausência de consenso ou de critérios claros a respeito do que, exatamente, separa (i) um elemento que é exigência de um verbo (um argumento *core*) de (ii) um elemento que apenas adiciona uma informação não necessária a este verbo (argumento *não-core*). Os exemplos 1, 2 e 3, ilustram o ponto:

(1) Vive no Piauí
(2) Fósseis foram localizados no Piauí
(3) Plantaram no Piauí as bases do projeto

A indicação de *local*, quando estamos diante de um verbo como "viver", é considerada acessória, mas diante de um verbo como "localizar" ou "plantar", é considerada essencial.

A distinção entre obrigatoriedade e opcionalidade entre complementos e núcleos (sejam eles verbais ou nominais) está longe de ser um ponto pacífico nos estudos linguísticos. Vilela e Koch (2001), por exemplo, reconhecem que a distinção argumento/adjunto "tem merecido algumas reflexões e ainda não se chegou a uma conclusão definitiva" (p. 347). Trabalhos alinhados à linguística funcionalista defendem que a estrutura argumental está relacionada à frequência de ocorrência de um verbo em situações reais de uso linguístico, critério que, ao menos teoricamente, também é levado em conta no PropBank. No entanto, esta mesma abordagem que defende a frequência como critério aponta para a necessidade de um tratamento não categórico/não binário para a questão, em desacordo com as propostas de anotação de papel semântico.

Seguindo com Manning (2003), a distinção categórica entre argumentos e adjuntos "é o melhor que se pode fazer" no mundo binário das gramáticas formais tradicionais: as coisas precisam ser selecionadas (como argumentos) ou não. Se não forem, são livremente licenciadas como adjuntos, que em teoria deveriam poder aparecer com qualquer núcleo e ser repetidos qualquer número de vezes, ficando sujeitos apenas à compatibilidade semântica. Ainda conforme Manning, "certamente há alguma base para a distinção: uma rica literatura (revista em Radford 1988, Pollard e Sag 1987, Schütze 1995) demonstrou assimetrias entre argumentos e adjuntos em muitos domínios sintáticos, tais como a ordenação frasal e as possibilidades de extração." No entanto, continua Manning, modelos

categóricos de seleção sempre foram problemáticos. No contexto (e enunciados linguísticos sempre são proferidos em algum contexto), a língua é usada de forma mais flexível do que modelos categóricos sugerem. Mesmo Chomsky, nos seus primeiros trabalhos (1955:131), observa que "uma teoria linguística adequada terá de reconhecer graus de gramaticalidade". Muitas distinções de subcategorização apresentadas como categóricas na literatura linguística "são, na verdade, contra-exemplificadas em estudos de grandes corpora de uso da língua escrita."

Visando a um aprofundamento sobre a distinção entre argumentos e adjuntos verbais em língua portuguesa, retomamos e ampliamos a discussão apresentada em Souza e Freitas (2022) e no levantamento feito em Souza (2023, seção 5.2.1) no contexto da anotação UD entre argumentos preposicionados nucleares e oblíquos, obl:arg e obl, respectivamente.

Em um levantamento sobre a literatura relativa a argumentos e modificadores de verbos em gramáticas de língua portuguesa, Souza e Freitas (2022) mostram como esta é inconclusa: os critérios utilizados para distinguir argumentos e modificadores são ou insuficientes ou deixados de lado quando levam a resultados "inesperados". Entende-se que a distinção é adequada, os critérios de distinção é que falham**.**

Um critério de amplo uso – presente por exemplo em De Marneffe et al. (2021) para distinguir argumentos *core* de argumentos *não-core*, e presente também em Vilela e Koch (2001), dentre outras gramáticas – é o que se baseia na possibilidade de realizar certas operações sintáticas com elementos de um grupo, mas não com elementos de outro. Em De Marneffe et al. (2021), o critério é apresentado com a seguinte formulação (grifo nosso):

> *iii Valency-changing operations such as passive, causative, and applicative are **often** restricted to the promotion or demotion of core arguments.*

Por este critério, a frase 1 não pode se transformar na frase 2, e por isso "**três horas" não é um argumento core**/argumento numerado do verbo "durar" (os exemplos são de Vilela e Koch (2001)):

1. A aula durou <u>três horas</u>
2. * <u>Três horas</u> foram duradas pela aula

Outro critério usado para distinguir argumentos *core* e não-*core* é a conversão do argumento em oração adverbial ou coordenada por meio dos denominados pró-verbos "fazer", para os verbos de ação/atividade, e "acontecer/passar-se com", para os verbos de processo. Se a conversão for bem sucedida, como ocorre da frase 3 para as frases 4 (conversão em oração coordenada com "fazer") e 5 (conversão em oração adverbial), estamos diante de um argumento não-*core* (de um "adjunto"). Se a transformação não puder ser feita, como ocorre da frase 6 para a frase 7, estamos diante de um argumento *core*. Por este critério, portanto, **três horas" é um argumento core**/argumento numerado do verbo "durar", conclusão oposta àquela obtida com o critério anterior.

3. Ele estudou <u>três horas</u> para preparar a aula
4. Ele estudou para preparar as aulas e <u>fez isso durante três horas</u>
5. <u>Trabalhando arduamente três horas</u>, ele estudou para preparar a aula.

6. A aula durou <u>três horas</u>
7. *A aula durou e isto aconteceu durante três horas.

Segundo a análise de Vilela e Koch (2001), o fato de a frase 7 não ser uma boa frase da língua (uma frase "natural") mostra como o mesmo elemento ("três horas"), na mesma frase ("A aula durou três horas"), pode ser classificado como adjunto (critério da transformação em passiva) ou como argumento (critério da transformação em coordenada). Vilela e Koch (2001) não resolvem a questão, indicando apenas que *esses critérios não são os mais adequados para a frase*. Por outro lado, poderíamos questionar a partir de quais critérios a frase 7 não é uma boa frase da língua.

Voltando à afirmação de Vilela e Koch (2001) sobre a adequação/rejeição de critérios, enfatizamos que este posicionamento – que relativiza a adequação de critérios conforme os resultados deles derivados – não é exclusividade ou deslize dos autores, pelo contrário. É comum, na linguística que se baseia em testes de gramaticalidade e na intuição, responsável por avaliar a gramaticalidade das frases usadas nos testes, que se julgue a qualidade do critério pela capacidade que tem de produzir classificações de acordo com aquela que esperaríamos. E mesmo teorias linguísticas mais modernas não escapam desse hábito. No trecho citado de De Marneffe et al. (2021), o grifo em "often" sinaliza justamente o espaço para a indefinição deixado pelos critérios/testes: se desejamos, com os critérios, descobrir o que não sabemos – em nosso caso, diferenciar argumentos e adjuntos – como garantir que não estamos justamente diante da "exceção", do caso não coberto pelo "often"? Ou seja, **só sabemos avaliar se o critério se aplica se já sabemos de antemão se estamos diante de um argumento *core* ou não**. Porém, saber se estamos diante de um argumento core ou não é exatamente o que desejamos com a aplicação dos critérios (dos seis critérios elencados em De Marneffe et al. (2021) para distinção entre argumentos *core* e *não-core*, cinco fazem uso de "often" e um faz uso de "usually"). Este tipo de manobra retórica já foi notado e explicitado por Borges Neto (2013), quando faz uma crítica ao que chama de "uso protocolar" da gramática tradicional, que a leva a ser usada como "uma espécie de "descrição ateórica", suposta por todos e, simultaneamente, ignorada por todos enquanto teoria." (2013:5).

Ainda para Borges Neto (2011), esta dificuldade que modelos teóricos ditos "modernos" têm de se desvencilhar da tradição (e em nosso caso toda a discussão sobre argumentos *core* e *não core* não deixa de ser uma releitura da distinção entre elementos essenciais e acessórios das gramática tradicional, dos gregos e latinos, que não nos legou apenas uma terminologia técnica, mas uma teoria das línguas), tolhe "significativamente suas possibilidades de desenvolvimento" (2011:55).

Trazemos agora um exemplo do verbo "colocar", retirado da resenha de Souza (2023), e que ilustra não apenas a já referida dificuldade dos critérios linguísticos de distinguirem argumentos e adjuntos, mas também sinaliza como esta dimensão dos papéis semânticos esbarra no primeiro ponto abordado nesta seção: a necessidade de explicitar critérios relativos à separação de sentidos/frames nos Frame Files.

Elementos com valor locativo são tradicionalmente considerados adjuntos, uma vez que respondem a perguntas do tipo *Onde?* – e perguntas do tipo O*nde? Quando? Como? Por que? Para que?* são tradicionalmente associadas a advérbios, o que fica visível na taxonomia de argumentos modificadores (ArgM) proposta na anotação PropBank. No entanto, em certos verbos, como "colocar" ("Ela colocou o livro na estante"), o elemento do tipo "onde" é considerado argumento (*core*), e não um adjunto/modificador, diferentemente do que acontece em verbos como "descobrir" ("Ela descobriu o livro na estante"). Vilela e Koch justificam assim a classificação:

> "*Ora, a nossa intuição diz-nos que os 'locativos' não têm, relativamente aos predicados (colocar e descobrir), mesmo estatuto: em relação a 'colocar' constitui um complemento nuclear ou actancial, em relação a 'descobrir', é um circunstante.*" (VILELA; KOCH, 2001, p. 348)

Antes da análise, retomamos Borges Neto (2012): "Talvez analfabetos possam ter 'intuições' sobre a língua, linguistas relembram análises com que tiveram contato" (2012: 92).

Uma vez que o critério de perguntas (e seu valor semântico), assim como o critério das transformações sintáticas, é insuficiente para determinar a natureza da relação sintática entre o nome e o argumento, Vilela e Koch (2001) buscam "critérios suplementares" e retomam o critério de supressão: ao se suprimir um adjunto, a frase continuaria completa, diferentemente da supressão de um argumento. Assim, poderíamos dizer uma frase como 8, mas não uma frase como 9 (e por outro lado poderíamos questionar a partir de quais critérios a frase 8 é uma boa frase da língua):

8. Ela descobriu o livro.
9. *Ela colocou o livro.

Souza (2023) faz uma busca em corpus[62] procurando ocorrências do verbo "colocar", a fim de verificar empiricamente a exigência de um complemento locativo para "colocar". Uma das frases que não traz a informação locativa está abaixo:

10. Eles têm o monopólio do mercado e **colocam** o preço que quiserem.

Em 10, "colocar o preço" é parafraseável por "precificar", que não envolve local. A questão que se coloca, portanto, é se precisamos criar um novo frame. Em caso afirmativo, teríamos uma estrutura argumental para "colocar o preço nas alturas" e "colocar o preço no trabalho", e outra estrutura para "colocar o livro na estante", com consequências para as representações semânticas e para o aprendizado automático de informações semânticas.

Voltando ao trabalho de Manning (2003), existem elementos que claramente são argumentos (como sujeitos e objetos) e elementos que claramente são adjuntos (como tempo e localizações "exteriores"), mas há muita coisa entre uma classe e outra, e são esses elementos situados no meio do caminho que "**are often classified back and forth as arguments or adjuncts depending on the theoretical needs and convenience of the author**

[62] A busca foi feita sobre o material do projeto AC/DC (Santos, 2011).

(Babby 1980, Fowler 1987, Maling 1989, Maling 1993, Li 1990, Wechsler and Lee 1996, Przepiórkowski 1999a, Przepiórkowski 1999b)" (2003:302, grifo nosso).

A crítica de Manning (2003) sobre a (necessidade de) distinção categórica entre argumentos e adjuntos converge com a crítica de Borges Neto. Para Manning (2003:302), "*Looking more broadly, careful consideration of the assumed categorical distinction between arguments and adjuncts further argues for new foundations for syntax*."

Não está no escopo deste relatório defender uma abordagem probabilística para a distinção de argumentos e adjuntos como faz Manning (2003), mas vale lembrar que, quando os criadores do PropBank assumem a frequência como critério (adicional?) de distinção[63], parecem não se dar conta de que a frequência também não é um critério "naturalmente" categórico. A frequência de diferentes padrões de subcategorização varia bastante entre gêneros textuais (Roland e Jurafsky (1998) *apud* Manning (2003:10)). Ou seja, considerar seriamente esse critério tem como consequência a criação de *frames*/estruturas de subcategorização instáveis e dependentes de corpus, o que vai contra os interesses e princípios defendidos pelo PropBank.

Uma proposta alternativa seria evitar a distinção e discussão relativa a argumentos *core* e não-*core*, como foi feito na anotação sintática do corpus PetroGold (de Souza e Freitas 2023) e motivados pelo mesmo argumento: a discussão e distinção têm muito mais interesse e relevância para teorias linguísticas (uma vez que envolvem indicar os elementos que fariam parte das entradas do léxico mental, e que levariam à construção de frases bem-formadas, evitando a geração de frases agramaticais, no caso de argumentos *core*) que para o PLN.

Teorias linguísticas categóricas, de acordo com Manning (2003), e concordamos, se dizem mais poderosas do que são. Por um lado, se apresentam como capazes de distinguir uma categoria da outra estabelecendo limites rígidos sobre estruturas que, na realidade, quando confrontadas com as questões da convencionalidade e da criatividade humana, vemos que não têm fronteiras claras.

Por fim, a escolha da gramática de Vilela e Koch para exemplificar a distinção entre argumentos e adjuntos não é aleatória. O linguista Mario Vilela é um dos gramáticos de língua portuguesa mais associados à Teoria de Valências ou Gramática de Valências, tendo produzido duas obras sobre o tema: "Gramática de Valências" (Winfried Busse, Mário Vilela, Coimbra: Almedina, 1986) e "Gramática de Valências: Teoria e Aplicação" (Coimbra: Almedina, 1992). Ou seja, dentre as muitas abordagens que fenômenos linguísticos podem receber, estamos diante daquela que provavelmente dialoga de maneira mais direta com os pressupostos linguísticos da anotação de papéis semânticos.

[63] "Numbered arguments reflect either the arguments that are required for the valency of a predicate (e.g., agent, patient, benefactive), or if not required, those that occur with high-frequency in actual usage."

6.2. Sobre a capacidade de generalização da taxonomia PropBank

Os resultados de estudos sobre generalização das representações linguísticas ao estilo PropBank apresentados na seção 1.4 ensejam alguns questionamentos. Antes, porém, retomamos estes resultados em diálogo com o artigo de Bonial et al.(2018), que traz direcionamentos recentes do PropBank.

Bonial et al. (2018:743) relatam que "While automatic SRL performance is quite good for the detection of Arg0 and Arg1, performance on identification of higher-numbered arguments, 2–6, is relatively poor", em conformidade com os resultados obtidos por Yi et al. (2007) e por Merlo & Van der Plas (2009). Este bom desempenho quanto aos argumentos Arg0 e Arg1, no entanto, parece estar mais relacionado às regularidades entre posição sintática e tipo de argumento, como vimos na seção 5.1.3 e também em Merlo & Van der Plas (2009). Levando em conta os resultados de Yi et al.(2007), a equipe do PropBank reconhece que um alinhamento entre argumentos numerados e classes semânticas da VerbNet pode ajudar na generalização, sobretudo quando se trata de novos textos/gêneros/verbos:

> "Nonetheless, these issues can be addressed by converting higher-numbered arguments into semantic role labels that are generalizable across lemmas." (Bonial et al. 2018: 743)

> "The utility of VerbNet mapping for improving SRL on new genres has been demonstrated by Yi, Loper, and Palmer [2007]**,** who focused on Arg2. By subdividing the Arg2 instances into coherent subgroups based on the VerbNet labels, using them for training, and then mapping back to Arg2 for testing, the F-score for Arg2 increased 6 points for Wall Street Journal test data, and 10 points for Brown Corpus test data. As mentioned previously, the reason that such labels were originally avoided in the PropBank corpus was that there is still considerable debate as to what the **ideal** set of thematic roles should be." (Bonial et al., 2018:744, grifo meu)

Os resultados de Merlo & Van der Plas (2009) associados à nossa análise qualitativa (5.1.2. e 5.1.3.) também sugerem a dificuldade de generalização no AM para os papéis menos frequentes – a partir de Arg2 – , o que pode ficar invisibilizado quando consideramos uma medida global de desempenho no AM, sem analisar o aprendizado individual de cada classe. Como sinalizam Merlo & Van der Plas (2009), a especificidade semântica dos papéis Arg3-Arg5 deixa os dados fragmentados, criando esparsidade e dificultando aprender a partir de exemplos (o que fica mais aparente na anotação dos papéis menos frequentes). Valores estado-da-arte para SRL ao estilo PropBank, como vimos em 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 simples, baseado em regras, consegue 83% – apenas 2% abaixo de valores estado-da-arte, considerando os cenários mais difíceis –, o que dá mais força ao argumento de que as generalizações facilitadas pela anotação estilo PropBank decorrem da sua forte correlação com a sintaxe, sendo capturadas por meio de regras. Se, como sugerem os diversos trabalhos mencionados aqui, não é possível generalizações consistentes a partir da taxonomia PropBank, uma pergunta

legítima é por que insistir nessa taxonomia, já que, ao menos aparentemente, a diferença entre um baseline simples e os melhores resultados estado-da-arte não estão muito distantes – ou seja, aparentemente, poucas coisas são realmente "aprendidas", e o objetivo primeiro do PropBank é servir como material de treino.

Voltando ao trecho citado de Bonial et al. (2018), os autores afirmam que a razão pela qual a anotação PropBank tem sido mantida (mais precisamente, afirmam que a razão pela qual as classes da VerbNet têm sido evitadas) é a existência de um debate acerca de qual seria o conjunto ideal de papéis. Ideal para quê, ideal em qual sentido, perguntamos. Se consideramos seriamente que o objetivo primeiro de um PropBank é servir como material de treino para o AM, a resposta sobre o conjunto ideal de classes está dada nos estudos elencados na seção 1.4, e os papéis da VerbNet – dentre aqueles que foram estudados – , ao que tudo indica, são mais adequados à generalização.

Por outro lado, como tentativa de melhorar o desempenho de classificadores de SRL, Bonial et al.(2018) elencam quatro estratégias que vêm sendo adotadas no PropBank:

> (i) desenvolvimento e melhorias no já referido SemLink;
> (ii) desenvolvimento de etiquetas semânticas para os papéis numerados, o que chamam de "function tags";
> (iii) unificação de verbos, substantivos e adjetivos relacionados morfologicamente (relacionando "destroy" e "destruction", por exemplo)[64].

Além disso, ainda que de forma não tão diretamente vinculada à performance do aprendizado de SRL,

> (iv) a ampliação dos elementos predicadores, incluindo nomes e adjetivos, e de predicações que envolvem construções com verbos leves/verbos suporte e expressões multivocabulares como *jump the shark*, *have in mind* e sintagmas preposicionais como *in love* (Bonial et al., 2018 e Pradhan et al., 2022).

Na busca pela generalização, e levando em conta nossos resultados parciais e os resultados da literatura sobre generalização de papéis semânticos, sugerimos outro caminho, que será apresentado na próxima seção. Antes, retomamos o argumento da taxonomia VerbNet como facilitadora da generalização (em oposição à taxonomia PropBank), que apesar de compartilhado por todos que investigaram a questão (abaixo, um trecho de Loper, Yi e Palmer (2007)) ainda não parece forte o suficiente.

> "By exploiting the mapping between PropBank and VerbNet, we can transform the data to make it more consistent, and to expand the size and variety of the training data. (...) Unlike the PropBank labels, the VerbNet labels are defined consistently

[64] "These mappings have enabled PropBank to now combine the Frame Files of etymologically related lemmas, unifying any conceptually similar Rolesets, allowing for greater generalization of their semantic behavior."(Bonial et al., 2018:742)

> across verbs; and therefore it should be easier for statistical SRL systems to model them. " (Loper, Yi e Palmer, 2007)

Apesar dos resultados de estudos sobre generalização convergirem para a adequação das classes VerbNet, as classes PropBank continuam sendo apresentadas como as mais adequadas pela equipe responsável pelo PropBank em seus trabalhos mais recentes, como Pradhan et al.(2022) – e ainda que, de maneira contraditória, Gunger & Palmer (2021) já tenham reconhecido os limites dessas mesmas classes no que se refere aos predicadores de baixa frequência ou de fora do domínio:

> The decision as to what constitutes a semantic role and the use of Penn Treebank as the syntactic scaffolding for the annotation contributed to high inter-annotator agreement, which led to higher performing machine learning models and fueled interest in the task. PropBank."(Pradhan et al.,2022, grifo meu)

> "Recent approaches to SRL have achieved large gains in performance through the use of pretrained text encoders like ELMo and BERT (Peters et al., 2018; Devlin et al., 2019). Despite these advances, performance on low-frequency predicates and out-of-domain data remains low relative to in-domain performance on higher frequency predicates." (Gunger & Palmer, 2021:51, grifo meu)

Os exemplos abaixo ajudam a entender por que é difícil generalizar (para argumentos diferentes de Arg0 e Arg1) com a anotação PropBank. Em todas as frases temos um elemento com a mesma ideia de *local* associado ao verbo, mas cada caso recebe um papel:

Plantou bananas em Cuba (Arg2)
Desapareceu em Cuba (Arg2)
Morou em Cuba (Arg1)
Viveu em Cuba (ArgM-loc)
Apareceu em Cuba (ArgM-loc)

Para enriquecer a análise, alguns dados do Porttinari-base:

- Das 4 ocorrências de "plantar", apenas 1 menciona local (Arg2).
- Das 12 ocorrências de "desaparecer", 1 menciona local (Arg2)
- Das 42 ocorrências de "viver"[65], 22 mencionam local (ArgM-loc).
- Das 20 ocorrências de "aparecer", 12 mencionam local (ArgM-loc)

Mais alguns dados preliminares do Verbo-Brasil e do Porttinari quanto à ideia de **local**:

- 8 frames/verbos local é Arg1;
- 26 frames/verbos local é Arg2;
- 4 frames/verbos local é Arg3,
- 1 frame/verbo local é Arg4,
- em 450 lemas distintos do Portinari local é ArgM-loc.

[65]A contagem já desconsidera as ocorrências de "viver" como auxiliar (viver.201) e de viver.02 , interpretar (na dramaturgia))

Associado ao ponto anterior está a assistematicidade semântica atribuída aos argumentos numerados. Em outras palavras, nem tudo o que é local (ou é, majoritariamente, local) é Arg2, como pode ter dado a indicar os exemplos de "plantar" e "desaparecer". Um levantamento rápido nos mostra, por exemplo, que

- *Local* pode ser Arg1 (comparecer...), Arg2 (plantar, instalar, acessar...), Arg3 (apontar...) e **ArgM-loc**
- *Beneficiário* pode ser Arg2 (facilitar; abastecer...); Arg3 (abrir...), Arg4 (arranjar...), além de **ArgM-adv**, por falta de opção melhor.
- *Causa* pode ser Arg2 (culpar, elogiar, cansar...) e Arg5 (ampliar-se...), além de **ArgM-cau**
- Modo/maneira pode ser Arg1 (andar), Arg2 (diminuir, tratar, atrapalhar, ampliar..) e **ArgM-mnr**

Nesse contexto, do ponto de vista da representação semântica – interesse da anotação SRL – nos perguntamos qual o ganho em basear a tipologia de anotação no argumento "exigência do verbo".

Outra consequência da anotação baseada em frames sintáticos que pode contribuir para dificultar a generalização (sobretudo em modelos que incorporam informação semântica distribuída) é que verbos com sentidos similares podem ter frames de subcategorização distintos, como os pares *avisar-alertar*, e *causar-provocar*, nas frases abaixo

1. a CPPU enviara uma **notificação** à prefeitura (...) [**alertando|avisando**] que os **infratores estavam sujeitos a as penas previstas**
2. (...) as concessões feitas a grupos que [**causaram|provocaram**] tanto **sofrimento** ao **país**.

Para a frase 1., se o verbo for "alertar" (em inglês, *alert.01*), o "aviso" é codificado como Arg2; se o verbo for "avisar" (em inglês, *tell.01*), o "aviso" é codificado como Arg1.

Para a frase 2., se o verbo for "provocar", o sintagma "o país" recebe a etiqueta Arg2 que, no caso de "provocar", sinaliza a presença de um "beneficiário". Mas se a frase tiver sido construída com o verbo "causar", não temos o que fazer com "o país", pois o verbo "causar" não prevê um espaço argumental para elementos do tipo "beneficiário". Em consequência, somos obrigadas a escolher alguma das etiquetas ArgM, nenhuma delas adequada (na anotação, foi escolhida ArgM-adv).

A mesma "inconsistência" na anotação acontece quando **dois elementos/sintagmas têm o mesmo conteúdo semântico previsto em um argumento *numerado*.** Neste caso, o primeiro deles recebe o argumento numerado e o segundo recebe uma etiqueta ArgM. No exemplo abaixo (1), considerando o frame do verbo *sair.01 (deixar um lugar ou uma situação*,) apresentado no quadro (10)), "*em ao menos dois carros*" foi anotado como Arg4 e "atirando" foi anotado como ArgM-mnr. No entanto, em ambos os casos temos a maneira como os criminosos saíram, ou seja, o Arg4 poderia ser atribuído a qualquer um dos elementos, e se a ordem dos argumentos na frase fosse outra, a atribuição de argumentos seria outra (o que

é contraditório com princípios da anotação de papéis semânticos, cuja ideia principal é fornecer uma mesma representação semântica ainda que a estrutura sintática seja diferente.).

1. Os criminosos, usando touca ninja, **saíram** em ao menos dois carros por ruas de essas cidades atirando contra alvos escolhidos por eles .

Roleset id: **sair.01**, deixar um lugar ou uma situação, vncls: 51.1-1 51.2, Mapeamento para o inglês: **leave.01, emerge.01, pull.02, go.01** **Arg1**: *entidade saindo* (vnrole: 51.2-theme, 51.1-1-theme) **Arg2**: *lugar ou situação de origem (sair de)* (vnrole: 51.2-source, 51.1-1-location) **Arg3**: *lugar ou situação de destino (sair para)* (vnrole: -destination) **Arg4**: *forma ou estado em que arg1 sai* (vnrole: -attribute)

Quadro 10: Frame de sair.01 conforme Verbo-Brasil

A frase (2) abaixo é outro exemplo da mesma questão. Tanto “53 pontos” como “liderança do torneio” indicam *“resultado, estado ou posição em que Arg1 chega ao final* (vnrole: -attribute)”, descrição de Arg2 do frame de “terminar” (quadro 11). No entanto, apenas um deles pode receber a etiqueta Arg2, e então “na liderança do torneio”, apesar de indicar “posição em que Arg1 chega ao final”, é anotado como ArgM-mnr. Novamente, se a ordem da frase fosse diferente (“termina a 24ª rodada de o Brasileiro em a **liderança** de o torneio, com 53 **pontos**”), a anotação seria diferente, ferindo princípios de atribuição de papéis semânticos.

2. Com o resultado , o Corinthians **termina** a 24ª rodada de o Brasileiro com 53 pontos , em a liderança do torneio .

Roleset id: **terminar.03**, chegar ao final; resultar, vncls: , Mapeamento para o ingl?s: **Arg0:** entidade que chega ao final do evento (vnrole: -theme) **Arg1:** evento que termina (pode estar elíptico) (vnrole: -theme) **Arg2:** resultado, estado ou posição em que arg1 chega ao final (vnrole: -attribute)

Quadro 11: Frame de terminar.03 conforme Verbo-Brasil

Por fim, outra consequência (indesejável para a generalização) da atribuição de papéis semânticos ser baseada na distinção *core* e *não-core* é o **inchaço desnecessário de certas categorias de ArgM**, especialmente *ArgM-adv* e *ArgM-mnr*. Essas classes acabam admitindo sentidos/instâncias diferentes daquelas para as quais foram criadas, e tornam a generalização ainda mais difícil, justamente porque não há o que ser generalizado para além da forma. Este ponto já foi ilustrado a frase 2 (“causar sofrimento ao país”, no qual “ao país” recebe ArgM-adv, diferentemente de “provocar sofrimento ao país”, no qual “ao país” recebe Arg2).

6.2.1. Em defesa dos ArgM

Na busca pela generalização e diminuição de dados escassos, dois caminhos têm sido explorados até aqui: utilização das classes VerbNet e unificação baseada na morfologia/etimologia dos elementos predicadores (verbos, substantivos e adjetivos). Nesta seção, sugerimos[66] um caminho alternativo, não necessariamente incompatível com os anteriores, que chamamos provisoriamente de *generalização baseada na semântica*.

Tanto PropBanks quanto VerbNets e FrameNets partem de propostas vindas de teorias linguísticas para aplicação no PLN – isto é, trata-se de implementações mais ou menos diretas de teorias linguísticas –, o que faz com que certas decisões possam ser questionadas do ponto de vista de relevância que têm para tarefas do PLN. Uma delas é a distinção entre complementos argumentos (*core*) e complementos modificadores/adjuntos (*não-core*) dos verbos, que reproduz a distinção entre argumentos e adjuntos/modificadores (seção 6.1.3). Se levarmos às últimas consequências o objetivo SRL de encontrar *quem o que onde quando como* etc, esta distinção perde a relevância, uma vez que ideias como *quando*, *onde*, *por quê*, *para quê*, *como* etc podem ser expressas tanto por complementos considerados *core* como por complementos considerados *não core* (6.2). Se, além disso, nos damos conta de que os limites da distinção entre complementos *core* e *não core* são reconhecidamente um ponto de debate nos estudos linguísticos (seção 6.1.3), nos perguntamos por que não abrir mão dessa distinção, o que permitiria uma mesma classificação para todas as ocorrências sublinhadas nas frases abaixo. Com isso, transformamos 3 instâncias singulares (dados esparsos) em 3 ocorrências de uma mesma classe (local?), criando condições mais propícias à generalização.

Ela vive em Cuba → ArgM-loc
Ela mora em Cuba → Arg1
Ela desapareceu em Cuba → Arg2

Anotação é um processo humano de construção (artificial) de igualdades entre diferentes elementos, que pode servir tanto para que seja possível (i) contar (e investigar) fenômenos que nos parecem ser de um mesmo tipo (ou nos interessa que sejam considerados assim) quanto para (ii) construir ativamente, isto é, intencionalmente, igualdades a fim de facilitar a generalização (ponto de partida do AM), mas sem perder de vista o nível de informatividade que desejamos ter codificado (que seja aprendido).

Esta visão contrasta com a visão de Levin e Rappaport Hovav (2005: 36, grifos meus):

> *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. Semantic roles, then, can be viewed as labels for equivalence classes of arguments, and the goal of a theory of semantic roles is the identification of a set of semantic roles that is applicable to any argument of any verb.*
> *(...) Semantic roles distill from the perhaps even wider range of semantic detail those **facets of meaning which are grammatically relevant**. **The choice of** features or **roles is justified to the extent that they define equivalence classes which recur in rules of***

[66] Trata-se apenas de uma sugestão, e nada do que está indicado foi implementado.

> ***phonology or morphosyntax***. *Phonological features, however, generally appear in clusters defining particular phonemes, and, in addition, define natural classes of phonemes, while semantic roles, as they were initially used, are taken to be atomic elements. As we discuss in section 2.3.1, there have been moves to define semantic roles in terms of more basic semantic elements, which could be viewed as the real analogues to phonological features.*

O trecho sinaliza que (i) o compromisso dos papéis semânticos é com classes de argumentos "naturais", exemplificando o processo de considerar parte da *língua* noções que são parte de uma *teoria da língua*. "O modo como identificamos e classificamos os fenômenos linguísticos depende crucialmente do enfoque teórico adotado. Os fatos não dizem como querem ser abordados. As necessidades da teoria é que nos levam a uma ou a outra classificação" (Borges Neto, 2013). O trecho também explicita que (ii) na descrição dos papéis semânticos, interessam apenas os aspectos do significado que são gramaticalmente relevantes (e podemos entender gramática, aqui, como morfossintaxe), o que claramente está alinhado aos interesses de algumas teorias linguísticas, mas não necessariamente se alinha aos interesses do PLN quando decide usar papéis semânticos como uma maneira de representar o significado das frases. Por fim, o trecho indica que (iii) a escolha dos papéis é motivada pela existência de regras análogas em outros níveis de representação linguística, o que, novamente, pode não se alinhar aos interesses do PLN quando decide usar papéis semânticos.

Na história dos PropBanks, ArgMs e Args numerados são tratados de forma diferenciada por dois motivos: (i) o status sintático de "exigência verbal", ou de parte integrante do significado do item lexical, e por isso codificado no léxico; (ii) sua generalidade, isto é, possibilidade de ocorrer com praticamente qualquer verbo, diferentemente dos argumentos *core*, que teriam sentidos específicos.

Porém, é justamente a sua generalidade – possibilidade de ocorrerem com qualquer verbo e de aparecerem em diferentes gêneros textuais – que faz dos ArgM um elemento relevante no PLN, já que tendem a ser frequentes, demonstrando potencial para ampliar informações encontradas em textos e também para ampliar generalizações.

Se, como indicam Levin e Rappaport (2005:36), "*Semantic roles, then, can be viewed as labels for equivalence classes of arguments*", não há porque usar classes diferentes para o sintagma "em Cuba" em *Ela vive em Cuba*, *Ela mora em Cuba* e *Ela desapareceu em Cuba*, exceto a justificativa linguística ou, mais precisamente, a justificativa alinhada a abordagens linguísticas que vêem o léxico (e as entradas lexicais) como um objeto que contém apenas informações consideradas essenciais – especificamente aqui, essenciais para completar o sentido do verbo.

Retomando as conclusões de Merlo & Van der Plas (2009) sobre a especificidade semântica de Arg3-Arg5[67] que dificulta a generalização, temos o seguinte cenário, considerando a versão atual do Verbo-Brasil e em uma exploração bastante inicial:

Arg3 (258)	beneficiário (36)	instrumento (25)	local (7) endere?o (1)	causa (3)/motivo (3)	t?pico (2)	atributo (25) finalidade (14)
Arg4 (74)	beneficiário (9)	instrumento (4)	local/ponto de chegada/ponto de saída (18)/end point (14)			atributo (5) finalidade (1)
Arg5 (5)		instrumento (1)/medium (2)		causa (1)		

Quadro 12: Distribuição semântica dos papéis de Arg3, Arg4 e Arg5 no Verbo-Brasil[68]

Como vemos, uma parte considerável do sentido dos elementos codificados como ArgM de papéis Arg3-Arg5 tem um valor previsto/generalizável, podendo se encaixar ou em um ArgM dos papéis PropBank ou em algum papel da VerbNet – e já previsto no alinhamento Verbnet - PropBank via SemLink[69] (veja-se também as tabelas 3 e 4 de Merlo & Van der Plas, 2009).

Abdicar da distinção entre argumentos *core* e *não core* a fim de unificar esses casos pode ser uma alternativa para ampliar a generalização dos papéis semânticos. Agindo dessa maneira, tomamos o caminho de uma "generalização guiada pela semântica" em contraste com a "generalização guiada pela sintaxe", que se sustenta na descrição dos *frames sintáticos*, unidades alvo da anotação ao estilo PropBank. Por *generalização guiada pela semântica*, nos referimos à uma classificação que (i) iguala argumentos considerados *core* a argumentos considerados *não core*, anotando igualmente todos eles e (ii) busca um conjunto de classes unificado entre os papéis elencados na VerbNet e os ArgM elencados no PropBank. Além disso, salientamos que abrir mão de decisões motivadas por teorias linguísticas (como a distinção entre argumentos *core* e não *core* e o estudo sobre alternâncias sintáticas) não significa abrir mão de conhecimento linguístico[70].

[67] Sem contar com a definição de papéis que praticamente não têm identidade semântica, apenas descrição sintática, como o Arg3 de ocupar: "aquilo que é usado para preencher o Arg1, quando separado do Arg0 (ocupar com; ocupar + gerúndio)". Neste caso, qual representação semântica é criada?

[68] Os "erros" nos caracteres especiais apenas indicam como os elementos foram buscados na interface alternativa do Verbo-Brasil.

[69] Além do SemLink, um alinhamento entre etiquetas PropBank e VerbNet também foi feito pela equipe do AnCora (para espanhol e catalão), e está disponível em https://clic.ub.edu/corpus/sites/default/files/2022-02/etiquetes_func_arg_PT.pdf

[70] Originalmente, o PropBank foi criado não apenas com o objetivo de viabilizar o desenvolvimento de sistemas de PLN, mas também com o interesse de viabilizar o estudo de alternâncias sintáticas de

Segundo abordagens de viés funcionalista/empirista, o conceito de subcategorização está ligado à frequência**.** A generalização guiada pela semântica, que não distingue aprioristicamente (com base em testes e na "intuição") entre argumentos *core* e *não core*, operacionaliza esta outra forma de lidar com a subcategorização: assumindo que um dos interesses (linguísticos/teóricos) na distinção entre *core* e não *core* está, por exemplo, na identificação de proposições incompletas (quando faltaria um argumento *core*), a pergunta não é mais sobre o que **deveria estar presente** de acordo com uma determinada noção teórica, mas sobre o que **costuma estar presente** (em um determinado gênero/tipo de texto?), mas não está. Assim como na visão das abordagens formais, os limites não estarão sempre claros e precisarão ser explicitados (e talvez seja preciso definir a partir de qual frequência podemos considerar que algo seria "esperado" ou "costumeiro"), mas esta forma de abordar a questão oferece flexibilidade para lidar com a ideia de padrão (ou de "costume"), que pode ser ajustada conforme as necessidades e interesses. Desse ponto de vista, passa a ser mais informativo, por exemplo, a ausência de um papel do tipo *local* em uma proposição que envolve o verbo "viver" do que em uma proposição que envolve o verbo "plantar" (considerando os dados apresentados na seção 6.2, 52% das ocorrências de "viver" aparecem com *local*, e apenas 25% das ocorrências de "plantar" aparecem com *local*, a despeito de serem consideradas informações *não-core* e *core*, respectivamente). Do mesmo modo, em certos cenários (por exemplo, detecção de desinformação), a ausência da fonte de informação relativa ao conteúdo de uma proposição é um dado relevante, e este tipo de informação é indicado por meio de argumentos *não-core*. Em suma, o que é "essencial" depende da tarefa e domínio. Informações relativas a *causa*, *finalidade*, *quantidade*, dentre tantas outras, podem ser "essenciais" conforme o interesse.

Outra consequência prática da eliminação da diferença entre complementos *core* e *não-core* seria a possibilidade de tratamento uniforme para os casos em que dois elementos/sintagmas têm o mesmo conteúdo semântico previsto em um argumento *numerado*. Ou seja, na frase abaixo poderíamos ter uma anotação mais informativa e consistente com os sintagmas em destaque anotados da mesma forma (por exemplo, "maneira"):

1. Os criminosos, usando touca ninja, **saíram** em ao menos dois carros por ruas de essas cidades atirando contra alvos escolhidos por eles .

Ainda outra consequência é a **diminuição da dependência de um léxico** externo (os *Frames*) para a anotação, uma vez que a atribuição de papéis se daria pela semântica dos papéis, somente, seguindo a hierarquia de papéis já indicada na VerbNet (reproduzida no Anexo). Esta dissociação é uma vantagem para línguas que não dispõem de um recurso completo como a língua inglesa dispõe. Por outro lado, nada impede que um léxico similar (no que se

---

um ponto de vista quantitativo, investigando como e por que essas alternâncias sintáticas ocorrem (Palmer et al., 2005:72). Fillmore (1968:278), 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*". Desde o seu surgimento, portanto, PropBank colocou lado a lado interesses de teorias linguísticas – e especificamente interesses mais voltados para a sintaxe – e interesses aplicados do PLN, apostando que ambos são naturalmente convergentes.

refere aos *padrões sintáticos*) seja criado após a anotação, com os dados vindos diretamente do corpus. Do ponto de vista semântico, as classes da VerbNet, assim como os *Frames* do PropBank, partem de classes conforme sugeridas em Levin (1993). Tais classes (ou agrupamentos semânticos) foram criados obedecendo primeiramente critérios sintáticos, partindo da hipótese[71] de que o comportamento sintático seria reflexo de aspectos do significado.

> Verb classifications help NLP systems to deal with this complexity by organizing verbs into groups that **share core semantic and syntactic properties**. VerbNet (Kipper et al., 2008) is one such lexicon, which identifies semantic roles and syntactic patterns characteristic of the verbs in each class and makes explicit the connections between the syntactic patterns and the underlying semantic relations that can be inferred for all members of the class. (...)
> VerbNet is a lexicon of approximately 5800 English verbs, and **groups verbs according to shared syntactic behaviors**, thereby revealing generalizations of verb behavior (...). Thus, although **the basis of VerbNet classification is syntactic**, the verbs of a given class do share semantic regularities as well because, as **Levin hypothesized**, the syntactic behavior of a verb is largely determined by its meaning.
>
> (VerbNet Annotation Guidelines, s/d[72])

Segundo Levin (1993), a **opção de privilegiar a sintaxe** se deve à dificuldade de identificação de significados possíveis com base apenas na intuição ou na definição de dicionários[73] – dificuldade já mencionada em 6.1.2. Porém, o PLN já dispõe de maneiras eficientes de representar semanticamente os sentidos das palavras – as representações distribuídas – , que também se combinam harmoniosamente com a ideia de representações do significado não-categóricas, características apontadas por Kilgarriff (1997) e Manning (2022). Assim, uma alternativa, seguindo com a generalização guiada pela semântica, seria deixar a organização dos sentidos dos verbos, que pode ser feita de maneira independente do

---

[71] Vale notar que, enquanto hipótese, e especificamente para o português, foi refutada (ao menos parcialmente) em Lima (2007), que mostrou que verbos de um mesmo grupo semântico ("vitória" e "derrota") não necessariamente compartilham o mesmo comportamento sintático. O trabalho de Lima foi orientado pelo linguista e gramático Mario Perini, também crítico da hipótese de Levin (Perini, 2008). Em Costa & Freitas (2019), visando à criação de um léxico com verbos do "dizer" para tradutores, verificamos que a subclasse "say verbs", de Levin, é transversal à nossa classificação, que compilou quase 300 verbos com uso atestado de verbos de elocução. O trabalho não tinha como objetivo propor uma classificação geral dos verbos ou refutar Levin, mas apenas distribuir os verbos de elocução em grupos de sentido que fossem relevantes para tradutores, e com isso queremos apenas mostrar que, retomando Kilgarriff (1997), existem diferentes possibilidades para distribuir os verbos semanticamente, sendo o comportamento sintático, critério de Levin, apenas um deles. Além disso (e por outro lado) é importante notar que a ideia que tomar a sintaxe como ponto de partida para a semântica se justifica (para algumas teorias linguísticas) na medida em que a sintaxe é capaz de oferecer pistas formais da língua, diferentemente da semântica, e por isso é considerada (para algumas teorias linguísticas) o único caminho possível para a semântica.

[72] https://verbs.colorado.edu/verb-index/VerbNet_Guidelines.pdf

[73] "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 1993:14-15)

processo de anotação SRL, a cargo de modelos de linguagem. Ou seja, deixamos para as máquinas aquilo em que nós, seres humanos, somos reconhecidamente limitados (organização semântica de todas as palavras da língua); e assumimos como tarefa aquilo que ainda não podemos deixar para as máquinas (classificação dos participantes de proposições segundo certas classes semânticas). Podemos pensar no agrupamento de Levin (1993), seguido pela VerbNet e pelo PropBank, como o resultado de uma clusterização semântica humana baseada na sintaxe, e o que sugerimos aqui seria uma clusterização semântica automática com base em exemplos (derivados da anotação de frames no PBP).

Por fim, defender a generalização baseada em semântica (na anotação SRL) **não é equivalente a eliminar a sintaxe** do PLN. Pelo contrário, trata-se apenas de redimensionar o seu lugar no fluxo linguístico do PLN. A sintaxe **pode** não ser uma etapa necessária ou anterior no desenvolvimento de sistemas de PLN, mas certamente é uma maneira de organizar os dados linguísticos fundamental para seres humanos criarem conjuntos de dados padrão ouro com qualidade e agilidade (e a ferramenta ET, utilizada ao longo deste projeto de anotação, é um bom exemplo disso). Classificações sintáticas do tipo “sujeito” e “objeto”, como qualquer outra classificação[74], constroem igualdades que permitem generalizações, e por isso facilitam enormemente a tarefa de criação de datasets linguísticos, como mostram as regras de anotação utilizadas no processo de criação do PBP.

[74] Para abordagens linguísticas de viés universalista, no entanto, trata-se de classes especiais, “naturais” e “universais”.

Referências

ALVA-MANCHEGO, F.E.; ROSA, J.L.G. Semantic role labeling for brazilian portuguese: A benchmark. In: *Advances in Artificial Intelligence – IBERAMIA 2012, Lecture Notes in Computer Science*, v. 7637, p. 481–490. Springer Berlin Heidelberg, 2012.

ATTARDI, G. Experiments with a multilanguage non-projective dependency parser. In: *Proceedings of the Tenth Conference on Computational Natural Language Learning*, CoNLL-X '06, p. 166–170, Stroudsburg, PA, USA, 2006. Association for Computational Linguistics.

BICK, Eckhard. Automatic Semantic Role Annotation for Portuguese. In: P*roceedings of TIL 2007 - 5th Workshop on Information and Human Language Technology*. ISBN 85-7669-116-7, p. 1713-1716, 2007. Versão ampliada disponível em: https://edu.visl.dk/~eckhard/pdf/TIL2007_roles.pdf.

BICK, E. From Treebank to Propbank: A Semantic-Role and VerbNet Corpus for Danish. In: TIEDEMANN, J. (Ed.). *NoDaLiDa 2017 - 21st Nordic Conference of Computational Linguistics, Proceedings of the Conference*. Linköping: Linköping University Electronic Press, 2017. p. 202-210. Disponível em: http://www.ep.liu.se/ecp/131/023/ecp17131023.pdf.

BONIAL, C.; BONN, Julia; CONGER, Kathryn; HWANG, Jena D.; PALMER, Martha; REESE, Nicholas. E*nglish PropBank Annotation Guidelines*. Disponível em: https://verbs.colorado.edu/PropBank/EPB-Annotation-Guidelines.pdf, 2015.

BONIAL, C.; CONGER, Kathryn; HWANG, Jena D.; MANSOURI, Aous; ASERI, Yahya; BONN, Julia; O'GORMAN, Timothy; PALMER, Martha. Current Directions in English and Arabic PropBank. In: PUSTEJOVSKY, James; IDE, Nancy (Eds.). *The Handbook of Linguistic Annotations*. Springer, 2018.

BORGES NETO, J. A naturalização da gramática tradicional e seu uso protocolar. *Texto de conferência proferida durante o VIII Congresso da ABRALIN*, Natal/RN, 02/02/2013. Disponível em: https://docs.ufpr.br/~borges/publicacoes/para_download/naturalizacao.pdf, 2013.

BORGES NETO, J. Gramática Tradicional e Linguística Contemporânea: continuidade ou ruptura?. *Todas As Letras - Revista De Língua E Literatura*, v. 14, n. 1, 2012. Recuperado de: https://editorarevistas.mackenzie.br/index.php/tl/article/view/4549.

BRANCO, A.; CARVALHEIRO, Catarina; PEREIRA, Sílvia; SILVEIRA, Sara; SILVA, João; CASTRO, Sérgio; GRAÇA, João. A PropBank for Portuguese: the CINTIL-PropBank. In: *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey. European Language Resources Association (ELRA), 2012. p. 1516–1521.

CARNEIRO, Murillo G.; CUPERTINO, Thiago Henrique; ZHAO, Liang; ROSA, João Luís Garcia. Semi-supervised semantic role labeling for brazilian portuguese. *JIDM*, v. 8, n. 2, p. 117–130, 2017.

CARRERAS, Xavier; MÀRQUEZ, Lluís. Introduction to the CoNLL-2005 Shared Task: Semantic Role Labeling. In: *Proceedings of the Ninth Conference on Computational Natural Language Learning (CoNLL-2005)*, Ann Arbor, Michigan. Association for Computational Linguistics. 2005. p. 152–164.

COSTA, B.; FREITAS, C. Um léxico de verbos do dizer para tradutores - e considerações sobre a classificação dos verbos de elocução. *Calidoscópio*, v. 17, p. 494-515, 2019.

FONSECA, E.R.; ROSA, João Luís Garcia. An architecture for semantic role labeling on portuguese. In: *Computational Processing of the Portuguese Language - 10th International Conference, PROPOR 2012*, Coimbra, Portugal, April 17-20, 2012. Proceedings. p. 204–209.

EVANS, R.; ORASAN, C. Sentence Simplification for Semantic Role Labelling and Information Extraction. In: *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*, Varna, Bulgaria. INCOMA Ltd., 2019. p. 285–294.

DONG, X.L. Generations of Knowledge Graphs: The Crazy Ideas and the Business Impact. *Proc. VLDB Endow.*, v. 16, n. 12, p. 4130–4137, 2023. Disponível em: https://doi.org/10.14778/3611540.3611636.

DURAN, M.S.; ALUÍSIO, S.M. Propbank-br: a brazilian treebank annotated with semantic role labels. In: *Proceedings of the International Conference on Language Resources and Evaluation (LREC-2012).* 2012. p. 1862–1867.

DURAN, M. S.; SEPÚLVEDA-TORRES, L.; VIVIANI, M. C.; HARTMANN, N. S.; ALUÍSIO, S. M. Seleção e Preparação de Sentenças do Corpus PLN-BR para Compor o Corpus de Anotação de Papéis Semânticos PropBank-Br.V2. *Série de Relatórios Técnicos do Núcleo Interinstitucional de Linguística Computacional, NILC-TR-14-07*, 45 pg., 2014.

DURAN, M. S.; ALUÍSIO, S. M. Propbank-Br: a Brazilian Portuguese corpus annotated with semantic role labels. In: *Proceedings of the 8th Symposium in Information and Human Language Technology*, Cuiabá/MT, Brazil, October 24-26, 2011. p. 164–168.

DURAN, M.S. *Manual de anotação do PropBank-Br v2*. São Carlos: ICMC-USP, 2014.

DURAN, M. S.; ALUÍSIO, S. M. Automatic Generation of a Lexical Resource to support Semantic Role Labeling in Portuguese. In: *Proceedings of SEM 2015: The Fourth Joint Conference on Lexical and Computational Semantics*, 2015. p. 216-221.

DURAN, M. S. *Manual de anotação de relações de dependência: versão revisada e estendida*. São Carlos: ICMC-USP, 2022. Disponível em: https://repositorio.usp.br/item/003100539.

DURAN, M., LOPES, L., NUNES, M., & PARDO, T. (2023). The Dawn of the Porttinari Multigenre Treebank: Introducing its Journalistic Portion. In Anais do XIV Simpósio Brasileiro

de Tecnologia da Informação e da Linguagem Humana, (pp. 115-124). Porto Alegre: SBC. doi:10.5753/stil.2023.233975.

DURAN, M.S.; FREITAS, C. *Guia de Anotação de Papéis Semânticos seguindo o modelo PropBank no corpus Porttinari-base*. São Carlos: ICMC-USP, 2024.

FALCI, Daniel Henrique Mourão; SOARES, Marco Antônio Calijorne; BRANDÃO, Wladmir Cardoso; PARREIRAS, Fernando Silva. Using recurrent neural networks for semantic role labeling in Portuguese. In: *Progress in Artificial Intelligence, 19th EPIA Conference on Artificial Intelligence, EPIA 2019*, Vila Real, Portugal, September 3-6, 2019, Proceedings, Part II, p. 682–694, 2019.

FILLMORE, Charles J. Em favor do caso. Tradução de Lúcia M. Lobato. In: LOBATO, L. M. P. (Sel.). *A semântica na lingüística moderna: o léxico*. Rio de Janeiro: Francisco Alves, 1977. p. 275-361. Tradução de: The case for case. [1968].

FITZGERALD, N.; MICHAEL, Julian; HE, Luheng; ZETTLEMOYER, Luke. Large-scale QA-SRL parsing. In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics* (ACL), 2018.

FONSECA, E.R.; ROSA, J.L.G. A two-step convolutional neural network approach for semantic role labeling. In: *Neural Networks (IJCNN), The 2013 International Joint Conference on Neural Networks*. p. 1–7, 2013.

FREITAS, C.; SOUZA, E.; CASTRO, M. C.; CAVALCANTI, T.; FERREIRA DA SILVA, P.; CORRÊA CORDEIRO, F. Recursos linguísticos para o PLN específico de domínio: o Petrolês. *Linguamática*, v. 15, n. 2, p. 51-68, 2023. DOI: https://doi.org/10.21814/lm.15.2.412.

FREITAS, C. Estudos linguísticos e Humanidades digitais: corpus e descorporificação. *Gragoatá*, v. 22, n. 44, p. 1207-1227, 2017. DOI: https://doi.org/10.22409/gragoata.v22i44.33556.

FREITAS, C.; PARDO, T. (submetido) PropBanks e representações semânticas: o que temos, o que podemos e o que queremos. *Linguamática* (submetido)

GÓMEZ-RODRÍGUEZ, C.; SARTORIO, F.; SATTA, G. A polynomial-time dynamic oracle for non-projective dependency parsing. In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, p. 917–927, Doha, Qatar, October 2014. Association for Computational Linguistics.

GUNG, J.; PALMER, M. Predicate Representations and Polysemy in VerbNet Semantic Parsing. In: *Proceedings of the 14th International Conference on Computational Semantics (IWCS)*, p. 51–62, Groningen, The Netherlands (online), 2021. Association for Computational Linguistics.

HARTMANN, Nathan Siegle; DURAN, Magali Sanches; ALUÍSIO, Sandra Maria. Automatic semantic role labeling on non-revised syntactic trees of journalistic texts. In: Silva, J., Ribeiro, R., Quaresma, P., Adami, A., Branco, A. (eds) *Computational Processing of the*

Portuguese Language. PROPOR 2016. Lecture Notes in Computer Science, vol 9727. Springer, Cham. https://doi.org/10.1007/978-3-319-41552-9_20.

HARTMANN, Silvana; MÚJDRICZA-MAYDT, Éva; KUZNETSOV, Ilia; GUREVYCH, Iryna; FRANK, Anette. Assessing SRL Frameworks with Automatic Training Data Expansion. In: *Proceedings of the 11th Linguistic Annotation Workshop*, p. 115–121, Valencia, Spain, 2017. Association for Computational Linguistics.

HE, Luheng; LEWIS, Mike; ZETTLEMOYER, Luke. Question-answer driven semantic role labeling: Using natural language to annotate natural language. In: *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2015.

HE, Luheng; LEE, Kenton; LEVY, Omer; ZETTLEMOYER, Luke. Jointly Predicting Predicates and Arguments in Neural Semantic Role Labeling. In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics* (Volume 2: Short Papers), p. 364–369, Melbourne, Australia, 2018. Association for Computational Linguistics.

HE, H.; CHOI, J. Establishing strong baselines for the new decade: Sequence tagging, syntactic and semantic parsing with BERT. In: *The Thirty-Third International Flairs Conference, 2020*.

HAJIČ, Jan; CIARAMITA, Massimiliano; JOHANSSON, Richard; KAWAHARA, Daisuke; MARTÍ, Maria Antònia; MÀRQUEZ, Lluís; MEYERS, Adam; NIVRE, Joakim; PADÓ, Sebastian; ŠTĚPÁNEK, Jan; STRAŇÁK, Pavel; SURDEANU, Mihai; XUE, Nianwen; ZHANG, Yi. The CoNLL-2009 Shared Task: Syntactic and Semantic Dependencies in Multiple Languages. In: *Conference on Computational Natural Language Learning* (CoNLL), 13., 2009, Boulder. Proceedings [...]. Boulder: Association for Computational Linguistics, 2009. p. 1–18.

KILGARRIFF, A. I don't believe in word senses. *Computers and the Humanities*, v. 31, p. 91-113, 1997.

KILGARRIFF, A.; PALMER, M. Introduction to the Special Issue on SENSEVAL. *Computers and the Humanities*, v. 34, n. 1/2, p. 1–13, 2000. Disponível em: http://www.jstor.org/stable/30204787.

KIPPER, Karin; KORHONEN, Anna; RYANT, Neville; PALMER, Martha. Extensive Classifications of English verbs. In: *Proceedings of the 12th EURALEX International Congress*. Turin, Italy, September, 2006.

LEVIN, B.; RAPPAPORT HOVAV, M. *Argument Realization*. Cambridge: Cambridge University Press, 2005.

LEVIN, B. *English Verb Classes and Alternations: a preliminary investigation*. Chicago: The University of Chicago Press, 1993.

LI, T.; KAZEMINEJAD, Ghazaleh; BROWN, Susan; SRIKUMAR, Vivek; PALMER, Martha. Learning Semantic Role Labeling from Compatible Label Sequences. In: *Findings of the Association for Computational Linguistics: EMNLP 2023*, p. 15561–15572, Singapore, 2023. Association for Computational Linguistics.

LIMA, B. *Valência dos verbos de vitória e derrota em Português*. Belo Horizonte: UFMG, 2007. Dissertação de Mestrado.

LOPER, E.; YI, Szu-ting; PALMER, Martha. Combining lexical resources: Mapping between PropBank and VerbNet. In: *Proceedings of the IWCS*. Disponível em: https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=77366e5ce668cdb94e64c39ab64718da2237171f, 2007.

MANNING, C.D. Human Language Understanding & Reasoning. *Daedalus*, v. 151, n. 2, p. 127–138, 2022. DOI: https://doi.org/10.1162/daed_a_01905.

MANNING, C. D. Probabilistic syntax. In: BOD; HAY; JANNEDY (eds), *Probabilistic Linguistics*, MIT Cambridge, MA: MIT press, 2003. p. 289-341.

DE MARNEFFE, Marie-Catherine; MANNING, Christopher D.; NIVRE, Joakim; ZEMAN, Daniel. Universal Dependencies. *Computational Linguistics*, v. 47, n. 2, p. 255–308, 2021. DOI: https://doi.org/10.1162/coli_a_00402.

MIKULOVÁ, Marie; BÉMOVÁ, Alevtina; HAJIČ, Jan; HAJIČOVÁ, Eva; HAVELKA, Jiří; KOLÁŘOVÁ-ŘEZNÍČKOVÁ, Veronika, et al. *Annotation on the tectogrammatical layer in the Prague Dependency Treebank. Annotation manual*. Prague: ÚFAL MFF UK, 2005.

MERLO, Paola; VAN DER PLAS, Lonneke. Abstraction and Generalisation in Semantic Role Labels: PropBank, VerbNet or both?. In: *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, p. 288–296, Suntec, Singapore, 2009. Association for Computational Linguistics.

MOHEBBI, M.; RAZAVI, S.N.; BALAFAR, M.A. Computing semantic similarity of texts based on deep graph learning with ability to use semantic role label information. *Scientific Reports*, v. 12, p. 14777, 2022. DOI: https://doi.org/10.1038/s41598-022-19259-5.

MONACHESI, P.; STEVENS, G.; TRAPMAN, J. Adding semantic role annotation to a corpus of written Dutch. In: *Proceedings of the Linguistic Annotation Workshop*. p. 77–84. ACL, 2007.

MÚJDRICZA-MAYDT, Éva; HARTMANN, Silvana; GUREVYCH, Iryna; FRANK, Anette. Combining Semantic Annotation of Word Sense & Semantic Roles: A Novel Annotation Scheme for VerbNet Roles on German Language Data. In: *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, p. 3031–3038, Portorož, Slovenia, 2016. European Language Resources Association (ELRA).

NGUYEN, Long H. B.; PHAM, Viet H.; DINH, Dien. Improving neural machine translation with AMR semantic graphs. Mathematical Problems in Engineering, 2021. DOI: https://doi.org/10.1155/2021/9939389.

NIVRE, J. Non-projective dependency parsing in expected linear time. In: *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*: Volume 1 - Volume 1, p. 351–359, Stroudsburg, PA, USA, 2009. Association for Computational Linguistics.

NIVRE, J.; NILSSON, J. Pseudo-Projective Dependency Parsing. In: *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, p. 99–106, Ann Arbor, Michigan, 2005. Association for Computational Linguistics.

OEPEN, S.; KUHLMANN, Marco; MIYAO, Yusuke; ZEMAN, Daniel; CINKOVÁ, Silvie; FLICKINGER, Dan; HAJIČ, Jan; UREŠOVÁ, Zdeňka. SemEval 2015 Task 18: Broad-Coverage Semantic Dependency Parsing. In: *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, p. 915–926, Denver, Colorado. Association for Computational Linguistics, 2015.

OLIVEIRA, A. S. M. *Semantic Role Labeling in Portuguese: Improving the State of the Art with Transfer Learning and BERT-based Models*. 2020. Dissertação de Mestrado. Departamento de Ciência de Computadores, Universidade do Porto.

PAGANO, A.; RASSI, A.; PAGNO, A. A ordem e a função das palavras em uma sentença. In: CASELI, H.; VOLPE NUNES, M. G. V. (Eds). *Processamento de Linguagem Natural: conceitos, técnicas e aplicações em Português*. BPLN, 2023.

PALMER, M.; KINGSBURY, P.; GILDEA, D. The Proposition Bank: An Annotated Corpus of Semantic Roles. *Computational Linguistics*, v. 31, n. 1, p. 71–106, 2005.

PALMER, M.; BONIAL, B.; MCCARTHY, C. SemLink+: FrameNet, VerbNet and Event Ontologies. In: *Proceedings of Frame Semantics in NLP: A Workshop in Honor of Chuck Fillmore (1929–2014)*, p. 13–17, Baltimore, Maryland USA, June 27, 2014.

PAN, J. Z.; RAZNIEWSKI, S.; KALO, J.-C.; SINGHANIA, S.; CHEN, J.; DIETZE, S.; JABEEN, H.; OMELIYANENKO, J.; ZHANG, W.; LISSANDRINI, M.; BISWAS, R.; DE MELO, G.; BONIFATI, A.; VAKAJ, E.; DRAGONI, M.; GRAUX, D. Large Language Models and Knowledge Graphs: Opportunities and Challenges. *Transactions on Graph Data and Knowledge (TGDK)*, v. 1, n. 1, p. 2:1-2:38, 2023. DOI: https://doi.org/10.4230/TGDK.1.1.2.

PARDO, T. A. S.; DURAN, M. S.; LOPES, L.; DI FELIPPO, A.; ROMAN, N. T.; NUNES, M. G. V. Porttinari - A large multi-genre treebank for Brazilian Portuguese. In: *Proceedings of the XIV Symposium in Information and Human Language (STIL 2021)*, p. 1-10, 2021.

PEETERS, B. Setting the Scene: Some Recent Milestones in the Lexicon-Encyclopedia Debate. In: *The Lexicon-Encyclopedia Interface*. Leiden, The Netherlands: Brill, 2000. DOI: https://doi.org/10.1163/9780585474465_002.

PERINI, M. *Estudos de gramática descritiva: as valências verbais*. São Paulo: Parábola Editorial, 2008.

PRADHAN, S.; BONN, J.; MYERS, S.; CONGER, K.; O'GORMAN, T.; GUNG, J.; WRIGHT-BETTNER, K.; PALMER, M. PropBank Comes of Age—Larger, Smarter, and more Diverse. In: *Proceedings of the 11th Joint Conference on Lexical and Computational Semantics*, p. 278–288, Seattle, Washington. Association for Computational Linguistics, 2022.

PRADHAN, S.; MOSCHITTI, A.; XUE, N.; URYUPINA, O.; ZHANG, Y. CoNLL-2012 Shared Task: Modeling Multilingual Unrestricted Coreference in OntoNotes. In: *Joint Conference on*

EMNLP and CoNLL - Shared Task, p. 1–40, Jeju Island, Korea. Association for Computational Linguistics, 2012.

RAPP, R. Using Semantic Role Labeling to Improve Neural Machine Translation. In: *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, p. 3079–3083, Marseille, France. European Language Resources Association, 2022.

REISINGER, D.; RUDINGER, R.; FERRARO, F.; HARMAN, C.; RAWLINS, K.; VAN DURME, B. Semantic Proto-Roles. *Transactions of the Association for Computational Linguistics*, v. 3, p. 475–488, 2015. DOI: https://doi.org/10.1162/tacl_a_00152.

RODRIGUES, J.; GOMES, L.; SILVA, J.; BRANCO, A.; SANTOS, R.; CARDOSO, H. L.; OSÓRIO, T. Advancing Neural Encoding of Portuguese with Transformer Albertina PT-*. In: *Progress in Artificial Intelligence: 22nd EPIA Conference on Artificial Intelligence, EPIA 2023, Faial Island, Azores, September 5–8, 2023, Proceedings, Part I*. Springer-Verlag, Berlin, Heidelberg, p. 441–453. DOI: https://doi.org/10.1007/978-3-031-49008-8_35.

SAEED, J. *Semantics*. Cambridge: Blackwell, 1997.

SAMPSON, G. Evidence Against the grammatical/ungrammatical distinction. In: SAMPSON, G. *Empirical Linguistics*. London: Continuum, 2001.

SAMPSON, G. (2000). The Role of Taxonomy in Language Engineering. *Philosophical Transactions: Mathematical, Physical and Engineering Sciences*, 358(1769), 1339–1355.

SANTOS, D.; MOTA, C. Experiments in Human-computer Cooperation for the Semantic Annotation of Portuguese Corpora. *In Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10*), Valletta, Malta. European Language Resources Association (ELRA). 2010.

de SOUZA, E. *Construção e avaliação de um treebank padrão ouro*. Rio de Janeiro: PUC-Rio, 2023. Dissertação de Mestrado.

de SOUZA, E.; FREITAS, C. Explorando variações no tagset e na anotação Universal Dependencies (UD) para Português: Possibilidades e resultados com base no treebank PetroGold. In: *Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana (STIL), 14*. , 2023, Belo Horizonte/MG. Anais [...]. Porto Alegre: Sociedade Brasileira de Computação, 2023. p. 125-134. DOI: https://doi.org/10.5753/stil.2023.233980 .

de SOUZA, E.; FREITAS, C. Still on arguments and adjuncts: the status of the indirect object and the adverbial adjunct relations in Universal Dependencies for Portuguese. In: *Proceedings of the Universal Dependencies Brazilian Festival*, p. 1–10, Fortaleza, Brazil: Association for Computational Linguistics, 2022. Disponível em: https://aclanthology.org/2022.udfestbr-1.5.

de SOUZA, E.; FREITAS, C. ET: A workstation for querying, editing and evaluating annotated corpora. In: *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, p. 35–41. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, 2021. Disponível em: https://aclanthology.org/2021.emnlp-demo.5.

STRAKA, M.; HAJIC, J.; STRAKOVA, J.; HAJIC, Jr., J. Parsing universal dependency treebanks using neural networks and search-based oracle. In: *14th International Workshop on Treebanks and Linguistic Theories (TLT 2015)*, p. 208–220, Warszawa, Poland. IPIPAN, 2015.

SHI, P.; LIN, J. Simple BERT Models for Relation Extraction and Semantic Role Labeling. *arXiv:1904.05255*, 2019.

SOLAWETZ, J.; LARSON, S. LSOIE: A Large-Scale Dataset for Supervised Open Information Extraction. In: *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, p. 2595–2600. Online: Association for Computational Linguistics, 2021.

STOWE, K.; PRECIADO, J.; CONGER, K.; BROWN, S. W.; KAZEMINEJAD, G.; GUNG, J.; PALMER, M. SemLink 2.0: Chasing Lexical Resources. In: *Proceedings of the 14th International Conference on Computational Semantics (IWCS)*, p. 222–227, Groningen, The Netherlands (online). Association for Computational Linguistics, 2021.

SUN, K.; XU, Y.; ZHA, H.; LIU, Y.; DONG, X. L. Head-to-Tail: How Knowledgeable are Large Language Models (LLMs)? A.K.A. Will LLMs Replace Knowledge Graphs?. In: *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, p. 311–325, Mexico City, Mexico. Association for Computational Linguistics, 2024.

TENNEY, I.; DAS, D.; PAVLICK, E. BERT Rediscovers the Classical NLP Pipeline. In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, p. 4593–4601, Florence, Italy. Association for Computational Linguistics, 2019.

TENNEY, I.; XIA, P.; CHEN, B.; WANG, A.; POLIAK, A.; MCCOY, R. T.; KIM, N.; VAN DURME, B.; BOWMAN, S. R.; DAS, D.; PAVLICK, E. What do you learn from context? Probing for sentence structure in contextualized word representations. *International Conference on Learning Representations*, 2019. Disponível em: https://research.google/pubs/what-do-you-learn-from-context-probing-for-sentence-structure-in-contextualized-word-representations/.

TAULÉ, M.; MARTÍ, M. A.; RECASENS, M. AnCora: Multilevel Annotated Corpora for Catalan and Spanish. In: *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco. European Language Resources Association (ELRA), 2008.

WALLIS, S. Completing Parsed Corpora. In: ABEILLÉ, A. (Ed.). *Treebanks: Building and Using Parsed Corpora*. Dordrecht: Springer Netherlands, p. 61–71, 2003.

WANG, N.; LI, J.; MENG, Y.; SUN, X.; QIU, H.; WANG, Z.; WANG, G.; HE, J. An mrc framework for semantic role labeling. In: *Proceedings of the 29th International Conference on Computational Linguistics*, p. 2188–2198, 2022.

WEISCHEDEL, R.; PALMER, M.; MARCUS, M.; HOVY, E.; PRADHAN, S.; RAMSHAW, L.; XUE, N.; TAYLOR, A.; KAUFMAN, J.; FRANCHINI, M.; EL-BACHOUTI, M.; BELVIN, R.;

HOUSTON, A. *OntoNotes Release 5.0* LDC2013T19. Web Download. Philadelphia: Linguistic Data Consortium, 2013. DOI: https://doi.org/10.35111/xmhb-2b84.

WU, N.; HOU, H.; LI, H.; CHANG, X.; JIA, X. Semantic perception-oriented low-resource neural machine translation. In: SU, J.; SENNRICH, R. *Machine Translation. 17th China Conference, CCMT 2021, Xining, China, October 8–10, 2021, Revised Selected Papers*. Springer Singapore, 2021. Disponível em: http://sc.cipsc.org.cn/mt/conference/2021/papers/T21-1009.pdf.

YADAV, H.; HUSAIN, S.; FUTRELL, R. Are formal restrictions on crossing dependencies epiphenomenal?. In: *Proceedings of the 18th International Workshop on Treebanks and Linguistic Theories (TLT, SyntaxFest 2019)*, p. 2–12, Paris, France. Association for Computational Linguistics, 2019.

YI, S.-t.; LOPER, E.; PALMER, M. Can semantic roles generalize across genres? In: *Proceedings of HLT/NAACL-2007*, Rochester, NY, p. 22–27, Apr 2007.

YU, K.; SOFRONIEV, P.; SCHILL, E.; HINRICHS, E. The parse is darc and full of errors: Universal dependency parsing with transition-based and graph-based algorithms. In: *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, p. 126–133, Vancouver, Canada. Association for Computational Linguistics, 2017.

Apêndice 1: Frames que contêm mwe verbais

Frames que têm a indicação de “fixo” em algum dos seus argumentos:

abrir.401
bater.402
bater.403
bater.404
cair.401
chamar.401
cruzar.401
dar.102
dar.401
dar.402
dar.404
dar.501
dar.502
dizer.401
entender.501
falar.401
fazer.401

fazer.402
fazer.403
fazer.404
fazer.405
fazer.501
ficar.401
jogar.401
levar.401
levar.402
pegar..404
pegar.501
perder.401
por.401
por.402
por.403
ter.401
ter.402
ter.403
tocar.401
tomar.401
tomar.402
tomar.403
virar.401
voltar.401

Apêndice 2: Respostas ao estudo de concordâncias e discordâncias

Formulários 1 e 2

Relata sua amizade com Janot e afirma que o ex-procurador-geral chamava Dodge de "bruxa" em **conversas reservadas**			
Form 1 (total 5)		Form 2 (total 8)	
LOCAL	3	LOCAL	3 (37,5%)
TEMPO	2	TEMPO	1 (12,5%)
MANEIRA		MANEIRA	2 (25%)
FINALIDADE		FINALIDADE	
		TEMPO_LOCAL	2 (25%)

A proposta foi aprovada em 2011, mas até hoje não decolou.(monossêmica)			
Form 1 (total 5)		Form 2 (total 8)	
LOCAL		LOCAL	
TEMPO	5	TEMPO	8
MANEIRA		MANEIRA	
FINALIDADE		FINALIDADE	
		TEMPO_LOCAL	

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.**			
Form 1 (total 5)		Form 2 (total 8)	
LOCAL	4 (80%)	LOCAL	1 (12,5%)
TEMPO	1 (20%)	TEMPO	1 (12,5%)
MANEIRA		MANEIRA	
FINALIDADE		FINALIDADE	
		TEMPO_LOCAL	5 (62,5%)
		TEMPO_FINALIDADE	1 (12,5%)

Nesse encontro , encerra-se oficialmente o mandato do senador Aécio Neves .			
Form 1 (total 5)		Form 2 (total 8)	
LOCAL	3 (60%)	LOCAL	1 (12,5%)
TEMPO	2 (40%)	TEMPO	2 (25%)
MANEIRA		MANEIRA	
FINALIDADE		FINALIDADE	
		TEMPO_LOCAL	5 (62,5)

Os cartolas locais foram ajudados pela vitória **nas duas últimas edições da Copa América.**	
Form 1 (total 5)	Form 2 (total 8)

LOCAL	2	LOCAL	
TEMPO	3	TEMPO	2 (25%)
MANEIRA		MANEIRA	
FINALIDADE		FINALIDADE	
		TEMPO_LOCAL	6 (75%)

Em encontro, Bono *pergunta* a Macri sobre argentino desaparecido.			
Form 1 (total 5)		Form 2 (total 8)	
LOCAL	2	LOCAL	2 (25%)
TEMPO	3	TEMPO	1 (12,5%)
MANEIRA		MANEIRA	
FINALIDADE		FINALIDADE	
		TEMPO_LOCAL	5 (62,5)

Não consigo deixar de amar essa magnífica e adorável figura paterna, mas está puxadíssimo ter prazer **em qualquer conversa .**			
Form 1 (total 5)		Form 2 (total 8)	
LOCAL	4	LOCAL	3 (37,5%)
TEMPO	1	TEMPO	2 (25%)
MANEIRA		MANEIRA	2 (25%)
FINALIDADE		FINALIDADE	
		TEMPO_LOCAL	1 (12,5%)

Nesses encontros, o governador adota tom crítico à gestão de Temer em assuntos como privatizações, reformas econômicas e segurança.			
Form 1 (total 5)		Form 2 (total 8)	
LOCAL	4	LOCAL	2 (25%)
TEMPO	1	TEMPO	2 (25%)
MANEIRA		MANEIRA	
FINALIDADE		FINALIDADE	
		TEMPO_LOCAL	4 (50%)

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 .			
Form 1 (total 5)		Form 2 (total 8)	
LOCAL	3	LOCAL	2 (25%)
TEMPO	2	TEMPO	
MANEIRA		MANEIRA	2 (25%)
FINALIDADE		FINALIDADE	
		TEMPO_LOCAL	3 (37,5%)
		LOCAL_FINALIDADE	1 (12,5%)
TOTAL respostas	5		

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

Form 1 (total 5)		Form 2 (total 8)	
LOCAL	4	LOCAL	1 (12,5%)
TEMPO	1	TEMPO	2 (25%)
MANEIRA		MANEIRA	2 (25%)
FINALIDADE		FINALIDADE	
		TEMPO_LOCAL	1 (12,5%)
		LOCAL_MANEIRA	1 (12,5%)
		TEMPO_LOCAL_MANEIRA	1 (12,5%)
TOTAL respostas	5		

Ele disse que eu agora vivo de pijamas **em casa**. (monossêmica)			
Form 1 (total 5)		Form 2 (total 8)	
LOCAL	5	LOCAL	8
TEMPO		TEMPO	
MANEIRA		MANEIRA	
FINALIDADE		FINALIDADE	
		TEMPO_LOCAL	
TOTAL respostas	5		

Em alguns casos, manteve-se o caráter vitalício do poder.			
Form 1 (total 5)		Form 2 (total 8)	
LOCAL	3	LOCAL	3 (37,5%)
TEMPO	2	TEMPO	3 (37,5%)
MANEIRA		MANEIRA	2 (25%)
FINALIDADE		FINALIDADE	
		TEMPO_LOCAL	
TOTAL respostas	5		

Hernanes foi o primeiro jogador do São Paulo a falar **na discutida reunião com as torcidas uniformizadas da última quarta-feira**.			
Form 1 (total 5)		Form 2 (total 8)	
LOCAL	4	LOCAL	3 (37,5%)
TEMPO	1	TEMPO	
MANEIRA		MANEIRA	
FINALIDADE		FINALIDADE	
		TEMPO_LOCAL	3 (37,5%)
		LOCAL_MANEIRA	1 (12,5%)
		TEMPO_LOCAL_MANEIRA_FINALIDADE	1 (12,5%)

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 .			
Form 1 (total 5)		Form 2 (total 8)	
LOCAL	1	LOCAL	1 (12,5%)
TEMPO	4	TEMPO	2 (25%)
MANEIRA		MANEIRA	1 (12,5%)
FINALIDADE		FINALIDADE	
		TEMPO_LOCAL	1 (12,5%)

		TEMPO_LOCAL_MANEIRA	1 (12,5%)
		TEMPO_LOCAL_MANEIRA_FINALIDADE	2 (25%)

No final da reunião, surpreendentemente, o Anselmo passa a adotar um comportamento mais duro, pressionando para delação.			
Form 1 (total 5)		Form 2 (total 8)	
LOCAL		LOCAL	
TEMPO	5	TEMPO	6
MANEIRA		MANEIRA	
FINALIDADE		FINALIDADE	
		TEMPO_LOCAL	2 (25%)
TOTAL respostas	5		

Saí **derrotada** pouco depois das 14h. (monossêmica)			
Form 1 (total 5)		Form 2 (total 8)	
LOCAL		LOCAL	
TEMPO		TEMPO	
MANEIRA	5	MANEIRA	8
FINALIDADE		FINALIDADE	
		TEMPO_LOCAL	

Formulários 3 e 4

Com o quarto mandato, ela deve chegar aos 16 anos de poder.			
Form 3 (total 7)		Form 4 (total 7)	
LOCAL		LOCAL	1 (14,3%)
TEMPO	1 (14,3%)	TEMPO	1 (14,3%)
MANEIRA	4 (57,1%)	MANEIRA	1 (14,3%)
FINALIDADE		FINALIDADE	
CAUSA	2 (28,6%)	CAUSA	1 (14,3%)
		TEMPO_CAUSA	2 (28,6%)
		TEMPO_MANEIRA	1 (14,3%)

Nós respeitamos a PGR e temos a soberania de decidir **por maioria .**			
Form 3 (total 7)		Form 4 (total 7)	
LOCAL		LOCAL	
TEMPO		TEMPO	
MANEIRA	7 (100%)	MANEIRA	6 (85,7%)
FINALIDADE		FINALIDADE	
CAUSA		CAUSA	
		MANEIRA_CAUSA	1 (14,3%)

Na falta de notícias, o pessoal acaba <u>focando</u> também em eleições.			
Form 3 (total 7)		Form 4 (total 7)	
LOCAL		LOCAL	
TEMPO		TEMPO	
MANEIRA		MANEIRA	1 (14,3%)
FINALIDADE		FINALIDADE	
CAUSA	7 (100%)	CAUSA	5 (71,4%)
		TEMPO_CAUSA	1 (14,3%)

Em um cenário extremo, os bancos centrais <u>temem</u> que poderiam até perder o controle da base monetária .			
Form 3 (total 7)		Form 4 (total 7)	
LOCAL	2 (28,6%)	LOCAL	
TEMPO	2 (28,6%)	TEMPO	3 (42,9%)
MANEIRA	2 (28,6%)	MANEIRA	1 (14,3%)
FINALIDADE		FINALIDADE	
CAUSA	1 (14,3%)	CAUSA	
		LOCAL_CAUSA	1 (14,3%)
		TEMPO_LOCAL_CAUSA	1 (14,3%)
		TEMPO_LOCAL_MANEIRA	1 (14,3%)

Quem é que <u>sai</u> **para trabalhar** pensando em tomar um soco na cara ?			
Form 3 (total 7)		Form 4 (total 7)	
LOCAL		LOCAL	
TEMPO	1 (14,3%)	TEMPO	
MANEIRA		MANEIRA	
FINALIDADE	6 (85,7%)	FINALIDADE	7 (100%)
CAUSA		CAUSA	
		TEMPO_CAUSA	

Em alguns roteiros, como na Croácia, é preciso **pagar** <u>na maioria dos trajetos.</u>			
Form 3 (total 7)		Form 4 (total 7)	
LOCAL	7 (100%)	LOCAL	5 (71,4%)
TEMPO		TEMPO	1 (14,3%)
MANEIRA		MANEIRA	
FINALIDADE		FINALIDADE	1 (14,3%)
CAUSA		CAUSA	
		TEMPO_CAUSA	

No momento em que a nova moeda for criada, se os preços <u>forem convertidos</u> **no pico**, não será necessário um aperto monetário, uma recessão temporária?			
Form 3 (total 7)		Form 4 (total 7)	
LOCAL	1 (14,3%)	LOCAL	

TEMPO	5 (71,4%)	TEMPO	4 (57,1%)
MANEIRA	1 (14,3%)	MANEIRA	1 (14,3%)
FINALIDADE		FINALIDADE	
CAUSA		CAUSA	
		TEMPO_MANEIRA	2 (28,6%)

Ou seja: o Emmy de 2017 foi politizado **do começo ao fim**.			
Form 3 (total 7)		Form 4 (total 7)	
LOCAL	1 (14,3%)	LOCAL	
TEMPO	3 (42,9%)	TEMPO	2 (28,6%)
MANEIRA	3 (42,9%)	MANEIRA	2 (28,6%)
FINALIDADE		FINALIDADE	
CAUSA		CAUSA	
		TEMPO_LOCAL_MANEIRA	1 (14,3%)
		TEMPO_ MANEIRA	2 (28,6%)

Em homenagem a Che Guevara, Cuba reage às declarações de Trump.			
Form 3 (total 7)		Form 4 (total 7)	
LOCAL	2 (28,6%)	LOCAL	1 (14,3%)
TEMPO	2 (28,6%)	TEMPO	
MANEIRA		MANEIRA	1 (14,3%)
FINALIDADE	2 (28,6%)	FINALIDADE	
CAUSA	1 (14,3%)	CAUSA	2 (28,6%)
		TEMPO_LOCAL	1 (14,3%)
		MANEIRA_ FINALIDADE	1 (14,3%)
		FINALIDADE_CAUSA	1 (14,3%)

Eles escreveram **na casa** a frase "SP não está à venda" . (monossêmica)			
Form 3 (total 7)		Form 4 (total 7)	
LOCAL	7	LOCAL	7
TEMPO		TEMPO	
MANEIRA		MANEIRA	
FINALIDADE		FINALIDADE	
CAUSA		CAUSA	
		TEMPO_CAUSA	

Com mais capital privado, a Sabesp poderá conquistar mercados hoje explorados pelas empresas privadas.			
Form 3 (total 7)		Form 4 (total 7)	
LOCAL		LOCAL	
TEMPO		TEMPO	
MANEIRA	6 (85,7%)	MANEIRA	3 (42,9%)
FINALIDADE		FINALIDADE	
CAUSA	1 (14,3%)	CAUSA	1 (14,3%)
		TEMPO_MANEIRA_CAUSA	1 (14,3%)
		MANEIRA_CAUSA	2 (28,6%)

Com a liderança caindo em seu colo, Lewis Hamilton ditou as ações da corrida.			
Form 3 (total 7)		Form 4 (total 7)	
LOCAL		LOCAL	
TEMPO	1 (14,3%)	TEMPO	
MANEIRA	1 (14,3%)	MANEIRA	2 (28,6%)
FINALIDADE		FINALIDADE	1 (14,3%)
CAUSA	5 (71,4%)	CAUSA	4 (57,1%)
		TEMPO_CAUSA	

A equipe gaúcha perdeu a oportunidade de manter a distância para o líder **ao ser derrotada pela Chapecoense por 1 a 0**, em casa.			
Form 3 (total 7)		Form 4 (total 7)	
LOCAL		LOCAL	
TEMPO	3 (42,9%)	TEMPO	1 (14,3%)
MANEIRA		MANEIRA	
FINALIDADE	1 (14,3%)	FINALIDADE	
CAUSA	3 (42,9%)	CAUSA	4 (57,1%)
		TEMPO_CAUSA	1 (14,3%)
		TEMPO_MANEIRA	1 (14,3%)

Como avançar soluções de amplo espectro **nesse cenário**, com nosso sistema político-eleitoral?			
Form 3 (total 7)		Form 4 (total 7)	
LOCAL	5 (71,4%)	LOCAL	4 (57,1%)
TEMPO	1 (14,3%)	TEMPO	1 (14,3%)
MANEIRA	1 (14,3%)	MANEIRA	
FINALIDADE		FINALIDADE	
CAUSA		CAUSA	
		TEMPO_LOCAL	2 (28,6%)

Para se livrar do problema, dobrou a aposta e injetou mais R$ 2 bilhões na JBS .			
Form 3 (total 7)		Form 4 (total 7)	
LOCAL		LOCAL	
TEMPO		TEMPO	
MANEIRA		MANEIRA	
FINALIDADE	5 (71,4%)	FINALIDADE	7
CAUSA	2 (28,6%)	CAUSA	
		TEMPO_CAUSA	

Naquele ano, o tucano tinha 30 % de aprovação.			
Form 3 (total 7)		Form 4 (total 7)	
LOCAL		LOCAL	
TEMPO	7	TEMPO	7
MANEIRA		MANEIRA	
FINALIDADE		FINALIDADE	

CAUSA		CAUSA	
		TEMPO_CAUSA	

Anexo: Taxonomia de papéis da VerbNet

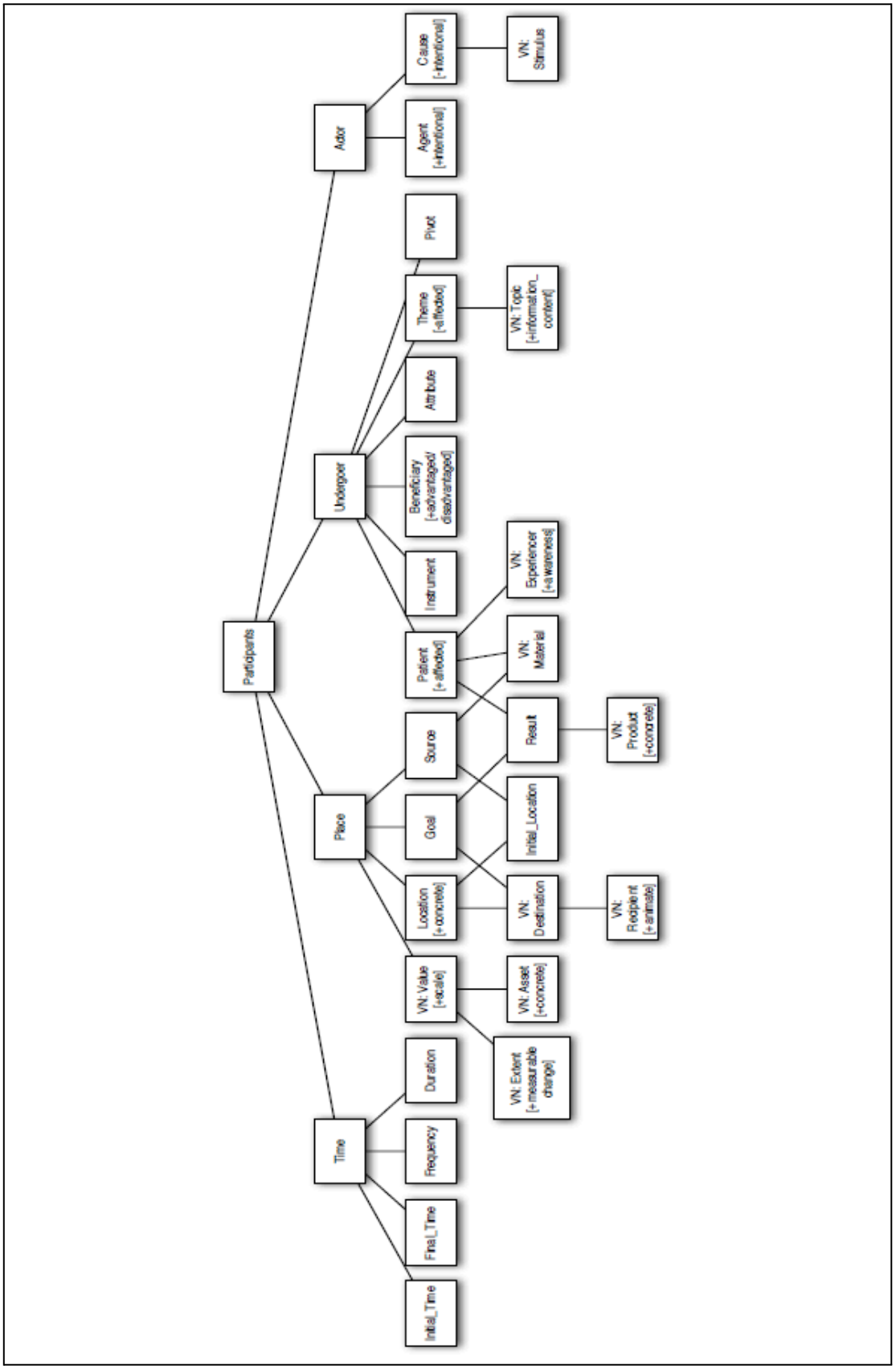


Figura: Taxonomia de papéis semânticos da VerbNet. Fonte: VerbNet annotation Guidelines (https://verbs.colorado.edu/verb-index/VerbNet_Guidelines.pdf)