UNIVERSIDADE FEDERAL DE JUIZ DE FORA
FACULDADE DE LETRAS
PROGRAMA DE PÓS-GRADUAÇÃO EM LINGUÍSTICA

Bárbara Dias Santos

Modelagem construcional de anáforas pronominais na FrameNet Brasil: contribuições para o mapeamento computacional da referência

Juiz de Fora
2022

Bárbara Dias Santos

Modelagem construcional de anáforas pronominais na FrameNet Brasil: contribuições para o mapeamento computacional da referência

Dissertação de mestrado apresentada ao Programa de Pós-Graduação em Linguística da Universidade Federal de Juiz de Fora como requisito parcial à obtenção do título de Mestre em Linguística. Área de concentração: Linguística

Orientador: Prof. Dr. Tiago Timponi Torrent
Co-Orientador: Prof. Dr. Ely Edison da Silva Matos

Juiz de Fora
2022

Ficha catalográfica elaborada através do programa de geração
automática da Biblioteca Universitária da UFJF,
com os dados fornecidos pelo(a) autor(a)

Dias Santos, Bárbara .
Modelagem construcional de anáforas pronominais na FrameNet Brasil: contribuições para o mapeamento computacional da referência / Bárbara Dias Santos. -- 2022.
122 p. : il.

Orientador: Tiago Timponi Torrent
Coorientador: Ely Edison da Silva Matos
Dissertação (mestrado acadêmico) - Universidade Federal de Juiz de Fora, Faculdade de Letras. Programa de Pós-Graduação em Linguística, 2022.

1. Modelagem construcional. 2. Linguística Cognitiva. 3. Semântica de Frames. I. Timponi Torrent, Tiago, orient. II. Edison da Silva Matos, Ely, coorient. III. Título.

Bárbara Dias Santos

Modelagem construcional de anáforas pronominais na FrameNet Brasil: contribuições para o mapeamento computacional da referência

Dissertação apresentada ao Programa de Pós-Graduação em Linguística da Universidade Federal de Juiz de Fora como requisito parcial à obtenção do título de Mestre em linguística. Área de concentração: linguística

Aprovada em 16 de dezembro de 2022.

BANCA EXAMINADORA

Prof.(a)Dr(a). Tiago Timponi Torrent - Orientador
Universidade Federal de Juiz de Fora

Prof.(a)Dr(a). Ely Edison da Silva Matos
Universidade Federal de Juiz de Fora

Prof.(a)Dr(a). Patrícia Fabiane Amaral da Cunha Lacerda
Universidade Federal de Juiz de Fora

Prof.(a)Dr(a) Márcia Machado Vieira
Universidade Federal do Rio de Janeiro

Juiz de Fora, 05/12/2022.

Documento assinado eletronicamente por **Ely Edison da Silva Matos**, **Técnico Administrativo em Educação**, em 16/12/2022, às 15:38, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Tiago Timponi Torrent**, **Professor(a)**, em 08/02/2023, às 14:47, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Marcia dos Santos Machado Vieira**, **Usuário Externo**, em 09/02/2023, às 10:03, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Patricia Fabiane Amaral da Cunha Lacerda**, **Professor(a)**, em 11/02/2023, às 10:23, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020.

A autenticidade deste documento pode ser conferida no Portal do SEI-Ufjf (www2.ufjf.br/SEI) através do ícone Conferência de Documentos, informando o código verificador **1065507** e o código CRC **9F9814DF**.

À minha bebê Helena e minha avó Penha - meu eterno amor.

AGRADECIMENTOS

Agradeço, em primeiro lugar, a Deus que esteve comigo em todos os momentos e me deu forças para continuar nesta jornada tão desafiadora rumo ao conhecimento.

Agradeço à minha família por todo suporte, em especial à minha mãe, que mesmo sozinha, com tantas dificuldades, possibilitou que eu estudasse em outra cidade e me ajudou a realizar tantos sonhos, como este, o mestrado. Agradeço ainda ao meu amado esposo Samuel, por ter sido minha fortaleza quando eu achei que não conseguiria continuar. Você foi parte essencial neste trabalho.

Ainda quero declarar aqui todo o meu amor pela minha bebê Helena e minha avó Penha, que estiveram comigo fisicamente no início da minha jornada acadêmica, e hoje permanecem eternizadas em meu coração. Obrigada por tanto amor e pelos ensinamentos de vida que me proporcionaram, conhecimentos esses que vão muito além desta dissertação de mestrado.

Agradeço também à Universidade Federal de Juiz de Fora, por todo aprendizado acadêmico e humano. Obrigada pela oportunidade de seguir com os estudos e pelo apoio financeiro durante o mestrado. Agradeço também ao Programa de Pós-Graduação em Linguística por todo o suporte durante toda a minha formação acadêmica.

Agradeço a todos os amigos que fizeram e fazem parte da FrameNet Brasil, cada um contribuiu, de uma forma especial, para a minha formação acadêmica e pessoal. Tenho muito orgulho de fazer parte dessa equipe e levo as amizades que fiz no laboratório para toda a minha vida.

Agradeço especialmente ao meu orientador Tiago Torrent e à professora Natália Sigiliano pela paciência, pela amizade e por todo o apoio dedicados a mim. Ainda agradeço ao meu co-orientador Ely Matos pela generosidade e por todos os ensinamentos computacionais compartilhados comigo e com meus colegas desde a graduação. Este trabalho não seria possível sem vocês três.

E, finalmente, agradeço aos meus amigos pelos conselhos e por tantos sorrisos.

E ainda que tivesse o dom de profecia, e conhecesse todos os mistérios e toda a ciência, e ainda que tivesse toda a fé de maneira tal que transportasse os montes, e não tivesse amor, nada disso me aproveitaria. (I Coríntios, 13:2)

RESUMO

O trabalho apresentado nesta dissertação se insere nos estudos desenvolvidos pela FrameNet Brasil (FN-Br). Por meio dos pressupostos teóricos da Semântica de Frames (FILLMORE, 1982) e da Gramática de Construções de Berkeley (KAY; FILLMORE, 1999), este trabalho tem como objetivos (i) fornecer o tratamento linguístico-computacional das construções anafóricas pronominais no português brasileiro; (ii) discutir a aplicação do modelo do Constructicon em métodos de extração de informação. Para atingir esses objetivos, o trabalho conta com uma metodologia dividida em: apresentação do corpus de análise, cadastramento de construções, análise de ocorrências, modelagem e experimento de reconhecimento de construções. Assim, este trabalho propõe uma representação linguística e computacional das construções anafóricas pronominais demonstrativas, possessivas, reflexivas, relativas, indefinidas, nominativas, oblíquas e de tratamento. Por meio da análise de dados em dois *corpora General e Natural Language Generation*, constatamos a diversidade dos padrões anafóricos pronominais em língua portuguesa. A partir disso, modelamos computacionalmente as construções por tipo de anáfora e definimos os elementos que compõem a estrutura: o antecedente e o pronome. Depois, limitamos morfossintática e semanticamente os elementos que constituem as estruturas por meio de constraints baseados em Universal Dependencies (UDs), ordem de constituintes e frames. Por fim, implementamos um sistema de reconhecimento de construções anafóricas pronominais e de seus antecedentes a partir das construções modeladas no *Constructicon* e quantificamos os dados obtidos. Os resultados apontam para o mapeamento correto dos elementos antecedentes pela aplicação do modelo em contextos que o antecedente localiza-se próximo ao pronome anafórico. Por outro lado, o modelo não obteve o resultado esperado em contextos que os antecedentes são sintagmas nominais complexos, entidades nomeadas, elipses ou estão localizados fora do limite da sentença. Tais limitações estão relacionadas a três fatores: às deficiências do modelo propriamente dito, à base de dados da FN-Br que não processa entidades nomeadas e, por fim, ao funcionamento do parser UD - que processa itens lexicais em uma sentença por vez. Os resultados reforçam a necessidade de ampliação da cobertura do Constructicon da FN-Br.

Palavras-chave: Gramática de construções. Semântica de Frames. Processamento de língua natural. Anáfora. Extração de informação.

ABSTRACT

The work presented in this thesis is included in the FrameNet Brasil studies (FN Br). Through the theoretical basis of Frame Semantics (FILLMORE, 1982) and Berkeley Construction Grammar (KAY & FILLMORE, 1999), it aims to (i) provide the computational and linguistic treatment required to model Brazilian Portuguese anaphorical constructions (ii) discuss the application of the Constructicon model to information extraction methods. To achieve these goals, this work proposes a linguistic-computational representation of the demonstrative, possessive, reflexive, relative, undetermined, nominative, oblique and treatment anaphorical constructions. By analyzing data in corpora, we noted some diversity related to pronominal anaphora patterns in Portuguese language. From that, we modeled the constructions according to the types of anaphora and we defined the elements which compose the structure: the antecedent and the pronoun. Furthermore, we specified all the elements which are part of the structure using linguistic descriptions – constraints based on Universal Dependencies (UDs), constituent order and frames. At last, we implemented a construction recognition system for pronominal anaphoras and their antecedents using the constructions we modeled in the constructicon tool. On the one hand, the results suggest a correct mapping of antecedents by applying the model, on the other hand they suggest some exceptions related to mapping some types of antecedents, such as complex noun phrases, named entities, ellipses or cases when they are located outside the sentence boundary. Those exceptions are related to three reasons: deficiencies related to the anaphora model itself, restrictions in the FrameNet database, which can't process named entities, and, lastly, restrictions in the UD parser, which processes lexical items in one sentence each time. The results emphasize the need to extend the coverage of the FN-Br Constructicon.

Keywords: Construction Grammar. Frame Semantics. Anaphora. Constructicon. Information Extraction.

SUMÁRIO

1 INTRODUÇÃO

Esta pesquisa se enquadra nas discussões e nos projetos desenvolvidos no laboratório FrameNet Brasil (FN-Br), no que tange ao processamento de língua natural e às suas aplicações. Nesse contexto, o trabalho procura desenvolver um modelo linguístico-computacional que interprete estruturas de dependência de longa distância, em específico, as anáforas pronominais.

A FN-Br é um laboratório que procura solucionar problemas de cunho linguístico-computacional diversos. As ferramentas de anotação textual disponibilizadas pela FN-Br não codificam certos padrões linguísticos como, por exemplo, as construções anafóricas. Isso acontece porque ainda não existe uma estrutura que mapeie linguístico-computacionalmente tais construções. Mediante isso, espera-se que a modelagem de uma construção de anáfora pronominal contribua para a recuperação de informações no modelo da FrameNet Brasil. Diante desse desafio, a pergunta que motiva esta dissertação é:

Como desenvolver e modelar construções anafóricas de forma que elas contribuam para a otimização de tarefas de compreensão de língua natural por máquina?

Diante dessa questão, este pesquisa tem como objetivos (i) fornecer a modelagem linguístico-computacional necessária para as construções anafóricas pronominais do PB na base construcional da FN-Br; (ii) implementar a construção anafórica no constructicon da FN-Br; (iii) discutir a implicação dessa modelagem em mecanismos de extração de informações através de um experimento computacional de identificação dos antecedentes de anáforas pronominais. Assim, a hipótese a ser testada neste trabalho pode ser formulada da seguinte forma:

A modelagem de construções anafóricas pronominais no Constructicon da FrameNet Brasil contribui para o processo de extração de informações em textos através da identificação dos referentes das anáforas pronominais.

Considerando os objetivos e hipótese propostos acima, esta investigação busca embasamento teórico na Semântica de Frames (FILLMORE, 1982), na Gramática das Construções de Berkeley, em especial no que concerne à satisfação de requisitos de valência (KAY; FILLMORE, 1999). Segundo essa proposta, as construções anafóricas funcionam segundo princípios de satisfação de valência entre o antecedente e o elemento anafórico fora da localidade sintática do antecedente via autorização.

A discussão proposta neste trabalho tem como objetivo entender e modelar as construções anafóricas pronominais no Constructicon da FrameNet Brasil para aplicação em ferramentas de extração de informações. Para isso, descrevemos nove construções anafóricas da família das construções anafóricas pronominais com todos os tipos de pronomes que compõem essa classe, com o intuito de modelar linguístico-computacionalmente por completo a categoria.

A partir da modelagem dessas estruturas, compreendemos que algumas dessas construções podem ser aplicadas em tarefas de localização de informações em textos, visando ao aprimoramento dessas ferramentas. Por outro lado, algumas sentenças demonstram limitações relacionadas à aplicação do modelo. Esses problemas se referem às restrições do parser UD, que funciona por meio da análise de uma sentença por vez, e do modelo de construções anafóricas respectivamente. Observamos a predominância em percentual de erros no experimento relacionados à localização de antecedentes fora do escopo da sentença (cerca de 30,15 %), à identificação dos antecedentes no frame de entidade (15,99%), ao mapeamento de variados nomes e pronomes no escopo da mesma sentença (15,73%) e, por fim, à localização de entidades nomeadas (14,72%), como será explicado com mais detalhes nos capítulos finais deste trabalho.

Para o estudo do fenômeno, esta pesquisa está organizado nos seguintes capítulos: no capítulo um, apresentamos a introdução deste trabalho de pesquisa. No capítulo dois, apresentamos os pressupostos teóricos e o contexto de desenvolvimento

deste trabalho, como a Gramática de construções de Berkeley (1999), o tratamento de fenômeno de referência de longa distância, a anáfora, e o recurso computacional desenvolvido na FrameNet Brasil, o constructicon. No capítulo três, descrevemos o campo de aplicação, nomeadamente, as técnicas utilizadas em tarefas de extração de informações. O capítulo quatro consiste na apresentação da descrição dos mecanismos anafóricos que ocorrem no português brasileiro (PB), segundo a teoria linguística. Nele, o foco recai sobre os processos anafóricos pronominais. No capítulo cinco, descrevemos a metodologia aplicada para a realização desta dissertação. O capítulo seis traz a modelagem das construções anafóricas. No capítulo sete, discutimos sobre a implementação das construções em tarefas de extração de informação e as implicações desse experimento para métodos de extração de informação. Por fim, apresentamos as conclusões resultantes ao longo de toda a pesquisa.

2 GRAMÁTICA DE CONSTRUÇÕES

Nesta seção, encontram-se os pressupostos teóricos referentes à abordagem conhecida como Gramática de Construções de Berkeley. Para apresentar esse modelo, discorremos sobre os princípios gerais que regem o paradigma conhecido como Gramática de Construções, passaremos pela Gramática de Construções de Berkeley, e, por fim, apresentaremos o tratamento do fenômeno de referência anafórica dentro do modelo. Dessa forma, pretendemos delimitar o modelo que é usado nesta pesquisa, o da Gramática de Construções Baseada em Unificação (KAY; FILLMORE, 1999).

2.1 A GRAMÁTICA DE CONSTRUÇÕES DE BERKELEY

Nesta seção, abordaremos as motivações para o surgimento do modelo da Gramática de Construções e o princípio de funcionamento do modelo. Ainda discutiremos os impactos que a gramática de construções proposta por Fillmore e Kay (1999) geram neste trabalho, principalmente no que se refere à implementação computacional do fenômeno em análise: as construções anafóricas pronominais.

Historicamente, a Gramática de Construções é um modelo que surgiu em conjunto com estudos dissidentes em relação à Teoria Gerativa, em especial, ao modelo padrão definido na obra *Aspects of the Theory of Syntax* (CHOMSKY, 1965) e suas derivações ao longo das décadas de 60 e 70. Ela faz parte daquilo que os linguistas chamam de virada construcional dos estudos de gramática (SALOMÃO, 2002). O paradigma da Gramática das Construções está inserido no que hoje a linguística intitula como Linguística Cognitiva (LC). A LC é considerada um programa de pesquisa flexível, composta por diferentes abordagens como, por exemplo, a Semântica de Frames (FILLMORE, 1982), a Teoria dos Espaços Mentais (FAUCONNIER, 1994), a Teoria das Metáforas Conceptuais (LAKOFF, 1979), entre outras.

Entre meados de 1980 e 1990, autores como Fillmore e Kay voltaram-se para fenômenos que eram considerados periféricos para o modelo de gramática da teoria gerativa, como, por exemplo, expressões idiomáticas. De acordo com esses autores,

esses fenômenos mobilizam estruturas gerais da gramática, ou seja, do funcionamento da língua como um aparato cognitivo e, por isso, deveriam ser tratados de forma central no domínio dos estudos gramaticais. Em todo o paradigma cognitivista, mais especificamente na abordagem da Gramática de Construções de Berkeley, o funcionamento da gramática, assim como as operações cognitivas do aparato humano, passaram a ser concebidas de maneira unificada. Dessa forma, não existem elementos com mais ou menos valor, mais centrais ou periféricos na Gramática de Construções.

O paradigma da Gramática de Construções se estendeu muito ao longo dos anos. Hoje existe um grande número de variações do modelo. Entre elas, encontram-se a Gramática de Construções de Berkeley – BCG – (KAY; FILLMORE, 1999), a Gramática Cognitivista de Construções – CCxG – (GOLDBERG, 1995; 2006), a Gramática de Construções Baseada em Signos – SBCG – (SAG, 2012), entre outras. Todas essas abordagens mantêm aspectos em comum e divergem em algumas questões.

Sob a perspectiva do paradigma construcionista, assume-se que as construções se constituem como pareamentos de forma e sentido que precisam ser aprendidos pelos falantes no processo de aquisição da linguagem. Elas são consideradas a unidade básica da língua e nelas os falantes apoiam-se para construir expressões linguísticas (FRIED; OSTMAN, 2004). As abordagens construcionais também negam a hipótese forte de composicionalidade da língua, defendida pela teoria gerativa, segundo a qual o somatório dos elementos que compõem uma estrutura seria o seu significado geral. Além disso, todas elas entendem a gramática como uma rede de construções que se relacionam entre si por herança, o que significa que compartilham propriedades diversas. O paradigma também defende a continuidade entre o léxico e a sintaxe, ou seja, existem princípios operacionais comuns entre esses dois elementos da gramática e, por isso, eles podem ser tratados com o mesmo aparato teórico-metodológico. As palavras passam a ser entendidas como construções menos complexas, mas que operam com o mesmo princípio de funcionamento de uma estrutura argumental mais complexa.

O papel do léxico no paradigma construcionista ganha um novo status, isto é, ele passa a fazer parte de todas as operações no interior da gramática, diferindo muito

do papel antes desempenhado no modelo gerativista, já que era garantido ao léxico um vasto número de operações de entradas lexicais.

Por último, o paradigma assume que as análises linguísticas são não-derivacionais, isso significa que, para esse modelo, não existem regras de transformação de uma forma para outra, uma construção não deriva de outra. Essa premissa aponta para a não existência de uma estrutura profunda. Mediante isso, o pareamento entre a forma de superfície e a estrutura semântica é tomado como direto, não havendo regras de transformação.

Diferentemente do modelo gerativo de língua vigente até então, a Gramática de Construções garante à semântica um papel principal nas operações do sistema da língua. O modelo também defende a continuidade entre a semântica e a pragmática e, assim, incorpora elementos de natureza pragmática na análise, como aspectos de topicalização, registro e polidez. Nas palavras de Fried e Ostman (2004, p. 12)

> Consequentemente, a Gramática de Construções enxerga as unidades linguísticas como associações particulares entre a forma e o sentido que precisa ser representada como tal, ao invés de deixar apenas tais associações com a operação de um conjunto de regras para a qual se combinam formas individuais.[1]

Passando-se às características de modelos específicos da Gramática de Construções, este trabalho elege o modelo conhecido como Gramática de Construções de Berkeley.[2] Nesse tipo de abordagem, as estruturas linguísticas, as construções, são descritas por meio de diagramas e as relações morfossintáticas e semânticas são apresentadas de forma mais interpretável e operacional para implementação computacional.

Dentro do paradigma da gramática das construções, a abordagem da Semântica de Frames (FILLMORE, 1982) é amplamente incorporada para a definição das

---

[1] Cf. Consequently, Construction Grammar sees linguistic units as particular associations between form and meaning that must be represented as such, rather than leaving such associations to the operation of a set of rules for how to combine individual forms.

[2] Desenvolvida por Goldberg (1995), a CCxG é talvez a abordagem construcionista mais utilizada nos trabalhos desenvolvidos sobre o Português do Brasil. Entretanto, é considerada menos formalista. Isso significa que ela é mais voltada para a língua em seu contexto de uso. Por isso, daremos menos enfoque para essa abordagem neste trabalho.

construções. O termo frame foi definido, no âmbito da semântica, por Fillmore (1982), da seguinte forma:

> Pelo termo “frame” eu tenho em mente qualquer sistema de conceitos relacionados de tal forma que para entender um dos elementos é necessário entender toda a estrutura na qual ele faz parte; quando uma das coisas nessa estrutura é introduzida em um texto, ou em uma conversa, todas as outras são automaticamente acionadas. (FILLMORE, 1982)[3]

Por meio da noção de frame defendida por Fillmore (1982), as construções são desenvolvidas. Nesse sentido, o frame estaria ancorado na estrutura da construção, isso indica que cada estrutura ativa determinados sistemas de conhecimento. Em resumo, as estruturas ativam determinados sistemas de conhecimento (os Frames) que, por sua vez, são formados por elementos que são ativados quando a construção está em uso na língua. Todo o sistema torna-se ativo quando a construção é usada.

Para entender melhor, o processo acontece da seguinte forma: a construção aciona um frame mais genérico e os demais itens lexicais que compõem a estrutura evocam frames mais específicos. Veremos essa situação com mais detalhes nos próximos capítulos, quando a família de construções da anáfora pronominal é apresentada. Essa perspectiva não é amplamente adotada por todos os vieses do paradigma construcionista. Para as gramáticas de construções baseadas em unificação como a BCG e a SBCG, nem todas as construções evocam frames. Por outro lado, para a CCxG, qualquer construção evoca um frame, mesmo que ele seja altamente abstrato (GOLDBERG, 2006).

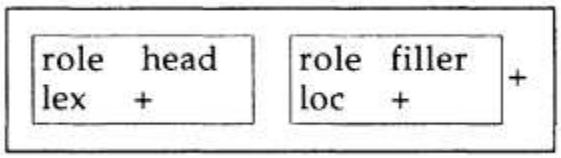
O modelo de gramática chamado BCG é baseado em processos de unificação. Dizer que esse modelo funciona por meio de processos de unificação significa afirmar que todas as operações, sejam sintáticas, semânticas ou lexicais, ocorrem no interior da mesma estrutura de processamento, no interior das matrizes. Além disso, novas matrizes podem ser geradas pela combinação de determinados constituintes (FRIED; OSTMAN, 2004).

---

[3] Cf.: By the term ‘frame’ I have in mind any system of concepts related in such a way that to understand any one of them you have to understand the whole structure in which it fits; when one of the things in such structure is introduced into a text, or into a conversation, all of the others are automatically made available.

As operações gramaticais são representadas por matrizes de atributo de valor (AVM). Esse tipo de notação indica informações gerais das construções, como, por exemplo, as características morfossintáticas, as exigências de valência dos elementos, o frame evocado pela estrutura e pelo núcleo e os papéis semânticos dos participantes. Essas propriedades podem adquirir valores diferentes de representação. Dessa forma, a representação ocorre por sinais binários [+, -], por presença ou ausência do traço, por papel semântico [agente], por hierarquia sintática [head], [dep], por caso [dativo], por propriedades lexicais [lform] etc. Para ilustrar de forma mais clara, observe a representação em matriz da construção núcleo-complemento do inglês na Figura 1.

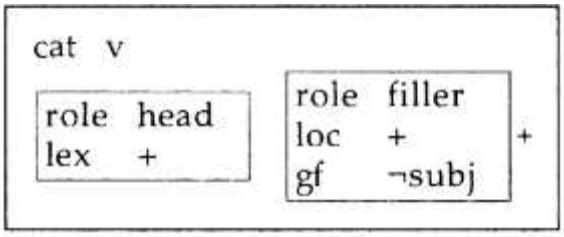
Figura 1: Construção Núcleo-Complemento em Inglês



Fonte: Fried & Ostman (2004, p.7)

A Figura 1 representa a construção núcleo-complemento (HC) da língua inglesa. Fried e Ostman (2004) explicam que essa construção especifica um sintagma que contém um núcleo lexical, que é representado por [lex +] que pode ser seguido por um ou mais constituintes. Na representação, cada caixa indica um constituinte da construção. No caso da construção HC do inglês, há a presença de dois constituintes, um é o núcleo e o outro é a estrutura dependente. Esse tipo de construção funciona como constituinte de outras estruturas, como é o caso da construção de sintagma verbal em inglês (Verb Phrase - VP), representada na Figura 2.

Figura 2: Construção de sintagma verbal (VP) em inglês



Fonte: Kay & Fillmore (1999, p. 8)

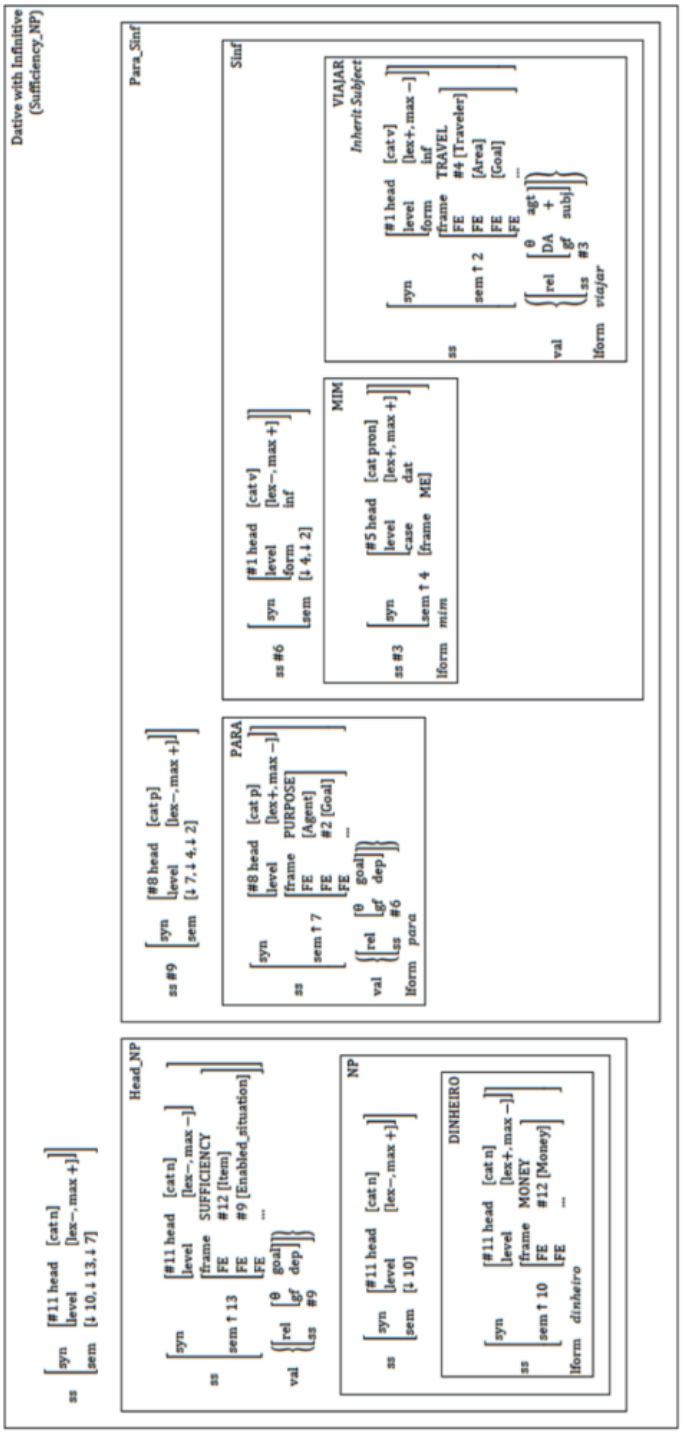
Na Figura 2, a AVM representa a construção de sintagma verbal do inglês. Na matriz, observamos a presença de dois constituintes, um é o núcleo [head] e o outro é a estrutura dependente [filler]. A construção é de natureza verbal, por isso o traço [cat v] aparece no topo da matriz e engloba os dois constituintes. Kay e Fillmore (1999) explicam que, quando uma construção herda de outra, ocorre um processo de compartilhamento de traços entre elas, com possíveis detalhamentos adicionais na construção herdeira. Na AVM representada pela Figura 2, a construção VP herda todas as características da primeira, construção núcleo-complemento, porém adquire traços adicionais como o traço de [cat v].

Além das relações de herança entre as construções, a BCG também considera as propriedades de valência no modelo. A valência é concebida como o conjunto de possibilidades argumentais que um predicador pode portar, isto é, ela está relacionada à quantidade e aos tipos de argumentos que completam o verbo. Ela também pode ser definida como as possibilidades combinatórias de itens lexicais individuais, em outras palavras, são as possibilidades combinatórias estabelecidas entre diferentes itens linguísticos como verbos, substantivos, preposições, adjetivos e seus complementos (FRIED; OSTMAN, 2004). Na BCG, os requisitos de valência verbal são importantes para a notação e para o funcionamento do modelo. A valência ganha notoriedade nessa abordagem porque, através dela, são definidos os papéis semânticos e as posições morfossintáticas de cada participante do frame evocado pela construção (FILLMORE, 2013). Tais noções podem ser observadas, com mais clareza, na representação proposta por Torrent *et al*. (2018) da construção dativo com infinitivo que licencia construtos como o exemplo (1).

(1) Ela deu dinheiro para mim viajar.

Na Figura 3, a representação indica os processos de unificação que ocorrem na construção dativo com infinitivo do PB (TORRENT, 2005). Torrent *et al* (2018) explicam que a construção em questão é constituída por um núcleo e uma sentença infinitiva regida pela preposição *para.*

Figura 3: Construção dativo com infinitivo



Fonte: Torrent et al. (2018, p. 112)

Em relação aos aspectos da organização da matriz, observa-se que, dentro de cada caixa da matriz, estão indicadas informações gerais da construção. Na parte

superior de cada caixa, encontramos o tipo sintagmático, isto é, se é um sintagma verbal (VP), uma sentença Para infinitivo (Para_Sinf) etc. Podemos observar que os constituintes da estrutura funcionam segundo princípios de hierarquia que são indicados pela presença ou ausência do traço de projeção máxima [max +], [max-]. Em relação às propriedades semânticas, na notação encontramos informações sobre os frames ativados por cada constituinte nuclear, os papéis participantes do frame, chamados de elementos de frame (EF) e os requisitos de valência que incluem os papéis semânticos ($\boldsymbol{\theta}$) e a função gramatical (GF). Na construção ilustrada pela Figura 3, os elementos nucleares que ativam frames são *dinheiro*, *para*, *mim* e *viajar*. Além disso, a construção EM SI evoca o frame de Suficiência. Por fim, as exigências de valência de cada item são indicadas entre colchetes.

Em resumo, através da representação das matrizes de atributos de valor, podemos verificar as operações sintáticas e semânticas que ocorrem no interior das construções, assim como os links de herança que ocorrem entre elas.

A descrição desenvolvida até aqui demonstra o alto grau de formalização dessa abordagem construcional. Em outras palavras, esse sistema de notação é considerado um mecanismo altamente implementável computacionalmente e, por isso, é a abordagem escolhida para o desenvolvimento de uma construção anafórica pronominal para aplicação em sistemas de extração de informação em textos e no *Constructicon,* objeto de pesquisa desta dissertação.

Antes de passarmos a essa discussão, entretanto, vejamos o que diz a BCG sobre as anáforas pronominais.

2.2 O TRATAMENTO DA ANÁFORA NA GRAMÁTICA DE CONSTRUÇÕES DE BERKELEY

Na abordagem da BCG (KAY; FILLMORE, 1999), o fenômeno de referenciação de longa distância é tratado segundo princípios de constituência. Os autores afirmam que dependências de longa distância existem entre um constituinte e um elemento que não é parte da estrutura constituinte da sentença em geral. Isso significa que as

construções estabelecem relações familiares entre si, isto é, uma construção pode especificar relações com partes de outras construções.

Através da análise da construção What´s X doing Y? (WXDY), exemplificada em (2) e (3), os autores apresentam propriedades dessa estrutura relacionadas às relações de dependência de longa distância que ocorrem entre os elementos que a compõem. De acordo com eles, os elementos *What* e *doing* mantêm uma relação de dependência entre si.

(2) What is this scratch doing on the table?

(3) What is it doing raining?

Atrelada a uma abordagem cognitivista da linguagem, a anáfora é definida por Fillmore em "On grammatical constructions" (1989). De acordo com ele:

> A anáfora é a relação entre duas expressões linguísticas, X e Y, da qual o Y pega sua interpretação da interpretação do X. A respeito dessa relação, o elemento de Y que corresponde ao elemento de X na interpretação completa da sentença ou do texto é chamado de anáfora, e o elemento de X cuja interpretação é compartilhada pela anáfora é chamado antecedente (FILLMORE, 1989, p.283).[4]

Tal definição descreve características universais do fenômeno anafórico. O autor divide as anáforas em dois grupos: o de anáfora pró-forma, que é o objeto de análise desta dissertação, e o de anáfora zero ou anáfora por elipse. Para explicar a anáfora pró-forma, Fillmore (1989) propõe que ela ocorre quando duas orações contêm elementos que se referem à mesma entidade e o lugar que deveria ser ocupado por uma forma lexical plena é preenchido por palavras funcionais, como os pronomes. Como observa-se na sentença (4).

(4) Eu gosto do Paulo e ele gosta de mim.

[4] Anaphora is the relationship between two linguistic expressions, X and Y, such that Y gets its interpretation from the interpretation of X. In respect to this relationship, that element of Y which corresponds to an element of X in the full interpretation of the sentence or text is called the anaphor, and the element of X whose interpretation is shared by the anaphor is called the antecedent.

Por outro lado, a anáfora zero ou elipse ocorre quando existe um entendimento anafórico entre duas orações, mas não há uma estrutura explícita na sentença que faça o link anafórico. Isso fica bastante evidente no exemplo (5).

(5) Paulo estuda Morfologia e eu, semântica.

No modelo da BCG, a anáfora é tratada segundo o sistema de incorporação de valências, no qual as condições de assimetria observadas entre um antecedente e a anáfora dependem da noção de comando de valência (v-comando) (KAY; FILLMORE, 1999). Isso indica que as relações estabelecidas entre os elementos de uma construção anafórica atendem a requisitos de valência entre os constituintes. Essa situação pode ser vista no excerto abaixo, no qual os autores descrevem as operações sintáticas que ocorrem no fenômeno de dependência de longa distância.

> Uma estrutura de traços ʎ v-comanda uma estrutura de traços distinta β se e somente se houver uma estrutura de traços ɣ de tal modo que ʎ seja um elemento da valência de ɣ e β esteja incluído na valência de ɣ. O v-comando é equivalente à versão original do c-comando alterada pela substituição da relação é-uma-filha-de pela relação é-um-elemento-de-valência-de. (KAY; FILLMORE, 1999).[5]

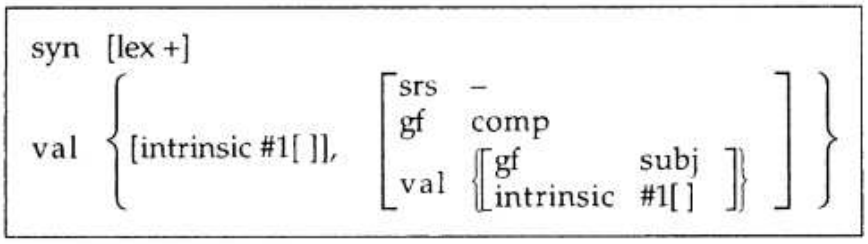
Os autores explicam as operações de valência entre os componentes da construção dentro de uma AVM. Na Figura 4, a construção de coinstanciação está representada. Na matriz, observamos exigências de valência de um predicador p com o valor correspondente do sujeito requerido do complemento de *p* controlado. Nesse tipo de construção, o predicador governa um complemento controlado (KAY; FILLMORE, 1999). As operações ocorrem de forma unificada em uma mesma matriz.

A AVM representada pela Figura 4 indica as unificações estabelecidas entre os constituintes da construção: os requisitos de valência, o sujeito e o complemento. Acerca da Figura 4, Kay e Fillmore (1999) comentam que, quando ocorre a unificação entre as exigências de dois constituintes, ocorre um processo matemático de função,

[5] A feature structure A v-commands a distinct feature structure B iff there exists a feature structure y such that a is a valence element of y and B is valence embedded in y. V-command is equivalent to the original version of c-command amended by the substitution of the relation is-a-valence-element-of for the relation is-a-daughter-of.

isto é, um elemento X torna-se função de outro Y. Assim, a semântica do constituinte X torna-se o argumento sujeito lógico da predicação, y(x). Essa operação matemática indica a unificação entre os constituintes de uma matriz.

Figura 4: Construção de coinstanciação



Fonte: Kay & Fillmore (1999, p. 23)

O fenômeno de dependência de longa distância – as construções anafóricas pronominais – também pode ser tratado sob a mesma ótica adotada pela BCG. Por meio dessa visão, entende-se que as construções anafóricas são definidas mediante ao cumprimento de exigências valenciais dos componentes dessa estrutura. Em relação às construções anafóricas pronominais, essas exigências são estabelecidas pelos sintagmas nominais da construção núcleo e são cumpridas pelas construções dependentes, isto é, pelas construções com o elemento anafórico.

A abordagem da BCG para a anáfora pronominal ainda não foi implementada para o Português Brasileiro e é um dos objetivos deste trabalho. Na seção a seguir, apresentaremos o Constructicon da FrameNet Brasil, recurso linguístico-computacional em que tal implementação será realizada.

2.3 O CONSTRUCTICON DA FRAMENET BRASIL

Nesta seção, será introduzido o Constructicon da FrameNet Brasil e seu aparato computacional. O detalhamento dos processos de modelagem e anotação de

construções, entretanto, será apresentado no capítulo de metodologia, de modo a evitar redundância no texto da dissertação.

Dentre as aplicações linguístico-computacionais desenvolvidas pelo laboratório da FN-Br, encontramos o Constructicon. O constructicon é definido por Fillmore (2008) como uma aplicação linguístico-computacional que comporta o repertório das construções de uma dada língua. Em outras palavras, a base de dados funciona como um repertório de construções para o português brasileiro e é gerida pela FrameNet Brasil WebTool.

Fillmore (2008) acrescenta que a criação do constructicon teve como objetivo o desenvolvimento de um aparato que desse conta de descrever as estruturas presentes na língua no que se refere aos aspectos semânticos e morfossintáticos que não eram contempladas no nível lexical, já modelado no Lexicon da FrameNet. Isso significa um aparato que contemplasse informações para além da valência básica dos itens lexicais.

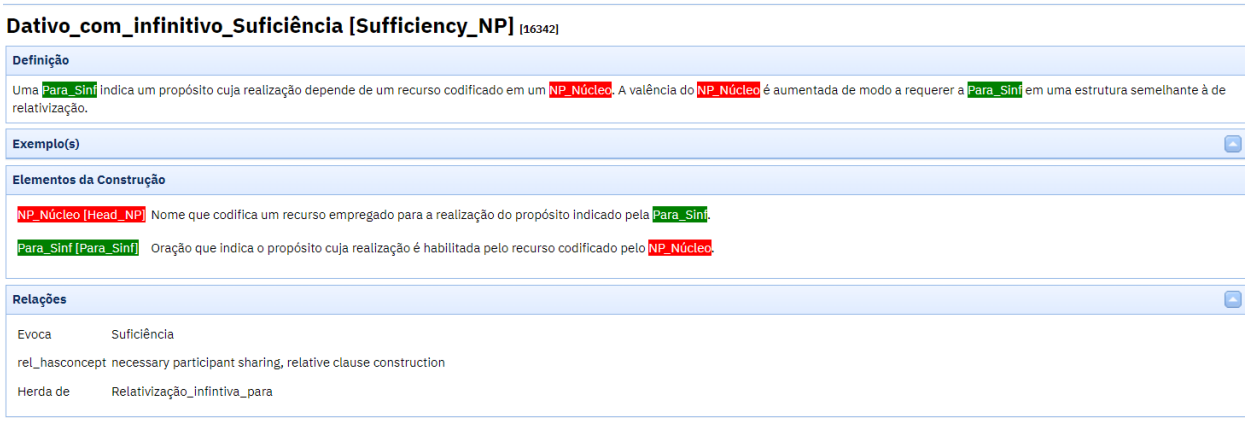
Com o objetivo de atender às demandas desenvolvidas dentro do domínio local da FN-Br, o constructicon foi implementado dentro do laboratório em 2010. Segundo Torrent *et al* (2018), o desenvolvimento de um constructicon teve como propósito geral contemplar esses fenômenos encontrados para além de aspectos semânticos e sintáticos das unidades lexicais no PB. Além disso, os autores acrescentam que o constructicon teve como meta inicial a análise da família de construções Para_infinitivo (Torrent, 2010), da qual a construção de Dativo com Infinitivo, estrutura vista na Figura 3, faz parte.

Ainda dentro do domínio da FN-Br, um dos desafios encontrados para o desenvolvimento do Constructicon foi a representação computacional de princípios de funcionamento e relações entre as estruturas construcionais. Sendo o Constructicon uma ferramenta para a descrição de construções, tornava-se necessária a modelagem de um aparato que desse conta de representar, de forma clara, as estruturas no nível linguístico-computacional (cf. DIAS-DA-SILVA, 1996).

No Constructicon da FN-Br, as construções são representadas por meio de seus elementos constituintes. Podemos encontrar ainda informações sobre as relações de herança entre construções e a evocação de frames por parte delas. Isso fica mais claro na Figura 5. Vemos a construção que licencia as estruturas de dativo com infinitivo de

suficiência no PB. Como já mencionado anteriormente, a representação é composta pela definição da estrutura, pelos elementos da construção e as relações que a construção mantém com outras estruturas. Na Figura 5, a estrutura Para_Sinf indica o propósito para que um recurso, codificado pela estrutura NP-Núcleo, será empregado. No PB a construção em questão evoca o frame de Suficiência e mantém relação de herança com a construção Relativização_Infinitiva_para.

Figura 5: Construção Dativo com Infinitivo de Suficiência



Fonte: Torrent (2009)

Torrent et al (2018) chamam a atenção para a distinção existente no funcionamento do constructicon desenvolvido pela FrameNet de Berkeley e por aquele da FN-Br. Os autores apontam que essa distinção está inserida no âmbito de criação e anotação das construções. Essas distinções são descritas de forma mais detalhada no trabalho de Lage (2013). A autora aponta que algumas informações relacionadas aos aspectos semântico-pragmáticos das estruturas não são contempladas na anotação lexicográfica de ULs. Já que por meio dessas análises obtemos apenas informações relacionadas ao padrão de valência dos itens lexicais.

A autora ressalta que o constructicon desenvolvido em Berkeley (2008) difere-se do constructicon desenvolvido pela FN-Br (2010) no que se refere à definição de elementos de construção (ECs). O primeiro define os ECs mesclando, de forma não estruturada, no próprio corpo do texto da definição, aspectos semânticos e

propriedades formais dos ECs. Por outro lado, o segundo formaliza cada conjunto de propriedades – semânticas e formais – dos ECs e, por consequência, das construções de modo a aprimorar a modelagem linguístico-computacional. Assim sendo, o Constructicon da FN-Br visa não só prover usuários humanos com análises de construções do PB, mas possibilitar que tais análises sejam, em alguma medida, legíveis por máquinas. A partir desse movimento, o Constructicon pode ser usado em diversas tarefas computacionais de tratamento da linguagem, tais como a de extração de informações, que será apresentada com mais detalhes no próximo capítulo.

Como foi observado neste capítulo, apresentamos um panorama sobre a teoria da Gramática de Construções com ênfase na Gramática de Construções de Berkeley. Além disso, discorremos brevemente sobre o tratamento do fenômeno de anáfora pronominal dentro da teoria. Ainda falamos sobre as motivações para criação do constructicon no âmbito da FN de Berkeley e a FN-Br. Em seguida, mostramos os objetivos e a definição da ferramenta. Além disso, diferenciamos as abordagens metodológicas aplicadas pelas duas FrameNets relacionadas à modelagem de ECs. Por fim, ressaltamos que a FN-Br tem o intuito de aprimorar diferentes ferramentas com interface entre o humano e a máquina, como métodos de extração de informações, tarefa contemplada nesta pesquisa.

3 EXTRAÇÃO DE INFORMAÇÕES

Neste capítulo, apresentamos um panorama da tarefa de extração de informação. Abordamos as bases do Processamento de Língua Natural e de sua evolução para a noção de Compreensão de Língua Natural, as diferentes técnicas utilizadas na tarefa de extração de informações, os desafios enfrentados nessa tarefa, assim como seu funcionamento.

3.1 INTRODUÇÃO À COMPREENSÃO DE LÍNGUA NATURAL

De acordo com Dias da Silva (1996), desde a criação dos computadores, havia a preocupação de estreitar as relações entre o homem e a máquina através do desenvolvimento de uma linguagem mais acessível, um código que fosse legível como interface entre ambos. Para alcançar esse objetivo, foram criadas algumas linguagens de programação, sendo a primeira delas a PROLOG, uma linguagem de programação declarativa. Assim que foi criada, a linguagem PROLOG continha muitos erros e realizava apenas funções primárias, puramente lógicas. Isso significa que os problemas eram descritos através de fatos e regras e, mediante isso, buscavam-se soluções possíveis, sem levar em consideração a dinamicidade das línguas naturais. Porém, mesmo com o visível avanço com o emprego das linguagens de programação, os computadores ainda eram ferramentas muito limitadas. Por consequência disso, aprimorar essas máquinas ainda era um grande desafio nessa época.

Com o objetivo de estreitar as relações entre o usuário e a máquina, os pesquisadores estavam determinados a procurar soluções mais robustas e inteligentes para a resolução desses problemas. A questão principal centrava-se na linguagem: era necessário criar estratégias que otimizassem a compreensão entre a máquina e o ser humano. Nesse momento, teve início uma busca por mecanismos mais sofisticados que dessem conta da ampla dinamicidade das línguas humanas. Outras linguagens foram desenvolvidas, além de sistemas capazes de interpretar e processar aspectos das línguas naturais. Nesse período, instaurou-se uma área de pesquisa voltada para o

estudo dessas relações entre homem e máquina, a área de Processamento de Língua Natural (PLN), que Dias da Silva (1996) descreve então como um campo de pesquisa bem heterogêneo e fragmentado.

A partir do final do século XX, essas máquinas foram adquirindo cada vez mais funções e passaram a desempenhar tarefas ainda mais complexas. Porém, o autor aponta que havia a preocupação de desenvolver meios para que essas tarefas fossem desempenhadas de forma mais inteligente. O grande desafio tornou-se buscar meios para que as máquinas compreendessem, mesmo que de forma menos complexa, as línguas humanas. Com esse objetivo, as ciências que formam a área de PLN uniram-se para estudar modelos ou sistemas que fossem capazes de integrar conhecimentos de diferentes áreas. Era preciso desenvolver estratégias que contemplassem aspectos computacionais e linguísticos. Isso significava a criação de linguagens ou modelos que fossem computacionalmente operáveis, mas que também mantivessem uma representação que seguisse a estrutura de uma língua natural.

Schank e Riesbeck (1981) apontam que, no início dos estudos em PLN, os pesquisadores tentavam implementar modelos baseados em regras, que enfatizavam a sintaxe, ou seja, a estrutura das línguas. Tais abordagens baseadas em regras eram, por sua vez, dependentes de listas de significado. Esses modelos, apesar de funcionais em certa medida, não contemplavam toda a complexidade das línguas. Isso significa que eles não eram funcionais em relação à compreensão e produção de sentido, ou seja, na representação do significado. Os pesquisadores perceberam que era mais fácil ensinar a máquina a aprender aspectos estruturais de uma língua, como, por exemplo, entender que, em português, o objeto vem depois do verbo, ou que existe oração sem sujeito, do que questões relacionadas à significação.

Por consequência disso, eles observaram que, para obter aprimoramentos mais significativos em PLN, de alguma forma, era preciso imitar a estrutura e os mecanismos do pensamento humano. Porém, era um grande desafio imitar o modelo de funcionamento da cognição humana, afinal, a cognição humana, como nos mostram os trabalhos em Linguística Cognitiva – veja-se, entre outros, Fillmore (1985), Lakoff (1987) e Fauconnier e Turner (2002) –, não funciona por meio de regras e listas

dicionarizadas. Aspectos do significado, que são facilmente depreendidos por nós, não são passíveis de serem compreendidos pelas máquinas a partir de tal metodologia.

Entende-se, assim, que as aplicações em Compreensão de Língua Natural (CLN) e PLN são uma tentativa de modelagem das capacidades cognitivas humanas. Para atingir esse objetivo, ao longo dos anos, muitos pesquisadores vêm desenvolvendo modelos como tentativa de se aproximar da complexidade da cognição humana. Ainda hoje, essas duas grandes áreas de pesquisa buscam otimizar a relação entre a máquina e o ser humano. Para isso, atuam em projetos que visem ao desenvolvimento de ferramentas em conjunto.

Com o objetivo de explicar a abrangência e as aplicações desenvolvidas em PLN e CLN, McShane (2019) apresenta alguns desafios e mitos que cercam essas duas grandes áreas de pesquisa.

Em um primeiro momento, a autora enumera problemas que envolvem as ferramentas de compreensão de língua natural em geral. Ela explica que, para um modelo de CLN funcionar minimamente, ele deve ser treinado para interpretar muitos aspectos paralinguísticos, como, por exemplo, a cooperação humana, o afeto, informações visuais e auditivas. Isso significa que esses modelos precisam ser programados para entender minimamente enunciados de forma não organizada e incompleta: elipses, implicaturas, ambiguidades lexicais e referenciais – estas últimas, um dos desafios encarados neste trabalho – atos de fala indiretos, linguagem não-literal, entre outros, são fenômenos cotidianos em qualquer língua humana e fazem parte da tarefa, bem como a identificação de referências de longa distância.

Todas essas questões são consideradas grandes desafios dentro do domínio da compreensão e processamento de língua natural. Uma delas será vista com mais detalhes abaixo, na seção 3.2 sobre Extração de Informações: a referenciação por anáfora pronominal.

Em um segundo momento, a autora aponta mitos que contribuíram para a falsa crença de que o PLN estatístico e a CLN baseada em conhecimento são áreas divergentes entre si, quando, na verdade, elas são complementares. Ela explica que cada uma das áreas de aplicação segue determinados métodos de análise, mas que

ambas são fundamentais para o desenvolvimento de tarefas linguístico-computacionais.

Em seguida, a pesquisadora discute sobre a abrangência do paradigma empírico do PLN e acrescenta que esse paradigma tem alcançado um bom desempenho no desenvolvimento de aplicações como a extração de conhecimento e tarefas de pergunta e resposta. Mas, segundo ela, apesar de todos os esforços para a otimização dessas tarefas por meio de anotação manual, ainda existem lacunas dentro de ferramentas do domínio do PLN que envolvem o tratamento de dependentes de longa distância e a ambiguidade.

Em resumo, apesar de todos os esforços para o desenvolvimento e otimização dessas duas áreas de pesquisa, ainda hoje, o maior desafio para o desenvolvimento de aplicações em PLN se concentra na escolha de um modelo de língua que contemple a semântica das línguas e não apenas seus aspectos unicamente estruturais. Ainda hoje é preciso desenvolver e aperfeiçoar modelos de língua que contam com um arcabouço teórico sólido, com objetivos bem traçados, parâmetros explicativos e descritivos capazes de auxiliar uma máquina a interpretar e compreender a linguagem humana. É nessa busca que se insere esta dissertação.

Ao longo do tempo, iniciou-se uma jornada incessante para a criação e o aperfeiçoamento de modelos linguístico-computacionais que contemplassem todos ou pelo menos alguns desses aspectos .

Em um primeiro momento, travou-se uma busca por modelos que contemplavam tarefas específicas e imediatas. Com o passar do tempo e o desenvolvimento de outras funcionalidades, o objetivo tornou-se em aperfeiçoar as funções dos modelos já existentes. Por meio disso, entende-se que, ainda hoje, o desenvolvimento de um modelo que seja operável em todos esses domínios é um grande desafio.

Por muito tempo, as tentativas de aprimoramento limitaram-se ao treinamento de algoritmos diversos. Isso significa que os algoritmos eram treinados para realizarem tarefas muito específicas. Além disso, eles necessitavam de supervisão humana em todas as atividades. Essa situação não era vantajosa nem mesmo operável porque dependia de interferência humana contínua no sistema.

A partir de todos esses esforços e o advento da inteligência artificial, os computadores passaram a realizar uma pluralidade de tarefas como, por exemplo, a tradução por máquina, a extração de informações, a autocorreção etc. Segundo Russel e Norvig (2009), os avanços em PLN culminaram com a implementação das redes neurais. Elas são um sistema de inteligência artificial simbólico que é utilizado para o reconhecimento de padrões. Esses sistemas passaram a contar com mecanismos de aprendizagem de máquina por meio dos quais eles tornaram-se capazes de melhorar sua performance a partir da experiência. Tal experiência surge através da extração de padrões de um *corpus* por meio de cálculos e inferências. De forma mais clara, a partir dos dados coletados, o modelo é treinado para, posteriormente, fazer inferências, ou seja, descobrir os padrões com base em características. A partir disso, percebemos que a relação estabelecida entre a máquina e o usuário tornou-se muito mais funcional. Isso é, o usuário passou a realizar muitas tarefas por meio da máquina.

Em relação aos aspectos metodológicos aplicados às redes neurais, os autores postulam que é possível destacarmos três modelos de aprendizagem: o modelo supervisionado, o não-supervisionado e o semi-supervisionado. No supervisionado, o modelo de dados e metadados etiquetados é desenvolvido por humanos em algum estágio qualquer. Por outro lado, no modelo não-supervisionado, o conjunto de dados não etiquetados agrupa dados semelhantes. Nesse último modelo, ocorre um processo de aprendizagem por meio da experiência de treinamento. Isso significa que o sistema aprende de acordo com as ocorrências e faz inferências. Por outro lado, em sistemas definidos como semi-supervisionados, também conhecidos como híbridos, ocorre uma integração dos dois modelos anteriores. Sendo assim, esses modelos de aprendizagem utilizam mecanismos orientados por dados de uma rede neural em conjunto com capacidades de manipulação de dados. Todas essas habilidades e aplicações são estudadas no cenário do PLN e da inteligência artificial (IA) e, em certa medida, contribuem para a realização de inúmeras tarefas, como já mencionado nos parágrafos acima.

Apesar de todos os esforços aplicados ao longo dos anos para o aprimoramento de modelos de aprendizagem no âmbito da IA (Deep Learning), ainda hoje as técnicas apresentam muitas limitações. No artigo “Deep Learning alone isn’t getting us to

Human-like AI", Gary Marcus (2022) aponta para o fato de que as redes neurais e os vários modelos de aprendizagem não funcionam como os neurônios biológicos humanos. Em suma, eles são modelos simplificados que imitam o funcionamento do cérebro humano, mas são menos complexos. Entendendo isso, o autor reafirma algumas restrições observadas em modelos de Inteligência Artificial.

Podemos citar como exemplo dessas limitações o fenômeno analisado nesta investigação. Essa situação acontece porque os modelos de redes neurais não contemplam tarefas como mapeamento de referentes de longa distância em decorrência das limitações da grande área de IA. Essas limitações estão relacionadas à escassez de modelos mediante à grande complexidade de alguns fenômenos da linguagem humana.

Com o objetivo de sanar algumas dessas lacunas, especialmente no que se refere ao mapeamento de co-referências em textos, este trabalho elabora um modelo linguístico-computacional com uma perspectiva cognitiva. Para isso, discutiremos sobre a tarefa de extração de informação em textos e como o modelo de construções anafóricas pronominais pode trazer benefícios e aprimorar essa tarefa. A aplicação será estudada com mais detalhes na próxima seção.

3.2 EXTRAÇÃO DE INFORMAÇÃO: FUNDAMENTOS E TÉCNICAS

A tarefa de extração de informações (EI) pode ser definida como qualquer método utilizado para filtrar informações específicas em um corpus. Basicamente, o mecanismo de extração de informação em textos propõe relações entre as várias informações em um texto. Marquéz *et al* (2008) definem extração de informações como uma subárea do PLN dedicada ao problema geral de detecção de entidades referidas em textos de língua natural. O autor aponta que os elementos que são os objetos de análise podem ser resumidos por meio das seguintes informações: "quem" fez "o quê" para "quem", "quando" e "onde".

Podemos encontrar aplicações de EI em diferentes contextos, como em empresas de tecnologias como a Google e até mesmo em pequenos grupos. Isso

implica em diferentes configurações para os sistemas de EI, já que essas aplicações podem desempenhar processos mais ou menos complexos. Tal situação será determinada de acordo com a natureza e a quantidade dos dados que serão filtrados. Apesar de existirem essas diferenças relacionadas aos níveis de complexidade do processo de extração, os métodos seguem sempre as mesmas etapas, como será visto com mais detalhes abaixo.

Piskorski e Yangarber (2013) apresentam um panorama da tarefa de extração de dados. Segundo os autores, as tarefas eram concentradas ao redor da identificação de nomes de entidades como pessoas, nomes de empresas e as relações entre eles. Porém, eles acrescentam que, nos últimos anos, devido uma maior disponibilidade de dados online, novas aplicações em EI apareceram. Em virtude disso, as técnicas utilizadas nessas ferramentas evoluíram muito ao longo dos anos.

Ainda hoje, as aplicações em EI enfrentam alguns desafios. Teixeira e Rodrigues (2014) afirmam que existem duas grandes questões a serem aprimoradas: a variedade e a expressividade das línguas naturais. A primeira delas refere-se, mais diretamente, às diversas formas de falar sobre a mesma coisa ou o mesmo assunto. Variações gramaticais, lexicais e anafóricas são fenômenos difíceis de serem formulados e interpretados computacionalmente. Essas variações podem ser observadas nos exemplos de (6) a (8). No enunciado (6), há a construção de uma forma passiva, em que os papéis são interpretados de acordo com o uso das preposições. No enunciado (7), há a presença de unidade lexical que gera uma metáfora, cabeça. O sentido metafórico da sentença não é facilmente compreendido pela máquina. Por último, no enunciado (8), ocorre uma referência por anáfora. Esses tipos de referência também representam dificuldades na interpretação por máquina porque ela precisa compreender que *Maria* e *ela* são a mesma pessoa, mas, ao mesmo tempo, que o nome feminino *linguística*, não pode ser tomado como o antecedente de *ela*. Esse último fenômeno é o objeto de pesquisa deste trabalho e ganhará maior atenção mais adiante.

(6) O dinheiro foi roubado da menina pelo ladrão.

(7) Paulo é o cabeça da empresa.

(8) A Maria estuda linguística e ela também trabalha.

O segundo desafio, considerado o maior para todas as aplicações em PLN, refere-se à ocorrência de muitas estruturas ambíguas e vagas nas línguas naturais. Tais estruturas geram variações na produção de sentido e, por isso, dificultam a interpretação por máquina. Os exemplos (9) e (10) ilustram casos que podem ser facilmente enunciados por um humano, mas dificilmente compreendidos pela máquina. No enunciado (9) existe a presença de um tipo de ambiguidade chamada de ambiguidade de escopo. Não é possível afirmar a qual elemento da sentença o sintagma *com óculos* se refere, se é ao *homem* ou à *menina*. Já no enunciado (10) não existem informações adicionais sobre a forma como Maria ama o João. Por causa disso, a sentença é vaga.

(9) O homem viu a menina com óculos.
(10) João ama Maria como ela o ama.

Ao longo do tempo, muitas abordagens foram desenvolvidas para aplicações em EI. Teixeira e Rodrigues (2014) apresentam os três tipos de abordagens, as quais são classificadas de acordo com a tecnologia empregada, com o grau de automação do sistema e com o tipo de input dos documentos. A abordagem pelo tipo de *input* é aplicada quando os documentos têm marcação de XML. Nesses casos, as informações seguem padrões de classificação chamados padrões alvo como, por exemplo, endereços, datas, nomes próprios etc. Para processar esses dados, os métodos de extração de informação identificam essas estruturas marcadas.

As abordagens relativas ao tipo de tecnologia sempre foram empregadas em sistemas de EI. Dentre elas, as abordagens baseadas em regras, chamadas abordagens de conhecimento projetado, utilizam regras codificadas por humanos. Nesse tipo de abordagem, ocorre uma correspondência entre padrões linguísticos e computacionais. Essas correspondências podem ser empregadas por etiquetas de entrada de dicionários, cadeias de texto e classes de palavras. Dentro da abordagem

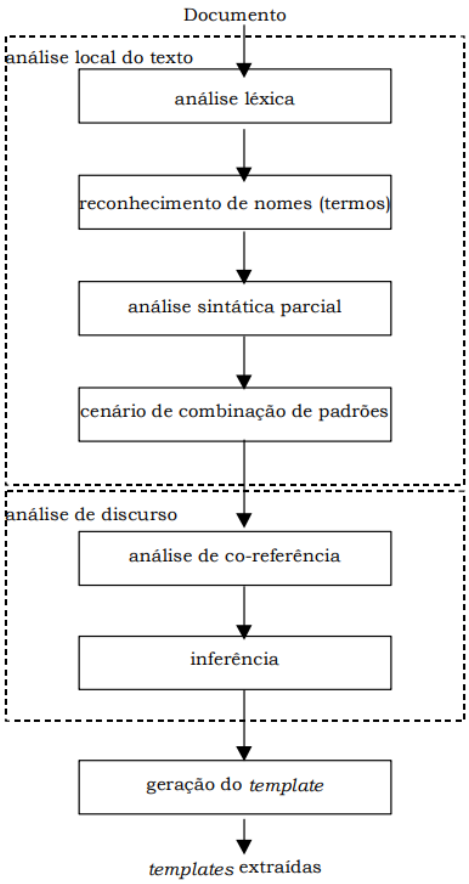
baseada em regras encontram-se o tratamento das construções anafóricas pronominais conforme apresentadas nesta dissertação.

Para explicar melhor, as estruturas anafóricas pronominais estão inseridas em uma abordagem baseada em regras porque elas seguem princípios linguístico-computacionais. Por causa disso, o que diferencia as estruturas anafóricas de outras exclusivamente baseadas em códigos é o nível de complexidade empregado para o desenvolvimento de cada uma. Isso acontece devido a duas situações: a primeira relaciona-se com a dificuldade de estudar esses pedaços de língua, as construções. A segunda está relacionada ao fenômeno de referência propriamente dito. Quando estudamos elementos de correferência, estamos analisando fenômenos linguísticos difíceis de serem formulados e interpretados computacionalmente, já que para o mapeamento dos antecedentes precisamos considerar regras em âmbitos semântico, morfossintático e computacional.

Por fim, existem também os métodos considerados híbridos, que aliam as duas abordagens anteriores. Eles combinam as duas estratégias de extração de dados e, por meio do aprendizado de máquina guiado por um algoritmo, melhoram a performance da tarefa.

Zambenedetti (2002) lista as etapas de funcionamento de um sistema de EI. Segundo ele, ocorrem os processos de análise léxica, reconhecimento de nomes, estrutura sintática, cenário de combinação de padrões, análise de co-referência, inferência e união de eventos. Essas etapas são repetidas infinitamente dentro da ferramenta através de um algoritmo de aprendizagem como podemos observar na Figura 6 abaixo.

Figura 6: Etapas do processo de Extração de Informações



Fonte: Zambenedetti (2002, p.28)

Como podemos observar na Figura 6 (ZAMBENEDETTI, 2002), o processo de EI funciona seguindo algumas etapas. Para explicar melhor, em um primeiro momento, dentro do módulo de análise local do texto, o documento que serve como input sofre o processo de análise lexical de cada elemento, também chamado de processo de tokenização. Depois disso ocorre o reconhecimento de entidades nomeadas. Nesse momento, os componentes são identificados segundo sua natureza semântica por meio de padrões, como pelo uso de Ltda, Jr. etc. Em seguida, ocorre a análise sintática parcial, também chamada de parsing. O parsing é uma técnica amplamente utilizada dentro do domínio do PLN, a qual, de acordo com os autores Othero e Menuzzi (2005),

consiste na classificação de sentenças. Isso significa que, por meio de programas chamados *parses,* as sentenças são desmembradas e classificadas morfossintaticamente. Em resumo, na técnica de parsing ocorre um processo de análise sintática. A partir disso, as estruturas linguísticas são interpretadas repetidamente em etapas.

O processo que se segue é o de cenário de combinação de padrões. Nesse momento, a ferramenta passa a identificar o domínio e o cenário em que o texto está inserido. A partir de etiquetas geradas, o sistema inicia a etapa de classificação semântica. Como exemplo disso, podemos observar a seguinte situação: em um texto que discute assuntos relacionados ao domínio do turismo, o sistema reconheceria tags relacionadas ao turista, ao lugar, ao tempo, à atração. Já em relação aos verbos, nesse cenário apareceriam tags relacionadas a verbos como viajar, passear, conhecer etc.

Já no módulo de análise do discurso, ocorrem os processos de análise de co-referência. Nessa fase, as anáforas são mapeadas e os elementos – referentes a antecedentes – são sinalizados. Nesse caso, o mapeamento de referências acontece por meio da localização do antecedente mais próximo. Essa situação contribui para o surgimento de ambiguidades na extração dos dados e ainda, erros no processamento dos dados.

Depois disso, os dados passam por um processo de inferência e união de eventos. Nessa etapa, o sistema abstrai padrões e obtém resultados usando a lógica infinitamente. A partir disso, ocorre a geração de etiquetas (templates) e elas são extraídas do texto. Isso significa que o modelo codifica os padrões de ocorrências mais prototípicos do documento e replica essas *tags* para as demais estruturas por analogia. Como mencionado anteriormente, isso acontece de forma infinita dentro da aplicação.

Como apresentado anteriormente, na etapa de reconhecimento de entidade nomeada, os dados são identificados por templates, ou tags. Nesse momento, as informações essenciais do texto são devidamente identificadas através de um algoritmo de aprendizagem, como, por exemplo, as pessoas envolvidas, o lugar, o horário etc. Atreladas a esses componentes, podemos encontrar anáforas pronominais que não podem ser interpretadas sem a identificação de seus antecedentes. Nesse momento, torna-se necessário o mapeamento dessas estruturas dentro e fora do escopo da

sentença através da utilização de um modelo linguístico-computacional que seja capaz de interpretar essas estruturas.

Tendo como base o que foi discutido nos parágrafos anteriores e os desafios apresentados até aqui, por meio da modelagem de construções anafóricas pronominais, contemplaremos a etapa de reconhecimento de co-referência dentro do processo de EI. Isso acontecerá porque, através da descrição do fenômeno de construções anafóricas, poderemos mapear os antecedentes que funcionam como referentes de longa distância dentro da ferramenta. Assim, a etapa de localização de referências anafóricas provida por pronomes pessoais, demonstrativos, possessivos, relativos, reflexivos, de tratamento e indefinidos será otimizada em métodos de EI.

Isso será possível devido à aplicação de um modelo computacional que interprete essas estruturas dentro da ferramenta. As bases e o desenvolvimento de tal modelo, objeto de pesquisa deste trabalho, serão apresentados, em detalhes, nos próximos capítulos. Em linhas gerais, porém, o modelo funciona da seguinte forma: a partir da identificação das referências formadas por construções anáforicas pronominais, os antecedentes serão devidamente mapeados com uma considerável redução de erros de localização entre antecedentes e anáforas pronominais. Além da otimização do processo de localização dos elementos, a aplicação do modelo contribuirá para a redução de ambiguidade no processo de extração através da geração de templates (categorias) em estruturas formadas por antecedentes que evocam frames de entidade e referentes pronominais.

A Tabela 1 mostra como essas relações são estabelecidas dentro de uma ferramenta de EI. De forma resumida, o processo acontece da seguinte forma: uma sentença entra no sistema, e, através de mecanismos de análise sintática, análise lexical, mapeamento de referentes, e inferências, os padrões resultantes são identificados e etiquetas são geradas. A Tabela 1 apresenta uma situação hipotética, com exemplos criados, para demonstrar como ocorreria o processo de extração de informações com a aplicação do modelo de construções anafóricas pronominais.

Tabela 1: co-referência e classificações

O Lucas e seu cachorro	Seu: anáfora pronominal Possessiva *Lucas:* Referente - evoca o frame de Entidade Relação: Pertencimento
O Lucas se machucou	Se: Anáfora pronominal Reflexiva *Lucas: R*eferente - Evoca o frame de Entidade Relação: reflexiva

Como observamos na Tabela 1, os *templates* (ou *tags*) identificam as categorias das estruturas presentes no texto. Vemos que, através das sentenças, obtemos determinadas categorias ou etiquetas de palavras chamadas templates. No enunciado *O Lucas e seu cachorro*, ocorre a presença de duas entidades *Lucas* e *cachorro* e, ainda, a presença de um elemento que funciona como anáfora pronominal, o seu. Através da aplicação do modelo, a tarefa de EI poderia mapear as relações de referência estabelecidas, neste caso, dentro do limite da sentença e, a partir disso, realizar as abstrações e geração de etiquetas.

Da mesma forma, no segundo enunciado, *O Lucas se machucou,* há a presença de uma entidade *Lucas* e uma anáfora pronominal *se*. Nesse caso, o sistema conseguiria localizar o referente por meio da aplicação do modelo. O *se* é um pronome reflexivo e aponta para o antecedente *Lucas*.

Apesar de alguns sistemas de EI já realizarem o mapeamento de referência com a análise da entidade mais próxima, eles possuem muitas limitações. Isso acontece porque, em muitos casos, a anáfora pronominal aponta para um antecedente fora do limite da sentença. O antecedente pode ser localizado na mesma sentença, na sentença imediatamente anterior, na sentença anterior, ou em sentenças anteriores.

Outro ponto importante está relacionado à variedade de anáforas pronominais existentes. Essas estruturas são responsáveis garantir ao texto sentidos variados relacionados à posse, entidade, ação sobre si mesmo etc. Por conta de tamanha variedade, torna-se necessário a delimitação e definição do fenômeno por meio do desenvolvimento de estruturas (ou códigos) passíveis de serem interpretadas por máquina. Isso será visto com mais detalhes nos próximos capítulos.

Como foi possível observar ao longo deste capítulo, foram apresentados os fundamentos teóricos e os princípios de funcionamento que norteiam as técnicas de Extração de Informação em textos. Ainda discutimos os benefícios do desenvolvimento de um modelo computacional que interprete as anáforas visando o aprimoramento da etapa de localização de referentes dentro da aplicação. Através de todo o arcabouço teórico apresentado relacionado ao processo de EI, entendemos que a modelagem de construções pronominais anafóricas pode otimizar o processo de localização de informações em textos de forma significativa.

Desse modo, partimos agora para a descrição dos mecanismos anafóricos encontrados no Português Brasileiro, contemplando os diversos tipos anafóricos com ênfase no tipo pronominal.

4 MECANISMOS ANAFÓRICOS EM PORTUGUÊS BRASILEIRO

Neste capítulo, apresentam-se os mecanismos anafóricos mais utilizados no Português Brasileiro. Essa descrição é um panorama do fenômeno anafórico e é constituída por diferentes abordagens sobre o processo de referenciação na literatura linguística existente. As descrições aqui expostas embasam a justificativa para o recorte do fenômeno investigado nesta dissertação: a anáfora pronominal.

4.1 PANORAMA DO FENÔMENO DA ANÁFORA

Para entendermos o funcionamento do fenômeno anafórico, antes é necessário discutirmos o processo de representação e referenciação na linguagem. A noção de referência tem sido amplamente discutida ao longo do tempo por diferentes perspectivas dos estudos da linguagem. Todas elas contribuem para o entendimento do processo de referenciação em sua totalidade. Assim como se verá adiante, passaremos brevemente pelo conceito de representação segundo a tradição filosófica e pela forma como ele foi primeiramente traduzido para a Linguística, pela visão do estruturalismo de Saussure. Em seguida, apresentaremos outras abordagens linguísticas do fenômeno da referência, essas de maior relevância para uma discussão mais ampliada da natureza do fenômeno. Discutiremos a forma como a referência é tratada em Lyons (1977), bem como sua contextualização no âmbito da Linguística Textual (KOCH; ELIAS, 2006; MARCUSCHI, 2010).

De acordo com Rojo (1997), as ideias de língua e representação sofreram mudanças ao longo do tempo. Para a tradição filosófica representada por Platão, Aristóteles e Heráclito, a linguagem teria a função de traduzir o mundo, assim, as palavras seriam reflexos dos objetos, representando-os. A autora acrescenta que a noção de representação ampliou-se com os estudos de Descartes, a partir dos quais a língua passou a ser concebida como um instrumento de representação das capacidades humanas que possibilitaria o conhecimento do mundo. Nesse momento

histórico-filosófico, as palavras e os objetos ganham a função de representar ou expressar as operações gerais do espírito humano.

Atrelado à noção de representação está o conceito de referenciação. Pode-se dizer que o ato de referenciar é um processo de identificação que se estabelece entre dois termos. Então, a partir disso, entende-se que o processo de representação liga-se à noção de referência quando diz respeito às relações de identidade entre duas coisas, sejam duas expressões linguísticas ou uma palavra e a coisa no mundo. Não se trata de termos sinônimos, mas complementares: A representação aponta para uma relação entre a palavra e o objeto a que ela se refere, já a referenciação aponta para uma relação de identidade entre duas entidades linguísticas que podem aparecer ou não no discurso. Dessa forma, as palavras e, especificamente, os nomes seriam referentes em sua essência, porque designam objetos no mundo.

Os contornos de tal complementaridade são desenhados de modo pouco claro na obra de Saussure (1916). Nela, postulam-se premissas basilares da teoria linguística, tais como a ideia de signo como uma entidade de duas faces (formada por um significado e um significante) e a noção de valor linguístico. O signo é entendido por meio da ligação entre o significado e o significante e pelas relações que ele estabelece com os outros elementos do sistema da língua. A identidade do signo é tratada apenas como a relação estabelecida entre um significado e o seu significante, ou seja, por meio de associações recíprocas. Por meio do signo, ocorrem representações arbitrárias do mundo, ou seja, ele é referenciado. Apesar de a postulação de Saussure ser muito importante para a linguística como ciência, a noção de referência é tratada de forma convencional no estruturalismo saussureano. Dessa forma, o signo linguístico refere-se ao mundo arbitrariamente, segundo acordos sociais.

Já no livro *Semantics* (1977), Lyons define o fenômeno de referência ao longo do tempo. O autor retoma o trabalho de Ogden e Richards (1923) quando apresenta a noção de referência defendida pela tradição filosófica. De acordo com eles, o termo referente é usado para qualquer objeto ou estado de coisas no mundo externo que é identificado por uma palavra ou expressão. Explicando melhor, o termo referência é empregado para o fenômeno de mediação entre uma palavra (ou expressão) e um referente. O autor acrescenta que a referência é a relação que se estabelece entre uma

expressão e qual ocasião particular de enunciação ela representa. Isso significa que a referência é uma expressão que representa algo que já apareceu anteriormente no discurso em conexão com a noção de significação.

Como será visto com mais detalhes ainda nesta seção, Lyons (1977) divide os tipos de referência para os sintagmas nominais em referência singular definida, sintagmas nominais definidos não referenciados, referência geral distributiva e coletiva, referência indefinida específica e não específica e referência genérica. Ele explica os tipos de referência para a língua inglesa, mas os exemplos também podem ser usados para explicar o fenômeno em língua portuguesa.

Lyons (1977) explica que a referência singular definida diz respeito às classes de indivíduos e a indivíduos. Segundo ele, no uso desse tipo de referência, podemos identificar um referente não apenas o nomeando, mas também provendo para o leitor/ouvinte uma descrição sobre ele com mais detalhes. Outra questão defendida pelo autor é que a existência de um referente pelo seu uso em uma descrição definida não implica, necessariamente, que a descrição é verdadeira para o referente ou ainda que ela é tomada como verdadeira pelo falante. O autor apresenta como exemplo a sentença abaixo. No excerto (11), o sintagma nominal “O homem alto” é uma descrição definida em um dado contexto de enunciação. Assim, o ouvinte é capaz de identificar o referente com base nas propriedades descritas para ele.

(11) O homem alto bem ali.

O autor continua a explicação e fala sobre os sintagmas definidos não referenciais. De acordo com ele, esse tipo de referência tem uma função predicativa de indicar algo sobre o indivíduo citado anteriormente. Nesse caso existe uma identificação entre os dois referentes presentes no exemplo (12)

(12) Dilma é a presidente do Brasil.

Para explicar a referência geral distributiva e coletiva, o autor apresenta a ambiguidade gerada por essa estrutura, como pode ser observado no exemplo (13).

Pode-se observar que a sentença apresenta dois significados distintos: (a) cada livro custa 10 reais distributivamente ou (b) todos os livros juntos custam 10 reais. Nesse caso, o enunciado pode ser interpretado por meio de uma referência coletiva ou distributiva.

(13) Esses livros custam 10 reais.

Antes de explicar a referência indefinida específica e não específica, Lyons (1977) define a diferença entre sintagmas nominais não-definidos e indefinidos do inglês. De acordo com ele, um sintagma não definido é qualquer sintagma nominal que não é um sintagma definido. Por outro lado, um sintagma indefinido é aquele que tem tanto um pronome indefinido ou um sintagma nominal introduzido por um artigo indefinido. A partir disso, o pesquisador apresenta ocorrências desses tipos de referência. A referência indefinida ocorre na presença de sintagmas nominais com o pronome indefinido como em (14). Nesse caso, o sintagma *uma menina* refere-se a um indivíduo indeterminado.

(14) Alguma menina entrou no banco.

O autor acrescenta que os pronomes indefinidos alguém/algum, ninguém/nenhum podem ser usados de forma específica ou não-específica. Além disso, sintagmas nominais introduzidos por algum/alguma também podem ser empregados das duas formas. Nos enunciados (15) e (16), essa situação pode ser percebida.

(15) Todo homem ama alguma mulher.
(16) Eu não quero namorar com ninguém.

Ambos os enunciados podem ser interpretados por meio de uma referência indefinida específica ou não-específica. Quando tomados como específicos, a referência pressupõe, ou implica na existência de um indivíduo que se enquadre nas

descrições. Caso contrário, se for não-específico, não há a presença de pressuposições ou implicaturas sobre a existência do indivíduo.

O linguista também explica o conceito de opacidade referencial. Para isso, ele cita Quine (1960) quando diz que construções, ou contextos, são opacos quando falham em preservar a extensionalidade sobre a substituição de expressões correferenciais singulares. Lyons acrescenta que essas expressões correferenciais em questão podem ser tanto definidas como não definidas. O autor ilustra esse fenômeno com o excerto (17). Segundo ele, a expressão “pelo Reitor” é ambígua, tem opacidade de sentido. Por esse motivo, ela pode ser construída referencialmente e atributivamente. Mas, mesmo à luz dessas interpretações, o Mr. Smith pode ou não saber quem é o Reitor.

(17) Mr. Smith está procurando pelo Reitor.

A opacidade referencial pode ser considerada um fenômeno normal da linguagem porque ela está relacionada com aspectos diversos no contexto de conversa, como, por exemplo, interpretação de inferências entre os interlocutores. Nesse sentido, quando não somos interpretados da maneira devida, reformulamos expressões, mudamos a estrutura para que não ocorra situações de opacidade referencial.

Por último, o autor explica a referência genérica e diz que esse tipo de referência tem chamado a atenção de muitos linguistas ainda hoje. Para explicar esse tipo de referência, o autor apresenta alguns exemplos, como pode ser observado em (18-20).

(18) O leão é uma besta amigável.
(19) Um leão é uma besta amigável.
(20) Leões são bestas amigáveis.

Esse tipo de referência genérica pode ser encontrada nas três proposições acima. Isso acontece porque, mesmo com o uso de artigos definidos ou indefinidos, todas as ocorrências podem se referir a um único indivíduo leão, como também a toda

classe de indivíduos. O autor acrescenta que as proposições genéricas são atemporais e não possuem aspecto verbal.

A partir disso, outros autores se debruçaram sobre o estudo do fenômeno de referência. A noção ampliou-se muito com o surgimento da Linguística Textual na década de 70. Muitos estudiosos da linguagem como Figueiredo (2003), Koch (2006), Marcuschi (2008) e Adam (2008) passaram a relacionar o fenômeno de referência aos processos de progressão e compreensão textual, ou seja, os mecanismos de referenciação facilitam as operações cognitivas gerais que o leitor precisa realizar ao ler um texto. De acordo com esses autores, o fenômeno de referenciação está relacionado a um processo de identificação estabelecido entre um referente e um antecedente cujas propriedades morfológicas, sintáticas e semânticas estão relacionadas. Figueiredo (2003) define que só os nomes ou expressões nominais têm uma tarefa referencial bem definida. Assim, a anáfora é, por natureza, um mecanismo de referenciação da linguagem humana.

Para resumir, nesta pesquisa, quando o termo referenciação for usado, ele estará relacionado à visão defendida por Lyons (1977), estabelecendo a definição a qual a referenciação é um mecanismo que ocorre no discurso entre duas unidades que mantém uma relação de identidade, sendo uma o antecedente e a outra a anáfora. De acordo com o objetivo proposto por este trabalho, o fenômeno de referenciação será analisado e desenvolvido visando sua aplicação computacional.

No Português Brasileiro, o fenômeno anafórico é, por natureza, uma ferramenta de referenciação textual. É uma operação que possibilita a progressão de um tema através da retomada de elementos previamente introduzidos no texto, um mecanismo de coesão no interior do texto. Koch e Elias (2017) explicam que a anáfora tem uma função fundamental na retomada ou progressão referencial de um texto. Por meio disso, ela possibilita a reativação de elementos que já apareceram no texto e sumarização das informações dos segmentos anteriores. Todas essas funções permitem uma organização macroestrutural do texto. Segundo as autoras, os mecanismos usados para essas estratégias podem ter um fundamento gramatical, como pelo uso de pronome, elipse, numerais, advérbios e locativos e, ainda, lexical

pelo uso de repetição de elementos, sinônimos, hiperônimos, nomes genéricos, expressões nominais etc.

De um modo geral, a literatura linguística apresenta as categorias anafóricas encontradas no Português Brasileiro. Segundo Koch e Elias (2005), Koch (2006) e Figueiredo (2003), as categorias anafóricas dividem-se em: anáfora pronominal, anáfora fiel, anáfora infiel, anáfora por nominalização, anáfora indireta, anáfora associativa e anáfora por elipse. Logo, essas classificações ocorrem de acordo com as características gramaticais que o elemento que funciona como anáfora possui. Nesse sentido, na anáfora pronominal o elemento que funciona como anáfora faz parte da categoria dos pronomes. Esse tipo de anáfora acontece com o uso de pronomes pessoais, oblíquos, reflexivos, indeterminados, demonstrativos, possessivos, relativos e advérbios pronominais. Vejam-se, por exemplo, as sentenças com anáforas pronominais em (21-28).

(21) Ontem eu vi a Bárbara e ela estava animada
(22) Comprei um carro e dei-o de presente para minha filha
(23) Vendem-se carros.
(24) Os vestidos serviram? Nenhum serviu.
(25) De quem é a bolsa? Essa é da Bárbara.
(26) Essa caneta é minha.
(27) O problema que eu disse não é fácil.
(28) Estou em Cataguases. Aqui faz um calor infernal!!

Assim como Koch e Elias (2017), Marcuschi (2008) também trata a referência pronominal como uma estratégia de coesão textual. De acordo com ele, a referência pronominal pode ser de natureza endófora, quando é resolvida na iminência textual e exófora, quando tem referência a um elemento contextual, externo ao texto. Dentro da referência endófora existem dois tipos, a anafórica e a catafórica.

A estratégia chamada de anáfora fiel ocorre com a repetição total ou parcial do referente. Nesse tipo de mecanismo, o elemento anafórico pode vir acompanhado de um pronome demonstrativo, como observa-se nas sentenças (29) e (30).

(29) O menino ganhou um prêmio, porém o menino não merecia.
(30) O menino ganhou o prêmio, porém este menino não merecia.

A estratégia conhecida como anáfora infiel ocorre quando existe um processo de substituição lexical no elemento anafórico, conforme (31) e (32). Nessa estratégia, a referenciação ocorre pelo uso de adjetivos, sinônimos, metáforas e paráfrases. Segundo Figueiredo (2000), nesse tipo de mecanismo anafórico, a correferência acontece quando o segmento de realidade designado tenha as propriedades requeridas ao mesmo tempo por uma e por outra unidade".

(31) O João tentou comprar um sapato e o bobo não estava com dinheiro.
(32) O João tentou comprar um sapato e o garoto não estava com dinheiro.

Já na anáfora por nominalização, acontece um processo de sumarização textual. O elemento anafórico refere-se a toda estrutura precedente no texto. Essa estratégia pode ser percebida em (33). O excerto (33) foi retirado do corpus do português brasileiro disponível na base de dados do *Sketch Engine*.

(33) A capacidade de mobilização em torno da equipa das Quinas , valeu ao BES a entrada no Livro de Recordes do Guinness , em 2006 , ao conseguir reunir cerca de 25.000 mulheres no estádio do Jamor , que formaram a maior bandeira humana de sempre. Um feito inédito , e ainda hoje recordado , que consolidou em definitivo a aposta do BES no futebol e no expoente máximo dessa paixão , a Selecção .

A estratégia de anáfora indireta, embora seja um mecanismo mais complexo e de difícil identificação, aparece em muitos gêneros do discurso. A identificação e interpretação desse tipo de anáfora envolvem diversos domínios cognitivos. Assim, a interpretação da anáfora indireta estaria relacionada à interpretação de todo o arranjo textual por meio de análise de pistas presentes na estrutura do texto, de imagens etc.

Koch (2002) afirma que as anáforas indiretas podem funcionar como âncora representações linguísticas de complexidade sintática, semântica e conceitual extremamente variável.

O mecanismo anafórico conhecido como anáfora associativa ocorre quando os elementos que fazem parte do processo de referenciação são unidades lexicais que compõem ao mesmo campo semântico. Nos exemplos (34) e (35), o termo “garçom” tem como referência o item lexical “restaurante”. Isso sugere que algumas anáforas podem ser facilmente associadas a determinadas cenas ou Frames.

(34) Entrei em um restaurante, e o garçom veio me atender.

(35) A praia estava vazia. Só tinha a orla e o mar.

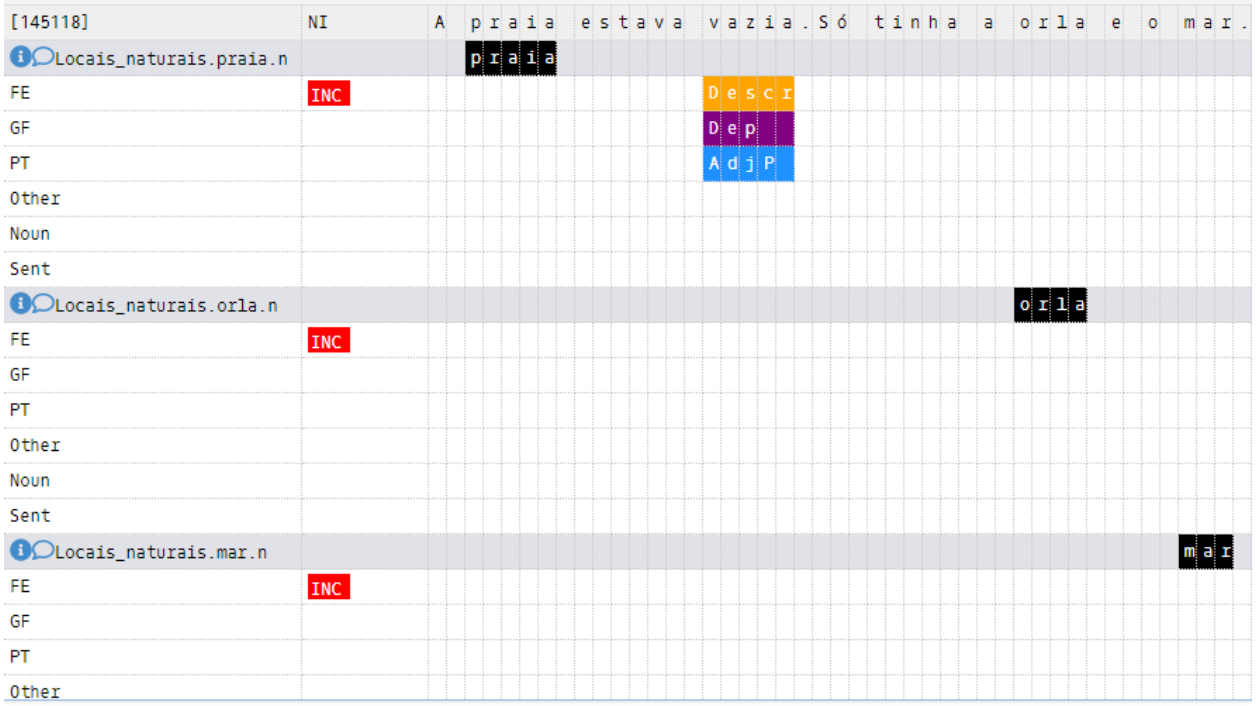
Por último, como já mencionado nos exemplos (34) e (35), há a estratégia de anáfora por elipse. Esse tipo de mecanismo acontece por meio de uma referenciação implícita na oração, isto é, quando não há a presença de um elemento que faça a ligação direta com o referente. Essa situação pode ser observada no exemplo (36).

(36) Paulo estuda morfologia e eu, semântica.

A partir das descrições acima, pode-se perceber que as estratégias anafóricas funcionam de acordo com princípios particulares e são formadas por itens lexicais e classes de palavras diversas. Essa diversidade exige, por muitas vezes, um estudo extenso e detalhado de aspectos linguísticos e computacionais que não cabem nesta dissertação. Por outro lado, algumas estratégias anafóricas, como a anáfora associativa e a fiel, já estão mapeadas na base de dados de recursos linguístico-computacionais como a FrameNet Brasil. Como pode ser observado na Figura 7, a anáfora associativa pode ser facilmente sinalizada na anotação lexicográfica por meio da repetição de frames evocados. A anotação lexicográfica funciona por meio da anotação de unidades lexicais específicas da sentença. Observa-se na Figura 5 a anotação das unidades lexicais *praia, orla* e *mar*. Percebe-se que as três unidades lexicais evocam o mesmo frame, o que garante a associação

entre elas. Em geral, na FN Br, todas as anotações lexicográficas das sentenças são feitas na ferramenta *Webtool 3.0.* A ferramenta funciona como um banco de dados no qual estão localizadas todas as informações sobre os frames, construções e unidades lexicais (ULs) acionadas por eles.

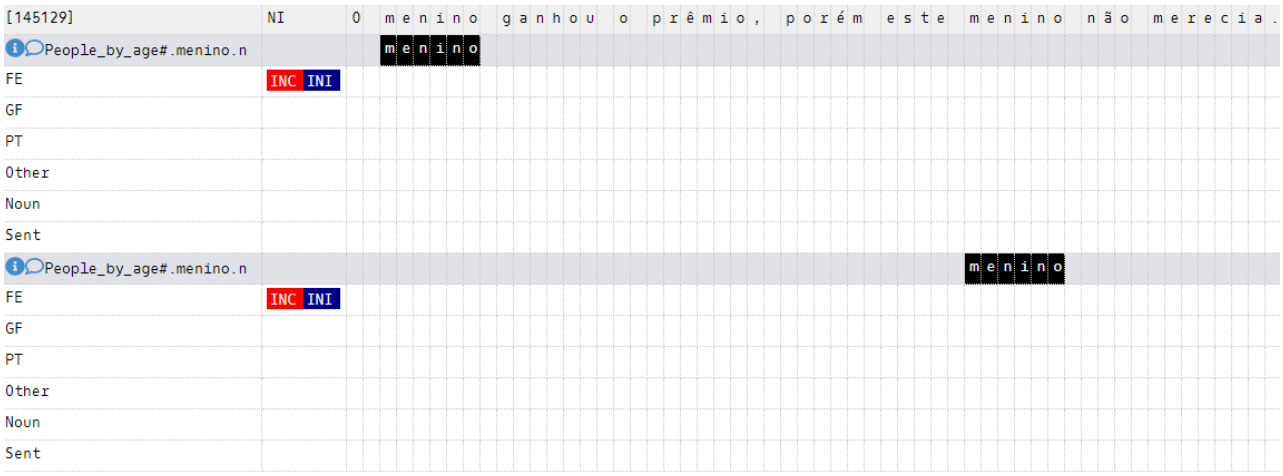
Figura 7: Anotação lexicográfica de anáfora associativa.



Fonte: Retirado de *Webtool 3.0*

Da mesma forma, a estratégia de anáfora fiel não é estudada neste trabalho porque o mecanismo ocorre por meio da repetição do item lexical anterior e, consequentemente, com a repetição do frame previamente anotado. Essa situação aparece na Figura 8 abaixo. A figura mostra a sentença *O menino ganhou o prêmio, porém este menino não merecia* anotada lexicograficamente. Observe que as duas unidades lexicais ativam o mesmo frame *People_by_age* e têm elemento de frame incorporado de *pessoa* e uma instanciação nula indefinida que corresponde ao elemento de frame *idade.*

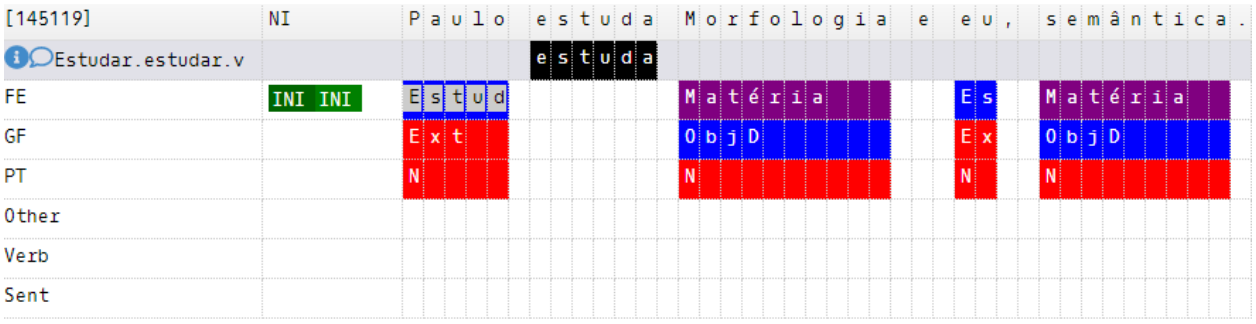
Figura 8: Anotação lexicográfica de anáfora fiel



Fonte: Retirado de *Webtool 3.0*

Como explicado acima, o estudo da estratégia anafórica por elipse não é analisado neste trabalho porque necessita de uma descrição robusta e de um mapeamento de zeros lexicais, o que, para um modelo baseado em anotação, é um problema. Do ponto de vista computacional, seria um grande desafio identificar os *gaps*. A Figura 9 esboça um exemplo de anotação lexicográfica da sentença *Paulo estuda morfologia e eu, semântica.* Observa-se na anotação que o verbo *estudar* foi anotado na sentença referente, mas ainda não existem ferramentas computacionais para identificá-lo na sentença anafórica por elipse.

Figura 9: Anotação lexicográfica de anáfora por elipse.



Fonte: Retirado de *Webtool 3.0*

Por último, na anáfora infiel e por nominalização, os itens lexicais usados evocam frames próprios, que podem ser muito diversos entre si, o que torna o reconhecimento dos elementos ativadores da construção de anáfora muito difíceis de serem identificados. O estudo minucioso dessas duas estratégias extrapolam o escopo de uma dissertação de mestrado.

Figura 10: Anotação lexicográfica de anáfora por nominalização.

Fonte: Retirado de *Webtool 3.0*

Tendo isso por base, a nossa proposta de estudar as construções anafóricas pronominais se justifica porque os pronomes atendem a requisitos computacionais e linguísticos necessários para a sua modelagem. Dessa forma, notam-se alguns argumentos para o estudo da anáfora pronominal. Entre eles, pode-se observar que, com a construção de um modelo computacional que interprete as construções anafóricas pronominais, essa classe passa a ser anotada. Por meio disso, contribui para o enriquecimento da base de dados da FN-Br. Além disso, existe uma relevância metodológica para a escolha da anáfora pronominal que é relacionada com as características da classe dos pronomes. Ela é relativamente fechada e pequena e, por causa disso, possibilita o seu estudo e tratamento computacional em uma dissertação de mestrado. Por fim, os pronomes têm formas raramente homônimas a outras classes, isso significa que existem poucas possibilidades de uma ativação equivocada da construção anafórica por outra palavra ou estrutura linguística.

Para finalizar, Koch e Elias (2003) apontam que as principais estratégias de referenciação textual no português Brasileiro acontecem por meio do uso de pronomes, de expressões nominais definidas, ou seja, pelo uso de um determinante definido seguido de um nome e não definidas, constituído por artigos indefinidos seguido de um nome.

Isso posto, passamos, a seguir, a uma revisão mais detalhada do mecanismo da anáfora pronominal.

4.2 ANÁFORA PRONOMINAL

Os pronomes são, por natureza, ferramentas de substituição e referenciação. Tais características podem ser percebidas na própria etimologia da palavra que o designa “pro+nome” que significa “substituição de um nome”. Conforme afirma Castilho (2010), deve-se ao estatuto semântico-discursivo dos pronomes a retomada ou antecipação de participantes por meio da foricidade, através de processos anafóricos e catafóricos.

Analisando esse fenômeno por um viés descritivo, Castilho (2010) apresenta os mecanismos anafóricos vigentes no Português Brasileiro por meio do processo de pronominalização. A análise do fenômeno de co-referência é caracterizada como uma função básica desempenhada pela classe dos pronomes. De acordo com o autor, do ponto de vista semântico-discursivo, os pronomes representam pessoas do discurso e permitem a retomada ou antecipação de participantes por meio de anáforas e catáforas. O autor acrescenta que os pronomes podem ser primitivos (ou primeiros), quando são resultados de processos de mudança linguística ao longo do tempo (eu, tu, si, este, esse e ele) e derivados, quando derivam dos primitivos (meu, teu, seu, nosso e vósso). Segundo ele, os pronomes podem assumir uma função demonstrativa (ou dêitica), quando apontam para uma coisa no mundo ou ainda uma função relativa (anafórica), quando fazem menção de algo que foi dito anteriormente no discurso. Além disso, os pronomes possibilitam duas relações de base: a da proximidade, quando a forma acompanha o nome e da substituição, quando ela substitui o nome. A propriedade da substituição é, por excelência, a mais importante para este trabalho.

Os pronomes são classificados segundo à função sintática que desempenham no interior da sentença, isso significa que suas categorias são definidas mediante às funções que desempenham em relação aos substantivos, como mencionado acima. Por meio disso, pode-se separar os pronomes nas seguintes classes: pronomes pessoais, possessivos, de tratamento, demonstrativos, relativos, interrogativos e indefinidos. Para a apresentação dos tipos pronominais do português brasileiro padrão, toma-se por base descrições encontradas nos trabalhos de Castilho (2010) e Cunha (2008), já que os autores apresentam o fenômeno pronominal anafórico por meio de uma abordagem linguística mais normativa-descritiva. O autor inicia com a apresentação de os pronomes pessoais porque eles relacionam-se com os outros tipos, ou seja, servem de base para a formação dos demais pronomes. Segundo ele, esse tipo de categoria pronominal pode assumir formas retas e oblíquas de acordo com a função sintática que eles assumem dentro da sentença. Assim, quando funcionam como sujeitos da oração, são chamados de pronomes retos, quando ocupam a posição normal de objeto ou complemento são chamados de pronome oblíquo.

Para a apresentação do pronome do tipo oblíquo, Castilho (2010) diz que esse subtipo de pronome também pode desempenhar uma função reflexiva quando as ações e os eventos expressos pelo verbo são aplicados sobre o referente do sujeito. Nessa situação, ocorre um processo de referenciação anafórica realizado pela anáfora representada pelo pronome reflexivo e pelo substantivo ao qual ele se refere. Os exemplos abaixo ilustram a anáfora pelo uso de pronomes reflexivos no português padrão. Em todas as sentenças, o pronome oblíquo com função reflexiva funciona como mecanismo de substituição e de economia linguística e indica que o agente da ação também tem o papel de paciente. Assim, em todos os exemplos abaixo ocorre o fenômeno de anáforas por pronominalização pelo uso do pronome oblíquo com função reflexiva. Perini (2010) afirma que o pronome oblíquo com função reflexiva é peculiar porque o papel temático atribuído ao sujeito pelo verbo e o papel atribuído ao objeto se aplicam ao mesmo referente, conforme (37-39).

(37) Eu me formei na faculdade.

(38) Tu te cortaste com o vidro.

(39) Ele se mudou para outro país.

Prosseguindo com o pronome possessivo, tradicionalmente, ele transmite uma relação de posse entre um possuidor e um objeto ou entidade possuída. Em outras palavras, esse tipo de pronome estabelece uma relação discursiva entre a entidade possuidora e a coisa possuída. Os possessivos também desempenham a função de referenciação em duas posições sintáticas, a posição canônica (anteposto ao substantivo) e a posição anafórica por elipse. Como será visto abaixo, os possessivos, por desempenharem funções semelhantes a um adjetivo e a um substantivo, concordam em número e gênero com os substantivos que acompanham. No excerto (40), o pronome possessivo refere-se à pessoa do discurso em relação ao objeto em sua posição habitual, quando precede o substantivo. No exemplo, o pronome *minha* tem um papel referencial porque aponta para a pessoa do discurso que fala no momento da enunciação. Dessa forma, o objeto bolsa é relacionado com uma entidade possuidora, neste caso, a primeira pessoa do discurso.

(40) A minha bolsa é marrom.

O processo de referenciação com pronomes possessivos também pode acontecer por meio do mecanismo de elipse. Conforme já mencionado, o fenômeno de elipse acontece por meio de um mecanismo de referenciação sem a presença de uma anáfora explícita. No exemplo (41), o pronome tem a função de referenciar o objeto que apareceu anteriormente no discurso. Nesse tipo de anáfora ocorre uma referenciação implícita na sentença porque o sintagma nominal anafórico é reduzido a apenas um dos elementos, representado pelo pronome possessivo.

(41) A blusa é minha.

Para resumir, os pronomes possessivos podem assumir múltiplos valores. A função mais empregada é a adjetival, isso significa que o pronome vem anteposto ao substantivo. Essa situação pode ser observada no excerto (42). Além disso, o pronome

possessivo também pode desempenhar a função de um substantivo. Nesse último caso, ocorre um processo de referenciação anafórica por pronome possessivo.

(42) A sua casa é enorme!

Os tipos pronominais chamados de pronomes de tratamento são palavras ou locuções usadas para referenciar pessoas do discurso que têm atributos de autoridade. Eles podem assumir também funções de referenciação a depender da posição que ocupam no interior da sentença. Desempenham funções e posições iguais aos pronomes pessoais desempenham. Como exemplo, observe a sentença (43) e (44). Pode-se perceber o pronomes de tratamento com função referencial anafórica na sentença (44). Outras expressões também são consideradas como pronomes de tratamento como em (45) e (46).

(43) Vossa Excelência julgou o réu.
(44) O juiz julgou todos os réus, vossa excelência encerrou todos os casos.
(45) O senhor doutor vai trabalhar hoje.
(46) O pai vai levar a mãe para a praia

Ainda dentro da classe dos pronomes existem os do tipo demonstrativo. A categoria dos demonstrativos relaciona-se diretamente com os pronomes pessoais. Os demonstrativos podem desempenhar duas funções básicas, funções dêiticas ou anafóricas quando, respectivamente, demonstram um objeto no mundo e quando apontam para algo que já apareceu anteriormente no discurso. Devido à sua função dêitica, eles são utilizados para determinar a proximidade espacial ou temporal da pessoa do discurso. As duas funções podem ser vistas no exemplo (47) e (48). Em (47) encontra-se um exemplo de pronome demonstrativo com função dêitica, isso significa que ele aponta para um referente é ancorado no momento em que o discurso é proferido, de informações extralinguísticas como gestos e olhares. Já no exemplo (48), há um exemplo de uso pronominal com função anafórica. Os excertos (47) e (48) foram retirados de um corpus do português brasileiro presente no Sketch Engine.

(47) - Mas não vilarejo , e foi então que meu guia , a quem chamo de Petrus (embora não seja esse o seu nome) , me disse : " Quem é esse deus Rama , afinal ?”

(48) Uma vez estabelecida uma dose que seja eficaz e tolerada , essa deve ser mantida .

Salta aos olhos a multiplicidade de formas dos pronomes do tipo relativo. Eles são, por natureza, palavras que se referem a termos que já apareceram no discurso. No que se refere à função de referenciação anafórica, os pronomes relativos enquadram-se como o subtipo mais proeminente da categoria. A estrutura de formação desse tipo de anáfora é a tradicional. Os pronomes relativos funcionam como elementos anafóricos e apontam para um antecedente que pode ser de natureza nominal, como, por exemplo, um substantivo, um adjetivo, um pronome, e ainda de outras naturezas como um advérbio ou até mesmo uma oração inteira. Tendo em vista a abrangência desse tipo de pronome. Os pronomes relativos também desempenham funções sintáticas diversas no interior da sentença de acordo com a posição que ocupam, como, por exemplo, de sujeito, objeto direto, objeto indireto, predicativo, adjunto adnominal, complemento nominal, adjunto adverbial, agente da passiva. Nas frases abaixo, (49-54), encontram-se, respectivamente, exemplares de cada tipo de anáfora pronominal relativa.

(49) O ácido undecilênico e o ácido benzóico são agentes fungistáticos que impedem o crescimento de bactérias e fungos , sendo indicados contra uso de TIORFAN (racecadotril) é contra-indicado para pacientes com conhecida hipersensibilidade aos componentes da fórmula .

(50) O menino que eu amo é brasileiro.

(51) A menina, que é alta, chegou.

(52) É bom que se diga que as crianças que passam fome nas ruas são um sério problema social , cuja resolução deve ser uma das prioridades máximas de qualquer governo (adjunto adnominal)

(53) Toda a louça onde a criança come deve ser bem limpa (adjunto adverbial)

(54) A Faculdade de Ciências Econômicas de Anápolis é uma entidade pública, criada pela Lei 3.430 , de 05 de julho de 1961 , do Governo do Estado de Goiás , por quem é ainda mantida até a presente data . (agente da passiva)

Ainda no estudo do mecanismo anafórico pelo uso de pronomes relativos, a referenciação também ocorre sem a presença de um antecedente. Nesse caso, acontece um processo chamado de referência genérica. Esse tipo de anáfora é encontrada com frequência em textos de cunho literário, como pode ser observado no exemplo abaixo. Em (55) o eu-lírico não pretende obter uma resposta à sua pergunta, mas o pronome tem uma função referencial genérica.

(55) Quem não equilibra trabalho com descanso , perde o entusiasmo , esgota sua energia , e não chega muito longe.

De acordo com os dados apresentados acima, observa-se que os pronomes relativos compõem a subclasse pronominal com maior diversidade de formas. Por esse motivo, entende-se que o estudo do fenômeno anafórico realizado apenas com esse tipo de pronome seria suficiente para fundamentar uma pesquisa de cunho científico.

Da mesma forma que os relativos, os pronomes indefinidos formam uma subcategoria com muitas variações em suas formas. Eles são utilizados para indicar algo indeterminado ou vago e são usados em terceira pessoa. Eles podem se apresentar por meio de uma única palavra ou uma locução pronominal. Esse tipo de pronome, canonicamente, acompanha os substantivos e funciona com valor adjetival nas sentenças. Os indefinidos também desempenham funções de referenciação, como observa-se no exemplo abaixo. Nos excertos (58) e (59), o pronome indefinido *nada* e *nenhum* funcionam como anáfora e têm como antecedente o que foi dito anteriormente no discurso, em (56) todo o restante da sentença e em (57) o carro. No caso dos pronomes indefinidos, percebe-se que o fenômeno de referenciação anafórica acontece com mais frequência nos pronomes com valor negativo como, por exemplo,

nenhum, nada. Normalmente, esses tipos de pronomes participam do processo de referenciação por elipse.

(56) Metade do tesouro do governo já havia sido colocado na bolsa do mendigo, e nada.
(57) Medido em carro , era uma droga - não dava para comprar nenhum.

Por fim, há a categoria dos pronomes do tipo interrogativo. Eles compõem a classe dos pronomes e, em geral, são usados para a formulação de perguntas. Em sua gramática, Cunha (2008) afirma que os pronomes interrogativos estão diretamente ligados aos pronomes indefinidos porque o processo de significação acontece mediante à resposta de uma determinada pergunta. Além disso, esse tipo de pronome pode referenciar coisas e pessoas. Nesse caso, a referência é catafórica, isto é, aponta para algo que aparecerá posteriormente no discurso. Por esse motivo, os pronomes interrogativos, exemplificados em (580) e (59) não fazem parte do escopo desta investigação.

(58) Quem é o senhor para vir até aqui perturbar a minha tranquilidade?
(59) Qual o melhor exemplo a seguir ?

A análise levada a cabo até aqui sustenta a proposta de estudo do fenômeno de anáfora pronominal, visto que esse tipo de mecanismo anafórico é de alta complexidade e abrange grande variedade de formas e diferentes princípios de funcionamento. Tendo isso como base, no que tange ao escopo deste trabalho, será dada maior ênfase aos mecanismos anafóricos pronominais.

Em outras palavras, os dados acima foram apresentados com o intuito de descrever e delimitar o fenômeno a ser estudado aqui, a anáfora pronominal. Por meio dessa análise, embasamos e definimos os métodos linguísticos e computacionais que serão necessários para o desenvolvimento de uma construção anafórica pronominal.

Neste capítulo, apresentamos um panorama do fenômeno da anáfora no PB. Tendo isso como objetivo, exploramos a noção de representação e referência

defendida por Lyons (1977). Depois, descrevemos diferentes estratégias anafóricas encontradas no português brasileiro como, por exemplo, a anáfora indireta, fiel, infiel, associativa, por elipse, por nominalização e a pronominal.

5 METODOLOGIA

Este capítulo apresenta as etapas para a modelagem das construções anafóricas do PB. Com esse propósito, ele é dividido entre as etapas de apresentação do corpus de análise, cadastramento de construções, análise de ocorrências, modelagem e experimento de reconhecimento de construções.

5.1 CORPUS

Para o estudo do fenômeno de referência de longa distância, antes foi necessária a delimitação do corpus de análise. Com esse objetivo, dois corpora foram selecionados dentro da base de dados da FN-Br, o *General* e o *Natural Language Generation*. O primeiro corpus é composto por 25 gêneros textuais sendo eles: artigo de opinião, autobiografia, carta do leitor, comando, conselho, conto, conto fantástico, crônica literária, editorial, ensaio, entrada de verbete, fábula, instruções de uso, lenda, notícia, receita, regulamento, relato de viagem, relato histórico, reportagem, resenha crítica, resumo, texto explicativo, texto expositivo e verbete de enciclopédia. O segundo corpus conta com textos retirados do *website Wikipedia* sobre a cidade do Rio de Janeiro.

Os corpora foram escolhidos segundo critérios de adequação linguística, variedade de gêneros textuais e aproveitamento da base de dados da FN-Br. Assim, o corpus General e o NLG foram selecionados por se adequarem às exigências dos três critérios. Os corpora frequentemente são utilizados para enriquecimento semântico e morfossintático de aplicações no laboratório através de anotação de texto corrido. Os corpora são formados, respectivamente, por dados provindos de gêneros textuais que circulam no ambiente escolar e por textos disponíveis na Wikipedia sobre a cidade do Rio de Janeiro.

5.2 ANÁLISE DE OCORRÊNCIAS DE ANÁFORAS PRONOMINAIS

Considerado o corpus de análise, o próximo passo consistiu na leitura dos textos e registro das ocorrências de anáforas pronominais. Em um primeiro momento, as

anáforas pronominais foram separadas segundo a sua função: adjetiva ou substantiva. Os dados ainda foram agrupados por tipo pronominal e localidade sintática. Assim, encontramos nos corpora pronomes com função pessoal, demonstrativa, possessiva, relativa, de tratamento e reflexiva. Por fim, classificaram-se os exemplos encontrados quanto à localidade sintática do antecedente da anáfora pronominal.

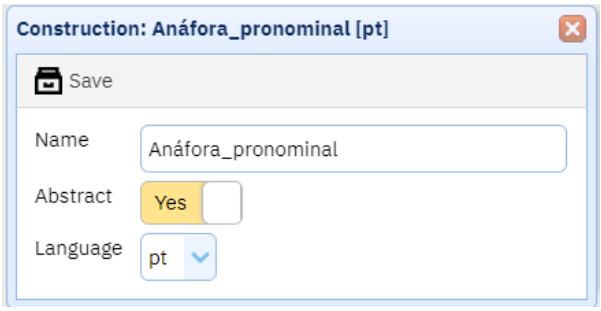
5.3 MODELAGEM DE CONSTRUÇÕES ANAFÓRICAS

Nas seções anteriores, foram mostradas as etapas de levantamento de corpora e classificação de ocorrências de anáforas pronominais. Ambas as etapas foram fundamentais para o processo de modelagem das estruturas estudadas nesta dissertação.

O processo de modelagem das construções anafóricas ocorreu através do cumprimento de etapas. Em um primeiro momento, através da análise do corpus, percebemos a predominância de uma estrutura abstrata, que porta características gerais presentes em todas os demais construtos. Essa estrutura mais genérica foi modelada como a construção de *Anáfora pronominal*. Mediante essas características, essa construção funciona como nó central das demais construções, isso significa que ela porta todas as informações morfossintáticas compartilhadas pela família de construções pronominais anafóricas. Por meio dela, as demais estruturas anafóricas foram desenvolvidas, como será visto no capítulo 6. A partir da *Anáfora pronominal*, os demais nós da rede anafórica foram modelados. A noção de rede construcional é postulada por Goldberg (1995). De acordo com a autora, as construções se organizam em rede e podem manter relações diferentes entre si.

A modelagem computacional de construções começa pelo cadastramento do nome da estrutura. Na Figura 11, vemos como isso ocorre.

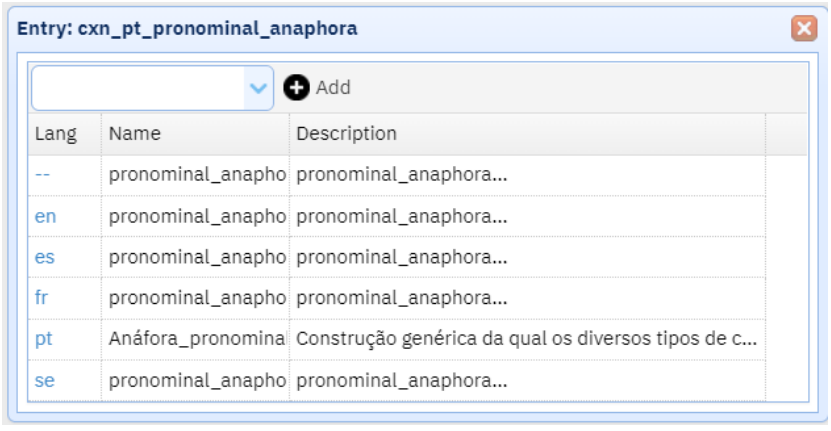
Figura 11: Edição de entrada da Anáfora_pronominal



Fonte: Retirado de *Webtool 3.0*

Para o cadastramento da estrutura, indicamos o nome no campo “name”. Já no campo “Abstract” indicamos a natureza da construção em relação ao seu potencial de licenciamento de construtos. Como dito anteriormente, percebemos que a construção genérica *Anáfora pronominal* é caracterizada como uma estrutura abstrata que agrupa propriedades de uma família de construções, mas que não é capaz de licenciar construtos na língua.

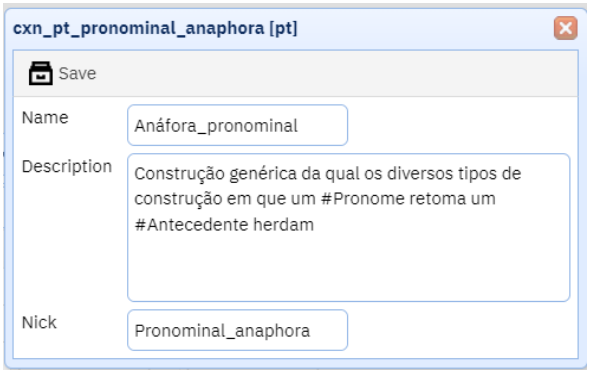
Figura 12: Edição de entrada por idioma



Fonte: Retirado de *Webtool 3.0*

Uma vez que a interface do Constructicon da FrameNet Brasil é multilíngue, na Figura 12 vemos o processo de edição da entrada da *Anáfora pronominal* na base de dados do sistema. Pela figura, percebemos que a descrição das estruturas pode ser realizada em cinco línguas: inglês, espanhol, francês português e sueco. Para a edição, clicamos na opção da língua portuguesa. Tal situação é vista com mais detalhes na Figura 13 abaixo.

Figura 13: Edição da definição da construção em português



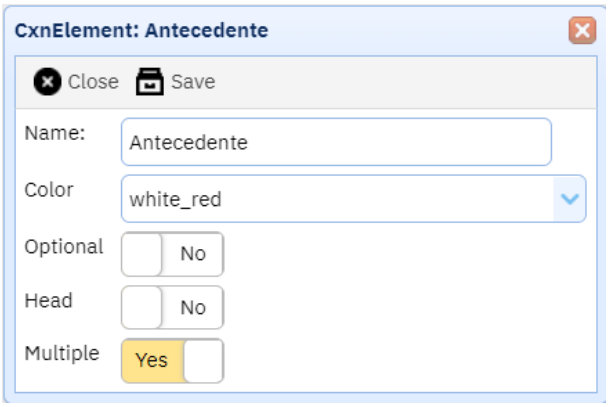
Fonte: Retirado de *Webtool 3.0*

Na Figura 13, ainda observamos a descrição da construção da *Anáfora pronominal* no português brasileiro. Para esse caso, mantemos o nome da construção em português, assim como a sua descrição. Por outro lado, o nick da estrutura é apresentado em língua inglesa. Esse processo de cadastramento é realizado com todas as estruturas e elementos da construção. A partir disso, os elementos da construção são modelados. Na *Anáfora pronominal*, assim como nas demais construções da rede anafórica pronominal, há dois elementos principais: o Antecedente e o Pronome.

Quando do cadastramento de cada EC, preenchem-se informações relativas ao seu nome, ao código de cor que será usado para identificá-lo na interface de anotação, além de três parâmetros binários que informam se (a) o EC é opcional ou não, (b) se é

o núcleo da construção e (c) se pode ser instanciado mais de uma vez na construção (vide Figura 14).

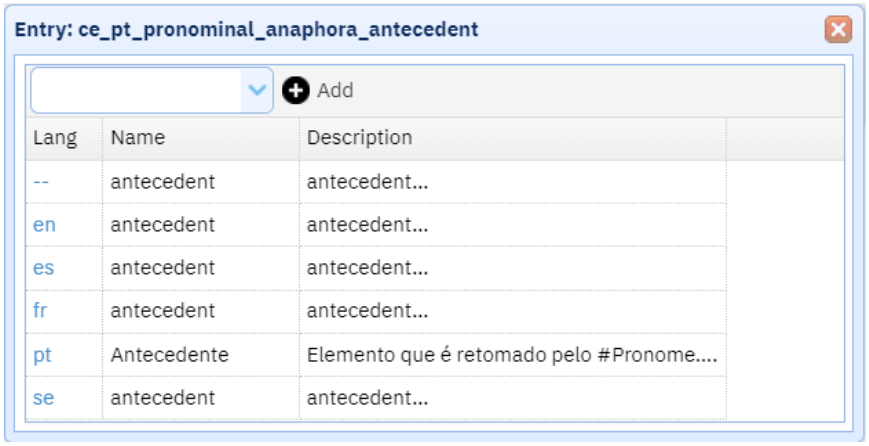
Figura 14: Edição da entrada do elemento de construção Antecedente



Fonte: Retirado de *Webtool 3.0*

Depois disso, temos acesso a outra tela com as informações do EC nas mesmas cinco línguas, como podemos ver na Figura 15. Novamente, ao clicarmos na língua para qual queremos realizar a edição, podemos alterar o nome do EC e sua definição.

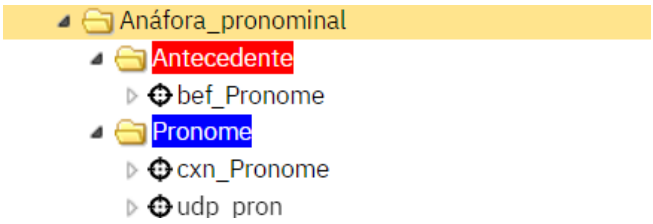
Figura 15: Edição de entrada por idioma



Fonte: Retirado de *Webtool 3.0*

Depois dessa etapa, que tem como foco maior a alimentação da base de dados com elementos úteis para usuários humanos (nome e definição da construção e dos ECs), partimos para a especificação dos constraints. Os constraints são elementos de modelagem cujo objetivo principal é o de formalizar traços da construção e dos ECs que sejam relevantes para a legibilidade por máquina. Por meio dos constraints, estabelecemos a identidade morfossintática e semântica da estrutura e dos seus elementos. Eles funcionam como limitadores da natureza morfológica, sintática e semântica da estrutura. Nesse caso, é possível especificarmos a constituência das construções, a posição de determinados elementos, as relações de herança estabelecidas entre a construção e demais estruturas, a natureza morfológica dos elementos etc. Assim, para cada estrutura da família de *Anáfora pronominal* estabelecemos limites, os constraints, como será visto com mais detalhes no próximo capítulo de modelagem das construções.

Figura 16: Restrições aplicadas à construção Anáfora_pronominal



Fonte: Retirado de *Webtool 3.0*

Na Figura 16, vemos como os constraints aparecem no sistema *Webtool 3.0* e ainda quais deles foram especificados para a estrutura *Anáfora pronominal*. Nessa construção, o elemento antecedente tem como constraint a informação de que ele precede um pronome. Isso acontece porque nessas estruturas todo antecedente

precede um pronome, esse último elemento funciona como anáfora pronominal. Logo abaixo, vemos os constraints que limitam o elemento construcional pronome. Nesse caso, foi necessário indicar que o elemento tem como construção genérica a cxn_pronome. Ademais, o Constructicon da FrameNet Brasil também utiliza como constraints traços (ou features) definidos no âmbito do projeto Universal Dependencies (UD). As UDs são traços que contribuem para o desenvolvimento de ferramentas linguístico-computacionais. Isso acontece porque, através da definição dessas características, determinamos a natureza gramatical de cada tipo de elemento.

Marneffe *et al* (2021) definem o projeto intitulado *Universal Dependencies* como uma teoria linguística para a anotação morfossintática entre várias línguas. Para entender melhor, o projeto propõe a criação de um inventário de traços universais capazes de serem aplicados para a anotação e representação de diversas línguas naturais. As UDs indicam as relações gramaticais, os traços morfológicos e as classes de palavras aplicáveis aos termos no processo de anotação e modelagem linguístico-computacional e, por meio disso, otimizam a tarefa através da adição dessas informações.

As características morfossintáticas e semânticas cadastradas para as construções são traços variáveis que são preenchidos para cada atributo, definido em termos de um constraint. A Tabela 2 traz um glossário dos tipos de constraints utilizados para a modelagem das construções de anáfora pronominal, bem como os valores aplicáveis a cada um deles.

Tabela 2: Glossário de constraints

Constraints	**Valores**
Construção que licencia o CE (cxn)	*Pronome* *Pronome Pessoal do Caso Reto* *Pronome Demonstrativo* *Pronome Pessoal do caso Oblíquo* *Pronome Possessivo* *Pronome Indefinido* *Pronome de Tratamento* *Pronome Reflexivo* *Pronome Relativo* *Sintagma Nominal*

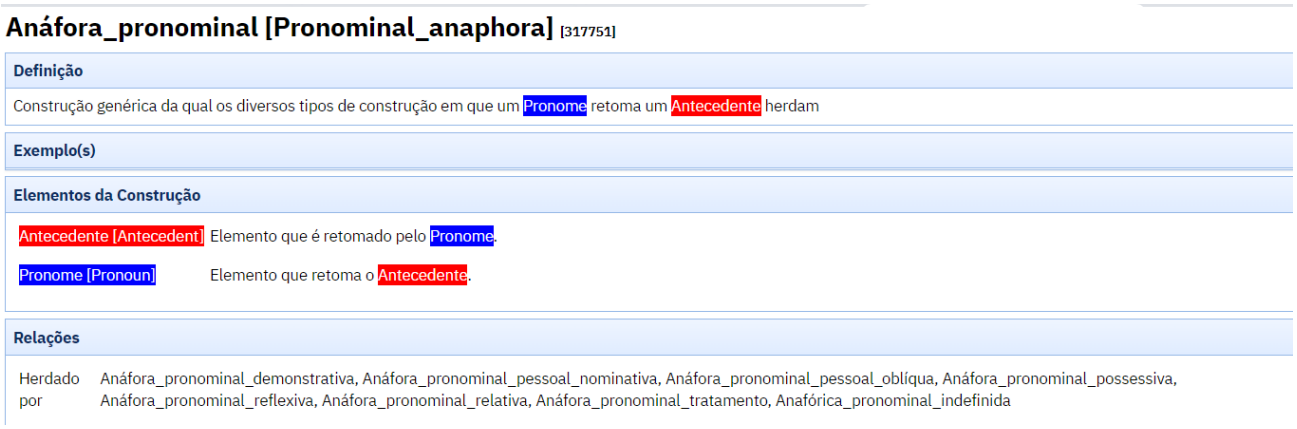
Ordem na sentença	*Antes de CE (bef_Xxxx)* *Depois de CE (aft_Xxxx)* *Imediatamente antes de CE (meet_Xxxx)*
Classe de palavra conforme as UDs (udp)	*Nome (n)* *Pronome (pron)*
Traços adicionais da classe de palavra conforme as UDs (udf)	*Caso nominativo (nom)* *Caso acusativo (acc)* *Caso dativo (dat)* *Caso acusativo (Acc)* *Caso nominativo (Nom)*
Tipo pronominal (udf_prontype)	*Demonstrativo (dem)* *Possessivo (poss)* *Reflexivo (refl)* *Indefinido (ind)* *Relativo (rel)* *Indefinido (Ind)* *Pessoal (Prs)*
Traços de concordância	*Número (nbr)* *Gênero (gdr)* *Pessoa (prs)*
Herança construcional (inh)	*Construção herdada pela construção modelada*
Família de frames do item lexical que preenche a posição na construção (fam)	*Frame*

Os traços apresentados na Tabela 2 estão inseridos em toda a família de construções *Anáfora Pronominal* estudadas neste trabalho. Ao longo da apresentação da modelagem de cada estrutura argumental, percebemos que essas características sofrem variações a depender da natureza pronominal de cada elemento construcional.

Agora que já entendemos os arranjos da construção genérica *Anáfora pronominal* em um nível computacional, podemos observar abaixo sua representação na base de dados da FN-Br.

Na Figura 17, vemos a representação da construção *Anáfora pronominal* na base de dados da FN-Br. Na parte superior, vemos o nome da estrutura em português e inglês. Abaixo, encontram-se a definição, os elementos da construção e, no final, as relações que a estrutura mantém com as demais da rede. A construção conta com dois elementos: o Antecedente e o Pronome.

Figura 17: Construção Anáfora_pronominal



Fonte: Retirado de *Webtool 3.0*

Como já discutido no capítulo 4, o fenômeno de referência de longa distância pronominal é composto por dois elementos centrais, o elemento antecedente que funciona como referente, e o pronome, que funciona como a anáfora. Da mesma maneira, essas estruturas são representadas computacionalmente na base de dados da FN-Br.

Por último, percebemos que a construção abstrata *Anáfora pronominal* estabelece relações de herança com as demais construções da rede das construções anafóricas pronominais já modeladas, as quais são apresentadas no capítulo 6.

5.4. EXPERIMENTO DE RECONHECIMENTO DE CONSTRUÇÕES DE ANÁFORA PRONOMINAL

Para testar a adequação da modelagem proposta à tarefa de extração de informações, um experimento de reconhecimento de construções anafóricas pronominais foi desenhado. Nele, processamos automaticamente cada uma das sentenças do corpus onde havia instâncias das construções da família de construções anafóricas.

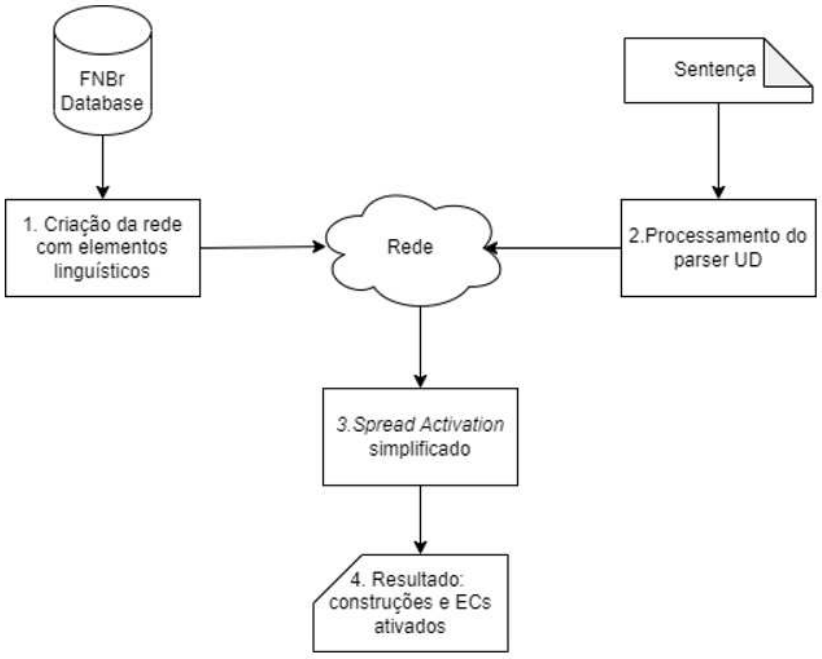
O processamento das estruturas ocorreu através do reconhecimento de traços semânticos e morfossintáticos de cada elemento construcional pelo programa. Em um

primeiro momento, o *parser* construcional identifica a estrutura sintática e as características morfológicas das palavras. A partir disso, ele percorre a rede de construções por meio das constraints modeladas no constructicon.

Uma vez que os traços associados às construções envolvem UDs e frames, o sistema proposto processa a sentença de entrada. A partir disso, ele obtém nós relativos aos lemas e às constraints. Essas informações são acrescentadas à rede e alimentam a base de dados. Em seguida, o parser aplica um processo de Spread Activation simplificado e as construções e elementos ativados são considerados como resultados do processamento das sentenças.

De forma resumida, o processamento conhecido como *spreading activation* funciona com a ativação de conceitos relacionados. Isso significa que, quando uma anáfora pronominal é acionada, todos os conceitos relacionados a ela são ativados, as construções e os elementos construcionais. De forma mais clara, observamos as etapas de funcionamento do sistema através do diagrama representado pela Figura 18.

Figura 18: Funcionamento do sistema de reconhecimento de anáforas



O sistema utilizado neste experimento funciona por meio da detecção desses constrains pre-modelados. Em outras palavras, ele procura todas as construções que têm os determinados constraints cadastrados. Essas limitações conhecidas como constraints serão apresentadas de forma mais detalhada no capítulo 6, no qual descrevemos a modelagem linguístico-computacional das estruturas anafóricas.

Seguindo esse princípio, para a localização dos elementos construcionais da família de Construções Pronominais Anafóricas, o *parser* operou por meio da identificação de constraints cadastrados nos elementos antecedentes. Assim, o parser procurou os elementos construcionais que utilizavam as construções pronominais anafóricas, já que ele só interpreta a rede de um modelo específico.

Neste capítulo, apresentamos as etapas que se referem à metodologia empregada para o estudo das Construções Anafóricas Pronominais. Com esse objetivo, apresentamos o corpus de análise que é formado por dois bancos de textos: o *General* e o *Natural Language Generation,* a análise de ocorrências, seguida pela descrição metodológica do processo de modelagem e do experimento de reconhecimento de construções anafóricas pronominais. Partindo da metodologia descrita neste capítulo, a seguir, veremos a modelagem das construções anafóricas no *Constructicon* em detalhes.

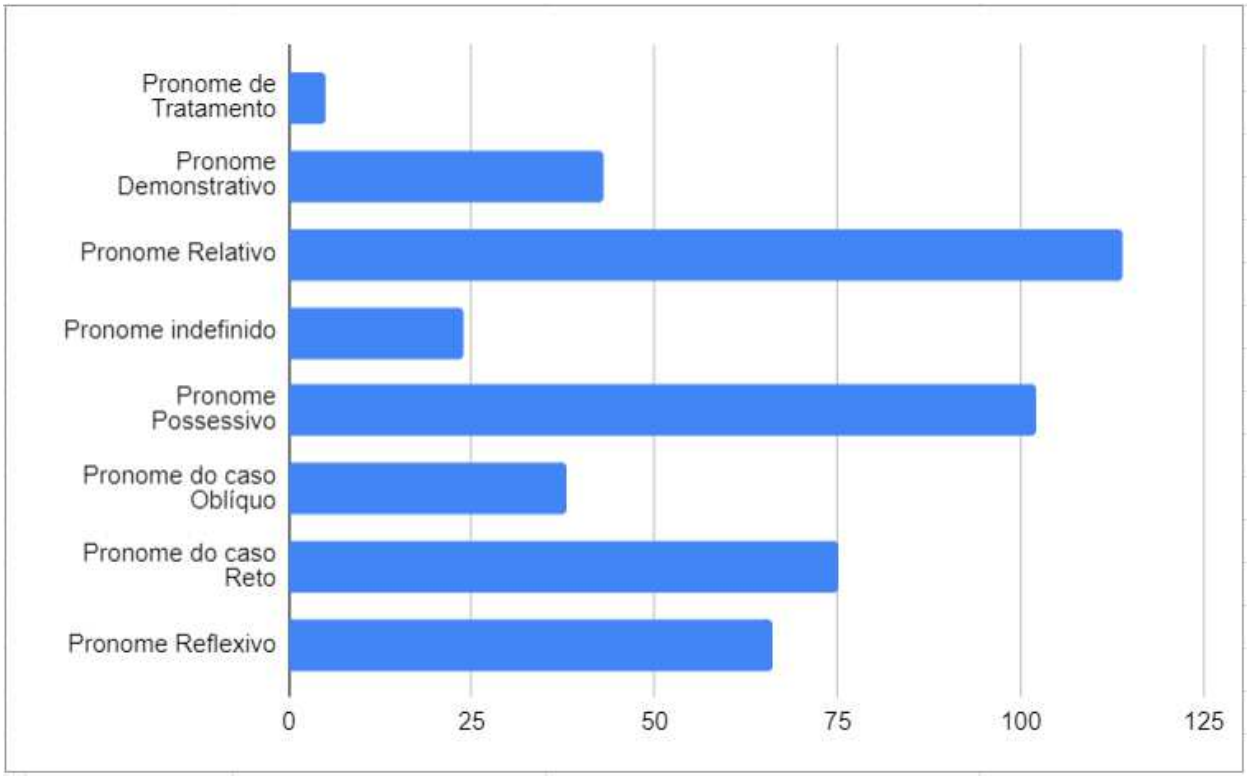
6 MODELAGEM DAS CONSTRUÇÕES ANAFÓRICAS NO CONSTRUCTICON

Neste capítulo apresentamos a modelagem computacional das construções anafóricas do português brasileiro. Tal modelagem resulta da análise do corpus por meio da identificação de instâncias de anáfora pronominal e classificação destas conforme as categorias de análise apresentadas no capítulo 5. Por fim, apresentamos o esboço da família de construções *Anáfora pronominal* por meio de um esquema em rede.

6.1 ANÁLISE DAS OCORRÊNCIAS DE ANÁFORA PRONOMINAL NO CORPUS

Ao todo, encontramos 469 sentenças com ocorrências de anáfora pronominal no corpus. Vemos no Gráfico 1 a quantidade de sentenças por tipo pronominal.

Gráfico 1: Análise de ocorrências de anáforas pronominais segundo o tipo de pronome



Dentro da classificação do corpus, em um primeiro momento, as anáforas foram separadas em duas categorias: anáfora pronominal adjetiva e anáfora pronominal substantiva. Além disso, os enunciados do corpus também foram classificados em relação à localidade sintática do antecedente em relação à anáfora. No que diz respeito à função anafórica desempenhada dentro de um enunciado, dizemos que uma anáfora é pronominal adjetiva quando o pronome não é o núcleo do sintagma e acrescenta informações ao antecedente. Por outro lado, dizemos que uma anáfora pronominal é substantiva quando o pronome funciona como núcleo do sintagma nominal. Assim, no grupo de anáforas do tipo pronominal adjetiva, temos 152 ocorrências encontradas. Já no grupo de anáforas do tipo pronominal substantiva, obtemos 317 ocorrências. Assim, os dados demonstram que, no português brasileiro, conforme o corpus utilizado, o fenômeno de referência de longa distância por pronome ocorre com predominância do padrão substantivo. Isso significa que o pronome funciona como núcleo do sintagma anafórico em sua maioria e não como informação adicional.

Tendo isso como base, dentro do contexto das anáforas pronominais adjetivas, encontramos 1 ocorrência de anáfora adjetiva por pronome de tratamento, com o referente (ou antecedente) localizado na mesma sentença. Houve ainda 33 ocorrências de anáfora adjetiva por pronome demonstrativo com o antecedente localizado no limite da mesma sentença e 18 ocorrências da anáfora pronominal adjetiva por pronome indefinido, localizadas na mesma sentença e, por fim, 100 ocorrências com pronome possessivo.Observamos o percentual de ocorrências de anáfora com função adjetiva por tipo pronominal no Gráfico 2.

Já no grupo de anáforas substantivas, 114 ocorrências são por pronome relativo, 67 por pronome reflexivo, 2 por pronome possessivo e 75 por pronome do caso reto. Ainda encontramos 6 ocorrências com pronome indefinido, 5 por pronome de tratamento, e 38 por pronome do caso oblíquo e 10 por pronome demonstrativo. Observamos um padrão relacionado à localização do antecedente, já que em todos os enunciados o antecedente está localizado em sentenças anteriores. O resumo percentual de cada tipo de anáfora por função substantiva pode ser observado no Gráfico 3.

Gráfico 2: Percentual de tipos pronominais com função adjetiva

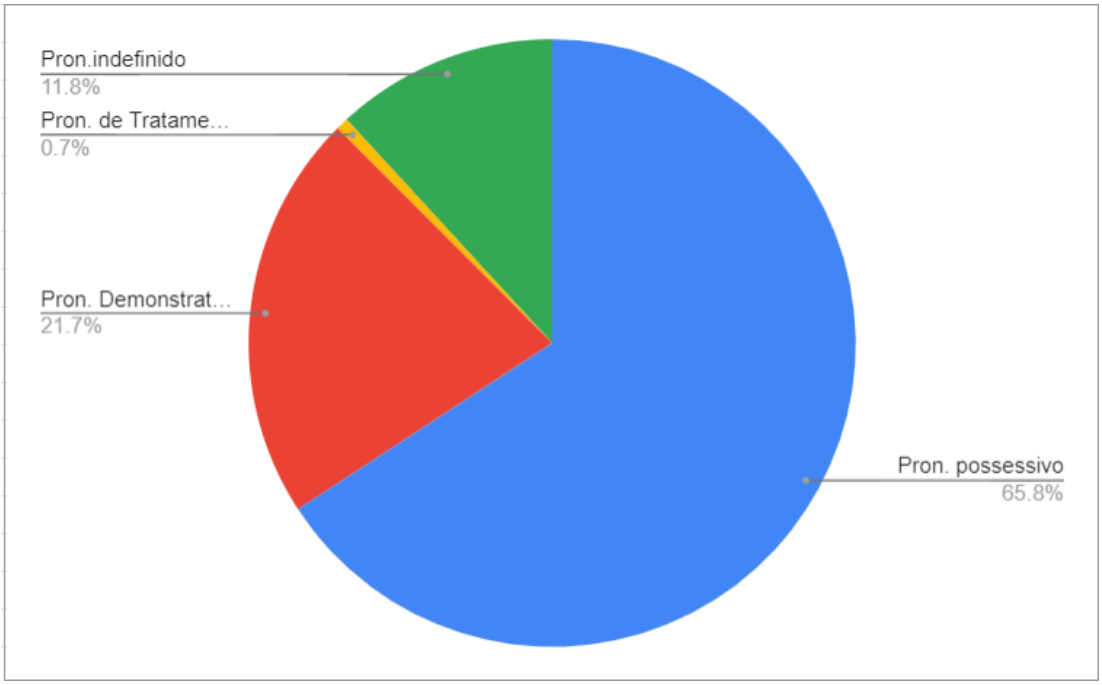
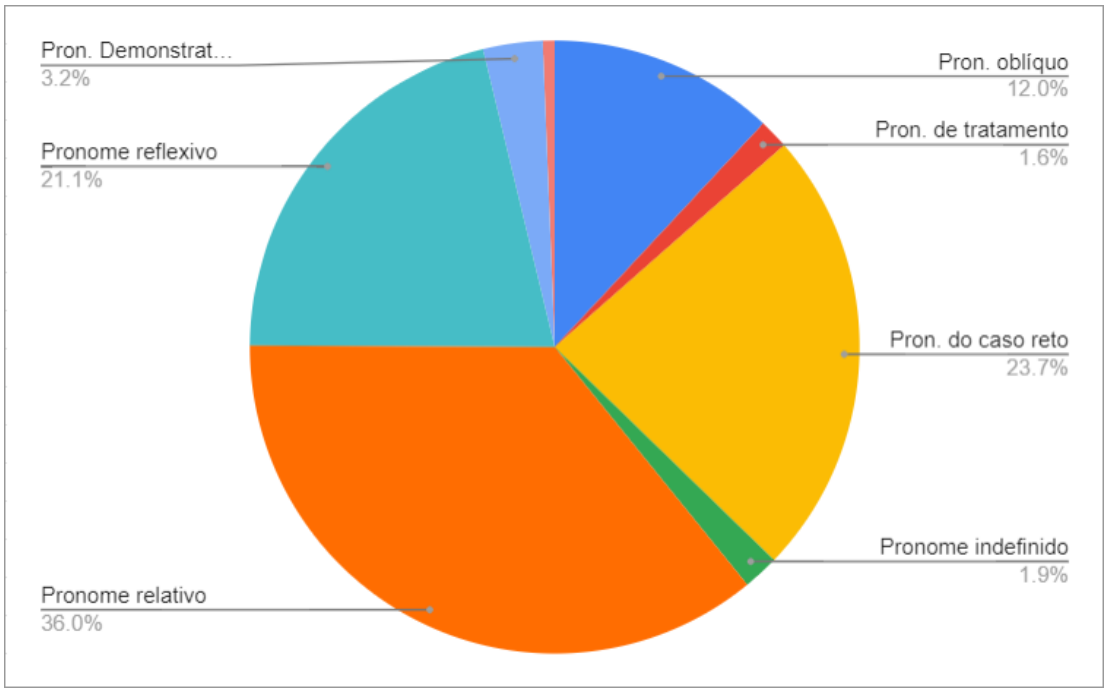


Gráfico 3: Percentual de tipos pronominais com função substantiva



Por meio de análise quantitativa, percebemos a prevalência de padrões anafóricos relativos, reflexivos e do caso reto com função substantiva. Já em relação às anáforas com função adjetiva, percebemos a predominância de padrões com o uso de pronomes possessivos e demonstrativos. A Tabela 3 traz exemplos de ocorrências no corpus por função e tipo pronominal.

Tabela 3: Ocorrências no corpus de anáforas com função adjetiva

Anáfora com função Adjetiva	Anáfora por pronome possessivo	A concorrência saudável beneficia o usuário, que terá mais opções de escolha, e, por que não dizer, os próprios taxistas, que certamente vão melhorar ***seus*** serviços.
	Anáfora por pronome demonstrativo	Muito antes de os brancos atingirem os sertões de Goiás, em busca de pedras preciosas, existiam por ***aquelas*** partes do Brasil muitas tribos indígenas, vivendo em paz ou em guerra e segundo suas crenças e hábitos.
	Anáfora por pronome de tratamento	Era uma vez, uma princesa, conhecida como ***sua alteza***, a Divinha.
	Anáfora por pronome indefinido	***Alguns*** astrólogos definem a astrologia como uma linguagem simbólica, uma forma de arte, ou uma forma de vidência, enquanto outros definem como ciência global e humana.

Anáfora com função Substantiva	Anáfora por pronome do caso reto	Toda vez que uma nova tecnologia é incorporada ao dia a dia dos cidadãos, ***ela*** rompe paradigmas, muda comportamentos, agita mercados e gera descontentamento em grupos e corporações que resistem a se adaptar aos novos tempos.
	Anáfora por pronome do caso oblíquo	“Não,“ exclamou o cordeiro, “eu ainda não bebia água, uma vez que o leite de minha mãe servia de alimento e água para ***mim***...“
	Anáfora por pronome Reflexivo	Meus pais eram ucranianos, que ***se*** conheceram e casaram no Paraná.
	Anáfora por pronome Indefinido	Embora outros jovens quisessem o amor da indiazinha, ***nenhum*** ainda possuía a condição exigida para as bodas, de modo que não houve disputa, e Potira e Itagibá se uniram com muita festa.
	Anáfora por pronome de tratamento	Correu até o Rei Macaco e lhe disse que encontrara um rico tesouro, mas, que nele não tocara, uma vez que por direito, pertencia a ***sua majestade***, claro, o Macaco.
	Anáfora por pronome Relativo	Na esperança de uma redução da pena de dois anos e meio de prisão domiciliar, ele vasculha obsessivamente sua documentação para fornecer à Justiça dados

		novos ***que*** contribuam para o esclarecimento dos fatos, mirando vingativamente em quem ele considera seus desafetos.

Além dessas classificações, também separamos aqueles casos em que o antecedente é o próprio autor do texto ou o leitor. Encontramos 44 ocorrências desses tipos de anáfora extra textual, como observamos na tabela 4 abaixo.

Tabela 4: Ocorrências no corpus de anáforas extra textual

Anáfora extra textual	Cursei a escola normal de Curitiba (atual instituto de educação do Paraná), diplomando-me em 1931.
	Você não terá sempre os verões livres.

Ainda analisamos a posição do antecedente em relação à anáfora. Dentro do corpus de análise, encontramos o montante de 159 ocorrências com o antecedente localizado fora do limite da sentença.

Por fim, encontramos 2 casos com antecedentes em elipse e 5 casos nos quais o antecedente se encontra inserido em um sintagma nominal muito complexo.

Na próxima seção, descrevemos os padrões de construções de anáfora pronominal modelados a partir da análise dos dados.

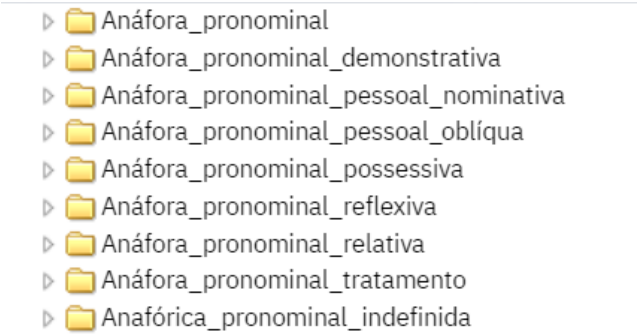
6.2 CONSTRUÇÕES MODELADAS

A partir da modelagem da construção genérica *Anáfora pronominal*, as demais construções foram desenvolvidas. Elas foram definidas de acordo com as categorias pronominais do português brasileiro, como observamos no Gráfico 1.

Na Figura 19, vemos a presença de nove construções anafóricas pronominais, das quais a primeira é uma estrutura genérica da rede, como será visto mais

detalhadamente na próxima seção, e as demais oito são variações das categorias pronominais do PB.

Figura 19: Lista de construções da rede Anáfora pronominal

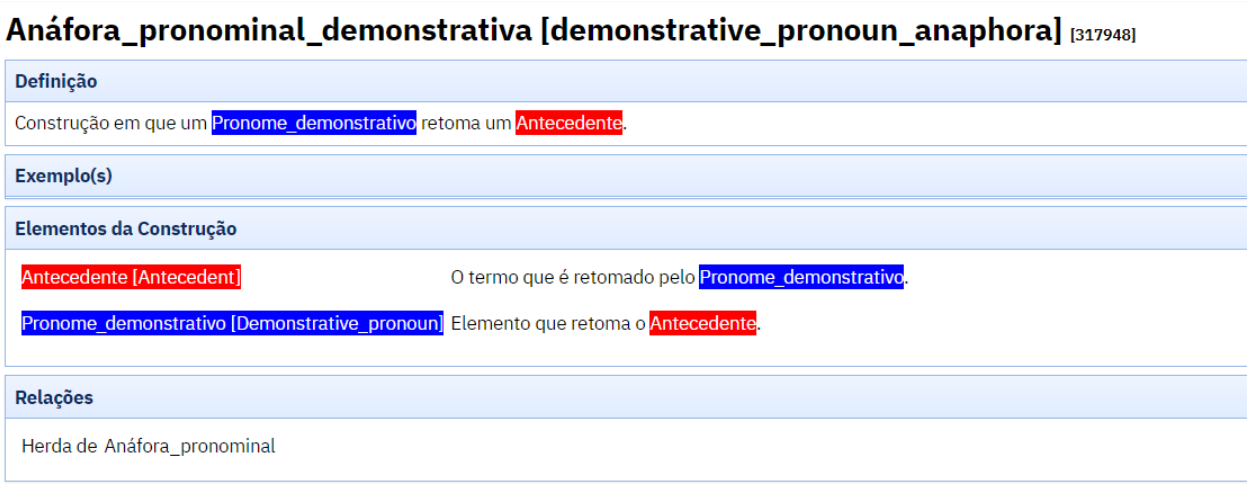


Fonte: Retirado de *Webtool 3.0*

A *Anáfora pronominal demonstrativa* segue os mesmos princípios da estrutura genérica. Ela é composta pelos elementos de construção (EC) pronome demonstrativo e antecedente. Nessa estrutura, o pronome demonstrativo resgata o antecedente que é formado pelo sintagma nominal. Na Figura 20, observamos a representação da estrutura na base de dados.

Para a estrutura pronominal anafórica demonstrativa representada na Figura 20, assim como para todas as construções anafóricas pronominais, sinalizamos as características morfossintáticas e semânticas para cada elemento construcional por meio de constrains, os quais incluem Universal Dependency (UD) features. Como já mencionado no capítulo anterior, os constrains funcionam como restrições morfossintáticas e semânticas que ajudam a definir a natureza de cada elemento construcional.

Figura 20: Construção Anáfora_pronominal_demonstrativa



Fonte: Retirado de *Webtool 3.0*

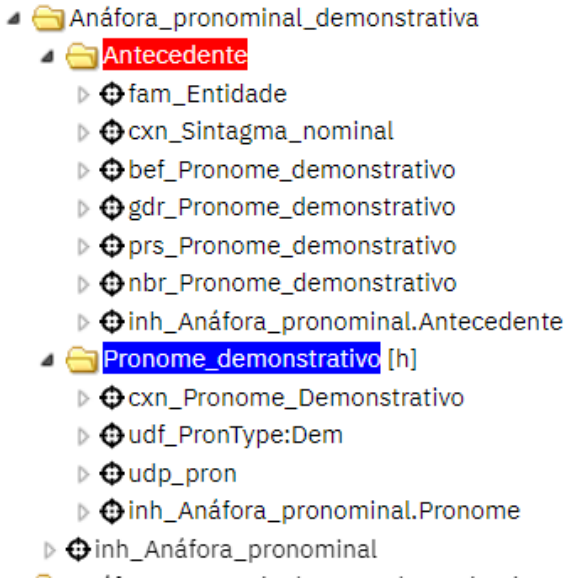
Na Figura 21, vemos a presença de constrains relacionados à posição sintática (bef), ao frame evocado (Entidade), às construções herdadas (inh), ao gênero, número e classe gramatical. Essas informações descrevem e modelam os elementos construcionais. Tendo isso como base, no EC antecedente, o *constraint bef* indica que o elemento vem antes de um pronome demonstrativo. A presença do constraint fam_Entidade significa que o antecedente será preenchido por uma unidade lexical que evoque um frame da família dos frames de Entidade da FN-Brasil, definindo sua natureza semântica. Além disso, a restrição Cxn_Sintagma_Nominal demonstra a construção que licencia o EC em questão. As informações de número, gênero e categoria gramatical indicam traços flexionais do elemento. Por fim, a presença do constraint inh_Anáfora_pronominal.Antecedente demonstra que o EC herda do EC equivalente na construção genérica de Anáfora_pronominal. Isso significa que a construção mantém relações de identidade e especificação com a construção mais abstrata Anáfora_pronominal.

Já no EC de Pronome_Demonstrativo, encontramos restrições relacionadas à construção que licencia o elemento, a Cxn_Pronome_Demonstrativo. As restrições também foram descritas para as categorias de classe de palavras conforme as UDs (udp), como a classe de pronomes e o tipo pronominal demonstrativo. Por último, vemos os *contraints* relacionados ao EC mais abstrato herdado pelo EC em questão:

Anáfora_pronominal.Pronome. É importante notar que, na construção genérica, o único constraint vinculado ao EC Pronome era udp_pronome, sem que houvesse qualquer especificação do tipo de pronome utilizado.

A incorporação dessas restrições permite com que esses itens sejam devidamente modelados pelo *constructicon*, de modo a alimentar a ferramenta do parser, porque essas aplicações funcionam seguindo mecanismos de identificação de traços morfossintáticos e semânticos.

Figura 21: Constrains de Anafóra_pronominal_demonstrativa



Fonte: Retirado de *Webtool 3.0*

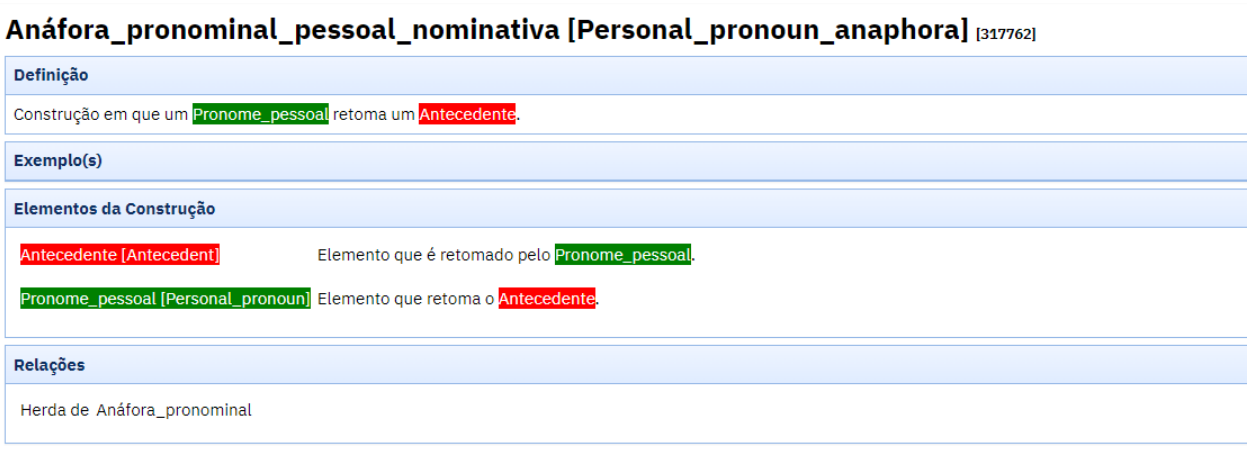
Prosseguindo para a *Anáfora Pronominal Pessoal Nominativa*, observamos o mesmo padrão estrutural, já que a estrutura é formada pelos ECs pronome pessoal e antecedente. Nesse caso, o pronome pessoal funciona como anáfora e retoma um outro elemento antecedente. Na Figura 22, podemos observar como isso é representado no sistema.

Em relação aos *constraints* (Figura 23), observamos certos padrões relacionados às restrições morfossintáticas e semânticas que se repetem em toda a família de construções anafóricas, como, por exemplo, a presença do constraints fam_Entidade, a qual significa que o antecedente será preenchido por um item lexical que evoca um frame do tipo Entidade. A restrição Cxn_Sintagma_Nominal demonstra a

construção que licencia o EC antecedente. Também indicamos as restrições de número, gênero e categoria gramatical. Vemos a presença de constrains relacionados à posição sintática (bef), que indica que o elemento vem antes de um pronome do caso reto. Além disso, a modelagem também conta com restrições relacionadas às construções herdadas (inh). Nesse caso, repete-se a herança do EC antecedente da construção abstrata (inh_Anáfora_pronominal).

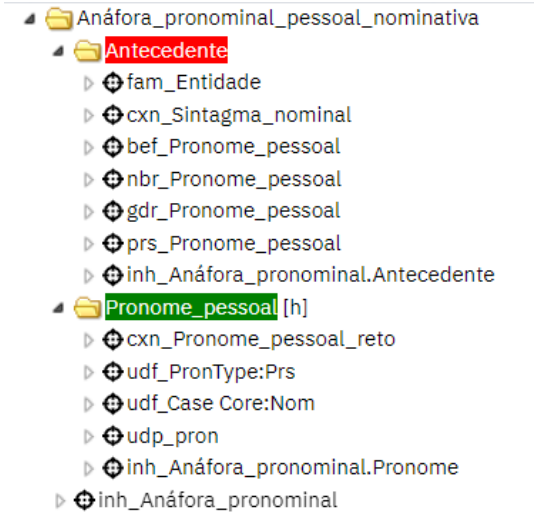
Já no EC de Pronome_pessoal, encontramos restrições relacionadas à construção que licencia o elemento, a Cxn_Pronome_pessoal. Também vemos as categorias de classe de palavras conforme as UDs (udp), como a classe de pronomes e o tipo pronominal pessoal, definido como nominativo, por ser este EC preenchido por pronomes pessoais retos. Tais informações se fazem presentes apenas nesta construção e não na abstrata.

Figura 22: Construção Anáfora_pronominal_pessoal_nominativa



Fonte: Retirado de *Webtool 3.0*

Figura 23: Constraints de Anáfora_pronominal_pessoal_nominativa



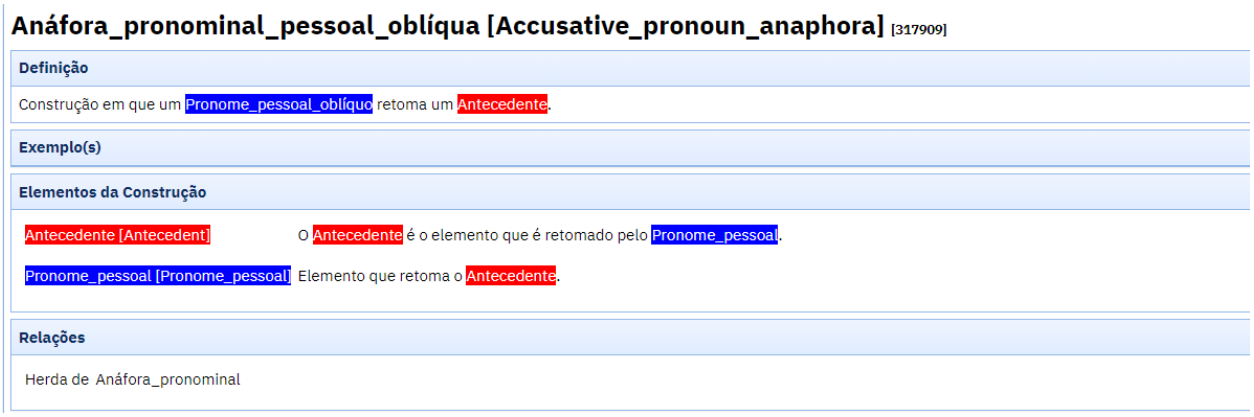
Fonte: Retirado de *Webtool 3.0*

Já na Figura 24, observamos a representação da construção *Anáfora pronominal pessoal oblíqua*. Ela é composta pelos ECs pronome pessoal oblíquo e antecedente. Essa estrutura mantém uma relação de herança com a construção anáfora pronominal, como podemos observar na parte inferior da Figura 24.

Da mesma forma, observamos certos padrões relacionados às restrições morfossintáticas e semânticas na modelagem da estrutura Anáfora Pronominal Oblíqua. Entre eles, estão a presença d*o constraint* fam_Entidade, a restrição Cxn_Sintagma_Nominal e as informações de número, gênero e categoria gramatical. Vemos também a presença de constraints relacionados à posição sintática (bef), às construções herdadas (inh), ao gênero, número e classe gramatical. Já no EC de Pronome_pessoal_oblíquo, encontramos restrições que diferenciam essa construção das demais. Primeiramente, temos a restrição relacionada à construção que licencia o elemento, a Cxn_Pronome_pessoal_oblíquo. Também vemos as categorias de classe de palavras conforme as UDs (udp), como a classe de pronomes e o tipo pronominal pessoal. Nessa estrutura, os *constraints* diferem em relação ao *PronType*_prs e aos *udf*_Acc e *udf*_Dat. Isso acontece, respectivamente, porque o pronome oblíquo faz

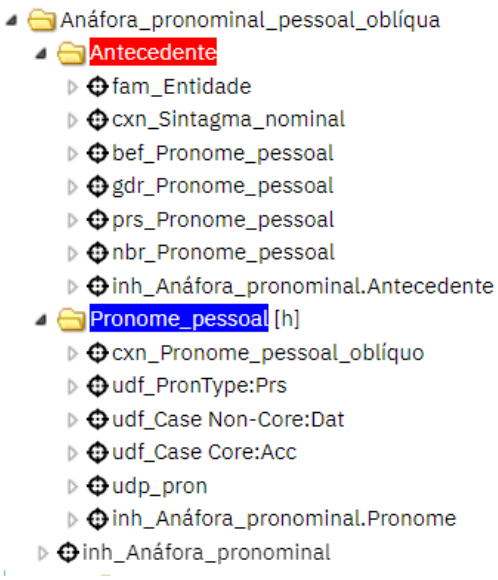
parte da categoria dos pronomes pessoais e ele tem caso acusativo ou dativo, uma vez que se trata de pronomes pessoais oblíquos, como podemos observar na Figura 25.

Figura 24: Construção Anáfora_pessoal_pronominal_oblíqua



Fonte: Retirado de *Webtool 3.0*

Figura 25: Constraints de Anáfora_pessoal_pronominal_oblíqua

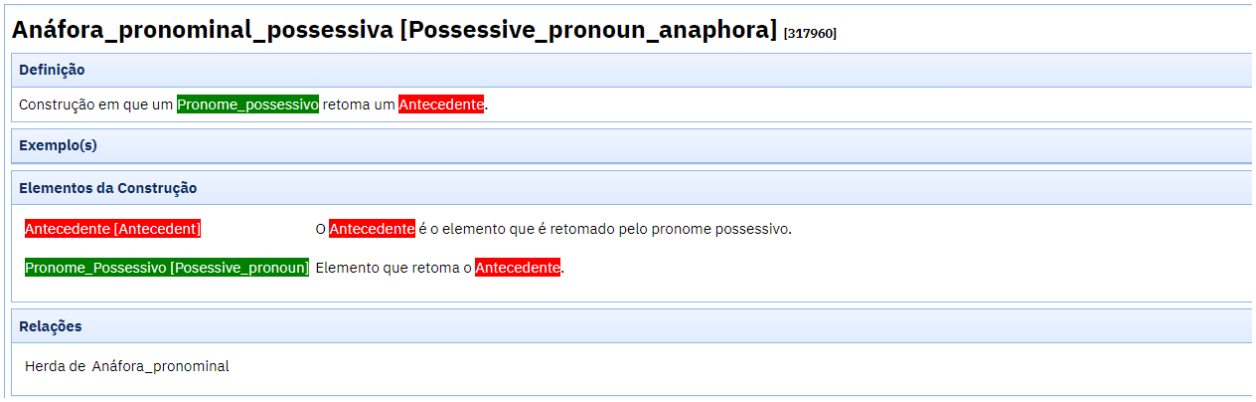


Fonte: Retirado de *Webtool 3.0*

Dando seguimento, observamos a construção *Anáfora pronominal possessiva* na Figura 26. Essa estrutura é formada por dois ECs, o pronome possessivo e o

antecedente e também mantém uma relação de herança com a construção genérica *Anáfora pronominal.*

Figura 26: Construção Anáfora_pronominal_possessiva

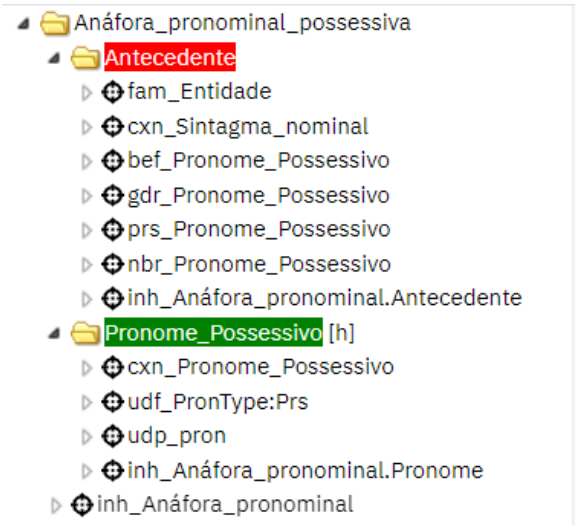


Fonte: Retirado de *Webtool 3.0*

Na anáfora pronominal possessiva, também observamos certos padrões relacionados às restrições morfossintáticas e semânticas. Entre eles a presença do constraints fam_Entidade, a restrição Cxn_Sintagma_Nominal e as informações de número, gênero e categoria gramatical para indicar a natureza morfológica do elemento Antecedente. Vemos também a presença de constraints relacionados à posição sintática (bef).

Já no EC de Pronome_possessivo, encontramos restrições relacionadas à construção que licencia o elemento, a Cxn_Pronome_pessoal_possessivo. Também vemos as categorias de classe de palavras conforme as UDs (udp), como a classe de pronomes e o tipo pronominal possessivo. Por último, vemos os *contraints* relacionados à construção herdada pelo EC, a Anáfora_pronominal. Tais restrições foram modeladas como podemos observar na Figura 27.

Figura 27: Constraints de Anáfora Pronominal Possessiva



Fonte: Retirado de *Webtool 3.0*

A família de construções *Anáfora pronominal* também é composta pela estrutura *Anáfora pronominal reflexiva*. A Figura 28 apresenta a construção cadastrada na base de dados. Ela é formada por dois ECs: pronome reflexivo e antecedente. Também observamos a relação de herança mantida com a construção *Anáfora pronominal.*

Figura 28: Construção Anáfora_pronominal_reflexiva

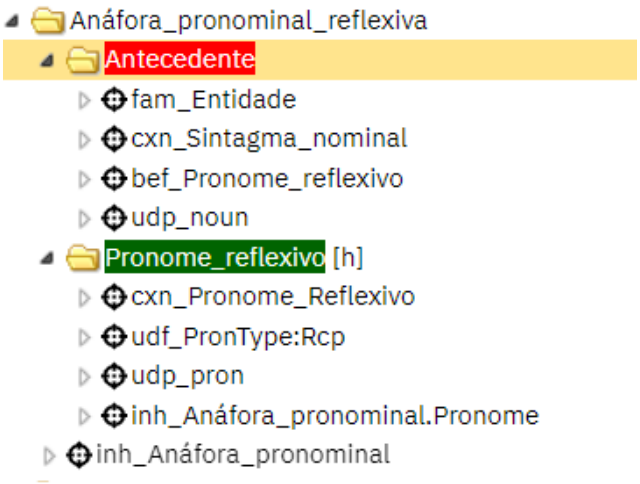
Anáfora_pronominal_reflexiva [Reflexive_pronoun_anaphora] [318055]

Definição
Construção em que um Pronome_reflexivo retoma um Antecedente.
Exemplo(s)
Elementos da Construção
Antecedente [Antecedent] — O Antecedente é o elemento que retoma o pronome reflexivo.
Pronome_reflexivo [Reflexive_pronoun] — Elemento que retoma o Antecedente.
Relações
Herda de Anáfora_pronominal

Fonte: Retirado de *Webtool 3.0*

Na anáfora pronominal reflexiva também observamos certos padrões relacionados às restrições morfossintáticas e semânticas, como já mencionado anteriormente: posição sintática (bef), construção herdada (inh), gênero, número e classe gramatical. Para o EC Pronome_reflexivo, encontramos restrições distintas relacionadas à construção que licencia o elemento, a Cxn_Pronome_pessoal_reflexivo, e também vemos as categorias de classe de palavras conforme as UDs (udp), como a classe de pronomes e o tipo pronominal reflexivo (vide Figura 29).

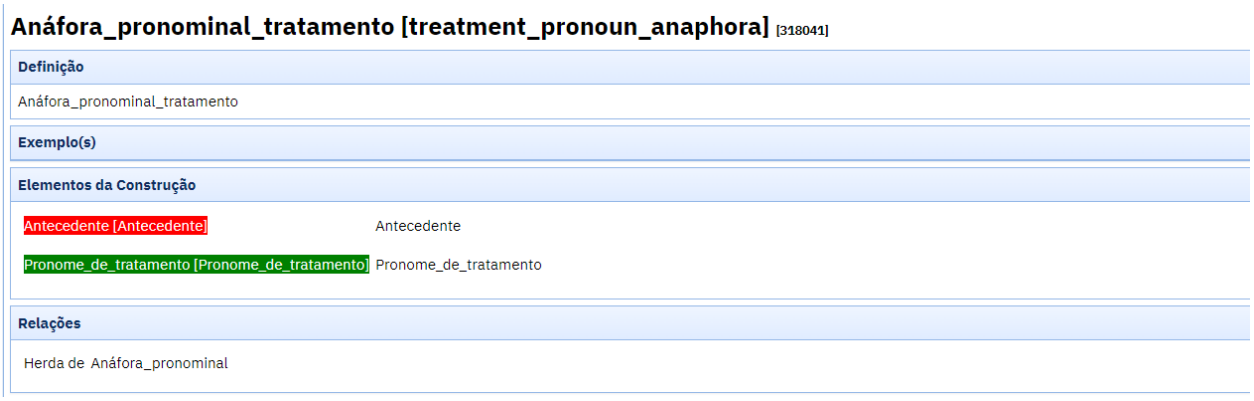
Figura 29: Constraints de Anáfora Pronominal Reflexiva



Fonte: Retirado de *Webtool 3.0*

Passando para o estudo da modelagem da *Anáfora pronominal tratamento*, da mesma forma, a estrutura mantém relação de herança com a construção da *Anáfora pronominal*. O que a difere das demais construções é o EC de pronome de tratamento. Podemos observar sua representação na Figura 30.

Figura 30: Construção Anáfora_pronominal_tratamento

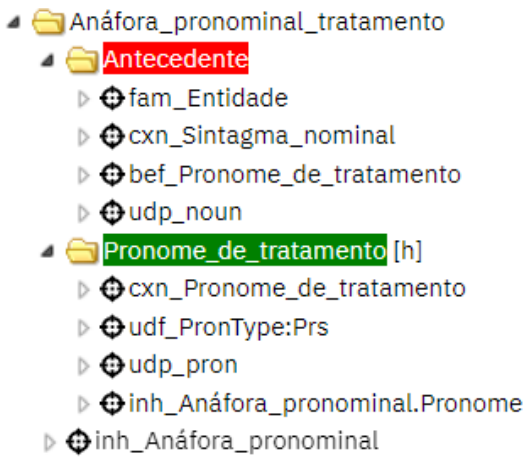


Fonte: Retirado de *Webtool 3.0*

Na anáfora pronominal tratamento, também observamos certos padrões relacionados às restrições morfossintáticas e semânticas. Vemos a presença de *constrains* relacionados à posição sintática (bef), às construções herdadas (inh), ao gênero, número e classe gramatical. Nesta construção, o constraints *bef* indica que o elemento vem antes de um pronome de tratamento, o *fam* indica que o elemento mantém relação com o frame de entidade e o udp_noun indica que o elemento é de natureza nominal. Além disso, o cxn_sintagma_nominal indica que o antecedente mantém uma relação de identidade com a construção de sintagma nominal. Por fim, a presença do constraints inh_Anáfora_pronominal demonstra que o EC herda da construção Anáfora_pronominal.

Já no EC de Pronome_de_tratamento, encontramos restrições relacionadas à construção que licencia o elemento, a Cxn_Pronome_de_tratamento. Também vemos as categorias de classe de palavras conforme as UDs (udps), como a classe de pronomes e o tipo pronominal de tratamento. Por último, vemos os *contraints* relacionados à construção herdada pelo EC, a Anáfora_pronominal.

Figura 31: Constraints Anáfora Pronominal de Tratamento



Fonte: Retirado de *Webtool 3.0*

Em seguida, na Figura 32, vemos a representação de mais uma construção da rede *Anáfora pronominal*, a *Anáfora pronominal relativa*. Ela é formada pelos ECs pronome relativo e antecedente e também mantém relação de herança com a construção abstrata *Anáfora pronominal*.

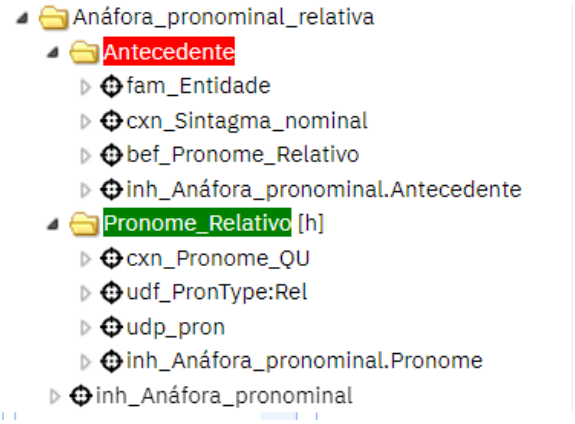
Figura 32: Construção Anáfora Pronominal Relativa

Fonte: Retirado de *Webtool 3.0*

Na anáfora pronominal relativa, vemos também a presença de constrains relacionados à posição sintática (bef), às construções herdadas (inh), ao gênero,

número e classe gramatical semelhantes aos anteriores. Já no EC de Pronome_relativo, encontramos restrições relacionadas à construção que licencia o elemento, a Cxn_Pronome_QU. Também vemos as categorias de classe de palavras conforme as UDs (udp), como a classe de pronomes e o tipo pronominal Relativo. Vemos essas informações na Figura 33.

Figura 33: Constraints Anáfora Pronominal Relativa

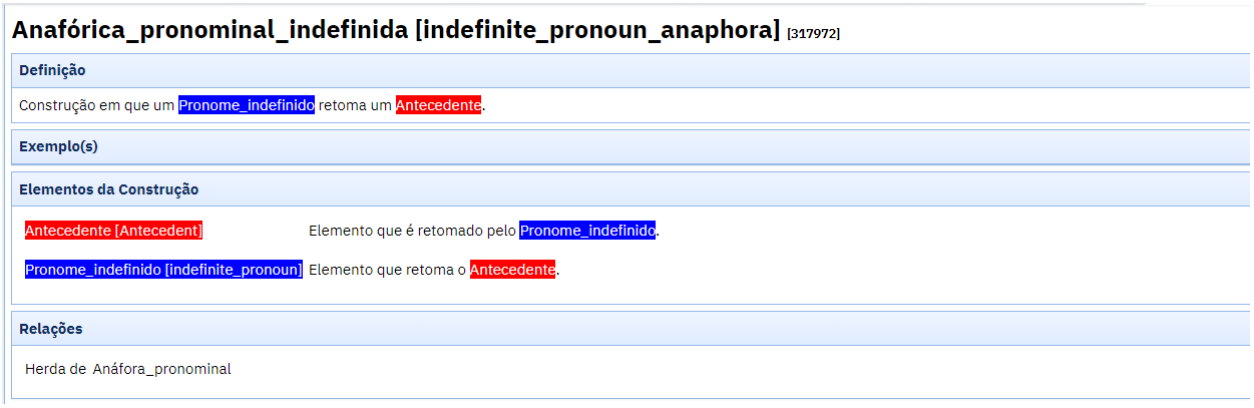


Fonte: Retirado de *Webtool 3.0*

Para finalizar, na Figura 34, observamos a representação da construção *Anáfora pronominal indefinida*. Da mesma forma que as demais estruturas da rede, essa construção é formada por dois ECs, pronome indefinido e antecedente e também mantém relação de herança com a construção *Anáfora pronominal.*

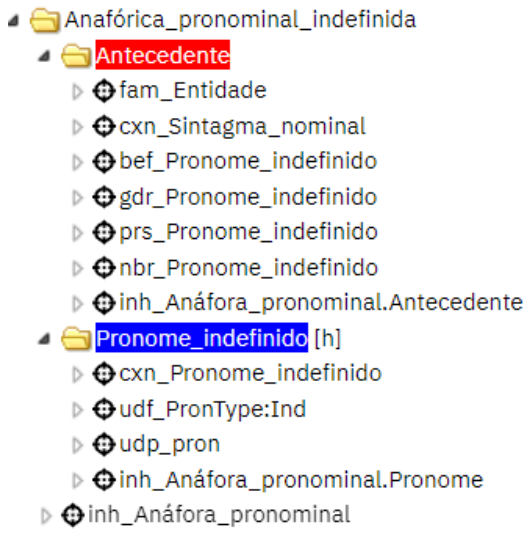
As limitações por *constraints* nessa estrutura (Figura 35) seguem o padrão comum às construções anafóricas pronominais e diferem no EC de Pronome_Indefinido, em que encontramos restrições relacionadas à construção que licencia o elemento, a Cxn_Pronome_Indefinido. Também vemos as categorias de classe de palavras conforme as UDs (udps), como a classe de pronomes e o tipo pronominal Indefinido.

Figura 34: Construção Anáfora Pronominal Indefinida



Fonte: Retirado de *Webtool 3.0*

Figura 35: Constraints Anáfora Pronominal Indefinida

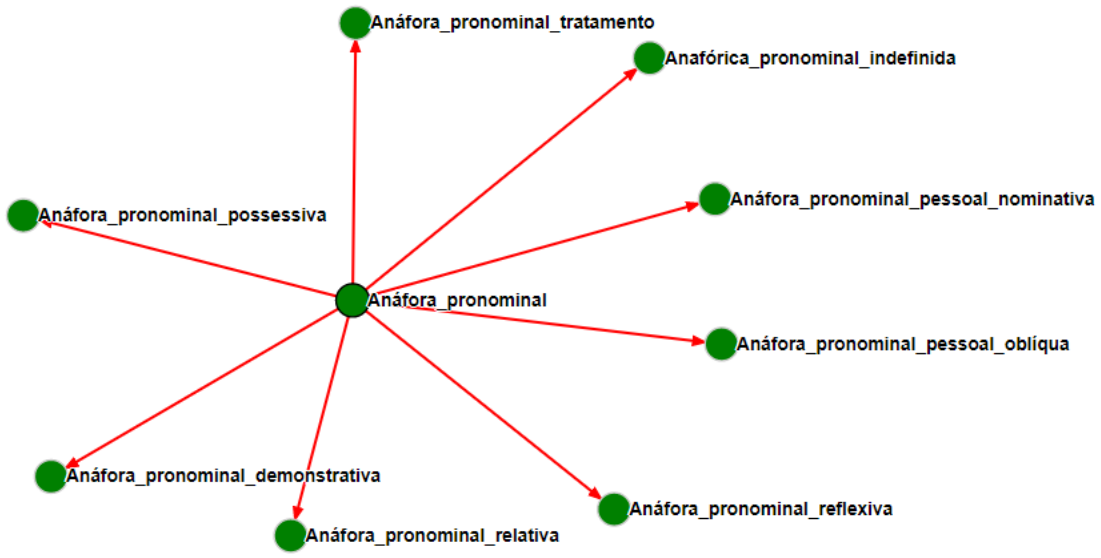


Fonte: Retirado de *Webtool 3.0*

No total foram modeladas nove construções anafóricas pronominais, sendo uma delas mais abstrata, a *Anáfora pronomina*l. Tendo isso como fundamento, na Figura 36, vemos a representação da rede de construções anafóricas na base de dados. Ela é formada por uma construção abstrata localizada ao centro, com as demais construções herdeiras ao redor. Através da Figura 36, observamos as relações de herança entre a

construção genérica, *Anáfora Pronominal* e as demais construções filhas, representadas pelos círculos verdes.

Figura 36: Rede de construções Anáfora Pronominal



Fonte: Retirado de *Webtool 3.0*

Neste capítulo, apresentamos a modelagem das construções anafóricas pronominais no constructicon. Isso inclui o levantamento de ocorrências com cada tipo pronominal sendo eles os tipos adjetivo e substantivo que incluem anáforas por pronome demonstrativo, indefinido, relativo, possessivo, pessoal do caso reto, pessoal do caso oblíquo, reflexivo e de tratamento. A partir desse ponto, as construções foram modeladas dentro da base de dados da FN-Br por meio da definição dos elementos construcionais e das restrições sintático-semânticas necessárias. Tais restrições podem ser observadas nas redes de cada tipo construcional.

Vimos as etapas de modelagem computacional de construções na base de dados da FN-Br, ou seja, a edição de entrada da construção, a definição da construção e dos ECs, as restrições aplicadas à construção, e, por fim, a representação da construção na base de dados. É a partir desse aparato computacional que as

construções podem ser implementadas em diversas ferramentas com interface linguístico-computacional como o constructicon, como apresentamos neste capítulo.

7 IMPLICAÇÕES DAS CONSTRUÇÕES ANAFÓRICAS NO PROCESSO DE EXTRAÇÃO DE INFORMAÇÃO

Este capítulo tem como objetivo discutir as aplicações e os impactos de construções anafóricas em ferramentas de extração de informação. Com esse propósito, abordaremos como a modelagem computacional de construções anafóricas pronominais pode ser aplicada a tarefas linguístico-computacionais, como a EI. Além disso, apresentamos a discussão do experimento na tarefa de extração de dados e os resultados quantificados.

Como vimos no capítulo anterior, cada tipo de construção anafórica pronominal foi modelada na base de dados da FN-Br. Através da modelagem dessas estruturas, nosso objetivo foi o de otimizar o processo de EI, trabalhando para que o modelo fosse capaz de localizar os referentes de longa distância. A hipótese a ser testada era a de que a modelagem de construções anafóricas pronominais no Constructicon da FrameNet Brasil contribui para o processo de extração de informações em textos através da identificação dos referentes das anáforas pronominais.

No restante deste capítulo apresentamos (a) a análise qualitativa preliminar, que nos permitiu testar a capacidade de identificação de construções anafóricas e seus antecedentes em exemplares das construções modeladas; (b) a análise quantitativa de desempenho do modelo e (c) a análise de erros, que busca identificar quais partes da modelagem ou do sistema de identificação de construções foi a responsável pela falha no experimento.

7.1 ANÁLISE QUALITATIVA PRELIMINAR: COMPORTAMENTO DO MODELO E DO PARSER PARA INSTÂNCIAS EXEMPLARES DE CONSTRUÇÕES ANAFÓRICAS

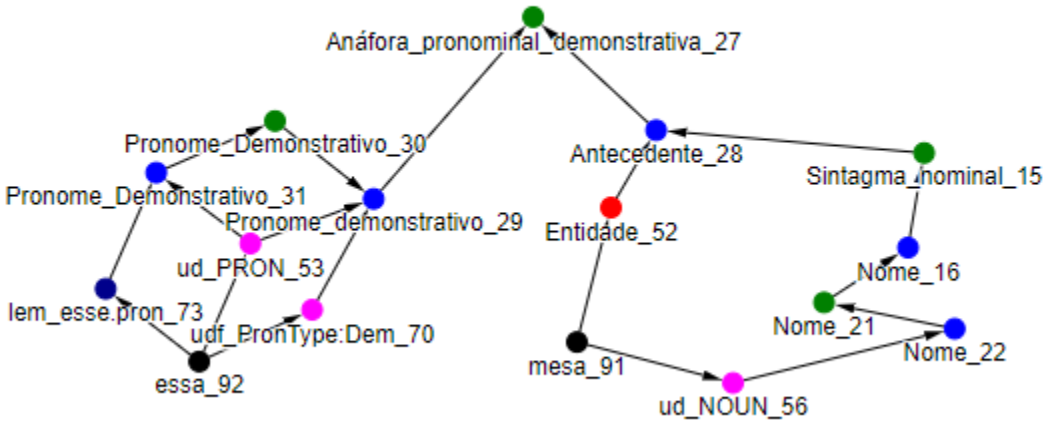
Os primeiros testes foram feitos com exemplos modelares de cada tipo de construção anafórica pronominal. Como resultado desse funcionamento, a ferramenta gerou redes semânticas que demonstram como a aplicação localizou os referentes em cada enunciado por meio do modelo proposto. As redes ilustram os componentes

construcionais: antecedente e a anáfora pronominal, assim como as restrições morfossintáticas definidas para cada um, como veremos com mais detalhes abaixo.

Primeiramente, para testar a modelagem de cada construção, alguns enunciados modelares foram selecionados. Para a construção *Anáfora Pronominal Demonstrativa,* selecionamos o enunciado (62). Na rede, o elemento *essa* funciona como anáfora, já o item lexical *mesa* funciona como antecedente. Para cada um dos itens lexicais, definimos limitações morfossintáticas. Vemos que a anáfora *essa* é um pronome do tipo demonstrativo e o antecedente mesa é um substantivo *(Noun)* que evoca um frame da família dos frames de Entidade. Ambos os elementos compõem a construção de Anáfora Pronominal Demonstrativa.

(60) Escolhida a **mesa** ideal, ***essa*** deve ser reservada com antecedência.

Figura 37: Rede de Anáfora Pronominal Demonstrativa



Fonte: Retirado de *Webtool 3.0*

Por meio da rede representada na Figura 37, vemos que o modelo proposto nesta dissertaçãopode ser considerado operacional porque o elemento anafórico *essa*

manteve uma relação de identificação com o referente *mesa.* Isso significa que a referência de longa distância foi mapeada pela ferramenta de extração de informações.

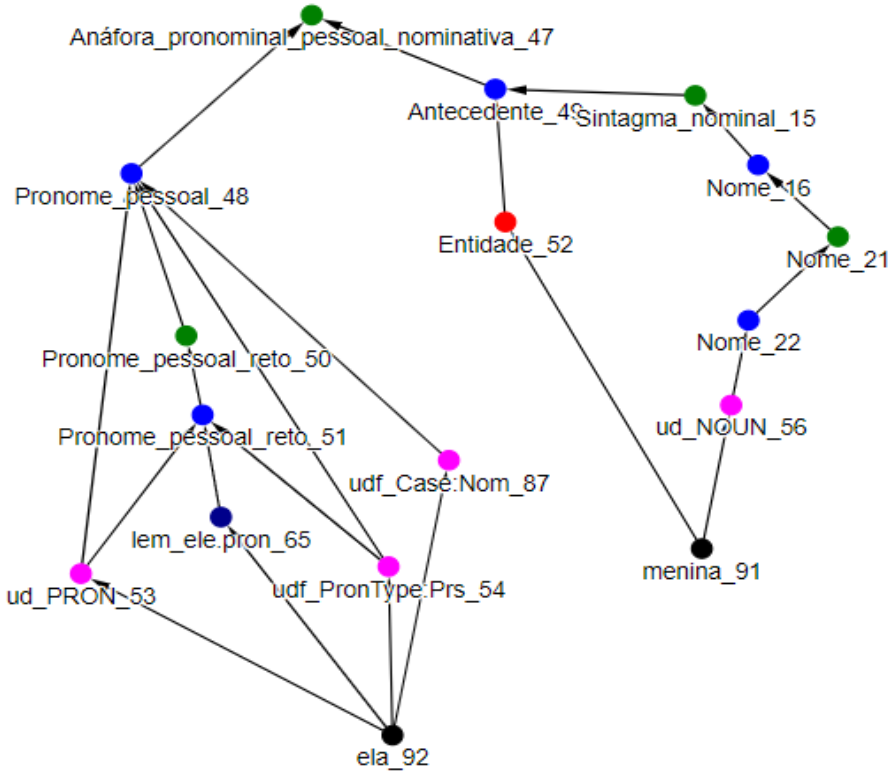
Com o mesmo propósito, realizamos testes com a construção nominativa. Para essa estrutura, selecionamos o enunciado (62) e ilustramos as relações semânticas mantidas entre os elementos que a compõem. Observe a rede gerada para o enunciado no Figura 38. O modelo também pode ser aplicado para o processamento do enunciado (62).

(61) Ontem eu vi a **menina** e ***ela*** estava animada

(62) Toda vez que uma nova **tecnologia** é incorporada ao dia a dia dos cidadãos, **ela** rompe paradigmas, muda comportamentos, agita mercados e gera descontentamento em grupos e corporações que resistem a se adaptar aos novos tempos.

Através da Figura 38, observamos as relações mantidas entre o *antecedente* e o elemento que funciona como *anáfora*. Vemos que o antecedente *menina* é definido como um substantivo, representado no diagrama por *Noun*, evoca o frame de *Entidade*. Em contrapartida, a anáfora pronominal *ela* é um pronome pessoal do caso reto com função nominativa e evoca o frame de *Pronome pessoal*. Ambos os elementos mantêm relação de constituência com a construção anáfora pronominal nominativa. Assim sendo, entendemos que o modelo também cumpre a função desejada para a construção anafórica pronominal nominativa. Afirmamos isso porque o elemento antecedente *menina* foi devidamente mapeado pela anáfora pronominal *ela*.

Figura 38: Rede de Anáfora Pronominal Nominativa

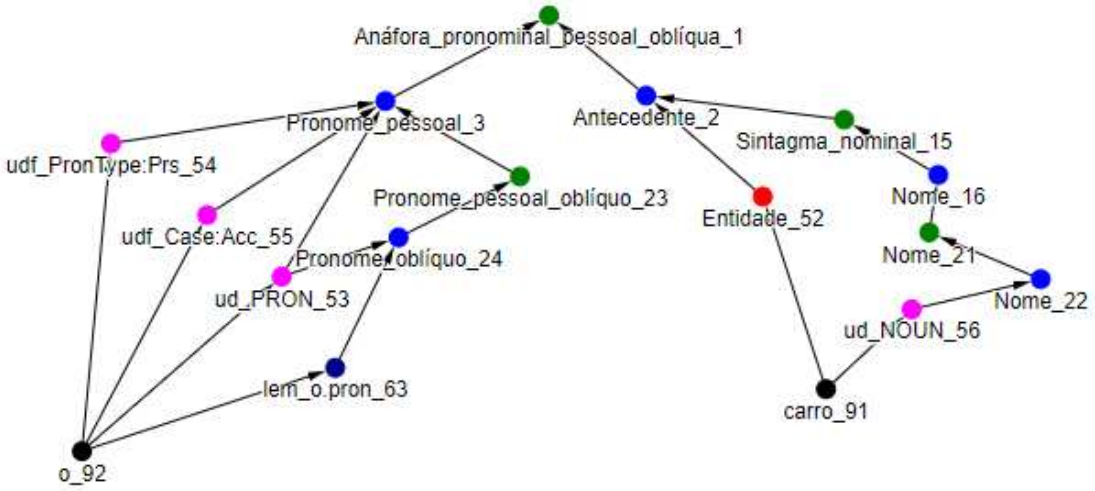


Fonte: Retirado de *Webtool 3.0*

Os testes também foram aplicados para a construção *Anáfora pronominal pessoal oblíqua*. Neste caso, o enunciado escolhido para testagem foi o (63). Como observamos na Figura 39. Nesse tipo de estrutura, o antecedente é o item ***carro*** e ele é retomado pelo pronome oblíquo ***o***. A rede representada pela figura 39 demonstra as relações estabelecidas entre os elementos construcionais antecedente e anáfora pronominal oblíqua.

(63) Comprei um **carro** e dei-***o*** de presente para minha filha.

Figura 39: Rede de Anáfora Pronominal Pessoal Oblíqua



Fonte: Retirado de *Webtool 3.0*

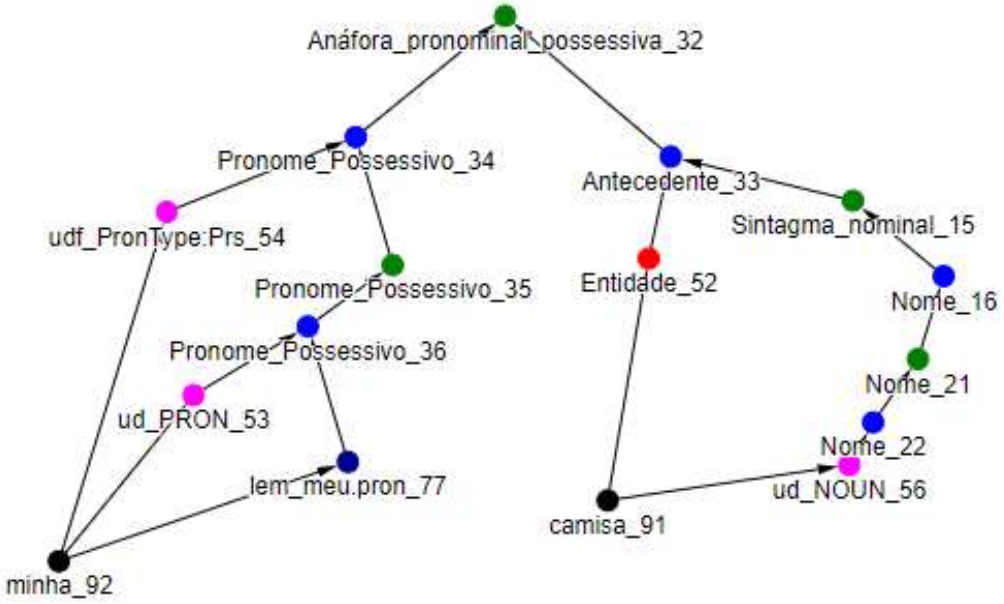
O experimento também processou as sentenças (64) e (65) com pronomes possessivos. O resultado do processamento desse tipo de anáfora pode ser observado na Figura 40.

(64) A **blusa** é ***minha***.

(65) A concorrência saudável beneficia o usuário, que terá mais opções de escolha, e, por que não dizer, os próprios **taxistas**, que certamente vão melhorar **seus** serviços.

Na Figura 40 também vemos as relações estabelecidas entre os elementos construcionais, assim como as características morfossintáticas e semânticas compartilhadas por eles relacionadas aos frames acionados e categorias morfológicas.

Figura 40: Rede de Anáfora Pronominal Possessiva



Fonte: Retirado de *Webtool 3.0*

As relações em rede semântica também são observadas para enunciados com anáfora pronominal reflexiva. O enunciado (66) está representado na Figura 41.

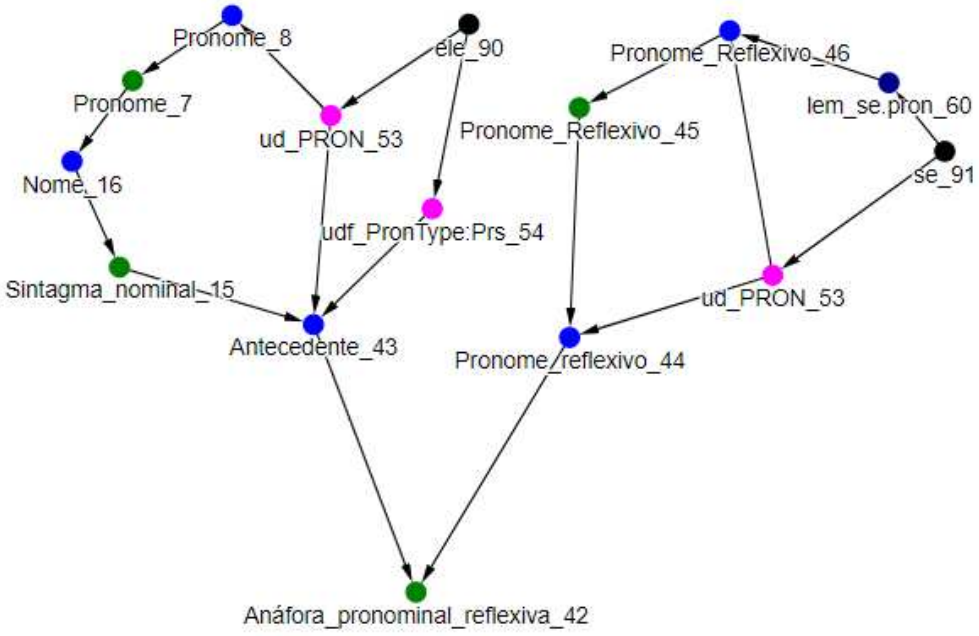
(66) **Ele *se*** mudou para outro país.

(67) Meus **pais** eram ucranianos, que **se** conheceram e casaram no Paraná.

Na rede da Figura 41, vemos as relações estabelecidas entre os elementos que compõem o enunciado, os constraints e as construções. A construção *Anáfora pronominal reflexiva* é formada por um elemento construcional antecedente e um pronome reflexivo. A partir desses dois elementos, a rede se amplia. Vemos as restrições morfossintáticas aplicadas ao antecedente. Ele pertence à categoria dos pronomes e, conforme definido para a construção Sintagma_nominal, a qual licencia o

EC Antecedente, pode ocupar a posição de núcleo nominal do SN. Por sua vez, o elemento anafórico compõe a categoria dos pronomes reflexivos.

Figura 41: Rede de Anáfora Pronominal Reflexiva

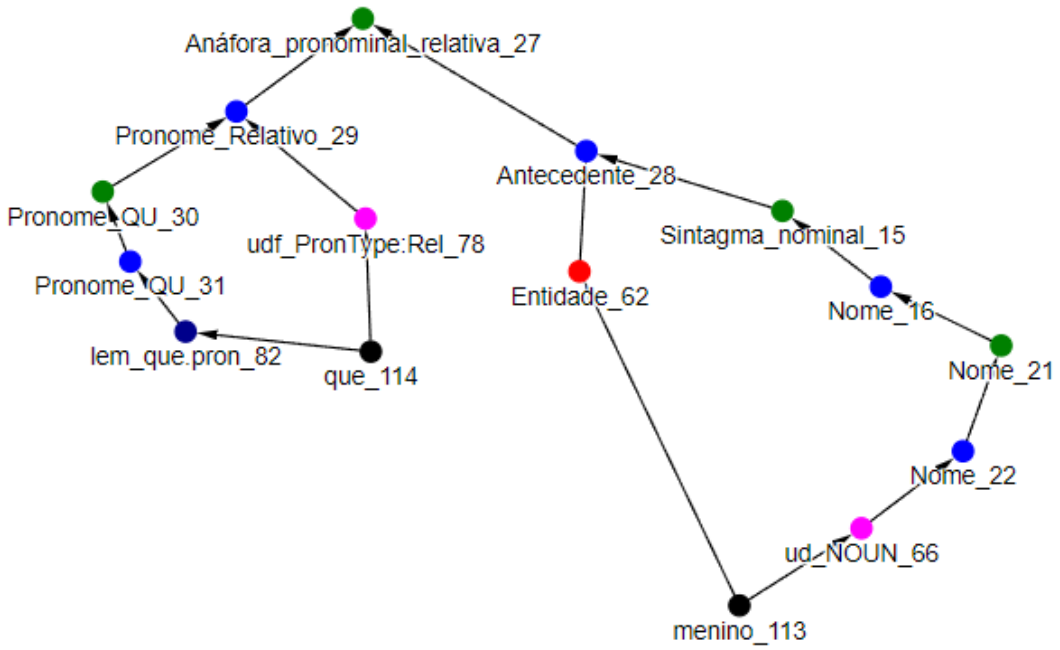


Fonte: Retirado de *Webtool 3.0*

A construção anafórica relativa pode ser representada por redes, como vemos no diagrama abaixo. No enunciado (68), podemos observar as relações entre o antecedente *menino* e o elemento anafórico *que*.

(68) O **menino *que*** eu amo é brasileiro.

Figura 42: Rede de Anáfora Pronominal Relativa



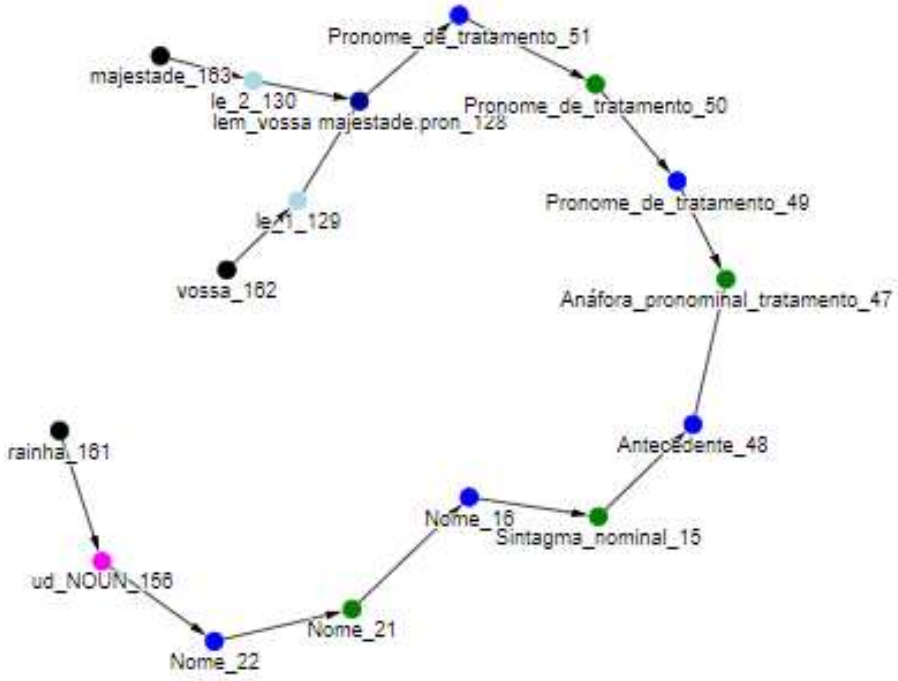
Fonte: Retirado de *Webtool 3.0*

Na Figura 42 também vemos as relações de restrição estabelecidas para cada elemento da construção. O elemento antecedente *menino* evoca o frame de *entidade* e tem como característica ser um *nome*. Por outro lado, o elemento que funciona como anáfora que evoca a construção Pronome QU (Marção, 2018) e tem como característica pertencer a classe dos *pronomes relativos*. Todas essas restrições e evocações fundamentam a construção *anafórica relativa*.

Em relação aos pronomes de tratamento, em um primeiro momento, acreditamos que a ferramenta não fosse capaz de interpretar essas estruturas devido à sua natureza *multi word*. Tal restrição não se manteve e os pronomes de tratamento, em sua grande maioria sintagmas nominais, também foram implementados com sucesso no experimento de teste. Observe o exemplo (69) e sua implementação na ferramenta na Figura 43.

(69) A **rainha**, ***vossa majestade***, celebrou seus 96 anos.

Figura 43: Rede de Anáfora Pronominal de Tratamento



Fonte: Retirado de *Webtool 3.0*

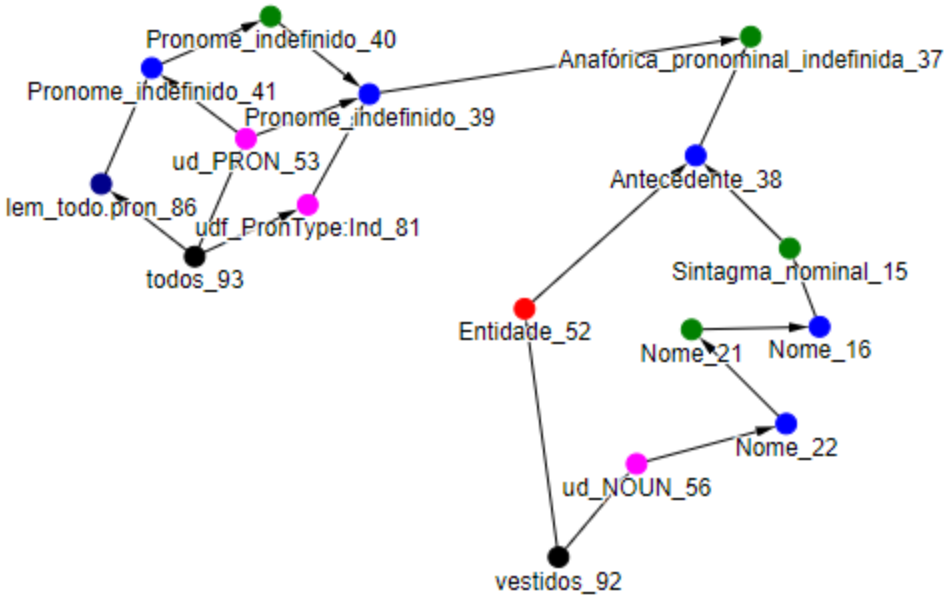
Da mesma forma, a construção de anáfora indefinida pode ser representada por uma rede semântica. Para essa ilustração, selecionamos o enunciado (70).

(70) Comprei muitos **vestidos** e ***todos*** serviram.

Nesse caso, o nome *vestido* funciona como o antecedente e o pronome *todos* como anáfora. Podemos ver as relações estabelecidas entre os elementos construcionais. O item lexical vestido evoca o frame de entidade e tem como restrição ser um nome. Já o pronome indefinido todo compõe a categoria dos pronomes e evoca a construção de pronome indefinido. Ambos os elementos construcionais formam a

construção anafórica indefinida, como podemos observar na parte superior da Figura 44.

Figura 44: Rede de Anáfora Pronominal Indefinida



Fonte: Retirado de *Webtool 3.0*

O mesmo não acontece com o pronome indefinido *nenhum*. Para esse tipo de pronome indefinido existe uma certa limitação no modelo porque ele não mantém concordância em número com o antecedente. Por esse motivo, o modelo não é capaz de mapear a relação de referência entre o elemento que funciona como antecedente e a anáfora pronominal. Essa situação pode ser percebida quando testamos o enunciado (71) na ferramenta de extração de informações, como vemos abaixo.

(71) Comprei cinco **vestidos** e ***nenhum*** serviu.

Para esse enunciado, apesar de a anáfora pronominal *nenhum* manter uma relação de referência com o antecedente *vestidos,* o pronome indefinido não pode ser

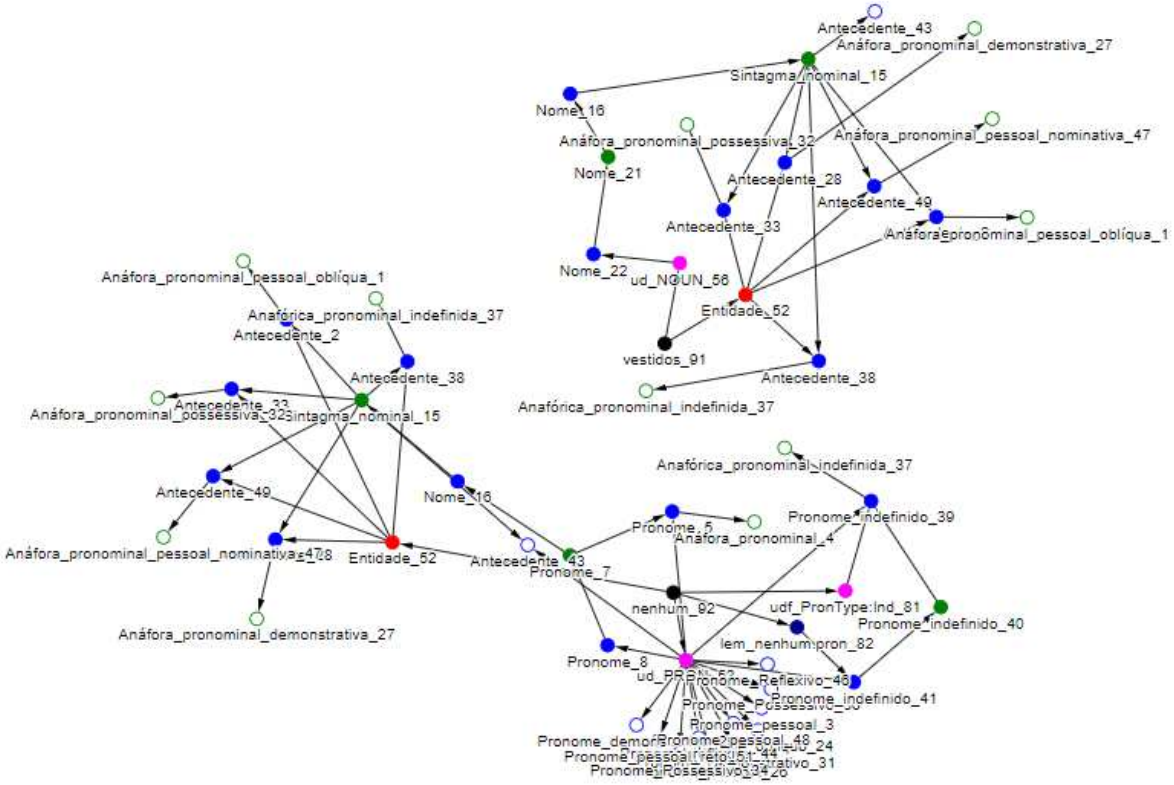
interpretado pela máquina. Esse não seria o caso se a anáfora fosse realizada com, por exemplo, *nenhum deles*.

A análise da sentença exemplar com anáfora pronominal indefinida chamou a atenção para uma potencial limitação do modelo. Ainda que a sentença escolhida para teste seja simples do ponto de vista de não haver, por exemplo, estrutura interveniente entre antecedente e anáfora. A ausência de um elemento que mapeie concordância na anáfora fez com que o antecedente não pudesse ser devidamente interpretado pelo modelo de construções anafóricas pronominais, objeto de pesquisa deste trabalho. Em resumo, a construção anafórica pronominal indefinida apontou algumas limitações para a aplicação do modelo.

Essas questões podem ser vistas na rede abaixo, através da Figura 45. Observamos que a rede ilustra uma incompatibilidade entre os traços associados aos elementos construcionais. Por esse motivo, não ocorre o mapeamento esperado entre os elementos *antecedente* e *referente* e a rede "não fecha".

Constatada a capacidade do modelo e do parser em identificar instâncias exemplares das construções de anáfora pronominal e de seus antecedentes, passamos à próxima fase do experimento, que consiste em submeter todas as sentenças identificadas no corpus ao parser e analisar o desempenho deste. Essa análise é apresentada na seção seguinte.

Figura 45: Anáfora indefinida nenhum

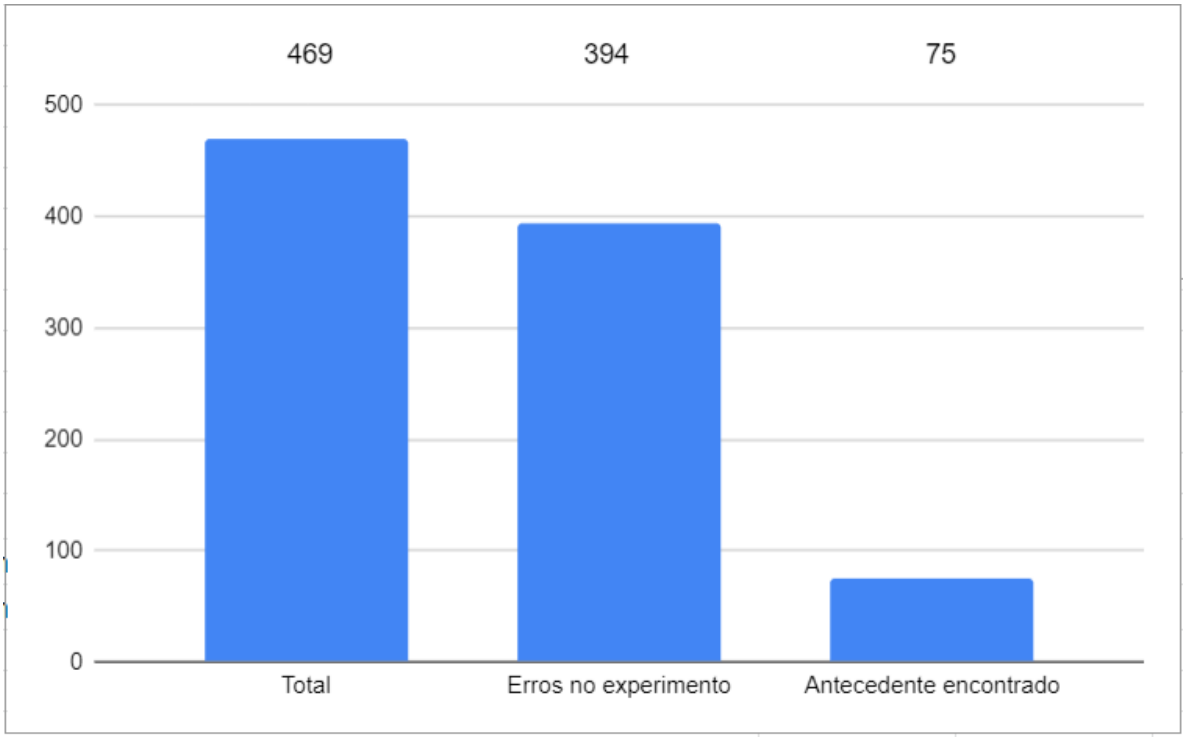


Fonte: Retirado de *Webtool 3.0*

7.2 ANÁLISE QUANTITATIVA: DESEMPENHO DO MODELO E DO SISTEMA PARA TODAS AS INSTÂNCIAS DO CORPUS

O resumo dos resultados do processamento de todas as sentenças do corpus na ferramenta pode ser observado no Gráfico 4. Através da análise das informações abaixo, entendemos o grau de eficiência do modelo de *Construções Anafóricas Pronominais* na tarefa de EI. Vemos o total de sentenças processadas, o total de antecedentes mapeados pelo modelo e os erros.

Gráfico 4 : Resultado do processamento de sentenças na ferramenta



Fonte: Criado pela autora

A partir dos testes, selecionamos um recorte dos resultados positivos mais prototípicos obtidos, como vemos na Tabela 5, bem como um dos resultados negativos mais prototípicos, mostrados na Tabela 6. Através da análise dos dados, percebemos padrões recorrentes que levaram à identificação, ou não, do antecedente no experimento.

Tabela 5: Resultados positivos no processamento das sentenças no experimento

Sentença	Tipo de construção	Antecedente mapeado	Motivo do bom desempenho
(1) Toda vez que uma nova tecnologia é incorporada ao dia a dia dos cidadãos, ela rompe paradigmas, muda comportamentos, agita mercados e gera descontentamento em	Anáfora_pronominal_pessoal_nominativa	tecnologia	Grande proximidade entre o antecedente pronome que funciona como anáfora. Não há outro nome para

grupos e corporações que resistem a se adaptar aos novos tempos.			confundir o parser
(2) A concorrência saudável beneficia o usuário, que terá mais opções de escolha, e, por que não dizer, os próprios taxistas, que certamente vão melhorar seus serviços.	Anáfora_pronominal_relativa	taxistas	Grande proximidade entre o antecedente pronome que funciona como anáfora. Não há outro nome para confundir o parser.
(3) O cupom fiscal será conferido pela agência promotora, que o disponibilizará o voucher cortesia no prazo máximo de 48 horas.	Anáfora_pronominal_relativa	agência	Grande proximidade entre o antecedente pronome que funciona como anáfora. Não há outro nome próximo para confundir o parser.
(4) Ele era o tipo raro de acadêmico que queria apenas isso: levar seu trabalho para as massas.	Anáfora_pronominal_relativa	acadêmico	Grande proximidade entre o antecedente pronome que funciona como anáfora. Não há outro nome para confundir o parser
(5) -Ei, janela, você faz o favor de se abrir um pouquinho só?	Anáfora_pronominal_tratamento	janela	Grande proximidade entre o antecedente pronome que funciona como anáfora. Não há outro nome para confundir o parser
(6) Enquanto realizava este lucrativo comércio, Portugal realizava no Brasil o extrativismo do pau-brasil, explorando da Mata Atlântica toneladas da valiosa madeira, cuja tinta vermelha era comercializada na Europa.	Anáfora_pronominal_demonstrativa	madeira	Grande proximidade entre o antecedente pronome que funciona como anáfora. Não há outro nome para confundir o parser
(7) E as lágrimas que desciam pelo seu rosto sem cessar foram-se tornando sólidas e brilhantes no ar, antes de	Anáfora_pronominal_relativa	lágrimas	Grande proximidade entre o antecedente e o pronome anafórico.

submergir na água e bater no cascalho do fundo.			Não há outro nome para confundir o parser

Tabela 6: Resultados negativos no processamento das sentenças no experimento

Sentença	Tipo de construção	Antecedente mapeado	Motivo da falha
(8) Aos 11 anos passei a me locomover "sobre rodas" (com o auxílio de cadeira de rodas), uma condição nova para mim, a qual logo me adaptei.	Anáfora_pronominal_relativa	rodas	Duas unidades lexicais de natureza substantiva próximas. Períodos longos.
(9) Mas se você procura Deus e quer um encontro mais íntimo com ele, visite Fátima, que vale a pena.	Anáfora_pronominal_relativa	Deus	Limitação em relação ao mapeamento de entidades nomeadas.
(10) Existe uma grande probabilidade de você vir a trabalhar para um deles.“	Anáfora_pronominal_tratamento	probabilidade	O parser classificou o pronome pessoal como sendo de tratamento.
(11) Só caso com quem fizer três adivinhações que eu não adivinhe e que adivinhe três que eu fizer.	Anáfora_pronominal_relativa	quem	O parser não reconheceu “adivinhações” como um nome que evoca o frame de Entidade.
(12) Tinha certeza que ele ia direto para a força e que, portanto, daqui para a frente a vaquinha era dela.	Anáfora_pronominal_pessoal_nominativa	vaquinha	Grande distância entre o antecedente e o pronome anafórico. O parser mapeou o substantivo mais próximo
(13) Aos 11 anos passei a me locomover “sobre rodas“ (com o auxílio de cadeira de rodas), uma condição nova para mim, a qual logo me adaptei.	Anáfora_pronominal_relativa	rodas	O parser ficou confuso e não reconheceu “condição” como um nome que evoca o frame de entidade.
(14) Meus pais eram ucranianos,	Anáfora_pronominal_	se	O parser

que se conheceram e casaram no Paraná.	reflexiva		reconheceu o pronome reflexivo como antecedente dele mesmo.
(15) Vovô radicou-se em Cruz Machado, onde papai trabalhava.	Anáfora_pronominal_relativa	Vovô	Limitação em relação ao mapeamento de entidades nomeadas.
(16) Na esperança de uma redução da pena de dois anos e meio de prisão domiciliar, ele vasculha obsessivamente sua documentação para fornecer à Justiça dados novos que contribuam para o esclarecimento dos fatos, mirando vingativamente em quem ele considera seus desafetos.	Anáfora_pronominal_relativa	documentação	Duas unidades lexicais de natureza substantiva próximas. Períodos longos. O parser ficou confuso.

Em contextos com anáfora pronominal reflexiva, no geral, os antecedentes foram mapeados de forma eficiente. Da mesma forma, a ferramenta processa as estruturas em contextos de anáforas pronominais relativas. Nesses casos, o experimento aciona a construção anafórica relativa na presença de outros tipos de construções, como a nominativa, demonstrativa, reflexiva etc. A natureza do antecedente também influenciou no processamento das sentenças. Observamos falhas na aplicação do modelo em sentenças em que o antecedente era composto de sintagmas nominais ou expressões nominais. Essa situação pode ser vista no enunciado 16 da Tabela 6. Outro aspecto que foi impeditivo para o processamento das estruturas pronominais anafóricas está relacionado à natureza do *parser* UD utilizado para o processamento das estruturas. Isso porque nesse tipo de aplicação, a análise é limitada à sentença em questão. Tendo em vista essa limitação, os antecedentes localizados fora do limite da sentença não podem ser mapeados pelo *parser* UD.

Além disso, dentro do total de sentenças do corpus, 44 sentenças não entraram no experimento de reconhecimento de referentes por algumas situações. A primeira delas se refere à natureza externa do antecedente. Essa situação foi observada em trechos que compõem textos autobiográficos e injuntivos. Nesses casos, o antecedente

é o próprio escritor do texto ou o leitor, a anáfora refere-se a um antecedente por inferência externa. Já a segunda situação está relacionada com a ocorrência de elipse, isto é, omissão de elementos na sentença, como, por exemplo, os antecedentes. Esses últimos dois casos impossibilitam o funcionamento do parser. Além disso, sentenças em que o antecedente era um sintagma nominal muito complexo ou estava em sentenças muito anteriores não foram incluídas no experimento, devido à limitação do parser UD em processar mais de uma sentença ao mesmo tempo, como mencionado anteriormente.

Através da análise dos resultados, observamos que ocorrem muitos erros na aplicação do modelo. Uma análise superficial, entretanto, focada apenas no total de erros, não aponta para as causas desses erros, ou, em outras palavras, não nos permite saber se o problema está no modelo proposto, no parser ou em ambos. Na sequência, olharemos para essas causas, localizando quais partes do modelo ou do sistema de identificação de construções estão envolvidas nos erros.

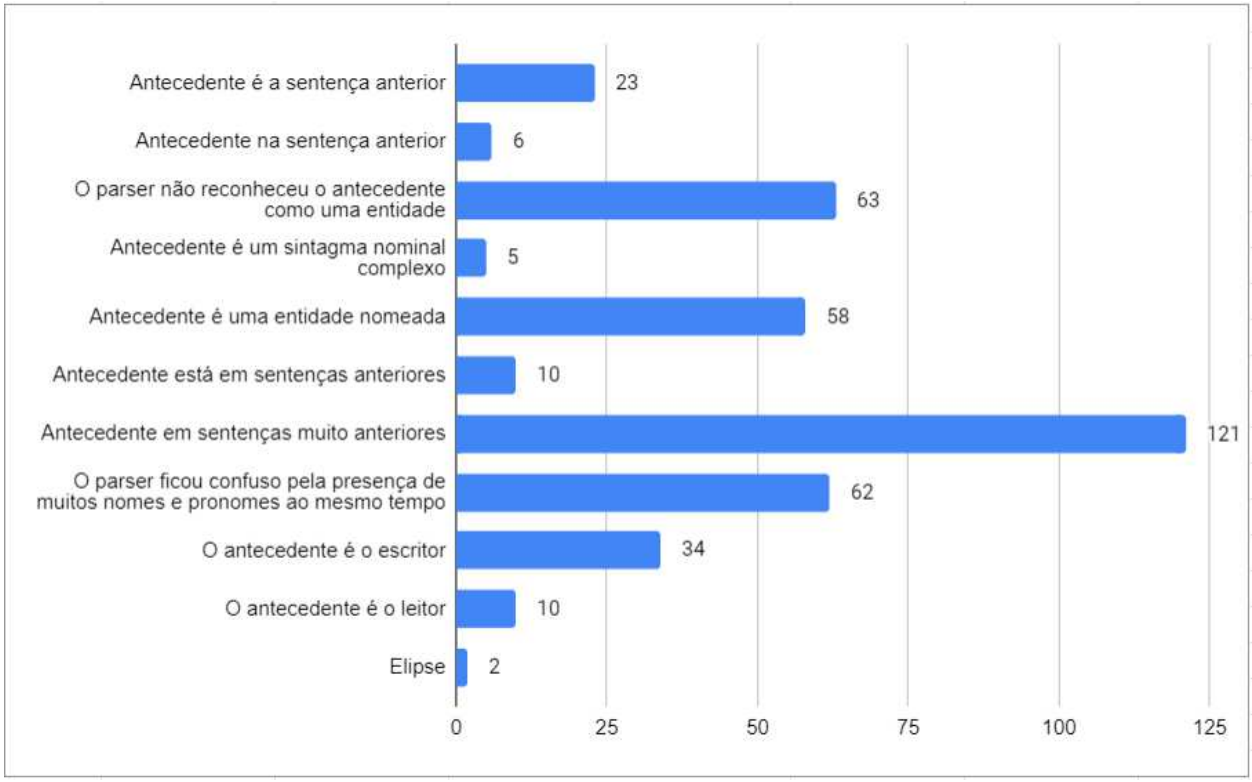
7.3 ANÁLISE DE ERROS DO EXPERIMENTO

Nesta seção nos debruçamos sobre uma análise qualitativa dos erros ocorridos no experimento. Como mencionado anteriormente, dentro do corpus existe o montante de 44 sentenças que não foram processadas no experimento devido à natureza externa do antecedente: o leitor ou o escritor. Além disso, observamos casos em que os antecedentes são sintagmas nominais muito complexos, e, por isso, trazem dificuldades de processamento pelo *parser* UD utilizado no experimento. Ainda temos 159 casos em que o antecedente se localiza fora do escopo da sentença, isso significa, em sentença anterior, em sentenças anteriores (2 sentenças anteriores à anáfora pronominal), ou muito anteriores ao pronome anafórico (3 ou mais sentenças anteriores à anáfora pronominal). Essa última situação impossibilita o mapeamento do antecedente pelo sistema, visto que o *parser* UD é limitado pela fronteira sentencial.

De forma mais detalhada, podemos dividir os erros de processamento em 8 tipos, segundo sua natureza linguístico-computacional. Todos os erros estão

relacionados à natureza, localização e complexidade morfossintática do antecedente. Por meio da análise de cada erro gerado pelo processamento de sentenças no *parser*, observamos os seguintes padrões: o antecedente é a sentença anterior, antecedente está na sentença anterior, o *parser* não reconheceu o antecedente como uma entidade, antecedente é um sintagma nominal complexo, antecedente é uma entidade nomeada, antecedente está em sentenças anteriores, antecedente está em sentenças muito anteriores e, por fim, o parser ficou confuso pela presença de muitos nomes e pronomes ao mesmo tempo. O resumo dessas ocorrências pode ser observado no Gráfico 5.

Gráfico 5 : Tipos de erro encontrados no processamento de sentenças na ferramenta parser



Por meio de análise dos dados observados no Gráfico 5, vemos que os tipos de erro mais proeminentes no experimento estão relacionados aos casos em que o antecedente está localizado em sentenças muito anteriores. Essa situação é responsável pelo maior percentual de erros gerados na ferramenta devido à limitação

computacional existente no *parser*, cerca de 30,71 % em relação ao total de erros. Os *parsers* UD encontram dificuldades em mapeamento de antecedentes em localidades sintáticas muito distantes dos elementos anafóricos e não realizam análises que ultrapassem o limite da sentença. Ainda em relação aos padrões de erro mais proeminentes no experimento, percebemos os casos em que o sistema ficou confuso pela presença de muitos nomes e pronomes ao mesmo tempo e, ainda, situações que não reconheceu o antecedente como uma entidade. Também observamos um considerável percentual de erros em casos que o antecedente é uma entidade nomeada. Essas situações, por sua vez, apontam mais para problemas no modelo criado no Constructicon da FN-Br. Isso porque o frame de Entidade usado para mapear a semântica dos antecedentes, por um lado, é bastante genérico e mais de um nome poderia ter sua semântica construída a partir dele. Por outro, ele não cobre toda a variedade de antecedentes presentes nos dados. Além disso, a base de dados da FN-Br não contempla as entidades nomeadas. Devido à vasta variedade existente de elementos nessa categoria, isso gerou também erros no experimento.

Por fim, vemos a ocorrência de determinados padrões com percentuais pouco expressivos em relação aos demais. Entre esses casos há aqueles em que o antecedente é a sentença anterior, em torno de 5,83% de casos de erros. Essa situação está relacionada tanto com a incapacidade do *parser* UD em cruzar a fronteira da sentença, quanto do modelo da FN-Br em identificar o antecedente no frame de Entidade, já que esse elemento é uma sentença inteira e o frame de Entidade é evocado por nomes. Os casos em que o antecedente está localizado na sentença anterior ou anteriores se correlacionam com a limitação da ferramenta da UD em violar o limite da sentença.

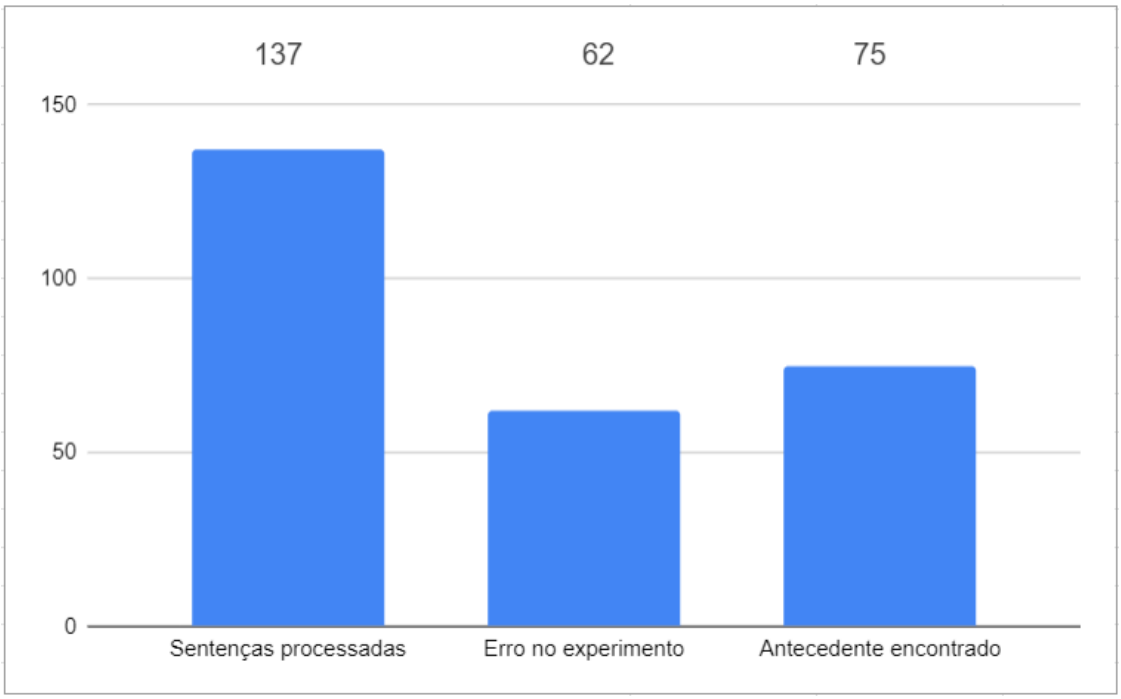
De outra forma, o parser UD também encontrou dificuldades para o processamento de estruturas em contextos com a inexistência de um material lexical explícito que identifique o antecedente de forma mais evidente. Nesses casos, encontramos a prevalência de ocorrências nas quais o antecedente é o escritor (8,63%), o leitor (2,53%) ou não aparece claramente no texto (elipse), cerca de 0,51%.

Para explicar melhor, todas as situações discutidas acima dificultam o mapeamento de antecedentes nas sentenças e, por esse motivo, entram no percentual

de limitações da ferramenta e do modelo de construções anafóricas pronominais no processamento das sentenças. No que concerne às limitações do modelo da FN-Br, temos 62 ocorrências ao todo, que equivalem a 15,74% dos erros. Já no que se refere às limitações do *parser* UD, temos um total de 332 ocorrências, que equivalem a 84,26% dos erros. Assim, o que se percebe é que a maior parte dos erros não se deveu à modelagem das construções em si, mas a limitações de um dos componentes do sistema de identificação de construções e seus antecedentes, nomeadamente o *parser* UD.

Dessa forma, excluindo-se os casos em que o erro de processamento está ancorado em uma limitação do *parser* UD, podemos avaliar o desempenho do modelo do Constructicon da FN-Br para a tarefa de EI. Visando um entendimento maior de análise dos dados, observe o Gráfico 6. Através dele observamos as sentenças que foram processadas e não entram no percentual de casos em que o *parser* não foi capaz de mapear devido suas limitações para a análise de outras sentenças. Em relação ao total de sentenças processadas, sem as limitações do parser por fronteira entre sentenças, vemos o total de 137 sentenças, das quais em 75 ocorrências (54,74%) o antecedente foi encontrado e 62 ocorrências (45,26%) o modelo não funcionou devido a erros de outras naturezas, como, por exemplo, antecedentes que são entidades nomeadas, sintagmas nominais complexos etc.

Gráfico 6 : Resultado do processamento de sentenças sem antecedentes em sentenças anteriores



Em resumo, neste capítulo discutimos as implicações das construções anafóricas pronominais em ferramentas linguísticas de Extração de Dados. Além disso, discutimos sobre as redes geradas por cada construção que compõe a família de construções anafóricas, as limitações do *parser* UD relacionadas à impossibilidade de mapear antecedentes fora do limite da sentença, antecedentes sem uma pista lexical e antecedentes que funcionam como entidades nomeadas. Por fim, demonstramos os resultados do experimento na tarefa de extração de informações em relação ao montante de enunciados processados, discutindo a eficiência do modelo. Notamos que a maior parte dos casos de falha no processamento pode ser atribuída a uma limitação do *parser* da UD. Essa limitação poderia ser minorada se outros tipos de *parsers* pudessem ser utilizados. No nosso entendimento, tais resultados negativos reforçam a necessidade de desenvolvimento de um constructicon para o Português do Brasil que seja capaz de analisar sentenças de diferentes graus de complexidade. A FN-Br tem se dedicado a esta tarefa ao longo da última década e pretende continuar desenvolvendo trabalhos que, como este, contribuam para ampliar a cobertura do *constructicon*.

8 CONCLUSÃO

Nesta pesquisa buscou-se estudar, descrever e explicar a modelagem linguístico-computacional da família de construções anáforas pronominais visando ao aprimoramento de ferramentas de extração de informações e o Constructicon da FN-Br.

Para a modelagem das construções anafóricas pronominais, em um primeiro momento, delimitamos o fenômeno de anáfora pronominal que seria objeto de estudo desta pesquisa. Após isso, cadastramos cada uma das estruturas na base de dados da FN-Br, através da definição de aspectos sintático-semânticos das estruturas para aplicação das restrições. As restrições foram definidas segundo frames, outras construções, ordem e UDs. Para isso, utilizamos os pressupostos teóricos da Gramática de Construções de Berkeley para embasar os traços sintático-semânticos de cada construção.

Ao longo dos capítulos, as construções foram descritas segundo suas características sintático-semânticas particulares. Algumas dessas características eram compartilhadas entre a família, já outras eram específicas de cada subtipo.

Modelamos nove construções que formam a família de construções Anáfora pronominal, sendo uma mais abstrata a *Anáfora pronominal,* herdada pelas demais construções: *Anáfora pronominal demonstrativa*, *Anáfora pronominal possessiva*, *Anáfora pronominal tratamento*, *Anáfora pronominal nominativa*, *Anáfora pronominal indefinida, Anáfora pronominal oblíqua, Anáfora pronominal reflexiva* e *Anáfora pronominal relativa*.

Para a comprovação da eficácia do modelo desenvolvido neste trabalho, as estruturas foram testadas. Com esse objetivo, implementamos cada uma das construções herdeiras na ferramenta de EI a fim de atestar a localização dos antecedentes pelos referentes anafóricos. Os testes realizados comprovaram a eficiência do modelo da FN-Br para uma parte das estruturas, exceto aquelas com antecedentes formados por entidades nomeadas, sentenças inteiras e sintagmas nominais mais complexos. Apesar disso, ele também revelou grandes limitações do *parser* UD em relação à presença de antecedentes localizados fora do limite da sentença em que se encontra o pronome.

Diante do que foi exposto até aqui, essa dissertação propõe os seguintes avanços:

(i) modelagem, a partir das discussões teórico-metodológicas e análise de corpora, de 9 construções anafóricas pronominais para o aprimoramento de ferramentas de extração de informação em textos;

(ii) o adensamento de ferramentas linguístico-computacionais na base de dados da FN-Br;

(iii) análise de erros do desempenho de ferramenta para EI com base em um modelo de construções.

Por meio dos resultados obtidos ao longo da pesquisa, observamos que a modelagem linguístico-computacional de construções anafóricas pode contribuir para a otimização de tarefas de compreensão de língua natural por máquina, através da identificação de referentes de longa distância por anáforas pronominais.

Tendo em vista, as limitações desta investigação, deixa-se abertura para investigações futuras visando ao aperfeiçoamento do modelo de Construções Anafóricas Pronominais e o estudo de outros tipos de anáforas dentro de um domínio mais amplo, multimodal. Ressalta-se, ainda, que a análise de erros aponta para a necessidade de desenvolvimento de novos *parsers* para o Português do Brasil, o que é indicativo da relevância de se adensar cada vez mais a base de dados do Constructicon da FN-Br.

REFERÊNCIAS BIBLIOGRÁFICAS

ADAM, J. M. **A linguística textual**. Editora Cortez, 2008.

CASTILHO, Ataliba, T. de. **Nova gramática do Português Brasileiro**. São Paulo: Editora Contexto, 2010.

CHOMSKY, N. **Aspects of The Theory of Syntax**. Cambridge, Mass.: The MIT Press, 1965.

CUNHA, C.; CINTRA, L. **Nova gramática do português contemporâneo.**

DIAS-DA-SILVA, B. C. **A face tecnológica dos estudos da linguagem: o processamento automático das línguas naturais**. Tese de doutorado-USP. Araraquara, 1996.

FAUCONNIER, Gilles. **Mental Spaces: Aspects of Meaning Construction in Natural Language**. Cambridge: Cambridge University Press, 1994.

FIGUEIREDO, O. M. F. G. **A anáfora nominal em textos de alunos: a língua no discurso.** Tese de doutorado em linguística e linguística aplicada. Universidade do Porto, 2000.

FILLMORE, Charles, J. **Frame Semantics**. In: THE LINGUISTIC SOCIETY OF KOREA. (org.). Linguistics in the morning calm. Seoul, Hanshin, 1982.

_____. **Frames and the Semantics of Understanding.** Quaderni di Semantica, Vol. 6, No. 2, pp. 222-254, 1985.

_____. **On grammatical Constructions**. University of California, 1989.

_____. **Berkeley Construction Grammar**. In: HOFFMANN, T.; TROUSDALE, G. Oxford handbook of Construction Grammar (Eds.). Oxford University Press, 2013.

FRIED, M.; OSTMAN, J-O. Construction Grammar: a thumbnail Sketch. In_____. **Construction Grammar in a Cross - Linguistic Perspective**. Amsterdam: John Benjamins, 2004.

KAY, P.; FILLMORE, C. J. **Grammatical Construction and linguistic generalizations: The What's X doing Y? construction**. In: Language, p. 1-33, 1999.

KOCK, I. G. V. **Desvendando os segredos do texto**. 5 ed. São Paulo: Cortez, 2002.

_____. **Introdução à Linguística Textual: Trajetória e grandes temas.** São Paulo: Martins Fontes, 2006.

KOCH, I. V.; ELIAS, V. M. **Ler e compreender os sentidos do texto**. São Paulo: Editora Contexto, 2017.

LAGE, L. M. **Frames e Construções: a implementação do Constructicon na FrameNet Brasil.** Dissertação de Mestrado. Universidade Federal de Juiz de Fora, 2013.

LAKOFF, G.; JOHNSON, M. **Metáforas da vida cotidiana**. Chicago: The University of Chicago Press, 1980.

LAKOFF, G. **Women, Fire and Dangerous Things: what categories reveal about the mind.** Chicago & London: University of Chicago Press, 1987.

LYONS, **Semantics.** Cambridge University Press, 1977.

MARCUS, G. Deep Learning Alone Isn't Getting Us to Human-Like AI. **Noema**, 2022. Disponível em: <https://www.noemamag.com/deep-learning-alone-isnt-getting-us-to-human-like-ai/> . Acesso em: 15 de Agosto de 2022.

MARCUSCHI, L. A. **Produção textual, análise de gêneros e compreensão**. São Paulo: Parábola Editorial, 2008.

Marneffe, M. C.; Manning, C.; Nivre, J.; Zemanet, D. **Universal Dependencies.** In: *Computational Linguistics*, ISSN 1530-9312, vol. 47, no. 2, pp. 255-308, 2021.

MÁRQUEZ, L. et al. **Semantic role labeling: an introduction to the special issue**. Comput Linguist 34(2):145–159, 2008.

MCSHANE, M. **Natural Language Understanding (NLU, not NLP) in Cognitive Systems.** In: Association for the Advancement of Artificial Intelligence. p. 43-56, 2017.

OGDEN, C. K.; RICHARDS, I. A. **The Meaning of Meaning**. London: Routledge & Kegan Paul, 1923.

PERINI, Mário A. **Gramática do português brasileiro**. São Paulo: Parábola Editorial, 2010.

PISKORSKI, J. YANGARBER, R. **Information extraction: past, present and future.** In: Multi-source, multilingual information extraction and summarization. Springer, Berlin, pp 23–49, 2013.

QUINE, W. V. **Word and Object**. Cambridge, Mass.: M.I.T. Press, 1960.

ROJO, R. H. R. **Linguagem: representação ou mediação?** Veredas, vol. 1, n.1, p. 41–49, 1997.

RUSSEL, S.J.; NORVIG, P. **Artificial Intelligence: A Modern Approach.** New Jersey: Prentice Hall, 2009 (3º Ed.).

SALOMÃO, M. M. M. **Gramática das construções: a questão da integração entre sintaxe e léxico.** Veredas, vol. 6, n. 1, p. 63–74, 2002.

SAUSSURE, F. **Curso de Linguística Geral**. 27ª ed. São Paulo. Cultrix: 2006 [1916].

SCHANK, R. C.; RIESBECK, C. K. **Inside Computer Understanding.** Lawrence Erlbaum Associates. (Ed.) 1981.

TEIXEIRA, A.; RODRIGUES, M. **Advanced applications of natural language processing for performing information extraction.** Editora Springer, Portugal, 2014.

TORRENT, T. T. **“O HOMEM VAI BOTAR UMA CASA PARA MIM MORAR” – uma abordagem sociocognitivista e diacrônica da construção de dativo com infinitivo.** Dissertação de Mestrado. Universidade Federal de Juiz de Fora, 2005.

TORRENT, T. T.; MATOS, E.; LAGE, L.; LAVIOLA, A.; TAVARES, T.; ALMEIDA, V. G.; SIGILIANO, N. **Towards continuity between the lexicon and the constructicon in FrameNet Brasil.** In: LYNGFELT, B.; BORIN, L.; OHARA, K. H.; TORRENT, T. T. (Orgs.). Constructional Approaches to Language. Amsterdam: John Benjamins Publishing Company, 2018.

FAUCONNIER, G. & TURNER, M. **Conceptual blending and the mind´s hidden complexities.** New York: Basic Books, 2002.

ZAMBENEDETTI, Christian. **Extração de Informação sobre Bases de Dados Textuais.** Dissertação de Mestrado. Universidade Federal do Rio Grande do Sul, 2002.