DOI: 10.14393/DL52-v16n4a2022-6

Modelação da valência verbal numa gramática computacional do português no formalismo HPSG

Modeling verb valency in a computational grammar for Portuguese in the HPSG formalism

Leonel Figueiredo de ALENCAR[*]
Alexandre RADEMAKER[**]

RESUMO: A HPSG é uma teoria gramatical lexicalista que propõe a formalização em paralelo das estruturas morfossintáticas e semânticas. Este trabalho descreve a implementação computacional de valências verbais numa nova gramática do português nesse formalismo. Essa gramática é relevante não somente para aplicações de compreensão textual, mas representa também uma contribuição à documentação formal das estruturas da língua, destacando-se pelo tratamento das construções de controle e alçamento. A gramática tem sido implementada incrementalmente por meio da sua aplicação a conjuntos de teste cada vez mais abrangentes. Com 278 entradas e um total de 215 lemas, o léxico verbal ainda é pequeno. No entanto, a hierarquia de tipos proposta modela as propriedades de 118 classes valenciais, das quais 57 são tipos que codificam classes de verbos. A gramática analisa 94% de um total de 581 sentenças gramaticais, apresentando, ao

ABSTRACT: HPSG is a lexicalist grammatical theory that proposes the parallel formalization of morphosyntactic and semantic structures. This work describes the computational implementation of verbal valences in a new Portuguese grammar in this formalism. This grammar is relevant not only for text understanding applications, but also represents a contribution to the formal documentation of the structures of the language, standing out for its treatment of control and raising constructions. The grammar has been incrementally implemented through its application to increasingly comprehensive test suites. With 278 entries and a total of 215 lemmas, the verb lexicon is still small. However, the proposed type hierarchy models the properties of 118 valence classes, of which 57 are types that encode verb classes. The grammar analyzes 94% of a total of 581 grammatical sentences, while showing low hypergeneration in a set of 167 ungrammatical examples.

[*] Doutor em Linguística, Professor Titular da Universidade Federal do Ceará e Professor Visitante da EMAp/FGV. ORCID: http://orcid.org/0000-0001-8148-6994. leonel.de.alencar@ufc.br
[**] Doutor em Informática, Professor Adjunto da EMAp/FGV e Pesquisador da IBM Research. ORCID: http://orcid.org/0000-0002-7583-0792. alexrad@br.ibm.com

mesmo tempo, baixa hipergeração em um conjunto de 167 exemplos agramaticais.

PALAVRAS-CHAVE: Linguística computacional. Engenharia da gramática. Análise sintática automática. Valência. Semântica computacional.

KEYWORDS: Computational linguistics. Grammar engineering. Syntactic parsing. Valence. Computational semantics.

1 Introdução

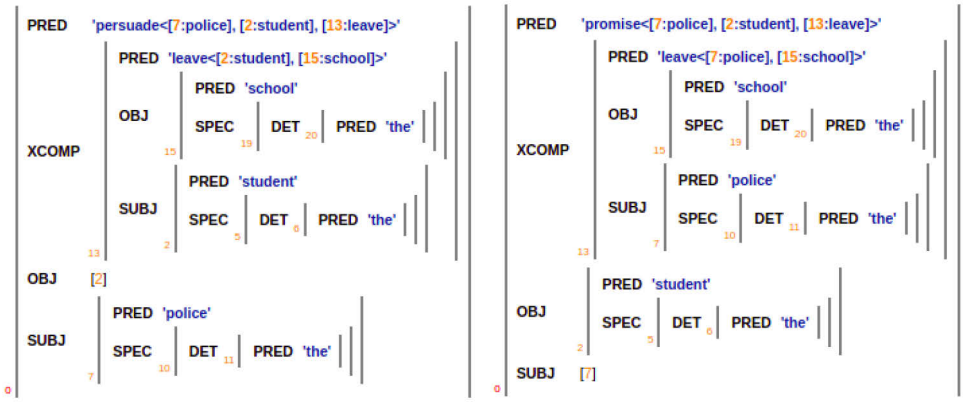
A análise sintática automática profunda é um componente de diversos sistemas comerciais de processamento de linguagem natural (doravante PLN). Essa tarefa consiste em construir, para as sentenças de um texto, representações suficientemente informativas para utilização por sistemas de compreensão textual. Um exemplo bastante característico é o Watson da IBM, sistema de resolução de perguntas representativo do estado da arte. Subjaz a esse sistema uma arquitetura híbrida, que conjuga abordagens estatísticas com simbólicas, da qual faz parte uma gramática computacional do inglês elaborada manualmente (FERRUCCI *et al.*, 2010; MCCORD; MURDOCK; BOGURAEV, 2012). Um dos principais marcos da inteligência artificial da última década, esse sistema, hoje amplamente empregado na área médica, venceu em 2011 dois campeões do *Jeopardy!*, um popular programa de perguntas e respostas da TV estadunidense.

É inegável que, em muitas aplicações de PLN, análises sintáticas rasas produzem resultados satisfatórios a um custo drasticamente menor do que o demandado pela implementação de analisadores profundos. No entanto, o maior esforço para construir esse tipo de componente é compensado pela maior riqueza de informações capaz de fornecer ao processamento semântico.

Para elucidar a distinção entre os dois tipos de análise sintática, comparem-se as diferentes representações de (1) e (2) nas Figuras 1 e 2.

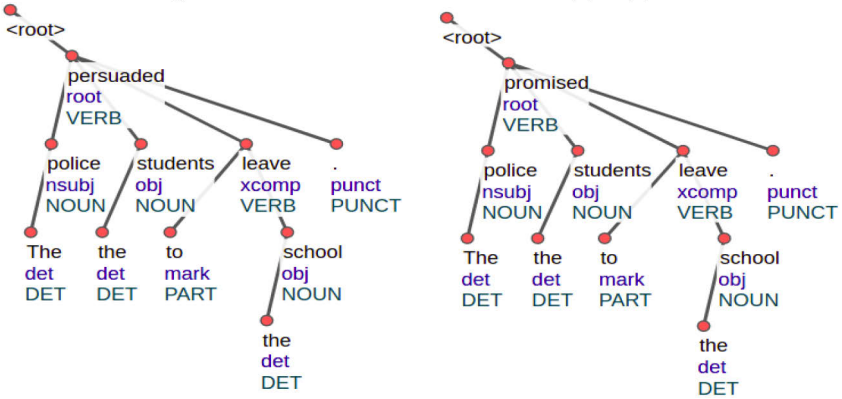
(1) The police persuaded the students to leave the school.
a polícia persuadir:PST;3SG os estudantes a deixar:INF a escola
A polícia convenceu os estudantes a deixarem a escola.'
(2) The police promised the students to leave the school.
a polícia prometer:PST;3SG os estudantes a deixar:INF a escola
'A polícia prometeu aos estudantes deixar a escola.'

Figura 1 – Análise sintática profunda de (1) e (2).



Fonte: gerada pelo programa XLE-Web (ROSÉN *et al.*, 2012) a partir da gramática do inglês no formalismo LFG.

Figura 2 – Análise sintática rasa de (1) e (2).



Fonte: gerada pelo analisador sintático estatístico UDPipe 2 com base no modelo *english-ewt-ud-2.6-200830* (STRAKA; STRAKOVÁ, 2017).

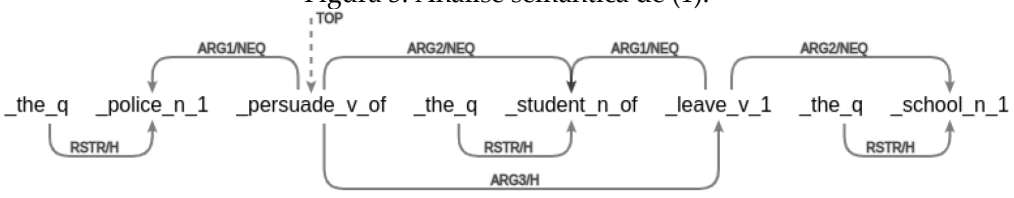
Nas duas representações da Figura 1, os verbos principais expressam uma relação de três lugares envolvendo duas entidades, expressas por *a polícia* e *os*

estudantes, e o estado de coisas referido pelo verbo encaixado. Esses três argumentos são realizados pelas funções sintáticas SUBJ (sujeito), OBJ (objeto) e XCOMP, denominado complemento predicativo ou complemento aberto, que corresponde ao complemento infinitivo. Este, por sua vez, possui dois argumentos, o segundo dos quais é a entidade referida por *a escola*. Em (1) e (2), o primeiro argumento do infinitivo, realizado em (3) na posição de sujeito, não é expresso nessa posição. Enquanto o primeiro argumento do verbo encaixado, na primeira análise da Figura 1, é a mesma entidade referida pelo objeto do verbo principal, na segunda análise é o sujeito do verbo principal que expressa esse argumento. O compartilhamento de informações entre o sujeito ou o objeto do verbo principal e o sujeito do verbo encaixado é expresso pelos índices entre colchetes. Por exemplo, o índice 2 em OBJ [2] da primeira análise da Figura 1 remete à estrutura do SUBJ do XCOMP, que recebe índice de igual valor.

(3) The police left the school.
'A polícia deixou a escola.'

A Figura 2 apresenta as análises de (1) e (2) conforme a teoria das Dependências Universais (doravante UD) (MARNEFFE *et al.*, 2021; NIVRE *et al.*, 2020). Salvo o lema verbal sob o nó <root>, os dois gráficos são idênticos, expressando as mesmas relações de dependência do verbo principal que nas análises da Figura 1, ou seja, sujeito (*nsubj*), objeto (*obj*) e complemento predicativo (*xcomp*). Uma diferença essencial entre as representações profundas no formalismo da LFG, na Figura 1, e suas contrapartes rasas em UD, na Figura 2, consiste em que estas últimas não especificam o sujeito do verbo encaixado.

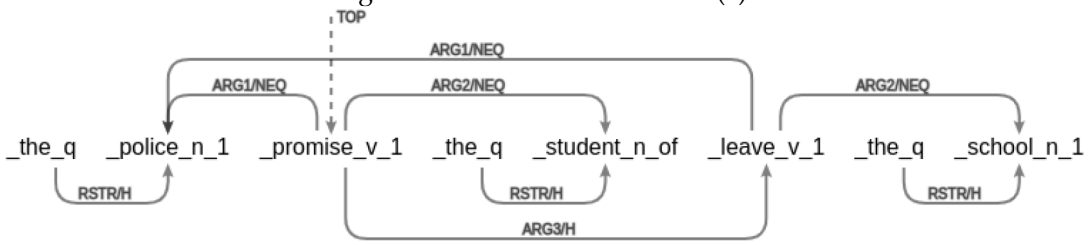
Figura 3: Análise semântica de (1).



Fonte: gerada pela ferramenta delphin-viz[1] a partir da gramática ERG 2018 (UW).

Além da LFG, outra teoria gramatical formal implementada computacionalmente e amplamente utilizada na construção de analisadores sintáticos profundos é a HPSG (SAG; WASOW; BENDER, 2003; MÜLLER, 2020). Uma vantagem importante de analisadores baseados nesse segundo modelo, sob a perspectiva da compreensão textual automática, é que integram a descrição sintática e a descrição semântica num único nível de representação. A Figura 3 e a Figura 4 são representações semânticas geradas pela *English Resource Grammar* (FLICKINGER, 2000) (doravante ERG), aparentemente a maior gramática computacional do inglês nesse formalismo. Conforme a Figura 3, o predicado *persuade* possui três argumentos, indicados por meio de setas rotuladas como ARG1, ARG2 e ARG3. Este último é o predicado *leave* que, por sua vez, tem dois argumentos, sendo o ARG1 destes o ARG2 de persuade. A Figura 4 é quase idêntica, exceto que o ARG1 de *leave* é o ARG1 do predicado *promise*.

Figura 4 – Análise semântica de (2).



Fonte: gerada pela ferramenta delphin-viz a partir da gramática ERG 2018 (UW).

Em (1) e (2), o primeiro argumento do verbo encaixado é determinado por uma função sintática do verbo matriz, numa relação denominada controle. Que função do verbo matriz controla o sujeito do verbo encaixado depende do tipo de verbo matriz. Para que um analisador sintático possa determinar esses sujeitos externos, é preciso

[1] http://delph-in.github.io/delphin-viz/demo/

que disponha de informações a respeito das relações de controle de cada verbo matriz. Isso, por sua vez, exige uma descrição exaustiva da valência verbal e sua codificação num formalismo computacional.

Observe que, em inglês, nenhuma característica formal distingue o controle do sujeito de (2) do controle do objeto de (1). Ambos os verbos governam um sujeito, um objeto direto e um complemento infinitivo. Desse modo, a mera anotação sintática de um corpus contendo sentenças com esses verbos aparentemente não seria suficiente para que um analisador estatístico determinasse o controlador no caso de verbos com essas propriedades de controle não presentes no corpus de treino.

O modelo Dependências Universais Estendidas (SCHUSTER; MANNING, 2016) propõe ampliar as anotações do projeto UD para uma representação sintática mais próxima da semântica. Para tanto, inclui, entre outras anotações adicionais, uma dependência rotulada *nsubj* entre o sujeito do verbo matriz e o verbo encaixado em sentenças como (4). No entanto, essa camada extra de anotação foi aplicada a poucas línguas, não incluindo o português (DROGANOVA; ZEMAN, 2019).

(4) Fred started to laugh.
'Fred começou a rir.'

Na release 2.8 da coleção *Deep Universal Dependencies* (DROGANOVA; ZEMAN, 2019), foram anotados no *UD_Portuguese-Bosque*, *treebank* do português do projeto UD (RADEMAKER *et al.*, 2017), mais de 2500 sujeitos externos (i.e., sujeitos de *xcomps*), dos quais a maioria é controlada pelo sujeito e o restante, pelo objeto do verbo matriz. Em português, contudo, não só o sujeito e o objeto do verbo matriz, mas também um objeto indireto podem controlar o sujeito de um verbo não finito encaixado, como evidencia (5).

(5) A mulher disse ao jardineiro para podar a árvore.

Outra limitação do esquema de anotação do projeto *Deep Universal Dependencies* é que não distingue entre os dois tipos de sujeitos externos exemplificados em (1), (2) e (5), por um lado, e (4), por outro. No primeiro grupo de exemplos, temos estruturas de controle, nas quais o sujeito externo é argumento semântico tanto do predicado expresso pelo verbo matriz quanto do predicado expresso pelo verbo encaixado. Pelo contrário, o sujeito externo de (4) é argumento semântico apenas do predicado do verbo encaixado, configurando uma construção de alçamento.

Em teorias lexicalistas como a LFG e a HPSG, controle e alçamento integram o domínio da valência, uma vez que constituem propriedades da estrutura argumental de determinados itens lexicais, notadamente verbos (POLINSKY, 2013; ABEILLÉ, 2021). Nas três últimas décadas, diversas iniciativas, sob diferentes perspectivas teóricas, voltaram-se ao levantamento das valência verbais do português, na esteira de Fernandes (1987), cuja primeira edição data de 1940. No entanto, esses trabalhos se furtam a uma modelação do controle e do alçamento, os projetos mais recentes se atendo à sentença simples.

Fernandes (1987) subdivide os seus mais de 12000 verbetes com base, primeiramente, em categorias de regência, que perfazem um total de oito, a saber, (i) intransitivo, (ii) transitivo, (iii) relativo, (iv) transitivo relativo (i.e., introduzido por preposição), (v) birrelativo, (vi) transitivo predicativo, (vii) predicativo e (viii) pronominal. A cada uma dessas subdivisões corresponde uma ou mais acepções, acompanhadas de abonações. Por exemplo, o verbo *convencer*, na acepção "obrigar com razões, argumentos (a reconhecer alguma coisa); persuadir" de (1), é transitivo-relativo. Não se indica como o sujeito do infinitivo que pode constituir o complemento relativo deve ser interpretado, trata-se de informação que precisa ser deduzida pelo usuário a partir das acepções dadas ou dos exemplos.

Borba (1991) oferece uma descrição bem mais detalhada de cerca de 6000 dos verbos do dicionário anterior, levando em conta não só aspectos morfossintáticos, mas também semânticos. Destaca-se pelo tratamento dos complementos oracionais, especificando não só a sua forma, mas também, ocasionalmente, relações de correferência entre o sujeito do verbo encaixado e um argumento do verbo matriz. No verbete de *impedir*, por exemplo, Borba (1991) afirma que (7) deriva de (6) pela transposição do sujeito da oração subordinada para complemento do verbo superior, repercutindo a análise transformacional do alçamento do sujeito (POLINSKY, 2013; ABEILLÉ, 2021). O verbete de *ouvir*, porém, carece de explicação dessa natureza, embora o verbo exiba comportamento análogo em (8) e (9). Analogamente, no verbete de *pedir*, consta que esse verbo significa "solicitar licença para" quando a oração principal e a infinitiva introduzida por *para* compartilham o mesmo sujeito, como em (10). No entanto, não explicitam qual seria o sujeito do infinitivo em (11) e (12).

(6) A polícia impediu que os ladrões assaltassem a joalheria.
(7) A polícia impediu os ladrões de assaltarem a joalheria.
(8) Bento ouviu bem que a velha gemia.
(9) [Pe. Lucas] ouvia o vento assobiar cada vez mais forte.
(10) [Telma] ligou direto, () pedindo para falar com Edith.
(11) Você pediu para não contar.
(12) Mamãe, peça ao Luiz para jantar conosco[2].

VerboWeb é um banco de dados sobre propriedades semânticas e sintáticas de verbos, com base na decomposição de predicados. Atualmente com 1418 lemas verbais e 1509 entradas, restringe-se a estruturas argumentais de até três argumentos realizados por sintagmas nominais e preposicionais (CANÇADO; AMARAL; MEIRELLES, 2017; CANÇADO *et al.*, 2021), limitação essa herdada das descrições de

[2] Exemplos de Borba (1991), dos quais (8)-(12) provêm de obras literárias.

diferentes classes verbais no mesmo quadro teórico que o antecederam e que incorporou, como Cançado, Godoy e Amaral (2013).

O projeto em andamento *Dicionário de valências verbais do português brasileiro*, tal como o VerboWeb, limita-se a sentenças simples, adiando para uma etapa futura fenômenos de valência envolvendo sentenças compostas, como os exemplificados em (1), (2), (4) e (5) (PERINI, 2016; PERINI *et al.*, 2019). Nesse dicionário, a valência de um verbo constitui-se de uma ou mais diáteses, concebidas como associações entre constituintes e papéis semânticos (PERINI, 2015). A versão atual abrange 635 verbos, aos quais se atribuem 326 diáteses ao todo. Controle e alçamento propriamente ditos não integram o quadro teórico subjacente, a noção que mais se aproxima desses construtos é a de emparelhamento de papéis semânticos, exemplificada em (13) (PERINI, 2015, p. 189–199). Na diátese de (14) associada a esse exemplo, considerada característica idiossincrática do verbo *achar*, emparelham-se os papéis dos sintagmas nominais complementos, porque entidade qualificada e qualidade implicam-se mutuamente, uma vez que esta predica sobre aquela. Uma desvantagem da representação de (14) em relação à entrada lexical desse verbo proposta por Abeillé (2021, p. 517–519) no quadro da HPSG é quebrar o paralelismo com o predicado de dois lugares de (15), paráfrase de (13), uma vez que implica que o verbo nesse exemplo é um predicado de três lugares. Conforme a abordagem de Abeillé (2021), pelo contrário, *achar* é um verbo de alçamento para objeto que expressa uma relação entre um experienciador (EXP) e um estado de coisas (SOA) tanto em (13) quanto (15), constituindo, portanto, em ambos os casos, um predicado de dois lugares, a diferença entre as duas variantes residindo na quantidade de complementos. Enquanto a segunda variante tem um único complemento, realizado como oração completiva, a primeira tem dois, um objeto direto e um complemento predicativo, cujo sujeito é compartilhado com o objeto direto.

(13) Marta acha Jim um idiota.
(14) VSubj>Agent V NP>*Qualified.thing* NP>*Quality*
(15) Marta acha que Jim é um idiota.

Visando permitir uma anotação sintaticamente profunda de corpora, entre outras aplicações que necessitem de uma compreensão textual mais refinada, iniciamos o desenvolvimento da PorGram, uma nova gramática computacional do português no formalismo HPSG. Cabe destacar que uma gramática computacional de uma língua não tem apenas utilidade prática na construção de softwares de processamento da linguagem. Cumpre, também, um importante papel na testagem de hipóteses linguísticas, constituindo uma forma de documentação de uma língua cuja consistência interna e validade empírica podem ser automaticamente verificadas (BENDER, 2010; BENDER; FLICKINGER; OEPEN, 2011). Diferentemente de sua congênere LXGram (COSTA; BRANCO, 2010), a PorGram é um projeto de software livre e de código aberto. Está ligada ao DELPH-IN Consortium, um movimento de implementação de gramáticas no formalismo da HPSG para várias línguas envolvendo grupos de pesquisa de vários países[3]. Código, documentação, conjuntos de teste, programas auxiliares desenvolvidos pela equipe da PorGram, tudo está disponível numa plataforma de software livre que integra diversas formas de interação social[4].

Neste trabalho, focamos a implementação da valência verbal, abrangendo a realização de argumentos tanto em sentenças simples, quanto em sentenças compostas, incluindo estruturas de controle e de alçamento. Na próxima seção, delineamos os princípios e conceitos fundamentais da teoria da HPSG relevantes para a compreensão do presente trabalho, o sistema *LinGO Grammar Matrix*, utilizado para

[3] https://github.com/delph-in/docs/wiki
[4] https://github.com/LR-POR/PorGram

construir parte substancial da PorGram, e a arquitetura da gramática. Na seção 3, descrevemos, em linhas gerais, o modelo de representação semântica gerada pela PorGram. Na seção 4, tratamos da implementação da valência, mostrando quais fenômenos puderam ser implementados por meio do questionário de customização e quais tiveram de ser codificados manualmente. Na seção 5, apresentamos os resultados da aplicação da PorGram sobre conjuntos de teste. Finalmente, na seção 6, sumarizamos os resultados alcançados e apontamos direções para os próximos passos a serem trilhados no desenvolvimento da gramática.

2 A teoria da HPSG, o sistema *Grammar Matrix* e a PorGram

A HPSG é uma teoria gramatical formal amplamente utilizada na implementação de analisadores sintáticos profundos. É uma teoria lexicalista, codificando parte substancial da informação gramatical no léxico, pelo que a formalização da valência assume enorme relevo. Diferentemente de teoria multiestratais como a LFG, a HPG é monoestratal. Em único nível de representação são codificadas informações sintáticas, fonológicas, semânticas, pragmático-discursivas etc.

A HPSG baseia-se na noção de signo, utilizando estruturas de traços providas de tipos para representar o significante e o significado tanto de palavras e sintagmas quanto de regras morfológicas e sintáticas. A unificação é a operação matemática fundamental desse formalismo. Essa operação permite construir representações mais complexas a partir de representações mais simples, assegurando que as informações sejam compatíveis entre si (FRANCEZ; WINTNER, 2012).

Estruturas de traços organizam as informações sobre as propriedades dos objetos linguísticos sob a forma de matrizes de atributos e valores. Por exemplo, podemos representar as propriedades de um item lexical como *amarela* por meio de

uma matriz com pares de atributos e valores como [CATEGORIA adjetivo], [GÊNERO feminino] e [NÚMERO singular]. Na HPSG, essas matrizes são providas de tipos, os quais desempenham duas funções. A primeira é restringir as estruturas de uma dada gramática, assegurando que contenham apenas atributos apropriados aos objetos que descrevem e restringindo os valores de cada atributo a determinados tipos. Por exemplo, o par [TEMPO presente] é apenas adequado na especificação de uma forma verbal, não de um adjetivo ou substantivo, ao passo que [TEMPO plural] não constitui uma especificação válida, pois *plural* não constitui subtipo do tipo tempo.

A segunda função dos tipos é formular generalizações sobre os objetos modelados, facilitando a descrição de classes, subclasses e instâncias específicas. Por exemplo, nas entradas lexicais dos verbos *convencer*, *forçar*, *proibir* e *impedir*, nas variantes dos exemplos abaixo, não precisamos descrever cada uma de suas propriedades individualmente:

(16) O diretor convenceu os estudantes a deixarem o local.
(17) Os seguranças forçaram os estudantes a deixar o local.
(18) A polícia proibiu os manifestantes de usarem capuzes.
(19) Os guardas impediram os manifestantes de invadir o prédio.

Observemos que (16)-(19) compartilham um conjunto de propriedades: (i) há um verbo principal numa forma finita e outro encaixado no infinitivo; (ii) o verbo principal expressa o tipo básico de evento de cada sentença (por exemplo, (18) descreve uma ação de proibição pela polícia e não o uso de capuzes por manifestantes); (iii) o objeto direto do verbo principal é o sujeito externo do verbo no infinitivo; (iv) o segundo complemento do verbo principal é um sintagma de complementador (CP, do inglês *complementizer phrase*). O tipo de complementador, porém, não é uniforme nas quatro sentenças, permitindo dividir os quadro verbos em dois subgrupos: enquanto convencer e forçar exigem como complementador a preposição *a*, *proibir* e *impedir*

exigem a preposição *de* (GABRIEL; MÜLLER, 2008, p. 39). Por outro lado, esses verbos se enquadram em duas classes diferentes relativamente a um outro critério, qual seja se o sujeito externo do infinitivo exerce papel temático do verbo matriz. Esse é o caso dos verbos de controle do objeto de (16)-(18), mas não de *impedir* em (19), classificado como verbo de alçamento para objeto. A propósito, o termo alçamento (*raising* em inglês) reflete a análise do fenômeno em modelos transformacionais da gramática gerativa, segundo a qual o objeto direto de (19) resulta do movimento de *os manifestantes* da posição de sujeito do verbo encaixado à posição mais elevada de objeto do verbo matriz. Utiliza-se esvaziado dessa acepção em modelos como a LFG e a HPSG, onde não há qualquer tipo de movimento de constituintes (FALK, 2001; SAG; WASOW; BENDER, 2003).

Uma maneira de modelar o comportamento dos verbos de (16)-(19) de forma elegante é codificar em tipos mais abstratos as propriedades gerais do grupo e as propriedades de subgrupos em tipos mais específicos, utilizando o mecanismo de múltipla herança na codificação dos tipos mais específicos. Por exemplo, podemos implementar um tipo *main-verb-lex* para dar conta da propriedade (ii) não somente desses quatro verbos, mas de qualquer outro verbo principal. A estrutura argumental dos verbos de (16)-(18) e o controle do sujeito externo do verbo encaixado pelo objeto direto do verbo principal podem ser codificados no tipo *ditrans-second-arg-control-verb-lex*, que, por sua vez, constitui subtipo de *ditrans-second-arg-control-lex-item*, que modela as propriedades análogas de itens lexicais de outras classes com a mesma aridade. A forma infinitiva do complemento oracional pode ser codificada no tipo *inf-ditrans-second-arg-control-verb-lex*, enquanto a exigência de uma preposição específica pode ser modelada pelos tipos *a-inf-ditrans-second-arg-control-verb-lex* e *de-inf-ditrans-second-arg-control-verb-lex*, exemplificados nas entradas de convencer e proibir da Figura 5. Analogamente, as propriedades de verbos de alçamento para objeto podem ser codificadas num tipo geral *ditrans-second-arg-raising-verb-lex*, propriedades essas

herdadas tanto pela variante de *impedir* em (19) quanto por *fazer* na variante causativa de (20). As particularidades de cada uma dessas variantes são modeladas em subtipos desse tipo geral, conforme as entradas da Figura 5, que determinam que o segundo complemento deve realizar-se como CP infinitivo encabeçado pelo complementador preposicional de, no caso de impedir, ou por um sintagma verbal nu, no caso de fazer.

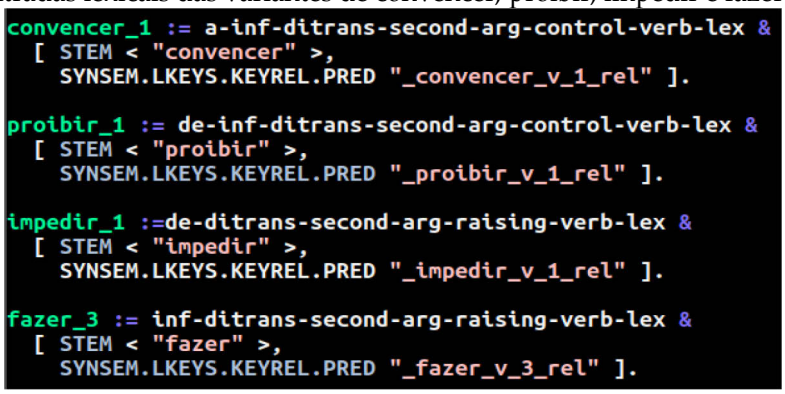
(20) a fumaça fez os agentes recuarem[5]

Como se pode constatar nas entradas lexicais da Figura 5, a organização dos tipos da gramática numa hierarquia com múltipla herança simplifica enormemente a codificação da valência. Na PorGram, como em gramáticas análogas, as entradas lexicais constituem definições de tipos, nas quais as informações se subdividem em quatro categorias. A primeira informação, antes do operador de definição ":=", é o tipo a ser definido, que funciona como identificador da entrada lexical. Por convenção, no caso de verbos, esse identificador constitui-se do lema com o sufixo _n, onde n indica a variante do verbo correspondente à definição do tipo à direita do operador ":=". Na Figura 5, a notação *impedir_1* indica que se trata de variante do verbo impedir distinta, por exemplo, de uma variante transitiva direta *impedir_2*, exemplificada em (21). A definição, por sua vez, constitui-se inicialmente de uma invocação de tipo, que codifica as propriedades da subclasse da qual o item em questão constitui instância, após o operador de unificação "&" seguem as propriedades idiossincráticas do item, que são o lema (STEM)[6] e o predicado semântico, que, convencionalmente, possui o mesmo índice do identificador.

(21) Os guardas impediram a invasão do prédio.

[5] Extraído de voto do TJ-SP de 15/06/2021 (https://tj-sp.jusbrasil.com.br/).
[6] O termo *stem* designa o radical de uma palavra, que, em inglês, corresponde ao lema. Em português, radical e lema não coincidem necessariamente.

Figura 5 – Entradas lexicais das variantes de convencer, proibir, impedir e fazer em (16)-(20).



Fonte: arquivo my-lexicon.tdl da PorGram.

Nessas duas entradas, temos uma amostra da linguagem TDL (do inglês *Type Description Language*), linguagem de descrição utilizada para elaborar a PorGram. Gramáticas especificadas nessa linguagem podem ser compiladas em analisadores sintáticos por meio de diversos sistemas, entre eles o LKB, que é o ambiente de desenvolvimento e testagem de gramáticas computacionais que utilizamos (COPESTAKE, 2002).

O desenvolvimento da PorGram desdobra-se em dois eixos. No primeiro, utilizamos o sistema de customização da *LinGO Grammar Matrix* (BENDER; FLICKINGER; OEPEN, 2002; BENDER *et al.*, 2010). Esse sistema gera o código TDL de um protótipo de gramática a partir de informações fornecidas pelo usuário sobre a estrutura da língua, sem necessidade de qualquer conhecimento da linguagem de descrição. Essas informações são elicitadas por meio de um questionário on-line dividido em 25 seções e submetidas a um sistema de validação, que emite avisos no caso de informações incompletas ou incompatíveis. Quatro seções visam especificar metadados, definir estratégias de toquenização, compilar conjuntos de teste etc. As demais cobrem diferentes aspectos da sintaxe e da morfologia e permitem especificar um número arbitrário de entradas lexicais para as principais classes de palavras. Cada seção cobre um amplo espectro de variação tipológica. Por exemplo, a seção sobre caso

permite especificar 9 padrões diferentes de marcação dos argumentos de verbos intransitivos e transitivos prototípicos, i.e., verbos monovalentes e verbos divalentes com estrutura argumental constituída de agente e paciente. Essa seção permite também especificar casos adicionais, não canônicos, para marcação desses argumentos.

A *Grammar Matrix* resulta de um esforço iniciado há duas décadas. Em sua primeira versão, teve como ponto de partida a ERG e a JACY (SIEGEL; BENDER; BOND, 2016), gramática de ampla cobertura do japonês. Esse núcleo inicial contemplava apenas fenômenos gramaticais universais (DRELLISHAK, 2009). Desde então, foram progressivamente adicionadas bibliotecas para lidar com fenômenos sujeitos a variação tipológica, por exemplo, caso e concordância (DRELLISHAK, 2009), tempo e aspecto (POULSON, 2011), regras morfológicas (GOODMAN, 2013), operações de mudança da valência (CURTIS, 2018), complementos oracionais (ZAMARAEVA; HOWELL; BENDER, 2019) e interrogativas parciais (ZAMARAEVA, 2021).

O segundo eixo de desenvolvimento da PorGram consiste na codificação manual na linguagem TDL, seja adaptando os tipos gerados automaticamente, seja criando novos tipos e entradas lexicais, de modo a analisar fenômenos gramaticais não abrangidos pelo sistema de customização. A Tabela 1 apresenta os componentes do código em TDL da gramática. Os 12 primeiros arquivos foram produzidos pelo sistema de customização, os três últimos foram codificados manualmente.

Os cinco primeiros arquivos não variam de uma língua para outra. Os dois primeiros constituem a espinha dorsal da gramática. Enquanto o segundo define 510 tipos disjuntivos para superclasses de palavras, como *+np*, que engloba substantivos e adposições, o primeiro fornece o arcabouço geral das estruturas gramaticais e semânticas, especificando as regras sintagmáticas e lexicais gerais, as classes de

valência etc. Como as especificações dos arquivos 3-5 não são relevantes no contexto do presente artigo, remetemos o leitor à documentação contida nos próprios arquivos.

Tabela 1 – Componentes em TDL da PorGram.7.

Número	Nome do arquivo	Tipos	Linhas	Palavras	Caracteres
	matrix.tdl	518	1957	8774	78278
	head-types.tdl	501	501	4947	24495
	labels.tdl	37	176	737	6188
	mtr.tdl	22	44	206	1113
	pet.tdl	0	21	48	415
	portuguese.tdl	363	1163	4960	60224
	lrules.tdl	8	8	24	268
	irules.tdl	90	270	540	5629
	lexicon.tdl	242	723	2894	24751
	rules.tdl	15	15	45	548
	roots.tdl	2	10	42	371
	portuguese-pet.tdl	0	22	50	438
	my-portuguese.tdl	179	487	2571	29023
	my-irules.tdl	135	424	1488	13754
	my-lexicon.tdl	198	606	2543	22736

Fonte: elaborada pelos autores.

Os arquivos 6-12 resultam do preenchimento do questionário de customização, refletindo particularidades gramaticais e lexicais do português. O arquivo 6 contém a maior parte das especificações gramaticais, os arquivos 7 e 8 se restringindo às regras lexicais. O arquivo 9 consiste de entradas lexicais.

Vejamos como se articulam as definições de tipos dos arquivos *martrix.tdl*, *portuguese.tdl* e *lexicon.tdl*. Por exemplo, a definição do tipo transitive-lex-item da Figura 6 abstrai da classe de palavra do item e da realização morfossintática dos argumentos, servindo como molde para qualquer item lexical biargumental. Desse

[7] Quantidades de linhas, palavras e caracteres computadas por meio da ferramenta wc do Unix, excluindo comentários e linhas em branco.

modo, pode ser utilizada como base para definir verbos transitivos canônicos tanto de línguas acusativas como o português, o alemão e o japonês quanto de línguas ergativas como o basco e o dirbal. Por outro lado, serve tanto para línguas que marcam os argumentos por meio de casos, como o alemão, quanto por meio de adposições, como o japonês.

Figura 6 – Definição do tipo *transitive-lex-item*.

```
transitive-lex-item := non-local-none-no-hcons & basic-icons-lex-item &
  [ ARG-ST < [ LOCAL [ CAT cat-sat,
                       CONT.HOOK [ INDEX ref-ind & #ind1,
                                   ICONS-KEY.IARG1 #clause ] ] ],
             [ LOCAL [ CAT cat-sat,
                       CONT.HOOK [ INDEX ref-ind & #ind2,
                                   ICONS-KEY.IARG1 #clause ] ] ]>,
    SYNSEM [ LKEYS.KEYREL [ ARG1 #ind1,
                            ARG2 #ind2 ],
             LOCAL.CONT.HOOK.CLAUSE-KEY #clause ] ].
```

Fonte: Imagem de parte do arquivo *matrix.tdl* no LKB.

Extrapolaria o âmbito deste artigo explicar detalhadamente toda a notação utilizada na definição do tipo da Figura 5 e de tipos análogos. É suficiente destacar que especifica a estrutura argumental (ARG-ST) do item como uma lista constituída de dois argumentos dotados de [INDEX ref-ind], ou seja, um índice referencial, próprio de signos nominais, por oposição a signos predicativos (BENDER; FLICKINGER; OEPEN, 2003). O valor do atributo HEAD, que especifica a classe de palavra de um dado núcleo, não é informado. Em vez disso, exige-se, por meio da especificação [LOCAL.CAT cat-sat] que o argumento constitua uma categoria saturada, ou seja, uma projeção máxima.

A partir das respostas que fornecemos no questionário de customização, o sistema gerou os tipos das Figuras 7-9. O primeiro é herdado tanto por verbos transitivos canônicos, como em (22), quanto não canônicos, como em (23), em que o objeto é marcado por uma preposição. Observe que esse tipo especifica a categoria de ambos os argumentos como *+np*, ou seja, preposição ou substantivo. O segundo tipo exige que o núcleo do primeiro argumento seja um substantivo no nominativo e o do

segundo, um substantivo no acusativo. Finalmente, a Figura 9 exibe a definição do tipo *trans-verb-lex*, herdado por verbos transitivos canônicos como *matar*, *amar* e *admirar*. Esses verbos herdam as especificações do tipo da Figura 8 e do tipo *noninh-refl-verb-lex*, que define a classe de verbos que não se submetem à regra lexical de afixação de um pronome reflexivo expletivo, de que trataremos mais adiante.

Figura 7 – Definição do tipo *transitive-verb-lex*.

```
transitive-verb-lex := main-verb-lex & transitive-lex-item &
  [ SYNSEM.LOCAL.CAT.VAL.COMPS < #comps >,
    ARG-ST < [ LOCAL.CAT.HEAD +np ],
             #comps &
             [ LOCAL.CAT cat-sat &
                         [ VAL [ SPR < >,
                                 COMPS < > ],
                           HEAD +np ] ] > ].
```

Fonte: Imagem de parte do arquivo *portuguese.tdl* no LKB.

Figura 8 – Definição do tipo *nom-acc-transitive-verb-lex*.

```
nom-acc-transitive-verb-lex := transitive-verb-lex &
  [ ARG-ST < [ LOCAL.CAT.HEAD noun &
                              [ CASE nom ] ],
             [ LOCAL.CAT.HEAD noun &
                              [ CASE acc ] ] >,
    SYNSEM.LOCAL.CAT.VAL [ SUBJ < [ LOCAL.CAT.HEAD.CASE-MARKED + ] >,
                           COMPS < [ LOCAL.CAT.HEAD.CASE-MARKED + ] > ] ].
```

Fonte: Imagem de parte do arquivo *portuguese.tdl* no LKB.

Figura 9 – Tipo *trans-verb-lex* e entradas lexicais com esse tipo.

```
trans-verb-lex := noninh-refl-verb-lex & nom-acc-transitive-verb-lex.
U:---  portuguese.tdl   17% L366   (TDL)

matar := trans-verb-lex &
  [ STEM < "matar" >,
    SYNSEM.LKEYS.KEYREL.PRED "_matar_v_rel" ].

amar := trans-verb-lex &
  [ STEM < "amar" >,
    SYNSEM.LKEYS.KEYREL.PRED "_amar_v_rel" ].

admirar := trans-verb-lex &
  [ STEM < "admirar" >,
    SYNSEM.LKEYS.KEYREL.PRED "_admirar_v_rel" ].

U:---  lexicon.tdl   36% L370   (TDL)
```

Fonte: Imagem de parte dos arquivos *portuguese.tdl* e *lexicon.tdl* no LKB.

(22) O gato matou uma ratazana.

(23) O cachorro obedece ao menino.

Os três últimos arquivos da Tabela 1 resultam de codificação manual. Alencar e Rademaker (2021, submetido à publicação) e Nunes, Rademaker e Alencar (2021) tratam das modificações manuais no terreno da morfologia flexional. Na seção 4, descrevemos as alterações em tipos gerados a partir do questionário de customização e os tipos completamente novos criados para suplantar as limitações do sistema no domínio da valência verbal.

3 O modelo da semântica de recursão mínima

Dentre os benefícios da análise sintática profunda, do tipo produzido por uma gramática no formalismo da HPSG, está a capacidade de produção de representações semânticas detalhadas, em paralelo e de forma composicional às análises sintáticas. Tais representações semânticas objetivam capturar o significado dos enunciados de forma canônica, abstraindo das diferentes variações sintáticas possíveis na linguagem natural. Considere as sentenças abaixo, todas expressam essencialmente o mesmo evento e os mesmos participantes, realizados sintaticamente de diferentes formas.

(24) Kim gave Sandy the green book.
Kim dar:PST;3SG Sandy o verde livro
'Kim deu o livro verde para Sandy.'
(25) Kim gave the green book to Sandy.
Kim dar:PST;3SG o verde livro para Sandy
'Kim deu o livro verde para Sandy.'
(26) The green book was given to Sandy by Kim.
o verde livro be:PST;3SG dar:PTCP para Sandy por Kim
'O livro verde foi dado para Sandy por Kim.'
(27) Sandy was given the green book by Kim.
Sandy be:PST;3SG dar:PTCP o verde livro por Kim

'Sandy ganhou o livro verde de Kim.'
(28) The green book, Kim gave Sandy.
o verde livro Kim dar:PST;3SG Sandy
'O livro verde Kim deu para Sandy.'

Para algumas aplicações de processamento automático de linguagem como extração de informações, inferência textual, resolução de perguntas, sumarização etc., é conveniente reconhecer que todas as sentenças expressam exatamente os mesmos eventos e que, a despeito de exercerem funções sintáticas distintas, o papel semântico das entidades envolvidas nos eventos é o mesmo. A lógica de primeira ordem (LPO) tem sido utilizada como linguagem de representação semântica, permitindo expressar a interpretação comum a diferentes construções sintáticas, como em (24)-(28). Por outro lado, permite representar as diferentes interpretações de uma mesma análise sintática de uma sentença ambígua como (29). Essa sentença possui duas possíveis interpretações em LPO, a saber (30a) e (30b). Note-se que a diferença de cada interpretação está principalmente na ordem dos quantificadores.

(29) All dogs chased a cat.
todos cães perseguir:PST;3PL um gato
'Todos os cães perseguiram um gato.'
(30) a. $\forall x\ (dog(x) \rightarrow \exists y(cat(y) \wedge chase(x, y)))$
b. $\exists y\ (cat(y) \wedge \forall x(dog(x) \rightarrow chase(x, y)))$
c. $\forall x\ dog(x) : \exists y\ cat(y) : chase(x, y)$
d. $\exists y\ cat(y) : \forall x\ dog(x) : chase(x, y)$

A representação em LPO, contudo, tem algumas limitações. A principal é não preservar a composicionalidade da estrutura sintática. Os dois sintagmas nominais quantificados de (29) não têm uma tradução para LPO independente, pois o sintagma nominal *all dogs* não pode ser associado a nenhuma subfórmula de (30a) ou (30b) isoladamente. Além disso, nas línguas naturais, uma variedade de outras expressões

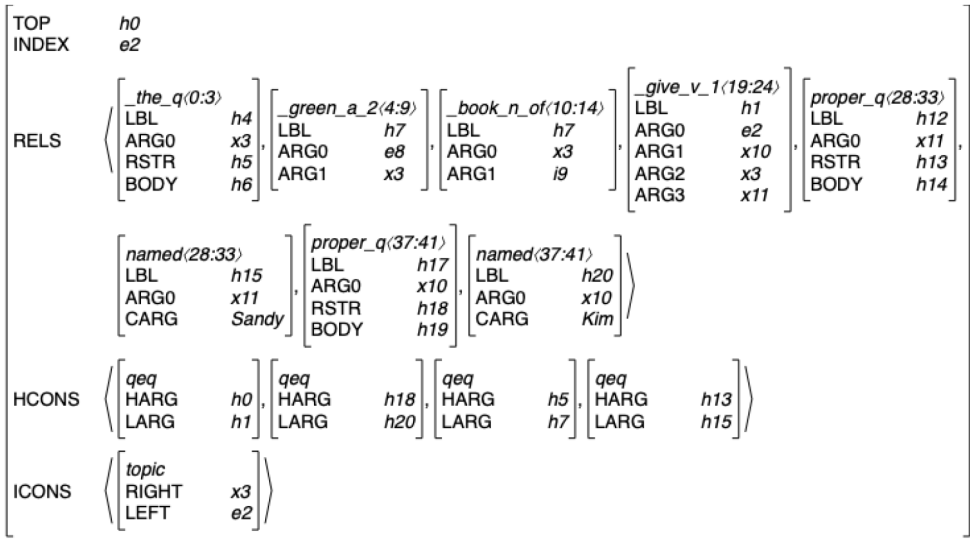
podem ser vistas como expressões quantificadoras, por exemplo, no inglês, *most* em (31), não representadas trivialmente com os quantificadores de LPO.

(31) Most dogs chased a cat.
a_maioria_dos cães perseguir:PST;3PL um gato
'A maioria dos cães perseguiu um gato.'

O que desejamos de uma representação semântica é um tratamento uniforme e composicional para expressões quantificadoras, incluindo as expressões não expressáveis com os quantificadores usuais de LPO. O uso dos quantificadores generalizados (WESTERSTÅHL, 2019), na forma de $Q\ x\ \alpha : \beta$, onde Q é um quantificador, x uma variável e α e β fórmulas em LPO, como em (30c,30d), constitui um passo nessa direção, oferecendo uma representação relacional dos quantificadores por meio de relações binárias entre subconjuntos do domínio. A expressão *all dogs*, por exemplo, denota o conjunto de todos os subconjuntos do domínio do qual todo cachorro é membro.

Finalmente, considere (32). A natureza binária da conjunção em LPO leva a uma ambiguidade espúria na representação, porque os possíveis agrupamentos (33a,33b) são irrelevantes para as condições de verdade do sintagma.

Figura 10 – MRS de (24)-(28).



Fonte: gerada pela ferramenta delphin-viz a partir da gramática ERG 2018 (UW).

(32) a furious white cat
um furioso branco gato
'um gato branco furioso'

(33) a. $\exists y(furious(x) \wedge (white(x) \wedge cat(y)))$
b. $\exists y((furious(x) \wedge white(x)) \wedge cat(y))$

As gramáticas produzidas pelo sistema *LinGO Grammar Matrix* adotam o formalismo da Semântica de Recursão Mínima (doravante MRS, do inglês *Minimal Recursion Semantics*) para representação semântica (COPESTAKE; LASCARIDES; FLICKINGER, 2001; BENDER; FLICKINGER; OEPEN, 2003; COPESTAKE *et al.*, 2005). A MRS não constitui uma teoria semântica, mas "uma linguagem de metanível para descrever estruturas semânticas em alguma linguagem de objeto subjacente" (COPESTAKE *et al.*, 2005, p. 282–283). Caracteriza-se por uma estrutura plana de predicações e por permitir subespecificação, isto é, que distinções semânticas permaneçam não resolvidas, de modo a permitir um raciocínio monotônico sobre tais representações semânticas parciais. A Figura 10 exibe a representação em MRS de (24)-(28).

A representação da Figura 10 é composta por 5 elementos principais: TOP, INDEX, RELS, HCONS e ICONS. No componente RELS temos a estrutura de predicados e argumentos expressa sob a forma de uma coleção de predicações (ou relações) n-árias vinculadas por variáveis (tipadas). Esta estrutura é caracterizada pela expressão "quem fez o que com quem" Apenas os signos linguísticos que contribuem com a semântica são representados, o verbo auxiliar da passiva, por exemplo, não contribuiu com nenhum predicado. Na figura, embora *book* (x3) seja o sujeito da passiva, corresponde ao ARG2 (elemento mais diretamente afetado pelo evento) do evento (e2) introduzido pelo predicado *_give_v_1*, cujo ARG1 (o causador volitivo de um evento) corresponde ao nome *Kim* (x9).

As predicações podem ser de superfície ou abstratas. Predicados de superfície seguem uma convenção de nomenclatura em que o símbolo é composto por três componentes separados pelo símbolo (_; U+005F): lema, POS (classe de palavra) e sentido. Por convenção, são marcados por um _ inicial e o campo de sentido é opcional. São introduzidos exclusivamente por entradas lexicais, cuja ortografia é uma forma (possivelmente flexionada) do campo lema no predicado. Na Figura 10, o predicado *_give_v_1* é um predicado de superfície.

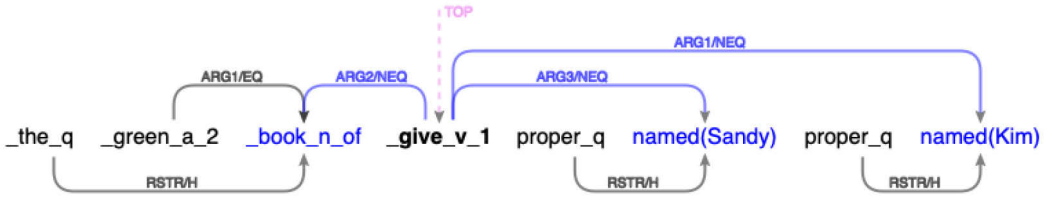
Predicados abstratos, por outro lado, são usados para representar a contribuição semântica de construções gramaticais ou entradas lexicais mais especializadas (como quantificadores implícitos, composição, nominalizações ou nomes próprios, no exemplo, *named*). Cada predicação é composta por: 1) uma relação ou predicado (por exemplo, _leave_v_1); 2) um rótulo LBL, cujo valor é sempre uma variável de escopo e servirá para agrupar predicações (*green* e *book* no exemplo); 3) um ARG0 com sua variável intrínseca; 3) ARG1. . . ARG4 correspondendo aos demais argumentos da relação, que podem ser preenchidos por variáveis sobre eventos (e), indivíduos (x), genéricas ou não realizadas (i ou u) ou de escopo (h, para predicados que recebem outras predicações como argumento).

Na representação da Figura 10, observamos três quantificadores generalizados que introduzem variáveis de indivíduos em seus ARG0 (instanciados pelo determinante *the* e dois determinantes implícitos para os nomes próprios). Estes quantificadores contêm dois argumentos RSTR (restrição) e BODY (escopo de aplicação).

Os componentes TOP e INDEX contêm como valores as variáveis h0 e e2, respectivamente. Em TOP temos a variável associada ao escopo mais externo ou abrangente da sentença e em INDEX o indivíduo principal da estrutura de predicados e argumentos, neste caso, o evento principal da sentença. Não iremos aqui detalhar o componente ICONS. Finalmente, o componente HCONS impõe restrições para os possíveis alinhamentos válidos para os quantificadores em uma possível resolução das subespecificações de escopo e ordem dos quantificadores. Por exemplo, o escopo h7, rótulo da predicação *book*, deve necessariamente estar embutido no escopo introduzido pela restrição (RSTR h5) do quantificador *the* que introduz a variável de indivíduo (x3), afinal, book predica sobre esta variável.

O formato de representação DMRS (COPESTAKE, 2009), exemplificado na Figura 11, constitui uma alternativa ao formato MRS na forma de um grafo livre de variáveis. De forma geral, as predicações são transformadas em nós e os argumentos, em arcos. As restrições de escopo (HCONS) são codificadas nos rótulos dos arcos, combinadas com os argumentos dos predicados (EQ para predicados com mesmo rótulo, no mesmo escopo, e NEQ para predicados em diferentes escopos) ou em arcos separados (RSTR/H).

Figura 11 – DMRS de (24)-(28).



Fonte: gerada pela ferramenta delphin-viz a partir da gramática ERG 2018 (UW).

4 Implementação da valência verbal na PorGram

A realização morfossintática de argumentos constitui um dos domínios gramaticais que mais variam de uma língua a outra. À diversidade tipológica soma-se uma grande variedade de padrões dentro de línguas específicas. O sistema de customização abrange um amplo leque de padrões em uma amostra bastante representativa de línguas das mais variadas famílias. Dada a complexidade dos fenômenos envolvidos, contudo, não poderia almejar a exaustividade de cobertura. Desse modo, a fim de tornar tratável, no âmbito do questionário de customização, a implementação das diferentes bibliotecas que interagem no tratamento computacional de sentenças com esses fenômenos, Drellishak (2009) limitou-se à realização dos argumentos de verbos intransitivos e de verbos transitivos como sintagmas nominais ou adposicionais. Verbos divalentes de alçamento para sujeito, cujo complemento é um verbo ou um sintagma verbal (34)-(36), classificados na *Grammar Matrix* como auxiliares, foram implementados posteriormente (POULSON, 2011).

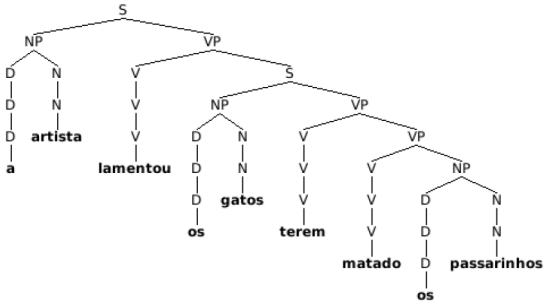
(34) Ele estava dormindo.
(35) Ele vai viajar.
(36) Ele tinha partido.

Analogamente, as biblioteca seguintes que permitiram a modelagem de complementos oracionais se restringiram a verbos divalentes (ZAMARAEVA; HOWELL; BENDER, 2019; ZAMARAEVA, 2021). Desse modo, pudemos, por meio do questionário, implementar sentenças como (37)-(42) com verbos regendo

complementos oracionais que expressam tanto proposições (37)-(39), quanto perguntas globais (41) e parciais (42). As Figuras 12 e 13 exemplificam os dois tipos de análise sintática gerados pela PorGram para as orações completivas de (37)-(41). Não obstante as discrepâncias estruturais entre as análises dessas duas figuras, ambos os exemplos compartilham a mesma representação semântica dependencial da Figura 14.

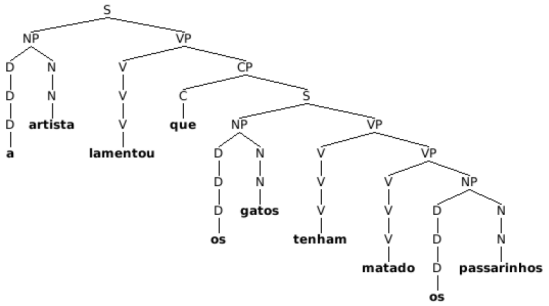
(37) Ela disse que ele estava dormindo.
(38) A Maria viu que as amigas choravam.
(39) A artista lamentou os gatos terem matado os passarinhos.
(40) A artista lamentou que os gatos tenham matado os passarinhos.
(41) Perguntei se ela estava triste.
(42) Ela perguntou quem tinha gritado.

Figura 12 – Árvore de (39).



Fonte: gerada pelo LKB a partir da PorGram.

Figura 13 – Árvore de (40).

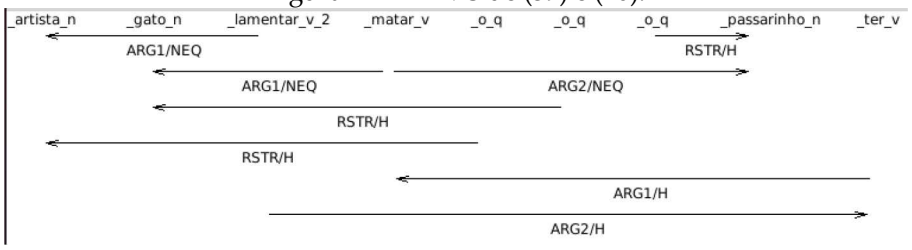


Fonte: gerada pelo LKB a partir da PorGram.

O questionário não contempla, portanto, sentenças com verbos trivalentes (43)-(47). Observe que, nesse último exemplo, temos um verbo de alçamento para objeto: o

sintagma os alunos ocupa a posição de objeto do verbo superior, mas constitui argumento semântico apenas do verbo encaixado (POLINSKY, 2013; ABEILLÉ, 2021).

Figura 14 – DRMS de (39) e (40).



Fonte: gerada pelo LKB a partir da PorGram.

(43) Ele doou uma bicicleta ao estudante.
(44) Ela contou ao estudante que o gato tinha morrido.
(45) Perguntei ao estudante se ele tinha dormido.
(46) Ela perguntou ao artista que gato o cachorro perseguiu.
(47) Ela fez os alunos chorar(em).

Uma outra limitação é que não permite a modelagem de estruturas de controle, como as exemplificadas abaixo:

(48) O artista tentou dormir.
(49) Ela prometeu ao pai estudar.
(50) Ela proibiu o artista de cantar.

Além disso, não oferece suporte para a implementação de verbos de alçamento para sujeito que regem complemento introduzido por complementador, exemplificados em (51)-(54), limitando-se àqueles que regem verbos ou sintagmas verbais nus como em (34)-(36).

(51) Ela tem que dormir.
(52) Temos de matar aquela ratazana.
(53) A estudante começou a sorrir.

(54) O cachorro tinha parado de latir.

Em que pesem essas limitações, o arquivo *matrix.tdl* inclui tipos abstratos para diversas outras classes valenciais, ainda não explorados pelo questionário de customização, que pudemos utilizar na codificação manual em TDL dos tipos necessários à implementação de (43)-(54), entre diversas outras construções.

A seguir, descrevemos, primeiro, as classes de valência implementadas por meio do questionário. Em seguida, tratamos das classes codificadas manualmente.

Na *LinGO Grammar Matrix*, os verbos se classificam em dois grandes grupos: verbos principais, subtipos de *main-verb-lex*, e verbos auxiliares, subtipos de *aux-lex*. Distinguem-se, primeiramente, pelos valores "+" e "-" do atributo AUX. Na PorGram, os auxiliares são verbos de alçamento para sujeito, constituindo subtipos de *subj-raise-aux*. Verbos de alçamento para sujeito, como em (55)-(59), têm como único argumento um VP não finito, não atribuindo, portanto, papel temático ao seu sujeito, que constitui argumento semântico do verbo encaixado.

(55) O gato tinha matado uma ratazana.
(56) O gato está perseguindo a ratazana.
(57) Ela precisa consertar o carro.
(58) Qualquer gato pode consumir esta ração.
(59) O artista começou a pintar o retrato.

Verbos de alçamento para sujeito distinguem-se de verbos de controle do sujeito, que ocorrem em construções superficialmente análogas, como em (60). Nesse caso, o sujeito da sentença é argumento semântico tanto do verbo matriz quanto do verbo encaixado. A diferença entre os dois tipos de verbos evidencia-se no contraste entre a gramaticalidade de (61) e (62), por um lado, e a agramaticalidade de (63), por outro (SAG; WASOW; BENDER, 2003, p. 376).

(60) O gato tentou agarrar a bola.
(61) O carro precisa ser consertado.
(62) O retrato começou a ser pintado pelo artista.
(63) *A bola tentou ser agarrada pelo gato.

Na *Grammar Matrix*, verbos auxiliares tanto podem ser dotados quanto desprovidos de um predicado próprio. Em português, integram o primeiro tipo, entre outros, verbos modais como *poder*, *dever* etc. e aspectuais como *começar* ou *parar* etc. No segundo tipo enquadram-se os auxiliares dos tempos compostos (55) e de locuções aspectuais (56). A existência ou não de dois predicados verbais em construções de verbo matriz e verbo encaixado reflete-se na possibilidade ou impossibilidade de modificação adverbial do verbo encaixado independentemente do verbo matriz, comparem-se os dois grupos de exemplos (64)-(69) e (70)-(73). No segundo grupo, o auxiliar não possui um predicado próprio que possa ser modificado pelo advérbio, contribuindo apenas com especificações de tempo, modo, pessoa e número para o significado da sentença, desempenhando, portanto, papel análogo ao da flexão verbal.

(64) Ele pode não ter chegado.
(65) Ele não pode ter chegado.
(66) O bebê parou de não querer comer.
(67) O bebê não parou de querer comer.
(68) Ela precisa muito trabalhar.
(69) Ela precisa trabalhar muito.
(70) Ele não tinha dormido.
(71) *Ele tinha não dormido.
(72) Ela não está trabalhando.
(73) *Ela está não trabalhando.

Numa versão anterior da PorGram, implementamos, por meio do questionário de customização, os dois tipos de auxiliares. No entanto, ao aplicar a gramática ao conjunto de teste, constatamos que sentenças como (74), com mais de um auxiliar,

sendo um deles desprovido de predicado, não eram analisadas, o que não ocorria com sentenças como (75), em que todos os auxiliares possuem um predicado.

(74) Ele está tentando trabalhar.
(75) Ele continua tentando trabalhar.

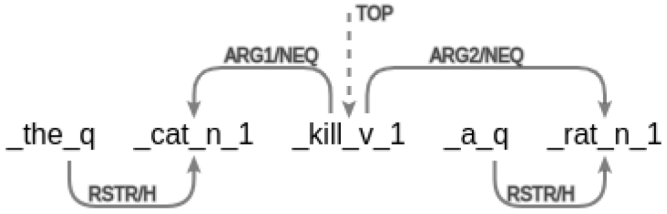
Essa assimetria decorre de uma limitação da *Grammar Matrix*, que não contempla a possibilidade de um auxiliar sem predicado tomar como complemento um VP nucleado por outro auxiliar (ZAMARAEVA, 2021, p. 320). Resolvemos provisoriamente esse problema, tal como Zamaraeva (2021) no seu fragmento de gramática da língua apinajé, atribuindo um predicado fictício (*dummy predicate*) a todos os auxiliares, indistintamente. Desse modo, o verbo auxiliar ter dos tempos compostos contribui com o predicado *_ter_v*, o verbo estar do progressivo com o predicado *_estar_v* etc. Reconhecemos que essa solução não é a ideal, uma vez que produz representações semânticas que discrepam estruturalmente das representações análogas produzidas pela ERG para exemplos do tipo de (76), sem que haja uma motivação linguística para tanto, como podemos constatar comparando as Figuras 15-18. No entanto, esses predicados fictícios podem ser assinalados de algum modo para que sejam ignorados no processamento semântico.

(77) the cat had killed a rat
o gato ter:PST;3SG matar:PTCP uma ratazana
o gato tinha matado uma ratazana
(78) the cat tried to catch the ball
o gato tentar:PST;3SG a agarrar:INF a bola
o gato tentou agarrar a bola

Na análise da Figura 15 para a tradução em inglês de (55), o auxiliar não contribui com nenhuma relação. No topo do gráfico dependencial, temos apenas o nó correspondente ao verbo principal, ou seja, *_kill_v_1*, de que constituem dependentes

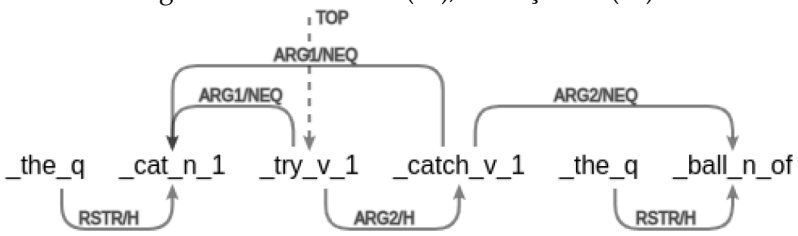
os nós correspondentes ao sujeito e objeto direto da sentença. Pelo contrário, o verbo matriz na análise da tradução de (60) na Figura 16 contribui com o predicado que ocupa o nó mais alto do gráfico dependencial, do qual o predicado correspondente ao verbo encaixado constitui o segundo argumento. Em ambas as análises da PorGram nas Figuras 17 e 18, o predicado do verbo matriz constitui o nó mais alto, quando somente deveria fazê-lo no segundo gráfico. Essa é uma deficiência que pretendemos sanar no futuro.

Figura 15 – Análise de (76), tradução de (55).



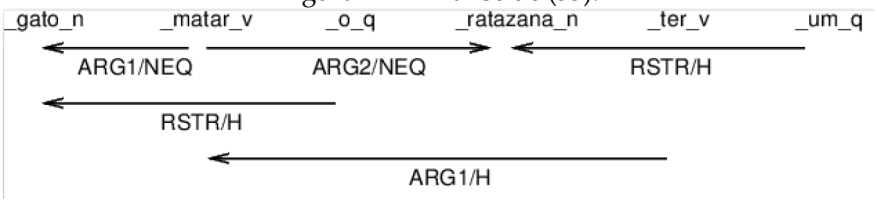
Fonte: gerada pela ferramenta delphin-viz a partir da gramática ERG 2018 (UW).

Figura 16 – Análise de (77), tradução de (60).



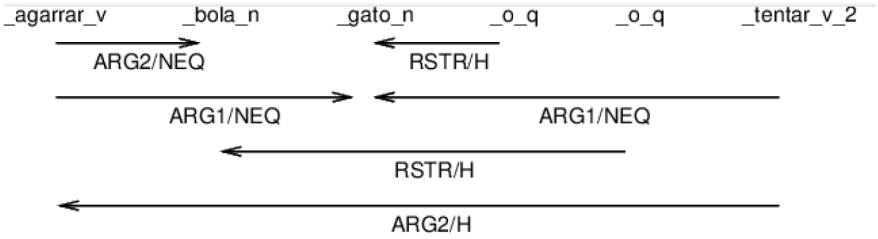
Fonte: gerada pela ferramenta delphin-viz a partir da gramática ERG 2018 (UW).

Figura 17 – Análise de (55).



Fonte: gerada pelo LKB a partir da PorGram.

Figura 18 – Análise de (60).



Fonte: gerada pelo LKB a partir da PorGram.

Os verbos principais implementados por meio do questionário dividem-se em dois grandes grupos: (i) verbos intransitivos e transitivos canônicos, i.e., verbos cujo segundo argumento realiza-se como um sintagma nominal ou clítico no acusativo; (ii) verbos com complemento oracional.

Utilizando o componente morfológico do sistema de customização, implementamos os verbos tradicionalmente chamados pronominais, incluindo tanto aqueles inerentemente reflexivos (78) quanto os facultativamente reflexivos (79). Por meio de uma regra lexical, esses verbos são sufixados com um reflexivo de forma obrigatória ou facultativa, dependendo do tipo do verbo. Essa regra não se aplica a verbos como *desaparecer* que não admitem o reflexivo expletivo (80).

(78) O artista queixou-se.
(79) A porta abriu(-se).
(80) *O gato desapareceu-se.

Quadro 1 – Estratégias de complementação oracional.

Número	Força iloc.	Compl.	Forma	Modo	Exemplos
1	proposição	que	finita	indicativo	(37)
2	proposição	(em|com|para) que	finita	subjuntivo	(84)-(88)
3	proposição	Ø	infinitivo flexionado	__	(39)
4	interrogação	se ou QU-	finita	indicativo	(41) (42)

Fonte: elaborado pelos autores.

O questionário permite especificar diferentes estratégias de complementação oracional verbal, área em que a língua portuguesa exibe grande variedade de padrões (MATEUS *et al.*, 1989, p. 265–279). O Quadro 1 sintetiza as estratégias implementadas por essa via. A segunda coluna especifica a força ilocucionária da oração completiva. A terceira especifica a forma do complementador, que pode ser foneticamente vazio ou consistir numa conjunção precedida ou não de uma preposição (*em*, *com* ou *para*) ou num elemento interrogativo QU-. A forma do verbo, na quarta coluna, pode ser finita ou o infinitivo flexionado. A estratégia 2 representa, na verdade, a consolidação de diversas estratégias, dependendo de cada verbo individual se o complementador pode ser precedido ou não por uma determinada preposição. Por meio da seção do léxico do questionário, utilizando essas estratégias, construímos tipos e entradas lexicais para todas as diáteses descritas por Mateus *et al.* (1989) envolvendo verbos divalentes cujo complemento é uma oração completiva, com as seguintes exceções: (i) verbos de controle; (ii) estruturas com inversão do sujeito (81), (iii) nominalizações (82) e (iv) construções com composição de predicados (83). Implementamos, também, a variante da construção causativa de (87) bem como a construção de (88), corriqueira em linguagem padrão, embora condenada por alguns puristas (MATOS, 2008; ARRAIS, 2017).

(81) Os professores acreditam terem os Centros recebido verba[8].
(82) As alunas lamentam o terem partido a jarra.
(83) Eu mandei escrever a carta aos alunos.
(84) Eu mandei que os alunos escrevessem a carta.
(85) O estudante insistiu em que a artista matasse a ratazana.
(86) O ruído fez que a criança chorasse.
(87) O estudante fez com que os cachorros perseguissem a ratazana.
(88) O estudante pediu para que a artista matasse a ratazana.

[8] Exemplos (81)-(84) extraídos de Mateus *et al*. (1989, p. 272-275).

Embora não descartemos que seja possível implementar as combinações de preposição e complementador de (85)-(88) de forma composicional por meio do questionário, optamos, na atual fase da PorGram, como mais simples, por codificar essas combinações sob a forma de locuções conjuncionais, ou seja, expressões de múltiplas palavras. No questionário, basta, para tanto, especificar *em que*, *com que*, *para que* etc. como lemas dos complementadores de diferentes estratégias de complementação oracional. Decerto se sacrifica, com isso, uma generalização linguística importante: verbos que licenciam uma oração completiva encabeçada por uma preposição via de regra também admitem um objeto preposicionado ou infinitivo com a mesma preposição, compare (89) e (90). No entanto, (91) e (92) mostram que há exceções. Por outro lado, (90) é uma estrutura de controle, não passível de implementação por meio do sistema de customização.

(89) Ela insistiu na viagem.
(90) Ela insistiu em viajar.
(91) Ela fez com que o filho viajasse.
(92) *Ela fez com a viagem do filho.

Por meio da codificação manual, recorrendo aos tipos abstratos definidos no arquivo *matrix.tdl*, implementamos entradas lexicais para todos os demais tipos de verbos que regem orações completivas descritos por Mateus *et al.* (1989), incluindo verbos de controle do sujeito (48) (49), do objeto direto (50) ou do objeto indireto (5) e verbos trivalentes com complementos realizados tanto por sintagmas nominais e preposicionais (43) quanto por orações (44), ressalvadas as exceções (ii)-(iv) mencionadas acima.

Também reformulamos manualmente as definições dos tipos *ind-cl-verb-lex* e *subj-cl-verb-lex*, referentes a verbos divalentes que regem orações completivas introduzidas pelo complementador que e exigem que o verbo encaixado esteja,

respectivamente, no modo indicativo e no subjuntivo. As definições geradas pelo sistema de customização estavam hipergerando, aceitando tanto sentenças gramaticais quanto suas contrapartes agramaticais, como em (93)-(96). A razão disso é que as estratégias de complementação oracional geradas automaticamente pelo sistema não permitem especificar o modo do verbo encaixado como uma propriedade lexical do verbo superior. Em vez disso, o verbo superior especifica um complementador que, por sua vez, determina o modo do seu complemento, ou seja, a oração encabeçada pelo verbo encaixado. O problema, nesse caso, é que um mesmo complementador licencia dois modos distintos, dependendo do verbo. Na reformulação manual, o verbo superior especifica o modo do verbo encaixado, permitindo que a gramática analise apenas os exemplos gramaticais de (93)-(96).

(93) Ela quer que ele durma.
(94) Ele viu que ela dormia.
(95) *Ela quer que ele dorme.
(96) *Ele viu que ela dormisse.

Implementamos, igualmente, de forma manual, verbos de alçamento para sujeito cujo complemento é encabeçado por um complementador, de modo a analisar sentenças com verbos modais e aspectuais como em (51)-(54). Para tanto, criamos tipos não só para esses verbos, mas também para os respectivos complementadores, os quais, com exceção de que com infinitivo do modal ter (variante mais informal da construção *ter de* + infinitivo), derivam de preposições. Como cada verbo exige um complementador específico, introduzimos o atributo COMP-FORM, cujos possíveis valores são tipos correspondentes às formas dessas preposições. Para analisar exemplos com o verbo ter modal de (51) e (52), criamos o tipo disjuntivo *que+de_comp* como supertipo de *que_comp* e *de_comp*. O verbo ter modal exige que a forma do

complementador do seu complemento seja esse supertipo, ao passo que um verbo como o aspectual parar *exige de_comp*.

O questionário possibilita incluir casos não canônicos para marcar argumentos nucleares, fenômeno comumente denominado *quirky* case na literatura anglofônica, como exemplificou Drellishak (2009) com um fragmento do alemão. Nessa língua, o objeto de verbos transitivos é canonicamente marcado com acusativo. No entanto, alguns verbos, como *helfen* 'ajudar', exigem um objeto no dativo. Drellishak (2009), porém, não implementou nenhum fragmento com a marcação não canônica por meio de adposições, também muito comum em alemão, em construções estruturalmente análogas a (97) e (98).

(97) Aquela artista depende de mim.
(98) O cachorro não obedece a mim.

Aparentemente, a implementação de sentenças como (97) e (98) por meio do questionário de customização é ou impossível ou excessivamente complexa. O sistema permite implementar tanto preposições marcadores de caso, i.e., categorias que constituem núcleos funcionais, semanticamente vazias, exercendo um papel meramente estrutural, quanto preposições que funcionam como núcleos lexicais, também chamadas preposições verdadeiras ou plenas (ZARING, 1991; MIOTO; SILVA; LOPES, 2005). Drellishak (2009) testou essa funcionalidade por meio de línguas em que todos os casos nucleares são realizados por adposições e línguas em que apenas um dos casos tem essa realização. No entanto, (97) e (98) não se enquadram em nenhuma dessas situações, uma vez que se trata de marcação não canônica.

A análise dos verbos (97) e (98) como verbos que regem caso conforma-se a análises de línguas como o francês, segundo as quais as preposições à e de marcam, respectivamente, os casos dativo e genitivo, exercendo função estruturalmente

análoga às flexões de caso correspondentes do latim (CARLIER; GOYENS; LAMIROY, 2013). A dificuldade da implementação desse tipo de análise via questionário decorre de que o caso expresso por uma adposição marcadora de caso é atribuído, também, ao seu complemento, como podemos constatar na Figura 19, onde o valor *#case* tanto do atributo *CASE* do núcleo adposicional quanto do núcleo nominal complemento indica identidade de valores. Com base nessa definição, portanto, se analisarmos a preposição *de* no exemplo (97) como marca de genitivo, o pronome *mim* será marcado igualmente com genitivo. Em (98), porém, esse mesmo pronome deverá estar no dativo. Essa duplicidade de marcação não nos parece razoável. Em vez disso, parece-nos mais sensato atribuir uniformemente o caso oblíquo à forma pronominal *mim* e às formas pronominais análogas em todas as construções em que funcionam como objeto de preposição (MIOTO; SILVA; LOPES, 2005), conforme o Quadro 2.

Quadro 2 – Pronomes pessoais no singular.

Pessoa	Nominativo	Acusativo	Dativo	Oblíquo
1	eu	me	me	mim
2	tu	te	te	ti
3	ele/ela	o/a	lhe	ele/ela

Fonte: elaborado pelos autores.

Desse modo, preferimos implementar manualmente as preposições marcadoras de caso e os verbos que regem complementos preposicionados. Para tanto, reformulamos a definição da Figura 19 como na Figura 20, onde o valor de CASE do núcleo adposicional e de seu complemento é o tipo case, pelo que não são necessariamente idênticos. Com base nesse tipo, criamos o subtipo *obl-case-marking-adp-lex*, cuja definição se encontra na parte inferior da Figura 20. As especificações desse subtipo, por sua vez, são herdadas pelas preposições marcadoras de caso, como veremos mais adiante.

Figura 19 – Tipo das adposições marcadoras de caso gerado pela *Grammar Matrix*.

```
case-marking-adp-lex := non-local-none-lex-item & raise-sem-lex-item &
  [ SYNSEM.LOCAL.CAT [ HEAD adp &
                             [ CASE #case,
                               MOD < > ],
                       VAL [ SPR < >,
                             SUBJ < >,
                             COMPS < #comps >,
                             SPEC < > ] ],
    ARG-ST < #comps &
             [ LOCAL.CAT [ VAL.SPR < >,
                           HEAD noun &
                                [ CASE #case,
                                  CASE-MARKED - ] ] ] > ].
```

Fonte: imagem de parte do arquivo *portuguese.tdl* no LKB.

Figura 20 – Codificação manual do tipo das adposições marcadoras de caso.

```
case-marking-adp-lex := non-local-none-lex-item & raise-sem-lex-item &
  [ SYNSEM.LOCAL.CAT [ HEAD adp &
                             [ CASE case,
                               MOD < > ],
                       VAL [ SPR < >,
                             SUBJ < >,
                             COMPS < #comps >,
                             SPEC < > ] ],
    ARG-ST < #comps &
             [ LOCAL.CAT [ VAL.SPR < >,
                           HEAD noun &
                                [ CASE case ] ] ] > ].

obl-case-marking-adp-lex := case-marking-adp-lex & [ ARG-ST.FIRST.LOCAL.CAT.HEAD.CASE obl ].
```

Fonte: imagem de parte do arquivo *my-portuguese.tdl* no LKB.

Verbos divalentes como obedecer e depender, que regem dativo e genitivo, respectivamente, herdam as propriedades dos tipos *dat-obj-verb-lex* e *gen-obj-verb-lex* da Figura 21. O tipo *prep-obj-verb-lex*, supertipo de ambos, constitui subtipo de *transitive-verb lex* da Figura 7, especificando adicionalmente que o objeto seja um sintagma adposicional. Enquanto *dat-obj-verb-lex* exige dativo, *gen-obj-verb-lex* requer genitivo. Ambos possuem subtipos com e sem um reflexivo expletivo, que se distinguem pelos prefixos *refl* e *nonrefl*.

Figura 21 – Alguns tipos de verbos de objeto preposicionado.

```
prep-obj-verb-lex := transitive-verb-lex &
                     [ SYNSEM.LOCAL.CAT.VAL.COMPS.FIRST.LOCAL.CAT.HEAD adp ].

dat-obj-verb-lex := prep-obj-verb-lex &
  [ SYNSEM.LOCAL.CAT.VAL.COMPS.FIRST.LOCAL.CAT.HEAD.CASE dat ].

gen-obj-verb-lex := prep-obj-verb-lex &
  [ SYNSEM.LOCAL.CAT.VAL.COMPS.FIRST.LOCAL.CAT.HEAD.CASE gen ].

nonrefl-gen-obj-verb-lex := noninh-refl-verb-lex & gen-obj-verb-lex.

refl-gen-obj-verb-lex := inh-refl-verb-lex & gen-obj-verb-lex.

nonrefl-dat-obj-verb-lex := noninh-refl-verb-lex & dat-obj-verb-lex.

refl-dat-obj-verb-lex := inh-refl-verb-lex & dat-obj-verb-lex.
```

Fonte: imagem de parte do arquivo *my-portuguese.tdl* no LKB.

Para verbos bitransitivos com complementos realizados por sintagmas nominais ou preposicionais, criamos vários subtipos do tipo abstrato *ditransitive-lex-item* do arquivo *matrix.tdl*. Verbos de transferência de posse (43), por exemplo, constituem instâncias do tipo *nom-acc-rec-ditrans*. Esse tipo especifica que os seus três argumentos, i.e., o sujeito e os complementos direto e indireto, sejam realizados por meio dos casos nominativo, acusativo e recipiente. Este último é uma espécie de supercaso, abarcando os casos dativo e alvo, realizados, respectivamente, pelas preposições *a* e *para*. Esse supercaso evita a necessidade de criar duas variantes verbais para cada verbo que participe da alternância exemplificada em (43), (44) e (49), por um lado, e (99)-(102), por outro. Verbos que não participam dessa alternância, como em (98) e (103), cujos complementos preposicionados devem ser encabeçado por *a* e *para*, instanciam tipos de verbos que exigem o caso dativo e o caso alvo, respectivamente.

(99) O artista doou uma bicicleta para o estudante.
(100) O artista contou para o estudante que o gato tinha matado uma ratazana.
(101) Perguntei para o estudante se ele tinha dormido.
(102) A artista prometeu para o estudante matar a ratazana.
(103) A artista desafiou o estudante para uma partida.

Também no campo dos verbos bitransitivos deparamos em português com uma grande diversidade de complementos oracionais, caracterizados por diferentes modos e formas verbais e uma ampla variedade de complementadores. Para analisar esses verbos, construímos diversos subtipos dos seguintes tipos abstratos do arquivo *matrix.tdl*: *ditransitive-lex-item*, *ditrans-first-arg-control-lex-item* e *ditrans-second-arg-control-lex-item*. Esses subtipos especificam as propriedades de verbos declarativos e de inquirição (44)-(46), de controle do sujeito (49), do objeto direto (16) e do objeto indireto (5).

O atributo COMP-FORM, criado para dar conta de exemplos como (51)-(54) com verbos de alçamento para sujeito, foi fundamental também para dar conta de verbos de controle cujo complemento oracional é encabeçado por complementador preposicional, como em (16) e (5), uma vez que cada um desses verbos exige um complementador específico. Utilizando esse atributo, implementamos entradas para outros complementadores preposicionais, como *para*.

Mateus *et al.* (1989) caracterizam como construções de controle apenas sentenças como (104)-(107) em que o verbo encaixado está no infinitivo não flexionado, cujo sujeito seria o pronome nulo anafórico PRO. Não haveria controle, contudo, dos sujeitos nulos de orações completivas no infinitivo flexionado ou numa forma finita, como em (108)-(115), não obstante serem correferentes do sujeito ou do objeto do verbo superior:

(104) Os artistas disseram ter ganho o festival.
(105) Nós acreditamos ter ganho o festival.
(106) As alunas lamentam ter partido a jarra.
(107) A Maria viu as amigas a chorar.
(108) As alunas lamentam terem partido a jarra.
(109) Os artistas disseram que ganharam o festival.
(110) Os artistas disseram terem ganho o festival.
(111) Nós acreditamos termos ganho o festival.

(112) Os alunos viram irem escorregar.
(113) Os alunos viram que iam escorregar.
(114) Eu autorizei os alunos a que escrevessem a carta.
(115) Eu autorizei os alunos a escreverem a carta.

Na PorGram, não restringimos a noção de controle a sentenças do tipo de (104)-(107), estendendo-a a alguns dos exemplos de (108)-(115). Esse segundo grupo de sentenças abriga estruturas superficialmente idênticas, porém heterogêneas do ponto de vista da valência verbal. Conforme Mateus *et al.* (1989), verbos declarativos, como *dizer*, e de atividade mental, como acreditar, ao contrário de verbos causativos, perceptivos e avaliativos de uso factivo, como *lamentar*, não licenciam orações completivas no infinitivo flexionado com sujeito na ordem canônica nos moldes de (39), exigindo a inversão exemplificada em (81). Esse contraste nos levou, por um lado, a atribuir a exemplos do tipo de (108) a mesma estrutura de (39), com a única diferença de que o sujeito da completiva é um pronome nulo *pro* (MIOTO; SILVA; LOPES, 2005, p. 245), não necessariamente correferente do sujeito do verbo matriz. Analogamente, os verbos superiores de (109) e (113) instanciam as mesmas variantes verbais de (37) e (38), respectivamente. Por outro lado, analisamos (110)-(112) como estruturas de controle, uma vez que não admitem a realização lexical do sujeito encaixado, necessariamente correferente do sujeito do verbo superior.

Também na análise de (114) e (115) divergimos de Mateus *et al.* (1989), que, embora reconheçam ser obrigatória a correferência nessa construção, não a tratam como estrutura de controle. Na PorGram, verbos como autorizar, em construções análogas a (114) e (115), instanciam os tipos *a-que-ditrans-second-arg-control-verb-lex* e *a-inf-ditrans-second-arg-control-verb-lex*. Ambos constituem subtipos do tipo *ditrans-second-arg-control-verb-lex*, que, como vimos na seção 2 a respeito de (16)-(18), codifica as propriedades comuns a todos os verbos trivalentes cujo segundo argumento

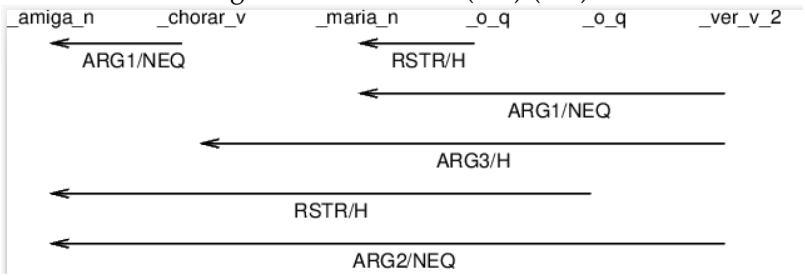
controla o sujeito da oração completiva que constitui o terceiro argumento desses verbos.

Para Mateus *et al.* (1989), subjaz a (116) estrutura paralela à de uma sentença como (39). Não se teria nesse exemplo, portanto, controle do objeto, uma vez que o sintagma as amigas não ocuparia a posição de objeto do verbo matriz, mas a de sujeito do verbo encaixado. Conforme essa abordagem, apenas haveria controle do objeto em (117), variante de (107) no português do Brasil. Por outro lado, (118), gramatical não só na variedade brasileira, mas também na europeia (GONÇALVES; CARRILHO; PEREIRA, 2016, p. 549), não é considerada por Mateus *et al.* (1989) plenamente aceitável no português europeu. Contrariamente a essa abordagem, analisamos *ver* em (116)-(118) como verbo de controle, dada a possibilidade de subida do clítico, ver (119)-(121)[9]. Na PorGram, a mesma variante de *ver* subjaz a (116)-(118), não obstante a variação de forma do verbo encaixado. Essa variante constitui instância do tipo *nonpast-nonfin-ditrans-second-arg-control-verb-lex*, que exige que o verbo encaixado tenha uma forma não finita e não passada, licenciando, portanto, tanto o gerúndio quanto o infinitivo flexionado ou não flexionado, mas excluindo o particípio passado. Essa análise, exemplificada nas Figuras 22 e 24, é adotada para todos os verbos perceptivos. Observe na Figura 22 que o predicado *ver_v_2*, atribuído à variante *ver_2*, possui três argumentos, correspondentes, respectivamente, aos sintagmas a *Maria* (ARG1), *as amigas* (ARG2) e *chorar* (ARG3) da Figura 24. Nas Figuras 23 e 25 temos as análises correspondentes da variante ver_1, que herda as propriedades do tipo *ind-cl-verb-lex*, próprio dos verbos que regem oração completiva no indicativo encabeçada por *que*. Como mostra a Figura 23, essa variante possui apenas dois argumentos. O sintagma *as amigas* constitui argumento apenas do verbo *chorar*.

[9] Para Abeillé (2021), a atribuição de caso em alemão permite enquadrar verbos perceptivos como de alçamento, questão que adiamos para uma próxima versão da gramática.

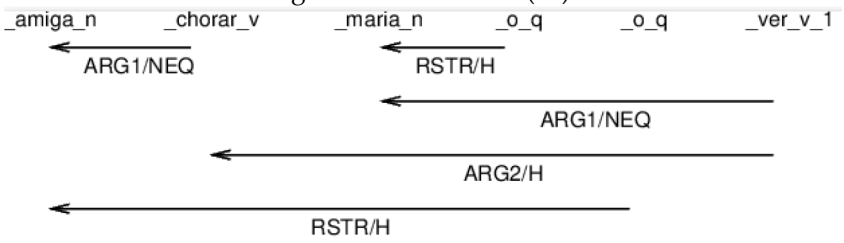
(116) A Maria viu as amigas chorarem.
(117) A Maria viu as amigas chorando.
(118) A Maria viu as amigas chorar.
(119) A Maria as viu chorarem.10
(120) A Maria as viu chorando.
(121) A Maria as viu chorar.

Figura 22 – DRMS de (116)-(118).



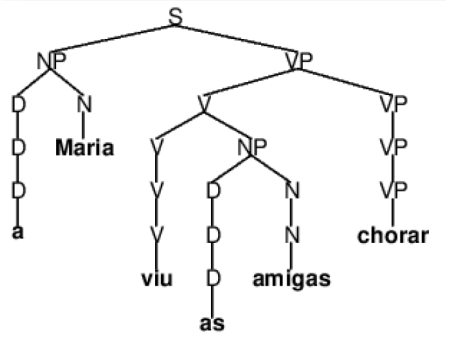
Fonte: gerada pelo LKB a partir da PorGram.

Figura 23 – DRMS de (38).



Fonte: gerada pelo LKB a partir da PorGram.

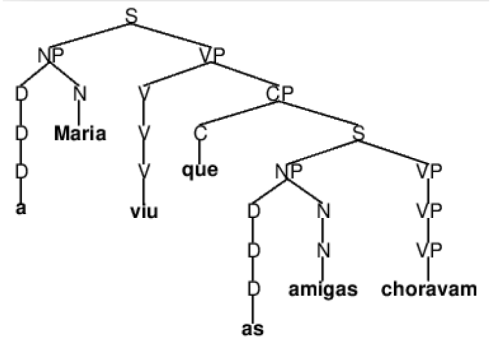
Figura 24 – Árvore de (118).



Fonte: gerada pelo LKB a partir da PorGram.

[10] Agramatical para Silveira (1997), essa construção ocorre na língua culta: "O intérprete da expedição os ouviu gritarem algo [...]." (SILVA; KOMISSAROV *et al.*, 1997)

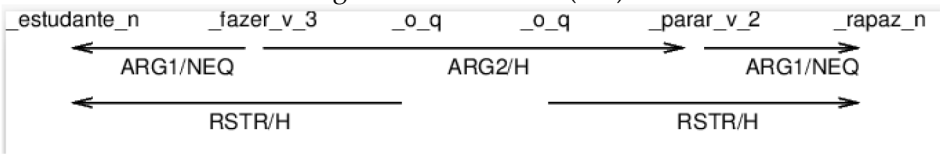
Figura 25 – Árvore de (38).



Fonte: gerada pelo LKB a partir da PorGram.

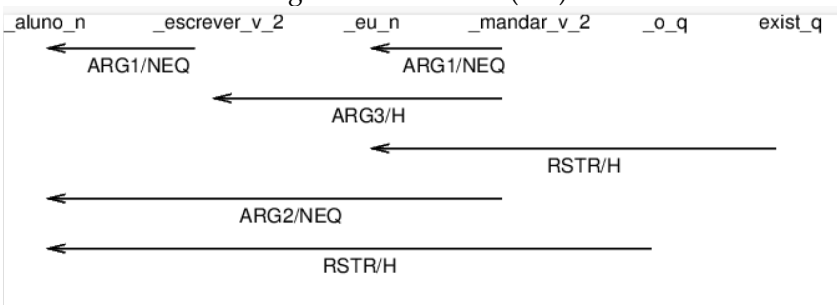
Outra divergência em relação a Mateus *et al.* (1989) consiste na análise dos verbos causativos, em exemplos como (122) e (123). Ambos os verbos licenciam a subida de clítico que realiza o agente do verbo encaixado em exemplos como (124). Ao que parece, esse exemplo seria agramatical para Mateus *et al.* (1989), que apresentam apenas o exemplo (131) com o agente do verbo encaixado no dativo. O caso acusativo do pronome em (124), ao nosso ver, evidencia que ocupa a posição de objeto do verbo superior. Não obstante essa semelhança superficial, as duas variantes verbais associam-se a tipos distintos na PorGram. Enquanto *mandar* no primeiro exemplo constitui um verbo de controle do objeto, *fazer* no segundo é um verbo de alçamento para objeto. Como podemos constatar nas Figuras 27 e 26, apenas no caso de *mandar* o objeto direto constitui argumento semântico do verbo superior. O contraste entre (126) e (127) mostra que *fazer*, ao contrário de *mandar*, não impõe restrições de seleção sobre o seu objeto (POLINSKY, 2013; ABEILLÉ, 2021).

Figura 26 – DRMS de (123).



Fonte: gerada pelo LKB a a partir da PorGram.

Figura 27: DRMS de (128).



Fonte: gerada pelo LKB a partir da PorGram.

Segundo Mateus *et al.* (1989, p. 275), verbos causativos como *mandar* e *fazer* "exprimem uma relação de causatividade entre um agente [...] e o estado de coisas descrito pela oração completiva", constituindo uma estrutura bioracional tanto em (84) e (86) quanto (122) e (123). Exemplos como (83), pelo contrário, teriam uma estrutura monooracional por conta da fusão do verbo superior e do encaixado em um predicado complexo. A flexão do infinitivo seria obrigatória nas variantes sem o complementador *que*, mas a sua ausência apenas reduziria a aceitabilidade da sentença, resultando em agramaticalidade apenas para alguns falantes (128). Ao que tudo indica, para Mateus *et al.* (1989), a subida de clítico somente é licenciada em estruturas monooracionais como (129)-(131). De fato, não apresentam exemplos análogos a (132) e (133), apesar de gramaticais não só no português brasileiro, mas também no europeu (GONÇALVES; CARRILHO; PEREIRA, 2016).

(122) Ela mandou os alunos escreverem uma carta.
(123) A estudante fez os rapazes parar.
(124) Ele a mandou comprar uma cerveja.
(125) O barulho os fez fugir.
(126) A cantora fez o chão tremer.
(127) *A cantora mandou o chão tremer.
(128) ?Eu mandei os alunos escrever.[11]

[11] Exemplos (128)-(131) extraídos de Mateus *et al.* (1989, p. 275-276), (132) e (133), de Gonçalves, Carrilho e Perira (2016, p. 551).

(129) Eu mandei-os escrever.
(130) Eu mandei-a escrever aos alunos.
(131) Eu mandei-lhes escrever a carta.
(132) A presidente da Assembleia mandou os deputados votar a lei.
(133) A presidente da Assembleia mandou-os votar a lei.

Na PorGram, tanto *fazer* quanto *mandar* aceitam as duas formas do infinitivo. Para dar conta dessa alternância, criamos o tipo disjuntivo *infl-or-not-inf*, supertipo dos tipos *inf* e *infl-inf*, correspondentes, respectivamente, ao infinitivo não flexionado e ao flexionado. Verbos como os modais e aspectuais constituem instâncias de tipos que especificam que a forma do complemento infinitivo é *inf*, ao passo que verbos como *lamentar* na variante de (39) herdam a especificação de que a forma da oração completiva é *infl-inf*.

Em suma, implementamos, por meio do questionário de customização da *Grammar Matrix*, não apenas os padrões canônicos de verbos monovalentes e divalentes, mas também verbos dessas duas superclasses com reflexivo expletivo e verbos que governam diferentes tipos de orações completivas. O sistema possibilitou, também, implementar diferentes tipos de verbos de alçamento para sujeito, classificados como auxiliares no quadro da *Grammar Matrix*, incluindo auxiliares de tempos compostos, modais e outros verbos que regem infinitivo, gerúndio e particípio. Por essa via, foram implementados, para verbos e auxiliares, 47 tipos, dos quais 27 são tipos com instâncias, que somam 140 entradas.

Por meio da exploração de tipos abstratos do arquivo *matrix.tdl*, superamos diversas limitações do questionário de customização. Em primeiro lugar, implementamos preposições como marcadores de caso, o que permitiu incluir no léxico verbos com objeto preposicionado. Em segundo lugar, implementamos verbos de alçamento para sujeito que regem complementos infinitivos encabeçados por complementadores, como os aspectuais *começar* e *parar*. Em terceiro lugar, ampliamos

a aridade dos verbos, incluindo verbos trivalentes com complementos realizados como sintagmas nominais, preposicionais ou oracionais, tanto finitos quanto não finitos, nus ou encabeçados por complementadores. Em quarto lugar, implementamos o controle não apenas do sujeito e do objeto direto, mas também do objeto indireto. Por último, implementamos verbos de alçamento para objeto. Esse esforço de codificação manual resultou em 71 novos tipos, dos quais 30 são tipos com instâncias. Ao todo, 138 entradas lexicais de verbos foram construídas por meio da manipulação direta do código em TDL, incluindo entradas cujos tipos foram codificados por meio do questionário. A maioria das classes valenciais implementadas possuem, além de uma ou mais entradas verbais, sentenças exemplificativas, constituindo, desse modo, um arcabouço para uma ampliação sistemática do léxico por meio da exploração de corpora e outros recursos disponíveis.

Os dicionários de valência mostram que a grande maioria dos verbos do português enquadra-se em múltiplas classes valenciais. No entanto, nossa preocupação principal na atual fase de desenvolvimento da PorGram não foi alcançar uma grande cobertura lexical. Desse modo, o léxico verbal da gramática contém tão somente 278 entradas para 215 lemas, a 80% dos quais foi atribuída uma única valência. Dos restantes, um total de 13% possui duas valências, e 6.5%, 3 ou 4.

O verbo *dizer* é o único com seis variantes, que instanciam seis tipos distintos, incluindo os exemplificadas em (5), (37), (110) e (134). Essa lista, contudo, não cobre toda a gama de construções documentadas em Fernandes (1987) e Borba (1991). Algumas dessas construções correspondem a tipos da PorGram, outras exigem a construção de novos tipos a partir dos existentes.

4 Avaliação

Nesta seção, apresentamos os resultados da avaliação da gramática, no que tange ao tratamento da valência verbal.

Na construção de gramáticas computacionais de línguas naturais, tipicamente se adota um ciclo de desenvolvimento com as seguintes etapas: (i) elaboração de um teste constituído de sentenças a serem analisadas pela gramática, (ii) implementação do código necessário para alcançar esse objetivo, (iii) aplicação da gramática ao teste, (iv) correção do código com base nos resultados, (v) reteste caso necessário, (vi) ampliação do teste e, finalmente, retorno à etapa (ii).

Como explicamos anteriormente, o código tem sido elaborado tanto automaticamente, por meio do sistema de customização, quanto manualmente por meio da edição em TDL. Em engenharia da gramática baseada no conhecimento, consolidou-se a utilização da técnica do fragmento (FRANCEZ; WINTNER, 2012; ZAMARAEVA, 2021). Em vez de se objetivar o esgotamento da análise de um dado fenômeno gramatical, define-se inicialmente um recorte que abarca apenas determinados aspectos. A implementação desse recorte visa a dois objetivos: (i) analisar sentenças que exemplificam os aspectos envolvidos, (ii) não analisar sentenças agramaticais que violam as restrições postuladas.

Para verificar em que medida os dois objetivos foram alcançados pela implementação, constroem-se dois conjuntos de teste: um conjunto de teste positivo, com sentenças gramaticais, e um conjunto de teste negativo, com sentenças agramaticais. As sentenças do conjunto de teste negativo derivam de sentenças do conjunto de teste positivo por meio da introdução de violações às restrições postuladas. Por exemplo, (139) contraria exigência relativa à forma do verbo encaixado na construção progressiva, observada em (138).

(138) Perguntei-me a quem o cachorro estava obedecendo.
(139) *Perguntei-me a quem o cachorro estava obedecido.

A PorGram, no estágio atual, resulta de dezenas de repetições do ciclo de desenvolvimento delineado. No LKB, não é necessário armazenar os dois conjuntos em arquivos separados, bastando prefixar as sentenças agramaticais com um asterisco. O *parser* ignora esse símbolo. Os atuais conjuntos de teste estão agrupados em dois arquivos, DEV-TEST e MAT-TEST. O primeiro arquivo foi construído incrementalmente, com o acréscimo de exemplos a cada repetição do ciclo. Constitui-se no momento de 631 exemplos, dos quais 139 são agramaticais.

O segundo arquivo contém 89 exemplos gramaticais e 28 agramaticais derivados das sentenças que exemplificam a descrição de Mateus *et al.* (1989) dos vários tipos de orações que funcionam como complementos verbais. Seguindo convenção de praxe, além do asterisco para marcar a agramaticalidade, Mateus *et al.* (1989) usam um ou dois sinais de interrogação para assinalar pouca aceitabilidade gramatical. Em gramáticas computacionais como a PorGram, porém, só é possível modelar julgamentos binários de gramaticalidade. Desse modo, convertemos todos os sinais de interrogação dos exemplos de Mateus *et al.* (1989) em asterisco, excetuando exemplos como (118) e (123), que, como vimos, são gramaticais.

Na adaptação das sentenças de Mateus *et al.* (1989), preservamos todos os aspectos estruturais relevantes para a avaliação da cobertura da PorGram em relação aos padrões de valência objeto da seção 4. No entanto, substituímos algumas construções, como, por exemplo, a voz passiva, e parte do vocabulário, dadas as presentes limitações da PorGram. Por outro lado, incluímos sentenças com a perífrase progressiva tanto na sua forma no português europeu padrão, quanto sua contraparte com gerúndio na variedade brasileira e em dialetos da europeia (GONÇALVES; CARRILHO; PEREIRA, 2016), comparem-se (107) e (117).

A Tabela 2 apresenta os resultados da aplicação da gramática sobre os dois arquivos de teste DEV-TEST e MAT-TEST, computando as quantidades de

verdadeiros negativos (TN), verdadeiros positivos (TP), falsos positivos (FP) e falsos negativos (FN). Os exemplos positivos consistem nas sentenças gramaticais, enquanto os negativos são as agramaticais. O índice R de cobertura (*recall*) da gramática em relação ao recorte gramatical e lexical pré-definido mede-se pela fórmula TP/(TP+FN), enquanto a precisão P é o resultado de TP/(TP+FP)). O índice FM, chamado Medida F (*F-Measure*), é a média harmônica de R e P (BIRD; KLEIN; LOPER, 2009). A hipergeração H da gramática, que consiste em analisar construções agramaticais como gramaticais, indicando que as restrições implementadas não são restritivas o suficiente, obtém-se pela fórmula FP/(TN+FP). Quanto menor esse índice, mais restritiva é a gramática.

Tabela 2 – Resultados da aplicação da gramática sobre os conjuntos de teste DEV-TEST e MAT-TEST.

	TN	TP	FP	FN	R	P	FM	H
DEV-TEST	130	480	9	12	0.98	0.98	0.98	0.06
MAT-TEST	21	66	7	23	0.74	0.90	0.81	0.25

Fonte: elaborada pelos autores.

Como podemos constatar na Tabela 2, a gramática alcançou uma alta cobertura em relação ao conjunto DEV-TEST, analisando 98% dos exemplos gramaticais, ao mesmo tempo que analisou apenas 6% dos exemplos agramaticais, resultando num índice FM de 0.98. A maior parte dos falsos positivos desse conjunto são sentenças declarativas com verbo não finito ou no subjuntivo. Esse problema se deve a que a *Grammar Matrix* não restringe o modo ou a forma verbal da sentença raiz. Quantos aos falsos negativos, (140) é o único exemplo referente à valência verbal. A regência de *insistir* nesse exemplo, consignada em Borba (1991), representa uma variação mínima da exemplificada em (85), codificada na PorGram como *insistir_3* com o tipo *em-que-cl-verb-lex*. Para que a gramática pudesse analisar (140) bastaria construir uma variante *insistir_5* com o tipo *para-que-cl-verb-lex*, analogamente a *pedir* em (88). De forma

alternativa, poderíamos implementar um tipo disjuntivo de forma de complementador que permitisse condensar as duas regências em uma única variante verbal, a exemplo do que fizemos para os verbos perceptivos.

(140) Ela insistiu para que o gato dormisse.

Do conjunto MAT-TEST, a PorGram analisou 74% dos exemplos gramaticais e 25% dos agramaticais, resultando num índice FM de 0.81. Ressalte-se que esse conjunto possui relativamente poucos exemplos agramaticais, uma vez que Mateus *et al.* (1989) não construíram sistematicamente, a partir das sentenças gramaticais, exemplos com desvios das regularidades modeladas, limitando-se a exemplos esparsos. Certamente, um maior número de exemplos agramaticais teria aumentado o índice FM e reduzido a hipergeração da gramática relativamente ao conjunto MAT-TEST.

Os falsos negativos do conjunto MAT-TEST exemplificam fenômenos ainda não implementados, incluindo particularidades do português europeu, como em (81)-(83) e (107).

As representações sintáticas e semânticas geradas pela PorGram para os falsos positivos do conjunto MAT-TEST revelam que se trata de sentenças talvez semanticamente anômalas, mas possíveis sintaticamente. A MRS de (141), por exemplo, apresenta a variante *dizer_v_2*, indicando que se trata do uso do verbo *dizer* como declarativo de ordem, ver (134).

(141) *Os críticos disseram que o filme ganhe o festival.

5 Considerações finais

Neste artigo, tratamos da implementação da valência na PorGram, uma nova gramática computacional do português no formalismo da HPSG. Ainda num estágio intermediário de desenvolvimento, a PorGram propõem-se a constituir, num futuro próximo, uma alternativa de software livre e de código aberto à LXGram. Também implementada em HPSG, mas com código proprietário, a LXGram é a única gramática do português de larga escala voltada à análise sintática profunda. Esse tipo de análise tem-se mostrado bastante relevante em tarefas de compreensão textual automática.

Tal como sua contraparte de código proprietário, a PorGram utilizou o questionário de customização da *LinGO Grammar Matrix* para construção automática do código em TDL de uma gramática inicial, modificado e expandido em etapas posteriores de desenvolvimento. Como outras gramáticas que compartilham a arquitetura da *LinGO Grammar Matrix*, a PorGram produz não apenas representações sintáticas, mas também semânticas, utilizando, para tanto, o formalismo da MRS, que supera diversas limitações da LPO.

Com esse sistema de customização, implementamos verbos intransitivos e transitivos prototípicos, ou seja, com sujeito no nominativo e complemento no acusativo, verbos com reflexivos inerentes bem como verbos divalentes que regem um amplo leque de tipos de orações completivas, caracterizados por diferentes combinações dos possíveis valores dos seguintes parâmetros: (i) força ilocucionária (proposição ou interrogação), (ii) forma do complementador (conjunção se ou pronome interrogativo, conjunção *que* precedida ou não de preposição ou foneticamente vazio), (iii) modo ou forma verbal (verbo no infinitivo flexionado, no modo indicativo ou no subjuntivo). O sistema permitiu igualmente implementar verbos de alçamento para sujeito que regem complementos no infinitivo, particípio ou gerúndio. Esses verbos incluem modais e auxiliares de tempos compostos e de perífrases aspectuais. Ao todo, 27 classes de verbos foram implementadas por essa via.

Por meio da manipulação direta do código em TDL, tomando como ponto de partida os tipos abstratos do arquivo *matrix.tdl* da *LinGO Grammar Matrix*, conseguimos superar diversas limitações do questionário de customização, que não contempla os seguintes fenômenos relacionados à valência verbal: (i) verbos com dois complementos (ou seja trivalentes), (ii) verbos de controle do sujeito e do objeto direto ou indireto, (iii) verbos de alçamento para objeto e (iv) verbos de alçamento para sujeito com complemento no infinitivo encabeçado por complementador. Para tanto, criamos diversos novos tipos de verbos e complementadores e redesenhamos a hierarquia de formas do verbo, de modo a dar conta da compatibilidades de determinados verbos, como os causativos e perceptivos, com mais de uma forma do verbo encaixado. Também implementamos manualmente preposições como marcadoras de caso e diversos tipos de verbos que exigem esse tipo de complemento, uma vez que a implementação desse fenômeno por meio do questionário se mostrou bastante difícil. Por outro lado, a maneira como o modo verbal das orações completivas introduzidas pelo complementador *que* é determinado na versão da gramática gerada pelo sistema revelou-se insuficientemente restritiva, deixando de bloquear exemplos agramaticais com o modo verbal incorreto. Isso nos levou a redefinir manualmente os tipos de verbos que exigem oração completiva encabeçada pelo complementador *que* num modo verbal específico. Ao todo, construímos de forma manual 71 tipos para codificação da valência verbal, dos quais 30 representam classes de verbos.

A PorGram foi testada em dois conjuntos de sentenças exemplificativas dos fenômenos abrangidos pelo recorte gramatical de que pretende constituir um modelo formal. Esses dois conjuntos de teste também contêm exemplos agramaticais que visam medir o quão restritivo é o modelo. O primeiro, com 492 sentenças gramaticais e 139 agramaticais, foi construído incrementalmente ao longo do desenvolvimento da gramática. No momento, são analisadas 480 das sentenças gramaticais e apenas 9 das construções agramaticais. Mostramos que o único falso negativo referente à valência

verbal poderia ser trivialmente corrigido incluindo uma variante do verbo *insistir* com um tipo já implementado na gramática.

Do segundo conjunto, derivado dos exemplos de Mateus *et al.* (1989), totalizando 117 exemplos, são analisados 74% das 89 sentenças gramaticais e 25% das consideradas agramaticais. A menor cobertura da PorGram em relação a esse conjunto resulta, sobretudo, dos diversos fenômenos que exemplifica não relacionados diretamente com a valência verbal, como, por exemplo, objetos clíticos, construções de clivagem e a perífrase progressiva do português europeu padrão, constituída de *estar a* e infinitivo. A formação de predicados complexos nessa variedade, porém, constitui fenômeno do domínio da valência não abarcado pela PorGram no momento, representando um dos desafios para uma versão futura da gramática. Quanto aos falsos positivos desse conjunto de teste, a um exame mais detalhado revelaram-se sintaticamente possíveis, embora semanticamente anômalos. Não indicam, portanto, hipergeração da gramática em sentido estrito.

A maior parte dos 57 tipos de verbos implementados, incluindo auxiliares, são representados por pelo menos uma entrada lexical e exemplificados com uma ou mais sentenças dos conjuntos de teste. Com 215 verbos codificados em 278 entradas lexicais, a PorGram cobre uma fração diminuta do léxico verbal português, do qual o número de lemas em circulação ativa na língua Borba (1991) estimaram em cerca de 6000, menos da metade dos verbetes de Fernandes (1987).

No entanto, a hierarquia de tipos proposta constitui uma infraestrutura para um léxico computacional de valências de uma profundidade gramatical que nos parece ímpar no domínio da língua portuguesa. De fato, modela a valência no domínio tanto da sentença simples quanto da sentença composta, distinguindo entre controle e alçamento. Pelo contrário, os dicionários de valência de grande cobertura disponíveis (FERNANDES, 1987; BORBA, 1991) não são formalizados, furtando-se a modelar as noções de controle e alçamento. Por outro lado, descrições mais aprofundadas de

grupos de verbos de determinados campos semânticos, como, por exemplo, verbos de mudança (CANÇADO; GODOY; AMARAL, 2013), restringem-se a sentenças simples, não abordando verbos com complementos oracionais. Analogamente, o projeto *Valências Verbais do Português Brasileiro*, ainda em andamento, não inclui esses padrões mais complexos de complementação (PERINI, 2016; PERINI *et al.*, 2019), controle e alçamento não integrando o quadro teórico subjacente a esse levantamento (PERINI, 2015).

Para concluir, destacamos os principais desafios que se apresentam para as próximas etapas de desenvolvimento da PorGram no terreno da valência verbal. Em primeiro lugar, falta construir tipos para verbos impessoais e verbos com sujeitos oracionais, adaptando às particularidades morfossintáticas do português os tipos abstratos correspondentes do arquivo nuclear *matrix.tdl* da *Grammar Matrix*. Em segundo lugar, é preciso ampliar o leque de descrições linguísticas dos padrões de complementação mais complexos a serem tomadas como base, visto que nos limitamos, no presente trabalho, à abordagem de Mateus *et al.* (1989), levando em conta julgamentos de aceitabilidade de Gonçalves, Carrilho e Pereira (2016) sobre algumas construções do português europeu. Em terceiro lugar, há que implementar fenômenos próprios do português europeu, como a formação de predicados complexos, a construção progressiva de *estar* com infinitivo encabeçado pelo complementador a ou a inversão do sujeito em orações completivas com infinitivo flexionado. Em quarto lugar, falta construir regras lexicais que modelem os processos sistemáticos de alteração da valência, de modo a poder analisar não somente sentenças na voz passiva, mas também simplificar a codificação do léxico. No momento, apenas uma regra lexical desse tipo está implementada, que é a adjunção de um reflexivo expletivo, aplicável de forma obrigatória ou facultativa sobre determinadas classes de verbos. Essa regra permite dar conta da variante incoativa de verbos como *abrir* por meio de uma única entrada lexical, ao invés das duas que seriam necessárias sem a

regra. Diversas variantes de verbos do léxico atual da PorGram poderiam ser geradas automaticamente por meio de regras análogas aplicadas sobre uma entrada lexical de base. Finalmente, é preciso povoar os tipos da gramática com instâncias, por meio da exploração sistemática dos recursos disponíveis, como *treebanks*, redes léxico-semânticas ou levantamentos das valências de classes semânticas individuais de verbos, como os mencionados.

Referências

ABEILLÉ, A. Control and raising. *In*: MÜLLER, S. *et al*. (ed.). **Head Driven Phrase Structure Grammar**: The handbook. Berlin: Language Science Press, 2021. p. 489–535.

ALENCAR, L. F. de; RADEMAKER, A. Cross-validating language resources for the development of a large-coverage computational grammar of Portuguese. **Language Resources and Evaluation**. Submetido à publicação.

ARRAIS, D. **Você sabe a diferença entre "pedir para" e "pedir que"?** 2017. Disponível em: https://exame.com/carreira/voce-sabe-a-diferenca-entre-pedir-para-e-pedir-que/. Acesso em: 21 nov. 2021.

BENDER, E. M. Reweaving a grammar for Wambaya: A case study in grammar engineering for linguistic hypothesis testing. **Linguistic Issues in Language Technology**, v. 3, p. 1–34, 2010. DOI https://doi.org/10.33011/lilt.v3i.1215

BENDER, E. M. et al. Grammar customization. **Research on Language & Computation**, v. 8, n. 1, p. 23–72, 2010. DOI https://doi.org/10.1007/s11168-010-9070-1

BENDER, E. M.; FLICKINGER, D.; OEPEN, S. The Grammar Matrix: An open-source starter-kit for the rapid development of cross-linguistically consistent broad-coverage precision grammars. *In*: **COLING-GEE '02**: Proceedings of the 2002 Workshop on Grammar Engineering and Evaluation. [S.l.]: [s.n.], 2002. p. 8–14. DOI https://doi.org/10.3115/1118783.1118785

BENDER, E. M.; FLICKINGER, D.; OEPEN, S. **MRS in the LinGO Grammar Matrix**: A practical user's guide. [S.l.]: [s.n.], 2003. Disponível em: http://faculty.washington.edu/ebender/papers/userguide.pdf. Acesso em: 25 set. 2021.

BENDER, E. M.; FLICKINGER, D.; OEPEN, S. Grammar engineering and linguistic hypothesis testing: Computational support for complexity in syntactic analysis. *In*: BENDER, E. M.; ARNOLD, J. E. (ed.). **Language from a cognitive perspective**: Grammar, usage and processing. Stanford: CSLI, 2011. p. 5–29.

BIRD, S.; KLEIN, E.; LOPER, E. **Natural language processing with Python**: analyzing text with the Natural Language Toolkit. Sebastopol: O'Reilly, 2009.

BORBA, F. da S. (org.). **Dicionário gramatical de verbos do português contemporâneo do Brasil**. 2. ed. São Paulo: Editora da UNESP, 1991.

CANÇADO, M.; AMARAL, L.; MEIRELLES, L. **VerboWeb**: classificação sintático-semântica dos verbos do português brasileiro. Belo Horizonte: UFMG, 2017. Disponível em: http://www.letras.ufmg.br/verboweb. Acesso em: 13 dez. 2021.

CANÇADO, M. et al. Banco de dados VerboWeb: um panorama do léxico verbal do PB. *In*: SIMPÓSIO BRASILEIRO DE TECNOLOGIA DA INFORMAÇÃO E DA LINGUAGEM HUMANA (STIL), 13 , 2021. Evento Online. **Anais** [...]. Porto Alegre: Sociedade Brasileira de Computação, 2021. p. 372-380. DOI https://doi.org/10.5753/stil.2021.17817

CANÇADO, M.; GODOY, L.; AMARAL, L. **Catálogo de verbos do português brasileiro**: Classificação verbal segundo a decomposição de predicados. vol. 1: Verbos de mudança. Belo Horizonte: Editora da UFMG, 2013.

CARLIER, A.; GOYENS, M.; LAMIROY, B. De: A genitive marker in french? *In*: CARLIER, A.; VERSTRAETE, J.-C. (ed.). **The genitive**. Amsterdam: John Benjamins, 2013. p. 141–216. DOI https://doi.org/10.1075/cagral.5.07car

COPESTAKE, A. **Implementing typed feature structure grammars**. Stanford: CSLI, 2002.

COPESTAKE, A. Slacker semantics: Why superficiality, dependency and avoidance of commitment can be the right way to go. *In*: CONFERENCE OF THE EUROPEAN CHAPTER OF THE ACL, 12 , 2009, Athens. **Proceedings** [...]. Athens: Association for Computational Linguistics, 2009. p. 1–9. DOI https://doi.org/10.3115/1609067.1609167

COPESTAKE, A. et al. Minimal Recursion Semantics: An introduction. **Research on language and computation**, Springer, v. 3, n. 2, p. 281–332, 2005. DOI https://doi.org/10.1007/s11168-006-6327-9

COPESTAKE, A.; LASCARIDES, A.; FLICKINGER, D. An algebra for semantic construction in constraint-based grammars. *In*: ANNUAL MEETING OF THE ASSOCIATION FOR COMPUTATIONAL LINGUISTICS, 39, Toulouse. **Proceedings** […]. Toulouse: Association for Computational Linguistics, 2001. p. 140–147. DOI https://doi.org/10.3115/1073012.1073031

COSTA, F.; BRANCO, A. LXGram: A deep linguistic processing grammar for Portuguese. *In*: PARDO, T. A. S. *et al*. (ed.). **Computational Processing of the Portuguese Language**. Berlin, Heidelberg: Springer, 2010. p. 86–89. DOI https://doi.org/10.1007/978-3-642-12320-7_11

CUNHA, C.; CINTRA, L. **Nova gramática do português contemporâneo**. Rio de Janeiro: Nova Fronteira, 1985.

CURTIS, C. **A Parametric implementation of valence-changing morphoplogy in the LinGO Grammar Matrix**. Dissertação (Mestrado) — University of Washington, Seattle, 2018. Disponível em: http://hdl.handle.net/1773/41814

DRELLISHAK, S. **Widespread but Not Universal**: Improving the typological coverage of the Grammar Matrix. Tese (Doutorado) — University of Washington, Seattle, 2009.

DROGANOVA, K.; ZEMAN, D. Towards deep Universal Dependencies. *In*: INTERNATIONAL CONFERENCE ON DEPENDENCY LINGUISTICS (DepLing), 5, 2019, Paris. **Proceedings** [...]. Paris: Association for Computational Linguistics, 2019. p. 144–152. DOI https://doi.org/10.18653/v1/W19-7717

FALK, Y. N. **Lexical-Functional Grammar**: An introduction to parallel constraint-based syntax. Stanford: CSLI, 2001.

FERNANDES, F. **Dicionário de verbos e regimes**. 35. ed. Rio de Janeiro: Globo, 1987.

FERRUCCI, D. et al. Building Watson: An overview of the DeepQA project. **AI Magazine**, v. 31, n. 3, p. 59–79, 2010. DOI https://doi.org/10.1609/aimag.v31i3.2303

FLICKINGER, D. On building a more efficient grammar by exploiting types. **Natural Language Engineering**, Cambridge University Press, v. 6, n. 1, p. 15–28, 2000. DOI https://doi.org/10.1017/S1351324900002370

FRANCEZ, N.; WINTNER, S. **Unification grammars**. Cambridge: Cambridge University Press, 2012.

GABRIEL, C.; MÜLLER, N. **Grundlagen der generativen Syntax**: Französisch, Italienisch, Spanisch. Tübingen: Niemeyer, 2008.

GONÇALVES, A.; CARRILHO, E.; PEREIRA, S. Predicados complexos numa perspetiva comparativa. *In*: MARTINS, A. M.; CARRILHO, E. (ed.). **Manual de linguística portuguesa**. Berlin: De Gruyter, 2016. p. 523–557. DOI https://doi.org/10.1515/9783110368840-022

GOODMAN, M. W. Generation of machine-readable morphological rules with human readable input. **University of Washington Working Papers in Linguistics**, v. 30, p. 1–34, 2013.

MARNEFFE, M.-C. de et al. Universal Dependencies. **Computational Linguistics**, v. 47, n. 2, p. 255–308, 2021.

MATEUS, M. H. M. *et al*. **Gramática da língua portuguesa**. Lisboa: Caminho, 1989.

MATOS, M. J. **"Pedir que" vs. "pedir para"**. 2008. Disponível em: https://ciberduvidas.iscte-iul.pt/consultorio/perguntas/pedir-que-vs-pedir-para24813. Acesso em: 11 nov. 2021.

MCCORD, M. C.; MURDOCK, J. W.; BOGURAEV, B. K. Deep parsing in Watson. **IBM Journal of research and development**, IBM, v. 56, n. 3.4, p. 3–1, 2012. DOI https://doi.org/10.1147/JRD.2012.2185409

MIOTO, C.; SILVA, M. C. F.; LOPES, R. E. V. **Novo manual de sintaxe**. 2. ed. Florianópolis: Insular, 2005.

MÜLLER, S. **Grammatical theory**: From transformational grammar to constraint-based approaches. 4. ed. Berlin: Language Science Press, 2020.

NIVRE, J. *et al*. Universal dependencies v2: An evergrowing multilingual treebank collection. *In*: LANGUAGE RESOURCES AND EVALUATION CONFERENCE, 12,

2020, Marseille. **Proceedings** […]. Marseille: European Language Resources Association, 2020. p. 4034–4043. Disponível em: https://aclanthology.org/2020.lrec-1.497. Acesso em: 29 dez. 2021.

NUNES, A. L.; RADEMAKER, A.; ALENCAR, L. F. de: Utilizando um dicionário morfológico para expandir a cobertura lexical de uma gramática do português no formalismo HPSG. *In*: SIMPÓSIO BRASILEIRO DE TECNOLOGIA DA INFORMAÇÃO E DA LINGUAGEM HUMANA (STIL), 13 , 2021. Evento Online. **Anais** [...]. Porto Alegre: Sociedade Brasileira de Computação, 2021. p. 11–18. DOI https://doi.org/10.5753/stil.2021.17779

PERINI, M. Construindo o Dicionário de Valências: problemas e resultados. **Scripta**, Belo Horizonte, v. 20, p. 148–167, 2016. DOI https://doi.org/10.5752/P.2358-3428.2016v20n38p148

PERINI, M. A. **Describing verb valency**: Practical and theoretical issues. Cham: Springer, 2015. DOI https://doi.org/10.1007/978-3-319-20985-2

PERINI, M. A. et al. **Valency dictionary of Brazilian Portuguese verbs**. Não publicado. 2019.

POLINSKY, M. Raising and control. *In*: DIKKEN, M. den (ed.). **The Cambridge handbook of generative syntax**: Grammar and syntax. Cambridge: Cambridge University Press, 2013. p. 577–606. DOI https://doi.org/10.1017/CBO9780511804571.021

POULSON, L. Meta-modeling of tense and aspect in a cross-linguistic grammar engineering platform. **University of Washington Working Papers in Linguistics**, v. 28, p. 1–67, 2011.

RADEMAKER, A. et al. Universal Dependencies for Portuguese. *In*: INTERNATIONAL CONFERENCE ON DEPENDENCY LINGUISTICS (DepLing), 4, 2017, Pisa. **Proceedings** […]. Pisa: Linköping University Electronic Press, 2017. p. 197–206.

ROSÉN, V. et al. An open infrastructure for advanced treebanking. *In*: HAJIČ, J. **META-RESEARCH Workshop on Advanced Treebanking at LREC2012**. [S.l.], 2012. p. 22–29.

SAG, I. A.; WASOW, T.; BENDER, E. M. **Syntactic theory**: A formal introduction. 2. ed. Stanford: CSLI, 2003.

SCHUSTER, S.; MANNING, C. D. Enhanced English Universal Dependencies: An improved representation for natural language understanding tasks. *In*: INTERNATIONAL CONFERENCE ON LANGUAGE RESOURCES AND EVALUATION (LREC), 2016, 10, Portorož. **Proceedings** [...]. Portorož: European Language Resources Association, 2016.

SIEGEL, M.; BENDER, E. M.; BOND, F. **Jacy**: An implemented grammar of Japanese. Stanford: CSLI, 2016.

SILVA, D. G. B. da; KOMISSAROV, B. N. et al. (org.). **Os diários de Langsdorff**. Campinas: Associação Internacional de Estudos Langsdorff, 1997. DOI https://doi.org/10.7476/9788575412459

SILVEIRA, G. **O comportamento sintático dos clíticos no português brasileiro**. Dissertação (Mestrado) — Universidade Federal de Santa Catarina, Florianópolis, 1997. Disponível em: https://repositorio.ufsc.br/handle/123456789/112183

STRAKA, M.; STRAKOVÁ, J. Tokenizing, POS tagging, lemmatizing and parsing UD 2.0 with UDPipe. *In*: **Proceedings of the CoNLL 2017 Shared Task**: Multilingual parsing from raw text to Universal Dependencies. Vancouver: Association for Computational Linguistics, 2017. p. 88–99. DOI https://doi.org/10.18653/v1/K17-3009

WESTERSTÅHL, D. Generalized quantifiers. *In*: ZALTA, E. N. (ed.). **The Stanford encyclopedia of philosophy**. Stanford: Stanford University, 2019.

ZAMARAEVA, O. **Assembling Syntax**: Modeling constituent questions in a grammar engineering framework. Tese (Doutorado) — University of Washington, Seattle, 2021. Disponível em: http://hdl.handle.net/1773/47087

ZAMARAEVA, O.; HOWELL, K.; BENDER, E. M. Modeling clausal complementation for a grammar engineering resource. *In*: SOCIETY FOR COMPUTATION IN LINGUISTICS (SciL), 2019, New York. **Proceedings** […]. [S.l.]: [s.n.], 2019. p. 39–49.

ZARING, L. On prepositions and case-marking in French. **Canadian Journal of Linguistics**, v. 36, p. 363–377, 1991. DOI https://doi.org/10.1017/S000841310001450X

Artigo recebido em: 29.12.2021 Artigo aprovado em: 28.02.2022