LEONARDO ZILIO

VERBLEXPOR: UM RECURSO LÉXICO COM ANOTAÇÃO DE PAPÉIS SEMÂNTICOS PARA O PORTUGUÊS

PORTO ALEGRE
2015

UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
INSTITUTO DE LETRAS
PROGRAMA DE PÓS-GRADUAÇÃO EM LETRAS
ÁREA: ESTUDOS DA LINGUAGEM
ESPECIALIDADE: LEXICOGRAFIA E TERMINOLOGIA
LINHA DE PESQUISA: LEXICOGRAFIA, TERMINOLOGIA E TRADUÇÃO: RELAÇÕES TEXTUAIS

VERBLEXPOR: UM RECURSO LÉXICO COM ANOTAÇÃO DE PAPÉIS SEMÂNTICOS PARA O PORTUGUÊS

LEONARDO ZILIO

ORIENTADORA: PROFª. DRª. MARIA JOSÉ BOCORNY FINATTO
COORIENTADORA: PROFª. DRª. ALINE VILLAVICENCIO

Texto de tese apresentado como requisito parcial para a obtenção do título de Doutor pelo Programa de Pós-Graduação em Letras da Universidade Federal do Rio Grande do Sul.

PORTO ALEGRE
2015

CIP - Catalogação na Publicação

Zilio, Leonardo
VerbLexPor: um recurso léxico com anotação de papéis semânticos para o português / Leonardo Zilio. -- 2015.
195 f.

Orientadora: Maria José Bocorny Finatto.
Coorientadora: Aline Villavicencio.

Tese (Doutorado) -- Universidade Federal do Rio Grande do Sul, Instituto de Letras, Programa de Pós-Graduação em Letras, Porto Alegre, BR-RS, 2015.

1. Linguística Computacional. 2. Papéis Semânticos. 3. Processamento de Linguagem Natural. 4. Corpus. 5. Linguagens Especializadas. I. Bocorny Finatto, Maria José, orient. II. Villavicencio, Aline, coorient. III. Título.

Elaborada pelo Sistema de Geração Automática de Ficha Catalográfica da UFRGS com os dados fornecidos pelo(a) autor(a).

Agradecimentos

Às agências de fomento e aos projetos de que participo. Ao convênio CAPES/Cofecub (processo 12537-12-8), representado no Brasil pela CAPES, pela bolsa concedida em meu estágio em Grenoble (novembro de 2012 a outubro de 2013), e ao CNPq (processo 142356/2011-5) pela bolsa de doutorado e taxa de bancada, que tiveram vigência de abril de 2011 até março de 2015. Ao projeto CAMELEON (CAPES/Cofecub 707/11) e ao Projeto RITA (Programa CAPES-STIC-AMSud, Edital 043/2014, Projeto 047/14).

À professora Drª. Maria José Bocorny Finatto, que me aguenta há muito tempo como orientando, sempre me aconselhando e trabalhando incansavelmente para que todos os trabalhos realizados fossem (e sejam) os melhores possíveis, topando sempre qualquer parada. Eu não tenho palavras suficientes para agradecer à altura por todo o esforço e dedicação dela durante estes mais de dez anos de cooperação.

À professora Drª. Aline Villavicencio, que aceitou coorientar esta tese, elaborada por um estudante de Letras, e não da Computação, cujos conselhos foram valiosíssimos no desenrolar deste trabalho, e que gerou uma oportunidade de estágio no exterior que foi algo sem paralelo.

Aos colegas de doutorado em Letras, Aline, Bianca e Fabiano, que sempre estiveram dispostos a trocar experiências e bater um papo descontraído.

Ao colega doutorando em Computação, Rodrigo Wilkens, que colaborou em várias tarefas deste doutorado e sempre esteve disposto a discutir e explicar pacientemente fenômenos óbvios que minha mente de linguista não compreendia. Parceiro de trabalho desde, pelo menos, 2010, foi com quem tive muitas discussões filosóficas e com quem debati o maior número de projetos futuros para a melhoria do PLN.

Ao colega, mestre em Computação, Adriano Zanette, por ter feito um trabalho fantástico na ferramenta de extração de estruturas de subcategorização e sua interface de anotação, e por ter me auxiliado no aprendizado de Python.

Ao professor Dr. Mathieu Mangeot, que me recebeu excepcionalmente bem durante meu estágio no Laboratoire d'Informatique de Grenoble, além de ter oferecido espaço para colocar o *corpus* do Diário Gaúcho na plataforma Jibiki.

Ao professor Dr. Carlos Ramisch, que me deu um enorme suporte em Grenoble e se tornou um grande amigo, além de ter auxiliado em muito no desenvolvimento de alguns experimentos desta tese e ter sido um excelente consultor de assuntos computacionais. Ele também entra para a lista de pessoas que me auxiliaram muito no aprendizado de Python.

Aos demais amigos que conheci em Grenoble, principalmente Paula, Lorreine e Antoine, que, juntamente com o supramencionado Carlos, fizeram de nossa estadia na França um período extremamente agradável, sempre com atividades, festas, jogos e jantas muito divertidos.

Aos meus amigos de todas as horas, seja nos bares, no clube de tênis ou nas mesas de *board games,* por me ajudarem a tirar um pouco o foco dos estudos durante alguns momentos e, com isso, garantir sempre uma energia renovada.

À minha amada esposa, Rafaela, minha colega de profissão, parceira, amiga e confidente, a quem eu devo minha sanidade durante esses quase 12 anos de convivência.

Ao meu irmão, Felipe, pela parceria no tênis e pelas várias conversas ao longo de vários anos acadêmicos que ambos compartilhamos.

À minha irmã e ao meu cunhado, por terem me agraciado com a honra de ser dindo de um afiliado muito querido.

À tia Gemilde, pelo reiterado apoio financeiro, que concorre diretamente com o CNPq.

Aos meus pais, pelo suporte e apoio incondicional durante toda a minha jornada acadêmica e extra-acadêmica.

Resumo

Esta tese propõe um recurso léxico de verbos com anotação de papéis semânticos, denominado VerbLexPor, baseado em recursos como VerbNet, PropBank e FrameNet. As bases teóricas da proposta são interdisciplinares e retiradas da Linguística de Corpus e do Processamento de Linguagem Natural (PLN), visando-se a contribuir para a Linguística e para a Computação. As hipóteses de pesquisa são: a) um mesmo conjunto de papéis semânticos pode ser aplicado a diferentes gêneros textuais; e b) as diferenças entre esses gêneros se destacam no ranqueamento dos papéis semânticos. O desenvolvimento do VerbLexPor se apoia em dois *corpora*: um especializado, com mais de 1,6 milhão de palavras, composto por artigos científicos de Cardiologia de três periódicos brasileiros; e um não especializado, com mais de 1 milhão de palavras composto por artigos do jornal popular Diário Gaúcho. Os *corpora* foram anotados com o *parser* PALAVRAS, e as informações de sentenças, verbos e argumentos foram extraídas e armazenadas em um banco de dados. O VerbLexPor tem 192 verbos e mais de 15 mil argumentos anotados distribuídos em mais de 6 mil sentenças. Observou-se que o *corpus* do Diário Gaúcho privilegia uma sintaxe direta e pouco uso de voz passiva e adjuntos, enquanto o *corpus* de Cardiologia apresenta mais voz passiva e um maior uso de INSTRUMENTOS na posição de sujeito, além de uma menor incidência de AGENTES. Foram realizados também alguns experimentos paralelos, como a anotação de papéis semânticos por vários anotadores e o agrupamento automático de verbos. Na tarefa de múltiplos anotadores, cada um anotou exatamente as mesmas 25 orações. Os anotadores receberam um manual de anotação e um treinamento básico (explicação sobre a tarefa e dois exemplos de anotação). Usou-se o cálculo de multi-π para avaliar a concordância entre os anotadores, e o resultado foi de $\pi = 0{,}25$. Os motivos para essa concordância baixa podem estar na falta de um treinamento mais completo. A tarefa de agrupamento de verbos mostrou que a sintaxe e a semântica são igualmente importantes para o agrupamento. Este estudo contribui para a área de Linguística, com um léxico de verbos anotados semanticamente, e também para a Computação, com dados que podem ser consultados e processados para diversas aplicações do PLN, principalmente por estarem disponíveis nos formatos XML e SQL.

Palavras-chave: Anotação de papéis semânticos, recurso léxico, PLN, Linguística de *Corpus*

Abstract

This dissertation aims at developing a lexical resource of verbs annotated with semantic roles, called VerbLexPor, and based on other resources, such as VerbNet, PropBank, and FrameNet. The theoretical bases of this study lies in Corpus Linguistics and Natural Language Processing (NLP), so that it aims at contributing to both Linguistics and Computer Science. The hypotheses are: a) one set of semantic roles can be applied to different genres; and b) the differences among genres are shown by the ranking of semantic roles. The development of VerbLexPor has two corpora at the basis: a specialized one, with more than 1.6 million words, composed by scientific papers in the field of Cardiology from three Brazilian journals; and a non-specialized one, with more than 1 million words, composed by newspaper articles from Diário Gaúcho. The corpora were analyzed with the parser PALAVRAS, and sentence, verb and argument information was extracted and stored in a database. VerbLexPor has 192 verbs and more than 15 thousand arguments annotated with semantic roles, distributed among more than 6 thousand sentences. We observed that Diário Gaúcho has a more direct syntax, with less passive voice and adjuncts, while Cardiology has more passive voice and more INSTRUMENTS for subjects, and fewer AGENTS. We also conducted some parallel experiments, such as semantic role labeling with multiple annotators and automatic verbal clustering. In the multiple annotators task, each of them annotated exactly the same 25 sentences. They received an annotation manual and basic training (explanation on the task and two annotation examples). We used multi-π to evaluate agreement among annotators, and results were $\pi = 0{,}25$. Reasons for this low agreement may be a lack of a thoroughly developed training. The verbal clustering task showed that syntax and semantics are equally important for verbal clustering. This study contributes to Linguistics, with a verbal lexicon annotated with semantic roles, and also to Computer Science, with data that can be assessed and processed for various NLP applications, especially because the data are available in both XML and SQL formats.

Keywords: Semantic role labeling, lexical resource, NLP, Corpus Linguistics

Índice de Figuras

Índice de Tabelas

Sumário

1 Introdução

Esta tese se propõe a um estudo interdisciplinar entre a Linguística e a Ciência da Computação. Três áreas que oferecem bastante espaço para interação entre Linguística e Ciência da Computação são a Linguística de Corpus, a Linguística Computacional e o Processamento de Linguagem Natural (PLN), de modo que discutiremos nesta tese alguns conceitos que pertencem a elas. A união de duas grandes áreas de estudo que se baseiam, por um lado, em áreas humanas e sociais e, por outro, em áreas exatas implica que os modos de ver um mesmo objeto (no nosso caso, a linguagem), por vezes, são bastante distintos, mas certamente o trabalho interdisciplinar pode beneficiar ambos os lados.

Ao longo desta tese, abordaremos de modo mais aprofundado algumas noções do PLN, porém, cabe fazer nesta introdução uma breve explicação da ideia central dessa área de estudos. É preciso deixar claro desde já também que, apesar de ser possível fazer uma distinção entre Linguística Computacional e PLN, consideramos ambos como a mesma área de estudos, apenas abordada de diferentes pontos de vista; e a Linguística de Corpus é tida como uma área originada na Linguística que serve de interface para o trabalho com a Ciência da Computação. O PLN e a Linguística de Corpus não são a mesma área, porém, têm alguns pontos teóricos (a busca de resultados em *corpora*) e práticos (o uso de ferramentas computacionais) que podem ser usados como uma interface no trabalho interdisciplinar. Nesta introdução, daremos uma ênfase maior ao PLN, pois é uma área não pertencente à Linguística, mas trataremos nesta tese também de pressupostos e pontos de vista teóricos da Linguística de Corpus.

A área do PLN emprega seus esforços para facilitar a interação entre o homem e o computador. Nesse âmbito, os avanços potenciais e já alcançados (principalmente em línguas como o inglês) se concretizam como um grande desenvolvimento na história do ser humano, sendo comparados por Branco et al. (2012) com "a invenção da imprensa por Gutenberg". No entanto, para que se continue a avançar no PLN, principalmente no que diz respeito ao processamento do português, é importante que haja um esforço colaborativo entre várias áreas do conhecimento, incluindo aí as duas grandes áreas da Ciência da Computação e da Linguística.

Na atual situação, apesar de sua importância e apesar de o português ser a quinta língua mais utilizada na Internet[1], a quantidade de recursos disponíveis que pode ser processada por computador ainda é pequena. Ainda estamos muito distantes de outras línguas, como inglês, francês e espanhol, que investem mais tempo e dinheiro no desenvolvimento de recursos e ferramentas para o processamento da linguagem (BRANCO, MENDES, *et al.*, 2012). Nosso estudo procura suprir parte dessa lacuna, oferecendo um recurso que poderá ser utilizado para o PLN e também contribuirá para a descrição do português do Brasil.

Neste estudo, mostramos que é possível trabalhar em conjunto e enriquecer cada vez mais os estudos interdisciplinares, fornecendo recursos que impulsionam não apenas o PLN, mas ampliam o conhecimento linguístico. Para tal, escolhemos como foco o desenvolvimento de um recurso léxico em português com informações de papéis semânticos. Esse recurso poderá ser utilizado tanto por ferramentas computacionais no auxílio ao PLN como será uma fonte de conhecimento sobre o português, tendo em vista que representará uma descrição da nossa língua. Desse modo, este estudo tem seu propósito tanto para a Linguística quanto para a Ciência da Computação.

A princípio, o próprio nome RECURSO LÉXICO EM PORTUGUÊS COM INFORMAÇÕES DE PAPÉIS SEMÂNTICOS pode parecer intimidante, porém, tentaremos esclarecer brevemente cada uma de suas partes para facilitar o entendimento do que vem a ser isso. Os pontos cruciais a esclarecer são os termos RECURSO LÉXICO, que a princípio é bastante amplo, e PAPÉIS SEMÂNTICOS, que é um tema já bastante estudado, tanto por linguistas quanto por cientistas da computação, e bastante controverso.

A definição do que é um recurso léxico, ou recurso lexical, é por vezes negligenciada, talvez por ser entendida como algo trivial. Por exemplo, o livro *Speech and Language Processing: An Introduction to Natural Language Processing, Speech Recognition, and Computational Linguistics*, de Jurafsky e Martin (2000), não fornece uma definição do que seriam recursos léxicos. Felizmente, alguns autores se preocuparam em esclarecer o que é isso e, de acordo com a definição de Nunes (2008), RECURSOS LÉXICOS, que a autora chama de léxicos computacionais, são "estruturas de dados, em formato digital e adequado para consultas eficientes, contendo informações sobre o léxico (conjunto de unidades lexicais) de uma L[íngua] N[atural]". Em outras

[1] Dados de 2013, retirados do site http://www.internetworldstats.com/stats7.htm, acessado em 17 de dezembro de 2014.

palavras, são dados linguísticos armazenados de um modo que possam ser consultados por uma ferramenta computacional. O fato de o nome recurso léxico estar vinculado mais ao tratamento computacional, como indica a definição, não impede seu uso para fins estritamente linguísticos, pois nesses recursos se encontra uma base para a descrição da língua ou da linguagem.

Resta então definir o que vêm a ser PAPÉIS SEMÂNTICOS. Essa é uma tarefa mais complicada, pois a definição não só é complexa, mas também é controversa e vem sendo debatida há muitos anos. Desse modo, reservamos a Seção 2.5 para discutir várias definições de papéis semânticos, sob diferentes pontos de vista, e quais as suas implicações para este estudo. No entanto, cabe nesta introdução fornecer uma breve explicação sobre o assunto. Os papéis semânticos podem ser vistos como uma descrição simplificada e abrangente do significado, sendo usados para apontar a função semântica dos sintagmas de uma oração, conforme exemplificamos a seguir:

1.a. O homem **bateu** no cachorro.

No Exemplo 1.a, o sujeito **O homem** desempenha um papel de AGENTE (ou ARG0), ou seja, de participante no evento que executa a ação, e o objeto indireto[2] **no cachorro** tem o papel de PACIENTE (ou ARG1), isto é, ele é o participante no evento afetado pela ação. Assim, a informação semântica fornecida se configura como uma indicação da função de significado dos sintagmas na oração. Não é um significado como aquele encontrado em dicionários, mas fornece mais informações sobre o significado do que apenas as classificações sintáticas de **sujeito** e **objeto indireto**.

Do ponto de vista sintático, esse mesmo Exemplo 1.a também pode ser visto como a união entre um sintagma nominal (SN – **O homem**), um verbo (V – **bater**) e um sintagma preposicionado (SP – **no cachorro**), de modo que podemos representar essa sentença como SN_V_SP[3]. Esse tipo de representação é chamado também de

[2] Há bastante discussão nas gramáticas acerca do nome que esse tipo de complemento preposicionado pode receber. Bechara (1999) o chama de complemento relativo, enquanto Neves (2000) usa o termo objeto indireto (que Bechara reserva para um tipo diferente de complemento). Ao longo do texto, usamos a nomenclatura mais tradicional, como adotada por Neves (2000).

[3] Como veremos mais adiante, ao longo desta tese, utilizamos uma notação diferente para as estruturas de subcategorização, principalmente devido aos fatores que achamos importante destacar nas sentenças. Assim, por exemplo, o SN sujeito é marcado na estrutura de subcategorização como SUBJ, pois achamos importante explicitar qual SN na sentença representa o sujeito, principalmente pelo fato de que é possível haver inversão de posição com o objeto direto e, além disso, é possível que o sujeito não seja um SN, mas sim uma oração reduzida.

ESTRUTURA DE SUBCATEGORIZAÇÃO (*subcategorization frame* – SCF), e é bastante usada para unir sob uma mesma categoria sentenças com diferentes elementos lexicais, mas com os mesmos elementos sintáticos. Por servir para reunir sentenças com sintaxe similar, usaremos frequentemente as SCFs para representar sentenças neste estudo, e são as SCFs que formam a base inicial para anotação[4] dos papéis semânticos, como veremos na Seção 5.3.

Dadas as breves definições apresentadas, podemos dizer que o estudo desenvolvido nesta tese envolve a criação de uma coleção estruturada de dados linguísticos de língua portuguesa que contém informações sobre o significado de orações. Demos prioridade ao português brasileiro escrito em diferentes gêneros: textos de artigos de Cardiologia e textos do jornal popular Diário Gaúcho. A escolha desses gêneros textuais teve por base uma representação das variantes especializada e não especializada da língua portuguesa, isto é, buscamos representar, com esses gêneros textuais, o uso técnico-científico do português escrito e o uso menos marcado do texto escrito representado por um jornal diário de caráter popular, dirigido para leitores com menor poder aquisitivo e hábito de leitura diferenciado.

Com base nesses dois *corpora*, desenvolvemos um método de anotação amostral que visava a anotar o maior número de verbos e sentenças possível, sem deixar de atentar para os diferentes significados dos verbos. Para isso, o método de anotação foi sendo modificado ao longo deste trabalho, conforme foram sendo realizados estudos-piloto que testaram e aprimoraram a metodologia. Desse modo, este trabalho apresentará, em forma de relato, as diversas etapas nas quais o trabalho foi sendo estruturado e modificado, até que chegássemos aos resultados que apresentamos ao final desta tese.

Passamos agora a detalhar os objetivos deste estudo, que se dividem em objetivo primário e objetivo secundário.

[4] A tarefa de anotação envolve acrescentar informações a determinadas seções de texto (por exemplo, palavras, sintagmas etc.) dentro de um *corpus*, fomentando a sua análise em termos lexicais, sintáticos, semânticos etc. Existem anotações realizadas automaticamente, como veremos mais adiante, quando discutirmos a utilização de *parsers*, e anotações manuais, as quais envolvem o acréscimo manual de informações a um *corpus*, geralmente por um especialista (como, por exemplo, um linguista).

1.1 Objetivo primário

Com base nos dados presentes em textos de Cardiologia e do Diário Gaúcho (que apresentaremos mais detalhadamente no Capítulo 5), temos o seguinte objetivo primário:

> *Desenvolver um recurso léxico com informações sobre papéis semânticos para o português.*

Para tal, precisamos extrair e/ou anotar as seguintes informações acerca dos verbos principais:

- A estrutura de argumentos sintáticos e a estrutura de subcategorização;
- A classificação sintática e semântica da estrutura de argumentos; e
- A quantidade de estruturas de argumentos observadas para cada verbo.

A partir dessas informações, será possível analisar quais tipos de estrutura de argumentos se associam a determinados verbos e qual a sua influência sobre o significado desse verbo. Essa influência pode ser relacionada à proposta semântica presente em Saussure (2006)[5], que prevê a identificação de um significado no eixo paradigmático e um no eixo sintagmático. No eixo sintagmático, o significado de cada elemento linguístico se constrói na relação estabelecida com os outros elementos presentes no texto. Da mesma forma, o significado do verbo se constrói na relação com seus argumentos.

Para deixar mais claro o objetivo, apresenta-se, a seguir, um exemplo bem simples, somente a título de ilustração do procedimento de criação do recurso léxico que almejamos. Para tal, tomamos o Exemplo 1.b, retirado do *corpus* de Cardiologia (grifo nosso):

1.b. Atualmente esse aparelho **pode ser encontrado** nas unidades de atendimento, porém sua interpretação depende de especialistas, que muitas vezes não se encontram presentes no momento do exame.

[5] Para mais informações sobre essa interpretação semântica dos estudos de Saussure (2006), consulte Bouquet (1997) e Zilio (2011).

A partir de sentenças presentes nos *corpora*, tais como o Exemplo 1.b, pretendemos retirar as informações de que existe um verbo principal (**encontrar**) e dois argumentos ligados a ele (**esse aparelho** e **nas unidades de atendimento**). De posse dessas informações, classificamos os argumentos de acordo com os papéis semânticos que se apresentam na estrutura de argumentos. Assim, o argumento **esse aparelho** seria classificado como TEMA, por representar um elemento que não é afetado pelo evento, mas apenas está presente nele, e o argumento **nas unidades de atendimento** seria classificado como LUGAR, pois representa o lugar onde o evento ocorre.

Percebe-se que esse formato suprime o papel AGENTE, que poderia ter sido explicitado se houvesse um agente da passiva na sentença. Esse tipo de classificação se chama de anotação de papéis semânticos. Neste estudo, a anotação será realizada por apenas um anotador humano, o autor desta tese. Porém, como veremos no Capítulo 7, realizamos também um experimento com essa anotação sendo feita por múltiplos anotadores, um grupo de estudantes de Linguística do PPG-Letras da UFRGS.

Como este estudo utiliza *corpus*, as informações de argumentos podem ser extraídas de vários contextos reais, de forma que, para cada verbo, haverá mais de uma estrutura de argumentos e mais de uma ocorrência de cada estrutura. Assim, pretendemos observar as diferentes estruturas de argumentos encontradas para cada verbo e analisar a relação de significação entre as estruturas argumentais e os verbos, o que servirá de insumo para uma anotação manual de papéis semânticos. O procedimento da anotação será mais bem explicado ao longo dos capítulos correspondentes às várias etapas da anotação (Capítulos 6 e 8), mas cabe aqui desde já apresentar rapidamente a interface de trabalho, a qual pode ser vista na Figura 1.1.

Na Figura 1.1, pode-se perceber que o trabalho está bastante concentrado na avaliação das estruturas sintáticas e dos elementos linguísticos presentes, e na anotação de papéis semânticos para cada argumento por meio de uma lista de rolagem (Figura 1.2). Desse modo, o resultado do estudo será um recurso léxico composto por uma lista de verbos com respectivos exemplos, argumentos e papéis semânticos.

Figura 1.1 – Exemplo da interface para anotação de papéis semânticos

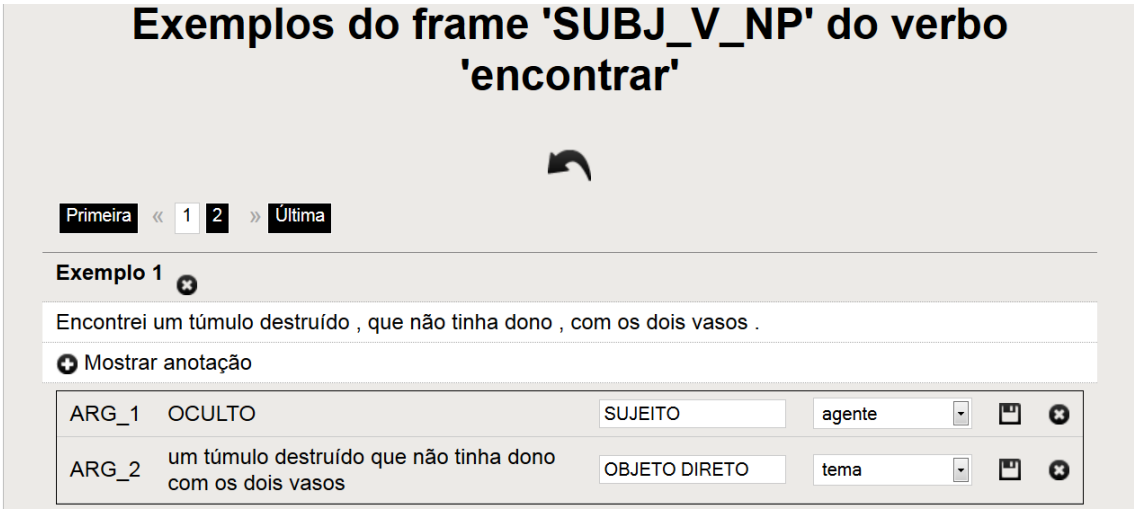
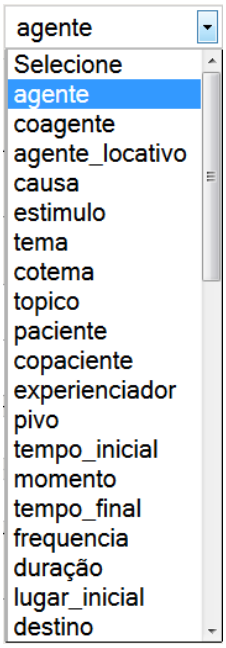


Figura 1.2 – Exemplo da lista de rolagem com os papéis semânticos



1.2 Objetivo secundário

Depois que o recurso léxico estiver pronto, com sentenças dos dois *corpora* semanticamente anotadas e uma lista com informações semânticas sobre verbos empregados em textos científicos de Cardiologia e no jornal popular, será possível averiguar outras informações decorrentes das informações adquiridas, ou mesmo realizar experimentos relacionados a elas.

Assim, nosso objetivo secundário é o seguinte:

> *Realizar uma comparação entre as sentenças e verbos nos gêneros textuais especializado e não especializado.*

Esse objetivo está vinculado também às hipóteses deste estudo, segundo as quais um conjunto de papéis semânticos pode ser empregado em diferentes gêneros textuais e as diferenças entre eles se dá no ranqueamento dos papéis, como será visto mais adiante (Seção 1.4.2), quando tratarmos das hipóteses.

Na Seção 1.1, apresentamos o Exemplo 1.b, com o verbo **encontrar**, que reproduzimos a seguir:

1.b. Atualmente esse aparelho **pode ser encontrado** nas unidades de atendimento, porém sua interpretação depende de especialistas, que muitas vezes não se encontram presentes no momento do exame.

Nesse exemplo, os argumentos do verbo **encontrar** foram classificados com os papéis semânticos TEMA e LUGAR. No Exemplo 1.c, retirado do *corpus* composto por artigos do jornal Diário Gaúcho, que será mais bem detalhado na Seção 5.1, observa-se que as palavras associadas ao verbo **encontrar** são bastante diferentes em relação ao que vimos no Exemplo 1.b, porém, no que diz respeito aos papéis semânticos, a sua configuração é a mesma (grifo nosso):

1.c. O pé direito do calçado **foi encontrado** no buraco da loja de celulares.

Nesse caso, o verbo **encontrar** também possui um argumento TEMA (**o pé direito do calçado**) e um argumento LUGAR (**no buraco da loja de celulares**).

Existem casos, porém, em que uma mesma estrutura sintática em diferentes tipos de texto (especializado e não especializado) apresenta significados diferentes para um mesmo verbo, como podemos ver nos Exemplos 1.d e 1.e, também extraídos dos *corpora* de Cardiologia e do Diário Gaúcho, respectivamente.

1.d. Se a decisão for pelo ensaio explanatório, a análise será feita de acordo com o tratamento que cada paciente **recebeu**.

1.e. O Real Madrid **recebe** o Roma.

Podemos perceber que, sintaticamente, os argumentos são os mesmos, com um sujeito e um objeto direto, porém, no Exemplo 1.d, os papéis semânticos são, respectivamente, de ALVO e TEMA. Já no Exemplo 1.e temos, respectivamente, AGENTE e TEMA. Desse modo, podemos ver que há diferença entre algumas sentenças, o que tem a ver com as diferenças semânticas das orações expressas em cada um dos gêneros textuais, mas queremos saber se essas diferenças são algo recorrente ou apenas um fenômeno isolado. É perceptível que essas diferenças podem ser uma simples questão de polissemia, em que os dois *corpora* apresentam o verbo com os dois significados. Contudo, pode ser que um *corpus* privilegie apenas uma das duas formas, de modo que seria uma marca de gênero textual, e não apenas de polissemia verbal.

A observação de diferentes significados em diferentes contextos torna possível estabelecer classes de significados e classes de elementos relacionados. Existe uma proposta parecida realizada já há bastante tempo por Harris (2002), que previa estruturas diferentes para textos especializados em relação à linguagem comum. Tendo os dois *corpora* anotados, será possível observar se há uma especificidade no uso de verbos nos textos de Cardiologia em relação aos textos jornalísticos do Diário Gaúcho, contribuindo para um maior reconhecimento terminológico dos verbos da área.

Agora que apresentamos os objetivos desta tese, apresentamos rapidamente alguns dos motivos que nos levaram à realização do trabalho.

1.3 Justificativa

O motivo que nos leva a escolher o desenvolvimento de um recurso léxico com anotação de papéis semânticos como foco e objetivo primário de nosso estudo é o fato de que esse tipo de recurso oferece insumos para uma série de aplicações tanto linguísticas quanto computacionais. Porém, não existe um número grande de recursos desse tipo para o português, e não existe nenhum com a metodologia que empregamos.

Na Linguística, esse tipo de recurso serve para a descrição do português (especializado e não especializado), tendo em vista que apresenta um catálogo estruturado de verbos com as respectivas informações sintáticas e semânticas. Sob esse ponto de vista, existem apenas três outros recursos que se apresentam de maneira semelhante: o PropBank.Br (DURAN e ALUÍSIO, 2011; DURAN e ALUÍSIO, 2012), a

VerbNet.Br (SCARTON, 2013) e a FrameNet Brasil (SALOMÃO, 2009); discutiremos esses três projetos mais adiante, no Capítulo 3, apontando também as diferenças que eles apresentam em relação ao recurso que descrevemos aqui.

No PLN, o recurso proposto pode ser empregado para a análise semântica de sentenças, o reconhecimento automático de significado e muitas outras tarefas associadas. Temos, por exemplo, trabalhos que usam informação semântica para resolução de anáforas (KONG e ZHOU, 2012), sumarização automática (YOSHIKAWA, IIDA, *et al.*, 2012), tradução automática (FENG, SUN e NEY, 2012; JONES, ANDREAS, *et al.*, 2012) etc.

Tomando por base os objetivos e a justificativa apresentados, passemos agora aos pressupostos, às questões de pesquisa e à hipótese desta tese.

1.4 Pressupostos, questões de pesquisa e hipóteses

1.4.1 Pressupostos

Nossos pressupostos se dividem em dois tipos: teóricos e metodológicos. Os pressupostos teóricos são as nossas principais visões sobre as teorias e escolas de pensamento que embasam este estudo. Já os pressupostos metodológicos são aqueles que tratam quase exclusivamente de nosso entendimento e/ou nossas restrições relativos a procedimentos práticos do estudo.

Assim, as premissas teóricas básicas deste estudo são as seguintes:

- **A linguagem humana é realizada através de textos, e esses textos podem se apresentar em suportes e formatos variados.**

Para efeitos desta pesquisa, restringimos a abrangência do termo *linguagem* à sua realização em textos escritos. Entendemos que a linguagem é a soma das palavras e sintagmas presentes nos textos e que ela está profundamente vinculada aos gêneros textuais e contribui para a sua definição, juntamente com os modos de dizer. Sendo que esses modos são, em última instância, também definidos na linguagem através de associações entre palavras. Sendo assim, a linguagem é entendida como a língua em uso.

- **A língua é uma estrutura que pode ser abstraída a partir da linguagem, e a linguagem se constitui como um sistema de relações probabilísticas.**

Com isso, queremos dizer que a língua é um sistema que pode ser estudado a partir da linguagem. É na linguagem que se encontram os elementos realizados da língua, e esses elementos se concretizam de acordo com determinadas probabilidades de ocorrência. Essa questão das probabilidades será discutida mais adiante, na Seção 2.1, quando tratamos de Linguística de Corpus.

É importante ressaltar que, embora entendamos que a língua possa ser abstraída a partir de textos, o que nos faz trabalhar com *corpora*, o foco deste estudo recai de fato sobre a oração, pois trabalharemos com elementos no entorno de verbos. Além disso, a metodologia empregada, como veremos mais adiante, não permite que observemos diretamente um contexto que vá além da sentença. Essas restrições são estritamente metodológicas e foram necessárias para que o estudo pudesse ser conduzido em larga escala com um bom custo-benefício.

- **O estudo da linguagem em *corpora* é eficaz para a compreensão de fenômenos linguísticos, pois permite que o linguista observe dados concretos.**

Os *corpora* são amostras da linguagem e, portanto, oferecem uma base para o estudo de fenômenos concretos da linguagem. Como já mencionamos acima, trataremos aqui exclusivamente da linguagem escrita, pois, como aponta Nunes **(2008)**, o PLN se restringe quase exclusivamente à escrita, deixando o tratamento da fala para outras áreas de estudo. A Linguística, como grande área, possui abordagens para o trabalho com texto falado, mas este não entrará em nosso estudo.

- **Os gêneros textuais são arquétipos de texto/discurso relativamente estáveis, como propôs Bakhtin (1997).**

Em nossa visão, muito mais bem articulada em Zilio (2009; 2012), entendemos que esses arquétipos são estabelecidos pelas comunidades discursivas envolvidas na comunicação, assim como propôs Swales (1990). Porém, diferentemente deste, assumimos, com Marcuschi (2002), que toda a forma de comunicação se dá por meio de algum gênero textual, e não apenas algumas delas. O tema dos gêneros textuais é complexo e gera muitas discussões acirradas, justamente por não ser algo trivial. Contudo, não é nossa intenção realizar um tratado sobre o assunto, de modo que apenas

apontamos para referências em que nossa percepção sobre o tema está mais bem exposta. Deixamos claro, porém, que entendemos as visões de Bakhtin (1997) e Swales (1990) como complementares, assim como o fazem Possamai e Leipnitz (2007), e que não discutiremos aqui as questões de diferença entre texto e discurso. Nesta tese, trabalhamos com dois gêneros em destaque: o artigo científico de Cardiologia e o texto jornalístico (composto por vários subgêneros). Esses dois gêneros também são tomados por nós como representantes de duas diferentes esferas da linguagem: respectivamente, a linguagem especializada e a linguagem não especializada.

- **Um texto é uma ocorrência comunicativa que atende a uma série de critérios, conforme estabelecidos nos princípios da Linguística do Texto, como propuseram Beaugrande e Dressler (2002) desde os anos 1980.**

Essa é uma definição básica, porém eficiente de texto, que está de acordo com nossa visão linguística. O texto é entendido como um todo de significado, formado por um conjunto finito e ordenado de orações coesas e coerentes que podem ser consideradas como signos linguísticos complexos **(HOFFMANN, 1988)**. Cremos que essa definição de Hoffmann, ainda que bastante correta em sua articulação, poderia ser ampliada para apontar as sentenças, e não as orações, como unidade formadora dos textos. As orações possuem limitadores que as sentenças não têm, e estas, por serem uma estrutura acima das orações, se enquadram como uma unidade melhor para a constituição de um texto.

- **O *corpus* jornalístico utilizado é um representante válido da linguagem não especializada.**

Os textos jornalísticos, em geral, apresentam certo nível de linguagem especializada, como o que ocorre em textos sobre Economia, Saúde, Esportes etc. Porém, os textos do Diário Gaúcho conseguem neutralizar grande parte dessa informação especializada. O objetivo do jornal popular, tal como é o Diário Gaúcho, é passar informações à população-alvo de um modo acessível sobre temas do seu cotidiano. Assim, a tendência do seu texto é a fuga das linguagens especializadas e o uso de um padrão de léxico e de gramática menos complexo, evitando-se estilos eruditos. Por esse motivo, esse material fornece contraponto eficiente para a relação entre linguagem não especializada e linguagem especializada.

Esses foram nossos pressupostos teóricos, o que seguem agora são os pressupostos mais voltados à metodologia:

- **Uma sentença é representada por uma palavra ou sequência de palavras seguida por um ponto final.**

Essa definição é bastante básica e, naturalmente, tem algumas exceções. Ela é uma definição metodológica. Do ponto de vista linguístico, uma sentença pode ser entendida como uma unidade complexa de significado, porém, neste estudo, uma sentença é entendida como uma unidade de texto que está delimitada de acordo com um determinado algoritmo de decisão que divide os textos em suas unidades menores. Em geral, uma sentença será como está indicado no pressuposto e terminará em um ponto final, porém, devido à natureza dos *corpora* é possível que uma sentença não acabe em um ponto ou mesmo que vá além dele, tendo em vista que o mesmo sinal gráfico do ponto final pode ser usado também para sinalizar abreviações, demarcar números etc.

- **Uma oração é uma sentença, ou parte de uma sentença, que é encabeçada por um verbo ou locução verbal.**

Diferentemente da definição de sentença, que é essencialmente metodológica, a definição de oração é mais linguística, mas preferimos colocá-la aqui por razões de fluxo textual. Como aponta Neves (2000), uma oração precisa ter um verbo para ser considerada como tal. Esse mesmo requisito não é feito para uma sentença, que pode perfeitamente ser constituída por apenas um sintagma nominal ou uma exclamação etc.

Em nosso estudo, por estarmos estudando fenômenos vinculados diretamente aos verbos, precisamos nos afastar um pouco da totalidade do texto e dar um *zoom* nas unidades que o formam. Anteriormente, utilizamos uma definição de Hoffmann (1988) para definir um texto como um conjunto de orações. Aqui, repetimos essa definição para chamar atenção ao fato de que trabalharemos diretamente com unidades menores que o texto, ainda que elas, em última instância, façam parte de um todo coeso e coerente. Em nossa metodologia de análise de *corpus*, essa unidade que é o texto acaba sendo dividida em suas unidades menores (unidades de análise), para que a anotação de papéis semânticos possa ocorrer em torno do verbo.

- **O verbo é um elemento central para a análise sintática e semântica de uma oração.**

O verbo, ou a locução verbal, é o elemento linguístico que une os demais elementos de uma oração, o que o torna também o centro para a determinação do significado da mesma. Essa visão será mais bem elaborada na Seção 2.3, ponto em que tratamos especificamente da importância do verbo para este estudo.

- **Havendo dúvida quanto à anotação sintática das sentenças, as informações fornecidas pelas ferramentas automáticas são consideradas corretas *a priori*.**

Aqui voltamos a ter um pressuposto estritamente metodológico. Nesta tese, as informações fornecidas pelo *parser* ou pelos demais sistemas computacionais que utilizarmos somente serão consideradas erradas se não houver nenhuma possibilidade de considerá-las corretas, ou seja, se não houver nenhuma dúvida quanto ao erro. Este pressuposto se refere principalmente à anotação automática do *corpus*.

- **A extração semiautomática de dados, sempre que possível, mesmo com seus problemas, é mais eficiente do que uma análise puramente manual.**

Este trabalho se apoia muito em ferramentas computacionais para o processamento de textos. Nossa opção pelo processamento automático em vez de uma análise completamente manual se deu justamente porque a análise manual, ainda que cuidadosa, está propensa ao erro aleatório, e esse erro aleatório muitas vezes gera mais problemas do que o erro de um sistema, que apenas vai errar onde o algoritmo não é robusto o suficiente para o caso em questão. Isso é ainda mais verdade quando tratamos de uma grande quantidade de textos, que é o nosso caso. Por isso, confiamos o trabalho pesado desta tese aos programas computacionais, e usaremos análise manual apenas para os casos em que não tivermos uma ferramenta disponível.

Os pressupostos acima variaram bastante em sua característica, sendo alguns mais teóricos, e outros de cunho mais metodológico. Ambos os casos são importantes, pois tratamos aqui do desenvolvimento de um recurso que requer um tratamento metodológico apurado e um embasamento teórico à mesma altura. Como não discutiremos profundamente cada uma das noções aqui apresentadas, embora algumas ainda tornarão a ser debatidas posteriormente, esta seção de pressupostos serviu para

nos posicionarmos teórica e metodologicamente em relação a questões mais abrangentes de Linguística e de Processamento de Linguagem Natural.

Agora passamos às nossas questões de pesquisa e às nossas hipóteses.

1.4.2 Questões de pesquisa e hipóteses

Dados os objetivos apresentados anteriormente, levantamos as seguintes questões de pesquisa:

- Como se caracterizam as estruturas argumentais de verbos do português brasileiro em textos de jornalismo popular?

Essa questão representa principalmente uma curiosidade em relação à configuração das orações em textos escritos em português. Estão envolvidas questões como o tipo de voz (ativa ou passiva), o uso de preposições, sujeitos ocultos vs. sujeitos explícitos. Para responder a essa questão, observaremos dados de um *corpus* de textos do jornal Diário Gaúcho levantados por um extrator de estruturas de subcategorização e da anotação de papéis semânticos, que faz parte do objetivo primário desta tese.

- Se existirem, quais são as diferenças que marcam textos especializados em relação a textos não especializados no que diz respeito às estruturas sintáticas e semânticas?

Essa questão decorre principalmente do objetivo secundário, de comparar os resultados encontrados nos dois *corpora*. As comparações realizadas serão tanto quantitativas quanto qualitativas e serão realizadas com diferentes tipos de informação: papéis semânticos, estruturas de subcategorização e categorias sintáticas. Respondendo a essa questão, poderemos traçar um paralelo entre os dois tipos de linguagem e gerar mais informações para os estudos de Terminologia e Lexicologia do português brasileiro. Ao responder a essa questão, observaremos diretamente a existência ou não de verbos terminológicos, algo que já se vem postulando em vários estudos (PICHT, 1987; MACIEL, 2001; BEVILACQUA, 2004).

Além dessas duas questões de pesquisa, temos duas hipóteses a serem verificadas, as quais apresentamos a seguir:

- ***Diferentes gêneros textuais podem compartilhar um conjunto de papéis semânticos descritivos genéricos.***

Tendo em vista que não existem, em nosso conhecimento, *corpora* especializados anotados com os mesmos papeis semânticos genéricos usados para *corpora* não especializados, a ideia subjacente a essa hipótese é de que, por mais que os textos apresentem um léxico diferente e também uma semântica diferente, por serem científicos ou de jornalismo genérico, a relação sintático-semântica e as funções semânticas que concernem aos verbos não serão diferentes entre os gêneros especializado e não especializado. A comprovação dessa hipótese decorrerá da anotação que realizaremos em dois gêneros textuais.

- ***O que define a especificidade dos domínios nos* corpora *estudados é o* ranking *dos papéis semânticos.***

Essa hipótese diz respeito ao que diferenciaria os gêneros textuais, e é complementar à primeira hipótese. Acreditamos que a principal diferença, no que diz respeito aos papéis semânticos nos gêneros textuais, é a frequência da associação deles aos verbos presentes nos diferentes gêneros. Assim, postulamos que os papéis semânticos nos textos do Diário Gaúcho apresentarão um *ranking* diferente daqueles que se encontram no *corpus* de Cardiologia.

2 Fundamentação Teórica

Dando sequência aos pressupostos, questões de pesquisa e hipótese que apresentamos anteriormente, neste capítulo, apresentaremos teorias e conceitos que sustentam nossos pressupostos e à luz dos quais realizamos este estudo. Começamos com Linguística de Corpus e, em seguida, apresentamos algumas informações sobre Linguística Computacional e PLN. Depois de apresentar as áreas de pesquisa das quais este trabalho toma sua principal fundamentação, passamos a tratar de conceitos que serão recorrentes neste estudo, discutindo aspectos que os tornam essenciais em nosso escopo. Ao final, apresentamos uma breve recapitulação com as principais informações do capítulo.

2.1 Linguística de *Corpus*

A Linguística de *Corpus* é uma área de estudos relativamente nova que se desenvolveu principalmente a partir dos anos 80, quando o computador se tornou comum na sociedade, ainda que se possam destacar estudos importantes anteriores a essa década[6]. A Linguística de *Corpus* postula que as investigações de linguagem devem ser feitas em aplicações reais da língua, preferencialmente em grandes extensões de textos (orais ou escritos), às quais chamamos de *corpus*[7].

Nesse âmbito, a linguagem é entendida como um sistema em que cada palavra tem uma determinada probabilidade de ocorrência dentro de um determinado contexto. Isso quer dizer que as palavras em um determinado contexto são previstas pelas demais palavras já presentes. Assim, a Linguística de *Corpus* é uma área bastante vinculada ao eixo sintagmático e ao modo como as palavras se articulam para formar um texto coeso, mas sem deixar de lado a semântica, que é entendida como a relação de coexistência com outras palavras. Desse modo, para a Linguística de *Corpus*, é o contexto que forma o significado. Esse entendimento de semântica foi formulado de maneira bastante

---

[6] O primeiro *corpus* digitalizado, por exemplo, foi o *corpus* Brown, concluído em 1964 (FRANCIS e KUCERA, 1964).

[7] Um *corpus* pode ser entendido como um conjunto de textos selecionados para um determinado fim e que se apresentam em formato processável por programas de computador. Existe uma série de critérios importantes para a montagem de um *corpus* e uma série de decisões que precisam ser tomadas para que o *corpus* possa ser utilizado da melhor forma possível. Para maiores informações sobre a montagem e o uso de *corpora*, consulte Berber Sardinha (2004).

inteligente em uma famosa frase de Firth (STUBBS, 1996, p. 35): “Diga-me com que palavras andas e eu te direi que palavra és”.

Um dos motivos que aproxima este trabalho da Linguística de *Corpus* é a crença de que os estudos sobre a linguagem (ou mesmo sobre a língua) devem ter como base alguma referência real. Sabemos que um *corpus*, por maior que seja, não apresentará todas as possibilidades da língua; porém, ele apresenta dados observáveis com os quais é possível adquirir informações sobre um fenômeno linguístico. Assim, os dados concretos servem como embasamento e prova de que os fenômenos linguísticos descritos ocorrem de fato na linguagem.

2.2 Linguística Computacional e PLN

A Linguística Computacional e o PLN andam lado a lado, por vezes sendo considerados similares (JURAFSKY e MARTIN, 2000; VIEIRA e LIMA, 2001). Outros autores distinguem as duas áreas, como, por exemplo, Dias da Silva (2006), que confere à Linguística Computacional dois *status* em seu trabalho: primeiro, a descreve como um rótulo utilizado por linguistas para trabalhos que se aproximam um pouco do domínio do PLN; e, mais adiante, a descreve como uma área da Ciência da Computação que se ocupa do estudo das linguagens formais e de programação e que “não deve ser considerada um desdobramento da Lingüística” (DIAS-DA-SILVA, 2006, p. 128). Othero (2006, p. 342) afirma que “a Lingüística Computacional pode ser didaticamente dividida em duas subáreas: a Lingüística de Corpus e o Processamento de Linguagem Natural (PLN)”. Othero e Menuzzi (2005, p. 25) dizem ainda que “a linguística computacional é a área da ciência linguística voltada para o tratamento computacional da linguagem e das línguas naturais”. Essas últimas afirmações apontam tanto o PLN quanto a Linguística Computacional como áreas da Linguística. Tendo em vista o fato de que o tratamento automático de linguagem vem sendo abordado há mais tempo e com maior comprometimento no âmbito da Inteligência Artificial, ramo da Ciência da Computação, cremos que subordinar ambas as áreas totalmente à Linguística parece ser uma ideia bastante controversa.

Não entraremos aqui no mérito de quem tem razão, nos limitamos a ponderar que, dados esses diversos pontos de vista, uma definição do que é Linguística Computacional ainda é tema para debate. Contudo, para os efeitos deste estudo, assumimos o posicionamento de Jurafsky e Martin (2000), que consideram que Linguística Computacional e PLN se ocupam do mesmo assunto, porém são

considerados diferentes apenas por se afiliarem a diferentes áreas. Para Jurafsky e Martin (2000, p. 9), "o processamento de fala e linguagem envolve várias áreas diferentes, mas que compartilham assuntos, em diferentes departamentos: linguística computacional no departamento de linguística, processamento de linguagem natural no departamento de ciências da computação". Como definição de Linguística Computacional, tomamos as palavras de Vieira e Lima (2001, p. 47): "a área de conhecimento que explora as relações entre lingüística e informática, tornando possível a construção de sistemas com capacidade de reconhecer e produzir informação apresentada em linguagem natural". Assim, cremos que a existência de dois termos é válida para ressaltar os diferentes pontos de vista pelos quais uma mesma área pode ser abordada, seja por linguistas ou por cientistas da computação, porém, conforme mencionamos anteriormente, usaremos os dois termos de maneira intercambiável, provavelmente pendendo mais para o uso do termo *PLN*.

Deixando de lado a discussão acerca das definições de Linguística Computacional e PLN, passamos agora a descrever alguns aspectos de estudos do ponto de vista das Ciências da Computação. As teorias dentro do PLN, ainda que existentes, estão mais voltadas aos fins concretos do que a uma discussão acerca do seu *ontos*. Em geral, o PLN se utiliza de teorias desenvolvidas em outras áreas (dentre as quais se encontra também a Linguística), mesclando-as com a Ciência da Computação para gerar *softwares* aplicados a soluções de linguagem. Para o PLN, é mais importante desenvolver um método que permita ao computador processar um texto e, por exemplo, responder a uma pergunta do que discutir quais são os elementos que fazem parte de uma resposta bem construída (ainda que isso provavelmente vá estar embutido na solução final). Isso não quer dizer que não existam estudos teóricos na área, basta observar os importantes trabalhos de Dias-da-Silva (1996; 2006), Jurafsky e Martin (2000), Lima, Nunes e Vieira (2007) e Rosa (2011) para comprovar a sua existência; porém, a grande quantidade de sistemas desenvolvidos, muitos deles presentes em nosso dia a dia, mostra que a teorização não é o principal foco da área.

Neste estudo, o PLN aparece principalmente no escopo e no método, além de ser a área que serviu como ponto de partida de trabalhos que influenciaram esta pesquisa. Quanto ao escopo, entende-se que o recurso léxico gerado poderá auxiliar no processamento do português, como já expomos anteriormente em nossa justificativa. Quanto ao método, o PLN está na base deste estudo, já que utilizamos ferramentas de etiquetagem e de extração de informação que são provenientes dessa área.

2.3 Verbo

Pode parecer estranho dedicar uma seção a um elemento linguístico como o verbo, porém, neste estudo, o verbo será a base. Como as discussões sobre os verbos no português são tão antigas quanto as primeiras gramáticas, não poderíamos deixar de comentar alguns dos principais trabalhos acerca desse elemento linguístico. Assim, utilizaremos esta seção para mostrar a importância dos verbos como organizadores de sentenças e orações.

No que diz respeito ao tratamento de verbos em geral, destacamos trabalhos da área de Lexicografia brasileira que abordaram a regência ou a valência verbal. Essas obras, além de apontar o significado dos verbos, como o fazem os dicionários comuns de língua, mostram algum elemento organizatório dos verbos, indicando que tipos de argumentos os verbos aceitam. Dentre essas obras, podemos citar o *Dicionário de verbos e regimes* (FERNANDES, 1963), o *Dicionário de regência verbal* (LUFT, 1996), o *Dicionário gramatical de verbos do português contemporâneo do Brasil* (BORBA, 1990) e o *Dicionário de usos do português do Brasil* (BORBA, 2002). Esses trabalhos, por mais que sejam exemplares na área, infelizmente não se preocuparam em gerar um recurso que pudesse ser utilizado para o processamento automático do português, tendo em vista que foram publicados apenas em papel e não disponibilizaram seu conteúdo de uma forma que pudesse ser utilizada por computador.

Na área de Terminologia, os verbos ocupam tradicionalmente uma posição secundária, dado que a maioria dos termos são substantivos ou têm um substantivo como elemento-base. Contudo, mais recentemente, a importância dos verbos começou a ser destacada, principalmente no que diz respeito às fraseologias especializadas, e começaram a se observar instâncias de verbos com valor terminológico. Por exemplo, Maciel (2001), que tratou em sua tese de especificidades de verbos performativos em textos jurídicos, aponta que os verbos no discurso jurídico são veiculadores de atos de fala. No que diz respeito à fraseologia[8], destacamos o trabalho de Bevilacqua (2004), que aborda unidades fraseológicas formadas por um núcleo verbal eventivo e um núcleo terminológico em textos sobre Energia Solar, no âmbito do Meio Ambiente. Podemos apontar também o recente trabalho de Alonso Campo e Renau Araque (2013), que, com base principalmente no trabalho de Lorente (2009), arrola unidades terminológicas

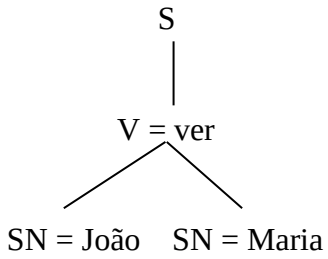
[8] Para maiores informações sobre fraseologias terminológicas, consulte Zilio (2009; 2012).

verbais em textos especializados de língua espanhola, discutindo a contribuição do contexto para a caracterização de um verbo como unidade terminológica. Por fim, o trabalho de Picht (1987) aponta a importância dos verbos para a Teoria Geral de Terminologia, uma teoria que sempre privilegiou os substantivos.

Como se pode ver pelas várias possibilidades de se trabalhar com verbos, não há como fugir de uma certa redundância ao abordar um objeto já bastante explorado. É preciso ficar claro também que não se pretende reinventar a roda, porém, como a linguagem é dinâmica, assim também se faz necessária uma renovação de tempos em tempos, seja na metodologia, seja no *corpus*. Em outras palavras, o trabalho aqui proposto visivelmente não parte do zero, não é o marco inicial dos estudos de verbos que, como aponta Neves (2013), podem ser vistos já nos estudos de Platão, mas visa a contribuir para essa temática por meio de informações renovadas que poderão ser utilizadas na Lexicografia, na Terminologia e no Processamento de Linguagem Natural.

No escopo deste projeto, o verbo é visto como elemento central na oração, de forma que esta sempre será estruturada tomando o verbo como cabeça, ou seja, sua estrutura parte do verbo para os demais elementos oracionais. Isso ocorre porque, assim como postula Neves (2013), com base em Tesnière (1959), entende-se que é o verbo que une e rege os demais elementos oracionais, é ele que coordena os elementos da oração, sendo o único elemento obrigatório desta, e está presente em quase todas as sentenças (e em todas as orações)[9]. No Exemplo 2.a, apresentamos (de forma simplificada) como uma sentença pode ser estruturada sintaticamente a partir do verbo.

2.a. João viu Maria.



Como se percebe, a sentença parte do verbo (V) para os sintagmas nominais (SNs). Essa interpretação toma como base a ideia de que, como já foi dito, o verbo estabelece uma

[9] Esse ponto de vista também é defendido por Fillmore (1967), como veremos mais adiante, no Capítulo 3.

relação sintático-semântica com os demais elementos da oração. Porém, não optamos por essa interpretação somente por seu viés teórico, mas também por uma questão prática e necessária a um tratamento computacional.

Assim, neste estudo, os elementos linguísticos de uma oração estarão sempre ligados ao verbo, formando uma estrutura de argumentos (ou uma valência verbal). Essa opção está vinculada à prática, pois, para obter estruturas de argumentos em uma grande quantidade de sentenças, o modo mais simples é usar um *parser* (*grosso modo* um etiquetador morfossintático e gramatical) para fazer automaticamente a análise sintática. Um dos melhores *parsers* do português (SANTOS e CARDOSO, 2007) é o PALAVRAS[10] (BICK, 2000), o qual adota o verbo como cabeça de sentença e de oração e é capaz de gerar representações em forma de árvores. Explicaremos o funcionamento de um *parser* em maiores detalhes na seção a seguir.

2.4 *Parsers*

Um *parser* é um programa de computador que faz uma análise automática de determinados elementos presentes em uma palavra, sentença, texto ou conjunto de textos. Os rótulos podem variar desde uma simples anotação morfossintática, com etiquetas que classificam as palavras em substantivos, adjetivos, objetos diretos, adjuntos adverbiais etc., até representações da estrutura hierárquica de uma sentença. Quando há essa representação da hierarquia, dizemos que o *parser* faz uma análise das dependências da sentença.

A principal aplicação de um *parser* para o trabalho aqui descrito é a identificação das dependências dos verbos que ocorrem nas sentenças dos conjuntos de textos sob exame, ou seja, o *parser* indica, a partir do verbo, quais elementos da sentença fazem parte do sujeito, quais fazem parte do objeto direto, quais fazem parte de adjuntos adverbiais etc. Um exemplo simplificado desse tipo de hierarquia é a estrutura sintática do Exemplo 2.a, o qual mostra os dois SNs se ligando ao V.

O *parser* que será utilizado, conforme mencionado anteriormente, é o PALAVRAS (BICK, 2000), o qual utiliza uma gramática de restrições (*constraint*

[10] De acordo com Bick (2000, p. 187-189), esse *parser* atinge um percentual de acerto de 96-97% no que concerne à árvore de dependências (organização hierárquica das funções sintáticas), ainda que, em nossa percepção, não cremos que os acertos em qualquer um de nossos *corpora* tenha chegado a esse valor tão elevado. Para uma descrição mais completa do funcionamento da ferramenta, consultar Bick (2000). O acesso ao PALAVRAS foi obtido através dos Projetos COMUNICA e CAMELEON (dos quais participamos), que detêm os seus direitos de uso.

grammar – CG)[11] para estruturar as sentenças. No Exemplo 2.b, mostramos a forma como o PALAVRAS etiqueta um texto:

2.b. João viu o cachorro.

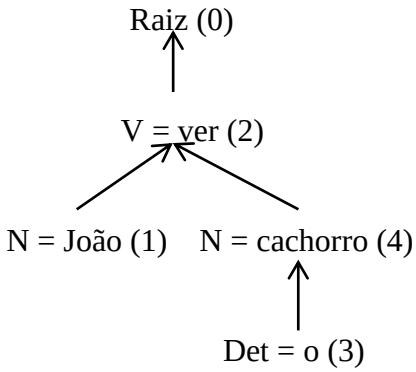
```
João        [João] <hum> PROP M S @SUBJ>  #1->2
viu         [ver] <vH> <fmc> <mv> V PS 3S IND VFIN @FS-STA  #2->0
o           [o] <artd> DET M S @>N  #3->4
cachorro    [cachorro] <Azo> N M S @<ACC  #4->2
$. #5->0
</s>
```

As etiquetas entre colchetes ([ ]) representam a forma lematizada de cada entrada lexical; as marcações entre colchetes angulares (< >) representam informações diversas, que podem ser: semânticas, gramaticais e/ou complementares para a organização interna do *parser*; em seguida, as etiquetas antes do sinal de arroba (@) são as informações gramaticais; e aquelas após a arroba (@) e antes da cerquilha (#), por vezes acompanhadas de um sinal de maior (>) ou menor (<), são as informações sintáticas de cada palavra[12]. Por ser um *parser* de dependências, a saída do sistema também apresenta números após a cerquilha (#) que indicam quem está ligado a quem (sendo que o primeiro número é o número da entrada lexical em questão, e o número após a seta é o número do elemento ao qual a entrada lexical está ligada), formando uma hierarquia.

Dessa forma, as etiquetas na oração-exemplo indicam que **João** (1) está ligado ao verbo **viu** (2), que **o** (3) está ligado a **cachorro** (4); e este está também ligado a **viu** (2). O verbo, por ser a cabeça da sentença, se liga a 0, que é a raiz. A representação gráfica abaixo pode auxiliar para uma melhor compreensão das relações de dependência apresentadas:

[11] Uma gramática de restrições (CG) utiliza regras para moldar a estrutura sintática, restringindo as opções de associação lexical conforme as regras adotadas. Um exemplo de regra seria que um DET (determinante: o, a, esse, essa *etc*.) seguido de um N (substantivo) forma um SN (sintagma nominal). A CG utilizada pelo PALAVRAS serve para construir uma estrutura sintática em forma de árvore cujo sintagma inicial é um sintagma verbal (SV), conforme pode ser visto nos exemplos mais adiante.

[12] Para maiores informações sobre as etiquetas utilizadas pelo PALAVRAS, pode-se consultar os seguintes sites: http://visl.sdu.dk/visl/pt/info/symbolset-floresta.html e http://visl.sdu.dk/visl/pt/info/symbolset-manual.html.



Quando um texto ou *corpus* é etiquetado, ele se apresenta como uma fonte riquíssima para pesquisas que vão desde a simples extração de sintagmas até a confecção de um dicionário ou a identificação de associações semânticas complexas com vistas a auxiliar, por exemplo, a tradução automática. Infelizmente, os recursos existentes para o português ainda são escassos, de forma que não há ainda um *parser* que identifique, por exemplo, papéis semânticos descritivos. O PALAVRAS, por exemplo, apesar de todos os seus recursos, não faz isso. Como aponta Zanette (2010), o desenvolvimento de *parsers* para a língua portuguesa está atrasado em relação a outras línguas, como o inglês, francês e espanhol. Além disso, muitas vezes, eles são de baixa precisão. Além disso, o PALAVRAS, que tem um bom desempenho, não é gratuito e sua licença normal de uso tem um custo bastante elevado. Esse é um dos motivos por que está no interesse deste estudo a colaboração para o desenvolvimento de *softwares* gratuitos que possam se alimentar do trabalho realizado, de modo que se tenha um maior aparato para o tratamento linguístico-computacional do português.

Agora que apresentamos informações sobre verbos e a organização de sentenças, e sobre *parsers*, podemos apresentar um exemplo mais completo, ainda que simplificado, do trabalho que desenvolvemos. O foco deste estudo é restrito à estrutura de argumentos de verbos em português e às relações sintático-semânticas que podem ser apreendidas a partir dessa estrutura de argumentos. Dessa forma, é relevante para este trabalho a identificação de que na oração-exemplo 2.a existe um sujeito que é **João**, um objeto direto que é **Maria** e um verbo que faz o vínculo entre esses dois elementos. Identificados esses elementos, o escopo deste estudo será registrar que, nesse mesmo exemplo, **João** é EXPERIENCIADOR e **Maria** é EXPERIENCIADO da ação. Essas classificações como EXPERIENCIADOR e EXPERIENCIADO são chamadas de papéis semânticos, algo que discutimos na próxima subseção.

2.5 Breves considerações sobre Papéis semânticos

Os papéis semânticos foram introduzidos na teoria linguística há milhares de anos, sendo o seu precursor o gramático indiano Panini (DOWTY, 1991; GILDEA e JURAFSKY, 2002; LEVIN e RAPPAPORT-HOVAV, 2005). Como comentamos rapidamente no Capítulo 1, os papéis semânticos representam a função semântica dos argumentos na oração: "os papéis semânticos distinguem [...] as facetas do significado que são gramaticalmente relevantes" (LEVIN e RAPPAPORT-HOVAV, 2005). Essas facetas do significado podem ser identificadas a partir da observação do léxico e da sintaxe (mas sem deixar totalmente de lado questões de semântica e pragmática). Porém, elas não são nem tão específicas quanto uma semântica lexical[13] (por exemplo, acepções em dicionários), nem tão amplas quanto uma semântica puramente sintática[14]. Em outras palavras, os papéis semânticos nem são tão semânticos para delimitar definições para cada palavra, mas também não são tão sintáticos a ponto de atribuir um mesmo papel para todos os sujeitos e objetos. Esse território intermediário entre semântica e sintaxe em que os papéis semânticos se encontram serve seu propósito para o processamento automático da linguagem e também para a classificação de verbos.

Para exemplificar o que são os papéis semânticos, tomemos como exemplo as sentenças a seguir[15]:

2.c. [João] abriu [a porta] [com a chave].
2.d. [A porta] abriu [com a chave].
2.e. [A chave] abriu [a porta].

Nas três sentenças acima, o verbo é sempre o mesmo (**abrir**), os sujeitos se alternam, mas sempre há um sujeito, e os demais elementos variam conforme a estrutura sintática

[13] Não almejamos aqui uma discussão profunda acerca do que vem a ser a semântica lexical, por isso, usamos a definição proposta por Vieira e Lima (2001), que é simples, mas útil para este estudo: "A semântica lexical considera as propriedades referentes a cada uma das unidades, ou seja, as palavras de uma língua, no léxico." Sabemos que aqui está um pouco desfocada a questão da semântica , tendo em vista que o termo *propriedades* é pouco específico. Ainda assim, essa definição é suficiente para distinguir entre uma semântica voltada para as unidades do léxico (semântica lexical) e uma semântica voltada para elementos mais abrangentes, como sintagmas ou orações.

[14] Em PLN, não é incomum a utilização de categorias sintáticas como sujeito e objeto direto como indícios de diferenciação semântica. Esse tipo de emprego da sintaxe para diferenciação semântica é o que estamos chamando de semântica puramente sintática.

[15] Os exemplos são inventados. Não provêm dos *corpora* envolvidos no estudo. Opta-se aqui por se usarem frases fictícias para simplificar o exemplo e permitir que o foco recaia sobre a explicação do que são papéis semânticos, sem envolver outras questões que poderiam surgir a partir de exemplos reais de uso.

do verbo permite. Os elementos a que chamamos atenção aqui, porém, não são os sintáticos, mas sim os semânticos. Em 2.c, **João** está executando uma ação e tem capacidade volitiva, o que lhe confere o papel de AGENTE (ou ARG0); **a porta** está sofrendo os efeitos dessa ação (está passando por uma modificação de fechada para aberta), o que caracteriza o papel de PACIENTE (ou ARG1); já **a chave** é o INSTRUMENTO (ou ARG2) utilizado pelo AGENTE para realizar a modificação no PACIENTE. Em 2.d, por mais que o sujeito agora seja **a porta**, ela não passa para uma função de AGENTE (ou ARG0), pois ela não está em condições de executar a ação de **abrir** e também não tem capacidade volitiva; assim, ela permanece como PACIENTE (ou ARG1), porque a ação está sendo executada por um elemento não divulgado na sentença. No Exemplo 2.e, o sujeito é **a chave**, mas, novamente, esta não é a executora da ação, ela permanece sendo apenas o INSTRUMENTO (ou ARG2) utilizado por um AGENTE implícito. Além das funções que cada argumento desempenha nos três exemplos apresentados, é possível entender a atribuição de papel por meio de tentativas de comutação dos elementos. Utilizando a estrutura sintática do Exemplo 2.c, que é a mais completa, podemos tentar trocar o léxico de lugar e teríamos seis permutações possíveis; apresentamos três delas a seguir:

2.f. A chave abriu João com a porta.
2.g. A chave abriu a porta com João.
2.h. João abriu a chave com a porta.

Os três exemplos acima mostram que os papéis semânticos não estão vinculados apenas à posição sintática, mas também a questões semânticas e pragmáticas envolvidas na comunicação. Ainda que os Exemplos 2.f, 2.g e 2.h sejam perfeitamente aceitáveis do ponto de vista sintático, eles não fazem sentido do ponto de vista semântico e pragmático. Nessas sentenças, se fôssemos obrigados a fazer uma análise de papéis semânticos, teríamos que aplicar a mesma estrutura que aplicamos a 2.c (AGENTE, PACIENTE, INSTRUMENTO), pois é a isso que nos leva a estrutura sintática e o significado do verbo envolvidos nesses exemplos, mas a dificuldade de aceitar esse tipo de exemplo como plausível também mostra que o valor dos elementos lexicais tem um peso na interpretação das sentenças e contribui para a classificação dos argumentos do ponto de vista dos papéis semânticos.

A partir desses exemplos, pode-se então perceber que os elementos sintáticos (sujeitos, objetos etc.) nem sempre têm uma semântica óbvia, sendo preciso acessar também alguma informação semântica ou pragmática para determinar sua real função semântica na sentença. Desse modo, discriminar os papéis semânticos desempenhados pelos elementos sintáticos em diversos contextos pode ajudar no processamento automático de textos. Poderíamos pensar no seguinte exemplo fictício: em um sistema de extração de informações hipotético, deseja-se conhecer o nome de todas as empresas compradas pela Google nos últimos 10 anos; para isso, não é suficiente detectar apenas orações com verbos de compra nas quais **Google** seja o sujeito, pois seriam ignoradas frases como esta: [Dois bilhões de dólares] compraram [a Android Inc.] [para a Google] [em 2005].

Nos exemplos fornecidos até aqui, apresentamos duas possibilidades de anotar os papéis semânticos: a forma descritiva (AGENTE, PACIENTE, INSTRUMENTO etc.) e a forma numerada (ARG0, ARG1, ARG2 etc.). As formas descritivas são a base para a VerbNet (KIPPER-SCHULER, 2005) e também para os vários projetos baseados na FrameNet (BAKER, FILLMORE e LOWE, 1998). Já a forma numerada foi proposta por Palmer, Gildea e Kingsbury (2005) ao desenvolverem o PropBank. Esses trabalhos serão discutidos no Capítulo 4.

2.5.1 Algumas questões sobre papéis semânticos

Na linguística moderna, os papéis semânticos ressurgiram com os trabalhos de Gruber (1965) e Fillmore (1967), posteriormente se desenvolvendo em trabalhos como os de Jackendoff (1990), Dowty (1991) e Levin e Happaport-Hovav (2005). Para o português, na teoria de papéis semânticos, podemos citar estudos de Franchi e Cançado (2003), Perini (2008), Cançado (2009; 2010), e Cançado, Godoy e Amaral (2012). Retomaremos alguns desses autores mais adiante, no Capítulo 3.

As principais discussões concernentes aos papéis semânticos giram em torno de questões como a quantidade de papéis necessários para representar uma linguagem natural e a subjetividade envolvida na atribuição dos papéis semânticos. Em particular, essas questões são discutidas por Levin e Rappaport-Hovav (2005), que tomam por base também a visão de outros autores citados anteriormente nesta tese. Em síntese, as autoras evidenciam a dificuldade de se estabelecer uma lista de papéis semânticos que não seja nem genérica demais a ponto de não apresentar diferenças suficientes entre os papéis, nem específica demais a ponto de que não se possam depreender generalizações.

A subjetividade é um fator que está constantemente presente nas discussões sobre semântica. Isso ocorre porque, em última instância, cada pessoa identifica um significado diferente (ainda que muitas vezes coincidente ou quase coincidente com o significado atribuído por outras pessoas) para cada texto com que se depara. Assim, existem discussões, por exemplo, sobre como as seguintes sentenças, retiradas de Kasper (2008), deveriam ser interpretadas:

2.i. *The cardinal loaded bottles on the wagon*.
[O cardeal colocou garrafas na carroça.]

2.j. *The cardinal loaded the wagon with bottles*.
[O cardeal colocou garrafas na carroça.][16]

A interpretação, conforme indicada por Jackendoff (1990), é de que, em 2.i, as garrafas não preenchem a carroça, enquanto em 2.j a carroça está completamente cheia. Porém, Fillmore (1968), citado por Kasper (2008), considerava que ambas eram sinônimas. Do ponto de vista dos papéis semânticos, se ambas veiculam o mesmo significado, então os papéis utilizados para os substantivos ***wagon*** e ***bottles*** serão os mesmos nas duas sentenças (assim como foi apresentado nos exemplos 2.c, 2.d e 2.f, em que **a porta** e **a chave** não mudam de papel semântico). Porém, se seus significados forem diferentes, então os papéis também vão diferir.

Em português, temos um exemplo parecido com o que foi apresentado para o inglês, porém, com o verbo *encontrar*:

2.k. O estudo encontrou a doença em 15 pacientes.
2.l. O estudo encontrou 15 pacientes com a doença.

Assim como nos exemplos 2.i e 2.j do inglês, a estrutura sintática das sentenças 2.k e 2.l apresentam diferenças claras devido ao emprego de diferentes preposições; porém, as duas sentenças podem ser consideradas paráfrases, principalmente em textos de

[16] A tradução em português infelizmente não faz jus à ambiguidade existente no inglês, pois não há um verbo que se aplique ao contexto para as duas sentenças com duas estruturas sintáticas.

Cardiologia, onde encontramos sentenças similares à segunda[17]. Desse modo, por um lado, as duas sentenças podem indicar que os pesquisadores encontraram a doença nos pacientes. Assim, o objeto encontrado, nas duas sentenças, é a doença, pois ela é que está sendo procurada, e não os pacientes (os pesquisadores sabem onde os pacientes estão). Por outro lado, a sentença 2.l pode indicar que, em uma busca, foram encontrados 15 pacientes que sofriam de uma determinada doença, de modo que o objeto encontrado, de fato, são os pacientes, pois eles estavam sendo procurados, e não a doença. A doença é apenas um atributo dos pacientes.

Do nosso ponto de vista, esse tipo de diferença parece só poder ser realmente averiguado a partir da observação do referente no mundo real, de um contexto mais amplo ou do gênero textual. Partindo apenas dessas frases escritas, uma pessoa pode interpretar o sentido das duas formas. Desse modo, há uma ambiguidade que só pode ser desfeita pelo conhecimento pragmático, por isso, cabe ao anotador ter ou inferir esse conhecimento e fazer uma anotação condizente com cada caso.

2.6 Estruturas de subcategorização

As estruturas de subcategorização, mais amplamente conhecidas por seu nome em inglês, *subcategorization frames*, são estruturas sintáticas mais abstratas do que as descrições normais de sujeitos, objetos e complementos. Segundo Messiant, Korhonnen e Poibeau (2008, p. 533), as "estruturas de subcategorização de predicados capturam as diferentes combinações de argumentos que um predicado pode ter no nível sintático", ou, como aponta Manning (1993, p. 235), "uma estrutura de subcategorização é uma ratificação dos tipos de argumentos sintáticos que um verbo (ou adjetivo) apresenta". Apesar de as definições fazerem menção ao nível sintático, as estruturas de subcategorização não descrevem, em geral, funções de elementos sintáticos, mas sim sua estruturação básica.

2.a. João viu Maria.

[17] Temos, por exemplo, em nosso *corpus* de Cardiologia, a seguinte sentença: Wilt e colaboradores[26], em 1996, publicaram um estudo de 4.155 pacientes com história de infarto agudo do miocárdio e encontraram 537 (12,9%) com quadro de aterosclerose difusa, sendo 353 (8,5%) com doença obstrutiva periférica e 215 (53,3%) com doença cerebrovascular.

No Exemplo 2.a, a classificação sintática seria: **João** = sujeito; **viu** = verbo; **Maria** = objeto direto. Porém, na classificação de estrutura de subcategorização, essa mesma sentença teria a seguinte análise: **João** = NP (do inglês, *nominal phrase*) ou SN (sintagma nominal); **viu** = V (verbo); **Maria** = NP ou SN. Se tivéssemos um caso com um objeto indireto ou um adjunto preposicionado, ele seria marcado como PP (*prepositional phrase*) ou SP (sintagma preposicional). Assim, as estruturas de subcategorização se apresentam em formatos como NP_V_NP e NP_V_NP_PP, ou, simplesmente, NP_PP (sem indicação da posição do verbo e, às vezes, também sem o sujeito). Com base nessas estruturas, é possível se obter uma boa indicação da estrutura sintática e do número de argumentos que um verbo admite.

O trabalho de Beth Levin (1993), que será discutido mais adiante, partiu do pressuposto de que verbos com uma semântica próxima compartilham estruturas sintáticas, sendo possível agrupá-los em classes semânticas com base apenas em seu comportamento sintático. Dado que as estruturas de subcategorização são um bom indicador da sintaxe das sentenças (podemos dizer que elas indicam a sintaxe de forma implícita), os estudos de PLN as têm usado para a classificação de verbos. Por serem relativamente fáceis de observar em grandes *corpora* analisados sintaticamente, as estruturas de subcategorização acabam servindo como substitutos de classificações sintáticas que identificam explicitamente sujeitos, objetos etc.

As estruturas de subcategorização já foram utilizadas para o agrupamento de verbos em diversas línguas, como alemão (SCHULTE IM WALDE, 2002), francês (MESSIANT, 2008; MESSIANT, KORHONEN e POIBEAU, 2008), inglês (PREISS, BRISCOE e KORHONEN, 2007) e italiano (IENCO, VILLATA e BOSCO, 2008). No Brasil, um trabalho pioneiro no reconhecimento automático de estruturas de subcategorização foi o de Zanette (2010), o qual será descrito no Capítulo 5. Um trabalho que usou essas estruturas para agrupar verbos automaticamente foi a dissertação de mestrado de Scarton (2013), cujos resultados estão expostos de modo resumido em Zanette, Scarton e Zilio (2012), e Zilio, Zanette e Scarton (2014).

2.7 Argumentos vs. Adjuntos

Na discussão sobre estruturas de subcategorização, na seção anterior, mencionamos seguidamente argumentos de verbos. Contudo, existe uma distinção linguística entre o que são argumentos e o que são adjuntos, e é preciso deixar claro qual é o nosso posicionamento acerca dessa distinção. Na noção de Jackendoff (2011),

os adjuntos representam modificadores semânticos, enquanto argumentos representam constituintes semânticos. Assim, a distinção entre argumentos e adjuntos seria uma questão de saturação semântica do verbo.

Segundo Franchi (2003, p. 157):

> "a adjunção se contrapõe à 'estrutura argumental', em que os predicadores atribuem na relação de irmandade os seus papéis temáticos, como uma estrutura não-temática, no sentido de que o adjunto não estabeleceria uma relação temática com o constituinte a que se adjunge ou, pelo menos, não recebe dele um papel temático".

Em outras palavras, Franchi (2003) defende que os adjuntos não recebem seu papel semântico em virtude do verbo. Essas noções tomam como base uma distinção que considera que os argumentos seriam sintagmas que saturam o verbo semanticamente, enquanto adjuntos apenas modificam a semântica da oração. A distinção entre argumentos e adjuntos é, em última instância, uma questão de predicação, é uma questão de determinar qual é o predicador de um determinado sintagma. Se entendermos que o verbo é o predicador de todos os elementos de uma oração, então todos os elementos dela são argumentos desse verbo. Contudo, se algum dos sintagmas da oração não tiver o verbo como seu predicador (em outras palavras, não receber do verbo o seu papel semântico), então esse sintagma não faz parte da estrutura argumental e, portanto, é um adjunto.

Essa distinção é, em teoria, bastante simples, pois basta observar se os sintagmas recebem do verbo o seu papel semântico. Porém, como aponta Cançado (2009), "a associação do [*status* de] argumento ao complemento de um verbo apresenta dificuldades, e a literatura sobre o assunto não é clara". Messiant (2008) também afirma que "não existem critérios linguísticos relevantes o suficiente para fazer uma distinção entre adjuntos e argumentos, não importando o contexto". Não usaremos aqui de representações sintáticas avançadas para distinguir entre argumentos e adjuntos, apenas apontamos, por meio de exemplos simples, algumas características que observamos.

2.b. João vendeu a casa a Maria *por R$50 mil*.

No Exmeplo 2.b, podemos observar que temos um sintagma destacado em itálico. Pode-se argumentar que esse sintagma é um adjunto, ou pode-se entender que ele é um argumento. No caso do verbo *vender*, a gramática tradicional indica que os

dois complementos possíveis para o verbo estão tomados pelo objeto direto *a casa* e pelo objeto indireto *a Maria*, restando um adjunto adverbial que indica o valor da venda. Pode-se também argumentar que esse mesmo adjunto pode ser substituído por um advérbio, como, por exemplo, *caro*, gerando 2.c. No entanto, por se tratar justamente de um verbo que envolve (normalmente) dinheiro, é de se esperar que haja uma quantia discriminada para completar o evento *vender*, de modo que esse sintagma que discrimina a quantia é visto frequentemente como argumento do verbo em questão, tendo em vista que ele está dentro do escopo do significado do verbo.

2.c. João vendeu *caro* a casa a Maria.

Neste estudo, como discutimos no Capítulo 8, reconhecemos que existem papéis semânticos que são potencialmente atribuídos apenas a adjuntos e que esses papéis não necessariamente dependerão do verbo em questão, mas sim de outros fatores, como, por exemplo, a preposição utilizada. Desse modo, a distinção feita em nosso recurso se dá através dos papéis semânticos empregados. Adjuntos recebem papéis semânticos específicos para adjuntos, que não são atribuídos pelo verbo, como aponta Franchi (2003).

Ainda assim, dada a distinção problemática e a dificuldade em estabelecer uma diferença clara entre argumentos e adjuntos, em nosso estudo, utilizaremos uma abordagem parecida com a de Cançado (2009), que atribui o título de argumento ao plano semântico e adjunto ao plano sintático. Em nosso caso, porém, o que faremos é, como já mencionamos, distinguir por meio de papéis semânticos os adjuntos e argumentos de um verbo. Porém, quando discutirmos a estrutura como um todo, não faremos uma distinção, de modo que, mesmo havendo um adjunto, a estrutura semântica de uma oração será referida como estrutura argumental. Desse modo, o título de argumento será aplicado de maneira genérica aos elementos presentes na estrutura semântica da oração, ainda que haja adjuntos em meio a essa estrutura. A distinção específica entre eles se dará por meio da anotação de papéis semânticos aplicada a cada caso.

2.8 Principais ideias discutidas no capítulo

Neste capítulo, vimos que a língua será considerada como um sistema probabilístico, em que as palavras dependem umas das outras para a formação dos

significados. Apesar de a Linguística de *Corpus* refutar a distinção entre língua e linguagem, cremos que a distinção proposta por Saussure (2006) ainda é válida. Se pensarmos que a linguagem é um sistema probabilístico, mas que não existe um ponto de onde tirar essas probabilidades, então ficamos com um elemento circular, já que não há um início. Se pensarmos que a língua é um sistema probabilístico, e que a linguagem é onde essas probabilidades se realizam e se atualizam, então temos uma separação em que um sistema (se) alimenta (d)o outro. Desse modo, quando estudamos a língua, partimos da linguagem, que é nosso objeto concreto e observável por meio de *corpora*. Assim, a língua será reconhecida a partir de um recorte da linguagem, tendo em vista que é praticamente impossível recolher todas as instâncias de linguagem existentes relacionadas a um idioma.

A Linguística Computacional (ou o PLN) entra como nossa área de foco, fazendo uma ponte entre a Linguística e a Informática. É nela que nos inserimos, pois estamos trabalhando com vistas a gerar resultados benéficos para ambas as áreas, assim como retiramos informações de trabalhos que advêm de ambas as áreas. Na Linguística, encontramos bases teóricas sobre verbos e papéis semânticos, como pode ser visto em nossa descrição sobre esses conceitos. Na Informática, encontramos recursos que permitem a realização do trabalho, como *parsers*, banco de dados estruturados e extração automática de informações. Porém, como apontamos, não apenas retiramos informações dessas áreas, como também devolvemos resultados. Para a Lexicografia e Terminologia, desenvolveremos uma descrição da língua portuguesa no que tange aos papéis semânticos, um assunto ainda pouco explorado concretamente, ainda que bastante debatido no mundo teórico. Para o PLN, entregaremos um recurso que pode ser processado por máquina, podendo ser empregado em sistemas de extração de informação, tradução automática, sumarização de textos etc.

Além das teorias, também apresentamos alguns conceitos e a forma como eles serão compreendidos nesta tese. Começamos apresentando nosso interesse pelos verbos, caracterizados como elemento central da oração e por isso elementos norteadores da anotação de papéis semânticos. O verbo principal será o elemento considerado para a organização dos dados. Em nosso recurso, ele funcionará como um lema em um dicionário. As informações de orações presentes nos *corpora* estarão vinculadas a ele, e é em relação a ele que os argumentos serão anotados, pois também a ele estarão vinculados (ainda que haja casos de papéis semânticos que serão anotados sem relação direta com o verbo, como é o caso dos papéis usados para adjuntos).

A apresentação do conceito de *parsers* neste capítulo se deu mais pelo fato de ele ser um elemento implícito em nosso estudo. Nossos *corpora* foram anotados com um *parser* e, portanto, cremos ser importante mostrar um pouco do seu funcionamento em geral, assim como mostrar como as sentenças estão representadas antes da extração dos dados. O mesmo podemos dizer das estruturas de subcategorização, tendo em vista que elas aparecem como elementos subentendidos neste estudo. Elas são o segundo elemento de organização de nosso recurso. Assim como temos os verbos como lema, as estruturas de subcategorização funcionarão como os indicadores de significado em um dicionário. Cada sentença do *corpus* que está vinculada a um verbo estará também vinculada a uma estrutura de subcategorização, que funciona como um segundo nível de organização de nosso recurso.

Antes da apresentação do conceito de estrutura de subcategorização, porém, discutimos um conceito central para este estudo, que foi o conceito de papéis semânticos. Esta tese visa ao desenvolvimento de um recurso léxico anotado com papéis semânticos, de modo que não poderíamos negligenciar esse ponto central. Os papéis semânticos serão compreendidos como uma função semântica dos sintagmas em uma oração. Eles ainda não dizem tudo o que o sintagma representa semanticamente, mas fornecem um indício desse significado que é mais completo do que a informação oferecida pela sintaxe. Existem muitas discussões teóricas acerca dos papéis semânticos, e muitos debates já foram travados ao longo dos anos, podendo o tema ser resgatado até períodos do mundo antigo, porém, daquilo que sabemos, apenas recentemente começamos a ter realizações concretas de dados estruturados com anotação de papéis semânticos. Assim, cremos que chegou um momento de sairmos um pouco do debate apenas no plano teórico e passar a debater os elementos concretos que se apresentam na linguagem. Aqui ainda não discutimos nossa lista de papéis, algo que deixaremos para comentar ao longo da tese, conforme formos expondo as diferentes perspectivas que assumimos. Por mais que tenhamos optado por uma lista definitiva para este estudo, essa lista passou por várias modificações ao longo desta tese, e essas modificações serão discutidas principalmente nos Capítulos 6 e 8. Desse modo, será possível ver como a lista foi sendo modificada até chegar ao seu estágio atual, mostrando como o trabalho prático é importante para desenvolvermos nossos questionamentos acerca dos papéis semânticos.

Por fim, debatemos rapidamente a questão de possíveis distinções entre argumentos e adjuntos. Após refletirmos sobre o assunto, à luz de teorias existentes e

dos debates que permeiam o assunto, optamos por simplificar as coisas, deixando a distinção entre argumento e adjunto para o plano da anotação de papéis semânticos, pois com eles, a distinção se torna um pouco mais clara, ainda que seja impossível haver uma distinção no plano sintático (tendo em vista que adjuntos adverbiais, no plano sintático, podem ser argumentos ou adjuntos no plano semântico).

Após esta breve retomada dos elementos expostos neste capítulo, partimos agora para um breve histórico das teorias de papéis semânticos.

3 Papéis Semânticos

Agora que já tratamos um pouco de alguns trabalhos existentes sobre papéis semânticos, dedicamos este capítulo às discussões teóricas e a um pouco da história dos papéis semânticos. Alguns dos problemas mais específicos já foram abordados na Seção 2.5, de modo que aqui trataremos mais especificamente de listas de papéis e dos diferentes pontos de vista teóricos empregados.

Quanto à história dos papéis semânticos, realmente não parece haver muito o que se contar no que diz respeito ao período anterior ao século XX. Conforme já mencionamos anteriormente, vários autores citam que os estudos de papéis semânticos remontam a milhares de anos, com a gramática de Panini, desenvolvida para dar conta do sânscrito. Essa gramática, assim como muitas outras posteriores, usava a perspectiva semântica para descrever a língua. Como o sânscrito era uma língua de casos morfológicos, uma descrição com base na semântica servia para fazer-se entender o que cada componente da oração representava e o porquê de sua declinação. Na sequência, certamente houve muitos outros casos de gramáticas que relatavam fatos linguísticos a partir de uma perspectiva semântica, tendo em vista que muitas das línguas antigas, assim como muitas modernas, usa(va)m casos morfológicos e, portanto, teriam um esclarecimento facilitado do ponto de vista semântico. Ainda assim, os autores recentes, pelo menos dentre os que lemos, não mencionam outros gramáticos antigos que tenham usado esse ponto de vista (exceto talvez por Fillmore [1967], que faz uma breve consideração histórica sobre o estudo dos casos). Mas também não é nossa intenção fazer uma exposição histórica que abrange desde o (Proto-)Indo-Europeu até os dias atuais.

Por isso, vamos dar um salto de muitos anos e chegar em Gruber (1965), que é um linguista interessado em descrever a língua do ponto de vista gerativo. Gruber tinha interesse em descrever uma forma de encontrar uma relação entre sintaxe e semântica que satisfizesse os princípios gerativos. Para tal, ele aponta uma série de relações que, posteriormente, seriam reconhecidas como papéis temáticos, papéis theta, estruturas de casos, casos profundos, papéis semânticos ou relações temáticas, entre outras denominações. As diferentes denominações tomam por base diferentes pontos de vista teóricos, porém, não chegam a ser uma distinção do fenômeno em si. Por exemplo, as denominações *papel temático* e *relações temáticas* têm a ver com a centralidade do papel TEMA para a organização dos demais papéis, e está vinculada às teorias de Gruber

e Jackendoff, sobre as quais falaremos neste capítulo. O nome *papel theta* vem diretamente da proposta gerativa de criação de uma descrição semântica vinculada à sintaxe. Os termos *casos profundos* e *estrutura de casos* estão relacionados à proposta de Fillmore, que estabelece uma certa relação entre casos morfológicos e estrutura profunda. Por fim, papéis semânticos fazem referência a uma abordagem mais funcional, em que os papéis são vistos como representativos da função semântica dos argumentos. Independente do nome que se escolha, o fenômeno em si é mais ou menos o mesmo, apenas visto de pontos de vista diferentes.

Retomando nosso relato, Gruber (1965) foi um dos primeiros linguistas modernos a usar papéis semânticos para descrever a linguagem. Em sua proposta, ele formula, por exemplo, a possibilidade de transformação numa estrutura prelexical entre os verbos *comprar* e *vender* em inglês (*buy* e *sell*). É ele também que propõe os padrões de fonte e destino, usando os padrões preposicionais *de-para* (*from-to*), que se aplicam a uma série de verbos, incluindo *comprar* e *vender*. Não entraremos aqui no mérito da proposta de Gruber (1965), tendo em vista que não nos interessam as possíveis relações transformacionais dos papéis semânticos, mas sim as relações superficiais que eles mantêm com os verbos. Aqui basta marcar que esse autor é reconhecido como um dos pioneiros da linguística moderna a usar papéis semânticos para esclarecer fenômenos linguísticos. Também é importante ressaltar que Gruber foi orientador do segundo trabalho, talvez ainda mais reconhecido, nesse campo: o trabalho de Fillmore (1967), que advogou o uso de estruturas de casos na descrição linguística.

Fillmore, em seu estudo *The case for case* (1967), lança uma primeira lista de papéis semânticos (aos quais chamava de casos). E talvez com isso tenha dado início a uma discussão, que não será encerrada em 2015, a respeito do tamanho adequado de uma lista de papéis semânticos, algo sobre o qual já discutimos, dentro do possível, no Capítulo 2. Mas Fillmore não estava preocupado em debater tamanho de listas, afinal, a sua era a primeira, e incluia os casos AGENTIVO, INSTRUMENTAL, DATIVO, FACTITIVO, LOCATIVO e OBJETIVO (FILLMORE, 1967, p. 46-47). Como se pode ver, essa lista deriva fortemente das estruturas de casos morfológicos, mas não era intenção de Fillmore tratar de estruturas de superfície. Assim como seu predecessor, Gruber, Fillmore se interessava pela estrutura profunda do caso. E, por isso, ele traz evidências também de outras línguas, chegando a mencionar até mesmo o trabalho do gramático bizantino Maxime Planude (FILLMORE, 1967, p. 18), que discutia a vinculação dos casos dativo, acusativo e genitivo aos diferentes tipos de movimento (respectivamente:

parado, movendo-se para algum lugar, movendo-se de algum lugar). Portanto, Fillmore apresenta sua tese para defender que os casos estão sim presentes na estrutura profunda.

Em sua proposta, Fillmore explicita que "uma sentença, em sua estrutura básica, é formada por um verbo e um ou mais sintagmas nominais, cada um associado ao verbo por meio de uma relação específica de caso" (1967, p. 41). Fillmore diz ainda que "cada relação de caso pode ocorrer apenas uma vez numa sentença simples" (1967, p. 41), ainda que possa ocorrer mais vezes em sentenças complexas. E esse é o primeiro ponto a que queremos chamar atenção, pois temos um certo desacordo com as ideias de Fillmore, principalmente por não assumirmos uma ideia de estrutura profunda. Se tomarmos o Exemplo 3.a, não temos como decidir, por critérios puramente semânticos, quem seria o único AGENTE da ação *dançar*, pois tanto *João* quanto *Maria* realizam a ação em conjunto. É claro que há um foco maior da sentença no sujeito, já que ele foi selecionado como tópico da oração, mas nada impede que os dois NPs troquem de posição. Porém, novamente, isso é uma questão de superfície, pois a estrutura profunda de Fillmore permite a estrutura "João e Maria dançam (juntos)". Como veremos em nossa lista de papéis semânticos, a qual será apresentada no Capítulo 8, temos papéis diferentes para classificar *João* e *Maria* em 3.a, mas é uma distinção apenas com base na sintaxe e no foco da oração, mas não na semântica.

3.a. João dança com Maria.

Assim, Fillmore (1967) desenvolve um trabalho intenso para vincular a semântica à sintaxe por meio dos chamados casos profundos. Em seguida, ainda perseguindo esse mesmo objetivo de vincular sintaxe à semântica, Jackendoff[18] retoma as ideias de Gruber e as adapta ao seu ponto de vista, chamando esses elementos semânticos de *relações temáticas*. Jackendoff (1990, p. 24) propõe que "a estrutura conceptual de um item lexical é uma entidade com zero ou mais espaços abertos para argumentos". O próprio autor reconhece a dificuldade de se implementar esse tipo de definição, tendo em vista que há muitas entidades em que se pode discutir quantos "espaços abertos" existem, mas a mantém como um guia ideal para o reconhecimento da estrutura conceptual. Como foi mencionado, Jackendoff retoma as ideias de Gruber e

[18] Infelizmente, não tivemos acesso às primeiras obras de Jackendoff, como o livro *Semantic Interpretation in Generative Grammar* (1972), e nossa visão de sua teoria provém de um estudo bastante posterior, chamado *Semantic Structures* (JACKENDOFF, 1990).

as elabora, defendendo que as representações gramaticais são diferentes das relações existentes no mundo (JACKENDOFF, 1990, p. 25-26). Assim, por exemplo, Jackendoff (1976, apud Perini (2008, p. 188-189)) aponta que em "*John stayed angry*", *angry* pode ser entendido como LUGAR (abstrato).

Esse tipo de análise, que desloca o foco dos papéis semânticos (ou, no caso, relações temáticas) para algo entendido como a língua, sem referência à realidade, parece ser uma forma artificial de tentar separar a sintaxe e a semântica da linguagem, tratando-as como níveis independentes das relações lexicais. Ainda que essa abordagem tenha seus méritos, por dar um tratamento teórico e formal bastante apurado nos exemplos do autor, nos parece que ignorar a influência do léxico e da pragmática sobre a semântica não é o melhor caminho para uma descrição da linguagem. Por vezes, é interessante tentar categorizar cada elemento da linguagem em um nível separado, apresentando morfologia, sintaxe, semântica e pragmática. Mas não se pode perder de vista que todos estão juntos na realização da linguagem. A semântica, por ser o fator que, juntamente com a pragmática, une os elementos de um texto para que este faça sentido e possa ser compreendido, não deveria ser interpretada de maneira isolada dos demais elementos, que certamente a influenciam. Assim, não concordamos com Jackendoff no que diz respeito a essa separação de semântica e realidade nos termos propostos. Acreditamos sim, que a semântica de uma sentença e, portanto, a determinação de papéis semânticos, deve ser entendida a partir da relação entre os elementos linguísticos presentes no próprio contexto, de acordo com a função semântica de cada elemento.

Seguindo adiante, temos uma proposta que visa a reduzir o grande problema da proliferação de papéis semânticos. Dowty (1991) propõe que, em vez de usarmos uma série de papéis semânticos para analisarmos as relações conceptuais, seria mais interessante dividirmos as posibilidades entre duas categorias distintas. Assim, o autor propõe a criação de dois protopapéis: o papel de PROTOAGENTE e o papel de PROTOPACIENTE. O PROTOAGENTE carrega consigo prototipicamente as noções causativas, volitivas, sencientes, de movimento e, possivelmente, de existência independente do evento em uma sentença (DOWTY, 1991, p. 572). Já o PROTOPACIENTE carrega os traços de mudança de estado, de tema incremental, de ser afetado por outro participante, de ser estacionário e, possivelmente, de não existir independentemente do evento (DOWTY, 1991, p. 572). Assim, Dowty defende que a

observação da semântica deve se dar em termos de prototipicidade, e não de uma especificação em várias classes diferentes.

A ideia de Dowty (1991) é interessante porque resolve vários problemas, como os de granularidade (consulte Seção 2.5) e de definição de fronteiras entre os papéis semânticos. O problema que essa proposta cria é o de vagueza. A existência de apenas duas categorias não é muito explicativa por si só. Teríamos muitos elementos enquadrados em cada uma delas, e o ganho para a descrição semântica das sentenças seria reduzido. Por mais que haja sustentação teórica para uma classificação simples com a proposta por Dowty, se quisermos uma descrição semântica mais ampla, que realmente reflita a semântica de uma sentença, seria necessário agrupar mais descritores que pudessem explicitar melhor as funções semânticas dos protopapéis nas sentenças, e isso acabaria retomando os problemas de definições e granularidade dos descritores.

Como temos visto até aqui, as discussões acerca dos papéis semânticos (independente do nome que se atribua a eles) têm girado apenas num plano teórico, tentando satisfazer uma teoria de vinculação entre a sintaxe e a semântica, mas nada de muito prático foi desenvolvido por esses autores, que se contentam em mostrar alguns exemplos para fundamentar suas explicações. Por outro lado, os autores que de fato se envolveram em desenvolvimentos práticos, como é o caso de Fillmore (um dos autores da FrameNet, sobre a qual discutimos no capítulo a seguir), mudaram de perspectiva. Isso não quer dizer que nada do que esses autores fizeram teve utilidade; pelo contrário, várias das propostas desses autores foram incorporadas, de uma forma ou de outra, em trabalhos que tiveram um cunho prático, como é o caso, por exemplo, da decomposição de predicados, teorizada por Gruber (1965) e desenvolvida por Jackendoff (1990), que foi utilizada pela VerbNet (KIPPER-SCHULER, 2005). Porém, quando temos um trabalho prático que requer a cobertura de vários verbos de uma língua, as decisões a serem tomadas podem oferecer problemas que as teorias ainda não haviam levado em conta. Esse foi um dos motivos que nos levou a tomar como base para o nosso estudo trabalhos que tiveram uma aplicação prática e não trabalhos puramente teóricos.

Aqui chamamos atenção para o fato de que não discutimos, neste capítulo, nenhum trabalho que tenha feito anotação de papéis semânticos descritivos em textos especializados. Nesse aspecto, temos conhecimento dos trabalhos que tomam a FrameNet como base, mas, nesse caso, os papéis semânticos empregados variam conforme o *frame*, de modo que eles não são genéricos e, por isso, não poderiam ser usados para uma comparação entre gêneros textuais, que é um dos nossos objetivos.

Assim, encerramos por aqui nossa apresentação histórica dos papéis semânticos e, no capítulo a seguir, apresentamos os trabalhos acerca de papéis semânticos e verbos que têm mais similaridade e/ou servem de base para o estudo que aqui apresentamos. Esses trabalhos que apresentaremos no capítulo a seguir têm, em maior ou menor intensidade, sustento nos trabalhos pioneiros que discutimos neste capítulo.

4 Trabalhos relacionados

Agora que já vimos alguns dos principais conceitos que nos sustentam e apresentamos algumas ideias teóricas sobre papéis semânticos, passaremos a apresentar trabalhos que se relacionam diretamente à anotação de papéis semânticos e à organização de verbos em léxicos, algo estreitamente relacionado aos objetivos desta tese. Começaremos este capítulo com o trabalho de Levin (1993) e, em seguida, prosseguiremos com a VerbNet (KIPPER-SCHULER, 2005), o PropBank (PALMER, GILDEA e KINGSBURY, 2005) e a FrameNet (BAKER, FILLMORE e LOWE, 1998).

4.1 Classes de Verbos

O trabalho de Levin (1993), que agrupou verbos de língua inglesa em classes e subclasses, é importante não só para o inglês, mas para a Linguística como um todo, pois agrupou verbos semanticamente próximos a partir de suas estruturas sintáticas. Apesar de haver várias críticas ao trabalho desenvolvido[19], Levin (1993) foi pioneira na área, principalmente pela magnitude do trabalho, de modo que merece destaque e consideração em estudos que abordem sintaxe e semântica associada a verbos.

Levin (1993) observou que, quando os verbos admitem as mesmas (ou quase as mesmas) alternâncias sintáticas, eles podem ser agrupados em categorias semânticas. Por exemplo, a partir da observação dos verbos *break*, *cut*, *hit* e *touch* e das suas possibilidades de alternâncias mediais, conativas e que envolvem partes do corpo, é possível analisar as diferenças semânticas entre esses verbos.

Para explicar rapidamente o que são essas alternâncias sintáticas, também chamadas de diáteses, e para ilustrar as diferenças entre esses verbos em inglês, reproduzimos os exemplos apresentados por Levin (1993, p. 6-7)[20]:

4.a. Margaret cut the bread. (Margaret cortou o pão.)
4.b. Janet broke the vase. (Janet quebrou o vaso.)
4.c. Terry touched the cat. (Terry tocou o gato.)
4.d. Carla hit the door. (Carla golpeou a porta.)

[19] Para uma amostra das críticas feitas ao trabalho de Levin (1993), consulte Perini (2008). O estudo de Lima (2007) também mostra como verbos de um mesmo grupo semântico não necessariamente apresentam as mesmas estruturas sintáticas.

[20] As traduções que colocamos entre parênteses nesta seção são literais e servem apenas para ilustrar a sintaxe das sentenças em inglês. Muitas das traduções não apresentam uma sintaxe possível no português.

Nas sentenças 4.a a 4.d, temos as formas transitivas diretas dos quatro verbos em inglês. Essa forma foi considerada a forma básica desses verbos. Agora vejamos como ficam essas sentenças na alternância medial (que passa o objeto direto da forma básica para a posição de sujeito e usa o verbo em sua forma intransitiva):

4.e. The bread cuts easily. (O pão corta facilmente.)
4.f. Crystal vases break easily. (Vasos de cristal quebram facilmente.)
4.g. *Cats touch easily.[21] (Gatos tocam facilmente.)
4.h. *Door frames hit easily. (Marcos de porta golpeiam facilmente.)

Como podemos observar, nas sentenças de 4.e a 4.h, os verbos *touch* e *hit* não permitem a alternância medial. Vejamos agora como fica a alternância conativa, na qual o verbo passa a ser intransitivo, e o objeto direto da forma básica é introduzido por preposição:

4.i. Margaret cut at the bread. (Margaret corta no pão.)
4.j. *Janet broke at the vase. (Janet quebra no vaso.)
4.l. *Terry touched at the cat. (Terry toca no gato.)
4.m. Carla hit at the door. (Carla golpeou na porta.)

Nas sentenças de 4.i a 4.m, vemos que os verbos *break* e *touch* não admitem alternância conativa. Por fim, passemos à alternância que envolve partes do corpo:

4.n. (a) Margaret cut Bill's arm. (Margaret cortou o braço de Bill.)
(b) Margaret cut Bill on the arm. (Margaret cortou Bill no braço.)
4.o. (a) Janet broke Bill's finger. (Janet quebrou o dedo de Bill.)
(b) *Janet broke Bill on the finger. (Janet quebrou Bill no dedo.)
4.p. (a) Terry touched Bill's shoulder. (Terry tocou o ombro de Bill.)
(b) Terry touched Bill on the shoulder. (Terry tocou Bill no ombro.)
4.q. (a) Carla hit Bill's back. (Carla golpeou as costas de Bill.)
(b) Carla hit Bill on the back. (Carla golpeou Bill nas costas.)

[21] O sinal * indica agramaticalidade.

Pelo que vemos nas sentenças 4.n a 4.q, apenas o verbo *break* não autoriza a alternância que envolve partes do corpo. Os resultados deste exemplo estão sumarizados na Tabela 4.1.

Tabela 4.1 – Comportamento dos verbos *break*, *cut*, *hit* e *touch*.

	Break	Cut	Hit	Touch
Medial	X	X		
Conativa		X	X	
Parte do corpo		X	X	X

A partir das sentenças-exemplo apresentadas e da sumarização presente na Tabela 4.1, podemos observar que, apesar de os quatro verbos serem transitivos, eles não autorizam os mesmos tipos de alternâncias sintáticas e, por isso, pertencem a quatro classes diferentes de verbos. O verbo *break*, por exemplo, compartilha as mesmas alternâncias de verbos como *crack* (rachar), *rip* (rasgar) e *shatter* (despedaçar), já o verbo hit está na mesma classe de *kick* (chutar), *whack* (bater), *bash* (espancar), e assim por diante. Além de perceber essa diferença na sintaxe, Levin (1993) também apontou que esses verbos apresentam diferenças em seus traços semânticos: o verbo *cut* envolve movimento, contato e mudança de estado; o verbo *hit* envolve contato e movimento; o verbo *break* envolve apenas mudança de estado; e o verbo *touch* envolve apenas contato.

Com base nessas observações de alternâncias sintáticas e de traços semânticos, Levin organizou mais de quatro mil verbos do inglês em um total de 193 classes e subclasses. Ao apresentar as classes, Levin contribuiu em muito para os estudos sobre verbos do inglês, pois determinados fenômenos aplicáveis a um verbo geralmente se aplicam também a toda uma classe.

Para o português, ainda não foi publicado um trabalho com a mesma magnitude do de Levin (1993)[22], porém, Cançado, Godoy e Amaral (2012) já apresentaram um projeto que intenta levar a cabo essa empreitada. O primeiro volume desse trabalho,

[22] Scarton (2013) realizou o agrupamento de verbos em classes, porém, partindo das classes em inglês e usando métodos semiautomáticos. Foram também publicados trabalhos isolados para uma ou algumas classes de verbos, como o trabalho de Lima (2007), mas desconhecemos a existência de um trabalho para o português que tenha a abrangência do trabalho de Levin (1993).

compreendendo verbos de mudança, já foi publicado (CANÇADO, GODOY e AMARAL, 2013) e compreende 862 verbos do português brasileiro, subdivididos em 4 classes e organizados de acordo com a teoria da decomposição de predicados.

A partir da seção seguinte, passamos a tratar de trabalhos diretamente relacionados à anotação de papéis semânticos. Nesses trabalhos, serão apresentadas formas diferentes de realizar a anotação e diferentes concepções de papéis semânticos.

4.2 VerbNet

Partindo das classes de Levin (1993), Kipper-Schuler (2005) desenvolveu um recurso léxico que ficou conhecido como VerbNet. A VerbNet contém as classes de Levin (1993) associadas a papéis semânticos que podem aparecer junto a verbos de cada uma das classes. No estágio atual da VerbNet (versão 3.2), foram utilizados efetivamente 30 papéis semânticos, partindo-se de uma lista inicial com 36 papéis[23].

Por partir das classes de Levin, a anotação de apenas 191[24] classes (na versão 1.0) já dava cobertura para 4.173 verbos. Em sua versão atual, com o acréscimo de outras classes de verbos, extraídas automaticamente a partir de *corpora*, já existe anotação para cerca de 5.800 verbos, divididos em 272 classes.

Os papéis semânticos utilizados na VerbNet são descritivos, ou seja, eles apresentam um rótulo que mais ou menos descreve a função dos participantes na oração (por exemplo: AGENTE, PACIENTE, EXPERIENCIADOR etc.). Esse tipo de papel se distingue dos papéis semânticos numerados, que veremos mais adiante, ao apresentarmos o PropBank. Além dos papéis semânticos, a VerbNet também apresenta restrições semânticas, como, por exemplo, *+location*, *-region*, *+object* etc. Essas restrições ajudam a especificar ainda mais o tipo de participantes que podem estar em um evento[25].

Para cada classe de verbos, a VerbNet apresenta informações de predicados semânticos. Conforme aponta Scarton (2013, p. 65), "os predicados semânticos fornecem as relações entre participantes e eventos, e são responsáveis por definir os

---

[23] Para maiores informações sobre os papéis semânticos utilizados na VerbNet atual, consulte a documentação fornecida no próprio site da VerbNet: http://verbs.colorado.edu/~mpalmer/projects/verbnet/VerbNet3.0ReadMe.doc.

[24] Ainda que o trabalho de Levin conte com 193 classes e subclasses, duas delas não puderam ser usadas na VerbNet.

[25] Quando mencionamos **evento** e **participantes**, estamos nos referindo ao verbo e aos elementos linguísticos (palavras ou sintagmas) vinculados ao verbo em uma oração.

componentes de significado de cada classe". Além dessas informações de papéis semânticos, a VerbNet contém informações temporais, indicando o momento em que o predicado é verdadeiro. Por exemplo: a expressão *motion(during(E), Theme)* indica que, durante o evento, o TEMA está em movimento[26].

Em relação ao nosso estudo, a VerbNet é o trabalho que mais tem elementos em comum. O ponto mais similar entre os dois trabalhos são os papéis semânticos descritivos, pois partimos da lista de papéis semânticos da VerbNet para chegar à nossa lista, como veremos mais adiante. Além disso, a apresentação da sintaxe e da semântica em nosso estudo é muito parecida com a da VerbNet. Algumas diferenças estão no fato de que não incluímos restrições semânticas em nossa anotação, nem apresentamos uma função temporal nos predicados semânticos; porém, por trabalharmos com *corpora*, apresentamos mais possibilidades de papéis semânticos para os verbos, enquanto a VerbNet apresenta exemplos inventados que nem sempre cobrem muitas possibilidades de apresentação dos verbos.

Para o português, além do nosso trabalho, pudemos acompanhar boa parte do estudo do estudo de Scarton (2013), que se propôs a transpor as anotações do inglês para o português aproveitando-se das conexões que existem entre a VerbNet (KIPPER-SCHULER, 2005), a WordNet (FELLBAUM, 1998) e a WordNet.Br (DIAS-DA-SILVA, 2005; DIAS-DA-SILVA, FELIPPO e NUNES, 2008). Desse modo, para as classes sinônimas entre a WordNet e a WordNet.Br, os papéis foram importados diretamente do inglês para os verbos em português. Esse trabalho foi pioneiro na criação de um léxico com anotação semântica descritiva para o português e se propôs como um passo inicial nessa área. Desse modo, já existe uma VerbNet.Br, porém, ela foi construída de modo semiautomático, podendo conter ruído[27], e apresenta apenas aquelas classes que são sinônimas ou quase sinônimas entre o português e o inglês.

A principal diferença que se deve ressaltar em relação ao trabalho de Scarton (2013) e este estudo é o fato de que Scarton usou o inglês como base e importou semiautomaticamente os dados que apresentam sinonímia ou quase sinonímia entre as WordNets do inglês e do português. O trabalho aqui apresentado parte do português e se

[26] Para maiores informações sobre os predicados semânticos na VerbNet, assim como informações mais detalhadas sobre o recurso léxico como um todo, consulte Kipper-Schuler (2005) e Scarton (2013).
[27] Por *ruído*, entendem-se informações que estão erradas, principalmente devido ao método automático de extração de informação. No caso específico da VerbNet.Br, é possível consultar os anexos do trabalho de Scarton (2013) para observar com detalhes os tipos de ruído apresentados no comparativo com o *gold standard*. Também apresentamos alguns breves exemplos de ruído na Seção 8.5.2.

baseia em uma anotação manual dos dados por um linguista. Assim, apesar de nosso estudo ser menos abrangente, ele apresenta uma menor propensão a ruído. Na Seção 8.5.2, fazemos uma comparação entre os dois recursos, assim como descrevemos com mais detalhes a VerbNet.Br.

4.3 PropBank

Continuando com a anotação de papéis semânticos, além de um recurso mais dicionarístico como a VerbNet, que apresenta classes de verbos e seus possíveis papéis, existe também o PropBank (PALMER, GILDEA e KINGSBURY, 2005), que apresenta sentenças de um *corpus* anotadas com papéis numerados. Esses papéis semânticos se aproximam muito da ideia de Dowty (1991) sobre o uso de papéis semânticos prototípicos. Assim, em vez de indicar se um participante é um AGENTE, PACIENTE, TEMA ou EXPERIENCIADOR, o PropBank indica que ele é um ARG0 ou ARG1. Os argumentos numerados se estendem de ARG0 a ARG4, mas existem também papéis semânticos específicos para adjuntos adverbiais (por exemplo: ARGM-LOC para adjuntos adverbiais de lugar, ARGM-TMP para adjuntos adverbiais de tempo etc.).

Apesar de esse tipo de opção representar uma facilidade para o anotador, que não precisa fazer distinções entre AGENTES e EXPERIMENTADORES, PACIENTES e TEMAS, entre outras, o resultado diminui a informação que se pode adquirir a partir da anotação. Como apontam Zapiran, Agirre e Màrquez (2008), "a interpretação dos papéis do PropBank são dependentes do verbo". Por exemplo, na sentença *João joga bola*, o sujeito do verbo *jogar* não é anotado como AGENTE, mas sim como ARG0, devendo ser interpretado como o papel semântico JOGADOR. Uma das vantagens do PropBank é que, por apresentar vários exemplos de cada um dos verbos anotados (por ser um *corpus* anotado), ele pode ser usado para treinar *softwares* de anotação automática de papéis semânticos, algo que a VerbNet, por ter um número restrito de exemplos, não permite.

O projeto SemLink (LOPER, YI e PALMER, 2007; PALMER, 2009) foi responsável por realizar a vinculação dos papéis semânticos da VerbNet às sentenças do PropBank. Desse modo, hoje já existem no PropBank sentenças anotadas também com papéis semânticos descritivos (AGENTE, PACIENTE etc.).

Assim como no caso da VerbNet, também existe para o português um projeto que se encarregou de desenvolver o PropBank.Br. Esse projeto, desenvolvido por Duran

e Aluisio (2011; 2012) já se encontra disponível[28] e contém mais de 5 mil instâncias anotadas. Apresentaremos mais informações sobre o PropBank.Br na Seção 8.5.1 desta tese, quando comparamos a anotação do PropBank.Br com a anotação do nosso recurso.

4.4 FrameNet

Por fim, existe ainda outro tipo de anotação de papéis semânticos, bastante difundida, que toma como base os cenários comunicativos, chamados de *frames*[29]. É assim que se estrutura a FrameNet (BAKER, FILLMORE e LOWE, 1998), um projeto que tem por objetivo anotar os papéis semânticos de cada participante de uma sentença em relação ao seu domínio e ao seu contexto. Por exemplo, os papéis semânticos do *frame* DECISÃO (Copa do Mundo) podem ser VENCEDOR, PERDEDOR, TORNEIO e FINAL[30].

Essa abordagem pode, em princípio, parecer um PropBank com papéis descritivos, porém, a verdade é que ela se baseia em cenários comunicativos, de modo que os papéis semânticos podem ser usados por mais de um verbo, desde que esses verbos compartilhem o mesmo cenário. Assim, os verbos *vencer* e *ganhar* podem compartilhar, por exemplo, os papéis semânticos VENCEDOR e PERDEDOR, desde que estejam no mesmo cenário comunicativo.

No Brasil, temos a FrameNet Brasil[31] (Salomão, 2009) utiliza essa mesma abordagem. Também temos anotações de *frames* de algumas áreas específicas, como, por exemplo, o Kicktionary_Br (CHISHMAN, SPADER e PADILHA, 2013) e a anotação de textos jurídicos (BERTOLDI e CHISHMAN, 2012).

As diferenças entre a VerbNet, o PropBank e a FrameNet estão principalmente na granularidade dos papéis. Os papéis da FrameNet são altamente específicos, pois se aplicam apenas a um determinado cenário comunicativo. Os papéis da VerbNet são menos específicos, tentando apresentar uma descrição de semântica que pode ser aplicada a qualquer contexto. Já o PropBank apresenta a solução mais abstrata, pois

---

[28]Disponível no site (acessado em 24/12/2014): http://www.nilc.icmc.usp.br/portlex/index.php/en/projects/propbankbringl.

[29] A palavra *frame* é bastante polissêmica. Neste artigo, tratamos de *subcategorization frames* (estruturas de subcategorização), como vimos anteriormente, e também de *frames* como os da FrameNet, que são compreendidos como domínios semânticos ou estruturas conceptuais (por exemplo, o *frame* DIRIGIR ou o *frame* JOGO DE FUTEBOL). Tentaremos deixar claro pelo contexto qual é o tipo de *frame* a que nos referimos.

[30]Exemplo retirado do site http://200.131.61.179/maestro/index.php/fnbr/report/frames?db=fncopa, da FrameNet Brasil (SALOMÃO, 2009).

[31] http://www.ufjf.br/framenetbr/.

apenas cinco papéis (ARG0 a ARG4) se aplicam a qualquer contexto, configurando-se como protopapéis.

5 Materiais

Após termos visto os trabalhos que estão relacionados a este estudo e as bases teóricas que nos guiam, neste capítulo, apresentamos alguns dos materiais utilizados. Como este estudo envolveu vários experimentos (anotação com múltiplos anotadores, agrupamento de verbos e dois estudos-piloto), sobre os quais discutiremos ao longo dos próximos capítulos, achamos por bem não mostrarmos aqui todos os materiais utilizados, tendo em vista que o capítulo poderia ficar muito confuso. Sendo assim, optamos por apresentar algumas informações diretamente nos capítulos referentes ao diferentes experimentos.

Dentre os materiais que foram utilizados neste estudo, selecionamos para apresentar neste capítulo os seguintes materiais:

- *corpora* que serviram de base para a anotação;
- ferramenta de extração de estruturas de subcategorização; e
- interface de anotação do recurso léxico.

Enquanto os *corpora* e a interface de anotação permaneceram inalterados ao longo do estudo, a ferramenta de extração sofreu algumas modificações, de modo que aqui apresentaremos as suas características básicas, mas reservamos para outros momentos uma descrição de determinadas configurações, conforme for surgindo a necessidade.

5.1 Corpora

Este trabalho realiza um contraste entre estruturas em textos especializados e não especializados; por isso, foram utilizados dois *corpora*, cada um representando uma esfera da linguagem. Para representar os textos especializados, selecionamos um *corpus* composto por artigos científicos da área da Cardiologia compilado por Zilio (2009; 2012). Para representar os textos não especializados, selecionamos o *corpus* de textos do jornal popular Diário Gaúcho, compilado no âmbito do projeto PorPopular[32]. Na Tabela 5.1, podemos ver a constituição dos *corpora* em relação ao número de palavras.

[32] Para maiores informações sobre o projeto e o *corpus*, acesse: http://www.ufrgs.br/textecc/porlexbras/porpopular/index.php. Os números atuais apresentados no site diferem dos números apresentados nesta tese porque nosso *corpus*, por ter sido compilado há algum tempo, não compreende a totalidade dos textos presente no *corpus* atual.

Tabela 5.1 – Tamanho dos *corpora*

Corpus	Nº de palavras[33]
Cardiologia	1.605.250
Diário Gaúcho	1.049.487

O *corpus* do Diário Gaúcho é composto por textos jornalísticos completos retirados da versão impressa do jornal ao longo do ano de 2008. Nele se encontram diversos subgêneros do texto jornalístico, e um dos elementos de destaque desse *corpus* é a sua orientação para indivíduos de menor poder aquisitivo e com pouco hábito de leitura, conforme explicam Finatto *et al.* (2011). Esse gênero de jornalismo popular tende ao uso de uma linguagem mais cotidiana, sem procurar ser rebuscado, erudito ou especializado demais, pois seu objetivo é passar informações claras a um público que pode não ter um hábito de leitura suficiente para acompanhar um texto mais técnico ou científico. Essa orientação específica e sua tendência, em teoria, a uma simplificação da linguagem é o que nos levou a eleger esse *corpus* como representante da linguagem comum.

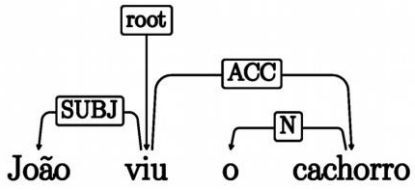
O *corpus* de Cardiologia é composto por 493 artigos científicos retirados de três periódicos brasileiros da área: os Arquivos da Sociedade Brasileira de Cardiologia (2005-2007), a Revista da Sociedade de Cardiologia do Estado de São Paulo (2005-2007) e a Revista da Sociedade de Cardiologia do Estado do Rio de Janeiro (2005-2007). Os artigos em questão são todos artigos originais, sem publicação prévia em outros meios de divulgação, e não estão entre eles outros tipos de artigos, como estudos de caso ou artigos de revisão.

Ambos os *corpora* foram analisados automaticamente pelo *parser* PALAVRAS (BICK, 2000) com árvores de dependências sintáticas[34]. Nessa anotação de dependências, o *corpus* anotado apresenta uma hierarquia de ligações entre os elementos sintáticos das sentenças. Isso pode ser visto no Exemplo 5.a, analisado com o *parser* PALAVRAS:

[33] Os números de palavras foram observados com a ferramenta WordSmith Tools, versão 4.0 (SCOTT, 2007).
[34] Para maiores informações sobre o *parser*, consulte a Seção 2.4.

5.a. João viu o cachorro.

João [João] @SUBJ> ***#1->2***
viu [ver] @FS-STA ***#2->0***
o [o] @>N ***#3->4***
cachorro [cachorro] @<ACC ***#4->2***
$. ***#5->0***
</s>



Na anotação do Exemplo 5.a, se observarmos os valores em negrito, após a cerquilha (#), é possível ver quais elementos estão ligados diretamente ao verbo e, com isso, definir os seus argumentos. O número antes do sinal "->" é o número da palavra, enquanto o número após o sinal "->" é o número da outra palavra à qual esta se liga. Assim, vemos que as palavras ***João*** e ***cachorro*** estão ligadas ao verbo ***viu***, e este está ligado a 0, que é a raiz. Com isso, cria-se uma árvore de dependências que tem um verbo ligado à raiz e os demais elementos ligados a ele. Além disso, após a arroba (@), está identificada a categoria sintática à qual pertence cada palavra da sentença. Essa estrutura é utilizada pelo extrator de estruturas de subcategorização (que será apresentado a seguir) para reconhecer automaticamente os argumentos dos verbos e suas categorias sintáticas, e os organizar em um banco de dados.

5.2 Extrator de estruturas de subcategorização

O extrator de estruturas de subcategorização (ZANETTE, 2010; ZANETTE, SCARTON e ZILIO, 2012; ZILIO, ZANETTE e SCARTON, 2012; 2014) é um *software* que, neste estudo, foi usado para realizar a preparação dos dados para a anotação. Como vimos no Capítulo 2, as estruturas de subcategorização podem ser compreendidas como uma forma simplificada de organização sintática. Essas estruturas são utilizadas pelo extrator de estruturas de subcategorização para organizar conjuntos de sentenças em uma mesma categoria, de acordo com sua base sintática. O funcionamento do extrator é razoavelmente simples. O sistema é dividido em quatro módulos (Leitor, Extrator, Construtor e Filtro) que apresentaremos individualmente a seguir.

Leitor. O módulo de leitura realiza exatamente o que o nome sugere: ele lê e reconhece cada uma das sentenças de um *corpus*, e a entrega para o módulo extrator. Este módulo é uma decisão de arquitetura do sistema que permite o uso de diferentes tipos de entrada (XML, texto, bancos de dados etc.).

Extrator. Para cada verbo finito reconhecido em cada uma das sentenças, o módulo Extrator extrai as dependências (ou seja, os elementos ligados ao verbo de acordo com a anotação do *parser*) e tenta classificá-las de acordo com o tipo de argumento, que pode ser:

- NP – sintagma nominal;
- PP[prep.] – sintagma preposicionado (a preposição que introduz o sintagma é apresentada entre colchetes);
- ADJP – sintagma adjetival.

Na verdade, esses são apenas alguns dos tipos básicos reconhecidos pelo sistema. Dependendo da versão do sistema, existem diferentes tipos de argumentos que foram sendo acrescentados ou subtraídos, conforme avançamos no estudo[35].

Ressaltamos que, apesar de usar os verbos finitos como base para a extração, este módulo reconhece se o verbo finito é auxiliar ou modal e busca automaticamente o verbo principal da oração, o qual é considerado como o verbo da oração e é passado para o próximo módulo. Além disso, o sujeito é considerado um argumento obrigatório pelo Extrator: se não houver um sujeito presente, o módulo assume um sujeito oculto. Essa decisão foi tomada para garantir que não houvesse estruturas de subcategorização diferentes para um mesmo verbo apenas devido à presença ou ausência de sujeito na oração.

Além de atribuir uma classificação para a estrutura de subcategorização, o módulo Extrator também reconhece a classificação sintática (sujeito, objeto direto, objeto indireto, adjunto adverbial etc.) de cada sintagma, com base nas informações do *parser*. Essa classificação sintática é utilizada para atribuir um valor de relevância para cada sintagma (por exemplo: 1 para sujeito, 2 para objeto direto, 3 para objeto indireto etc.), o qual poderá ser utilizado pelo módulo Construtor, como veremos a seguir. Por fim, com base nas informações sobre os verbos presentes na sentença, o módulo Extrator identifica se a oração está na voz ativa ou passiva, uma informação que

[35] Conforme formos apresentando os experimentos realizados nos próximos capítulos, as classificações possíveis serão explicitadas, mostrando como o extrator foi sendo modificado ao longo do estudo.

posteriormente é utilizada para distinguir entre estruturas de subcategorização que seriam iguais, exceto pelo tipo de voz.

Construtor. Após receber as informações do módulo Extrator, o Construtor monta a estrutura de subcategorização e organiza os argumentos em um banco de dados. A montagem da estrutura de subcategorização pode seguir dois parâmetros: relevância ou ordem. O parâmetro pode ser escolhido pelo usuário, de acordo com sua necessidade. O parâmetro relevância faz com que o Construtor monte a estrutura de acordo com o valor de relevância atribuído pelo módulo Extrator, ou seja, o sujeito sempre será o primeiro elemento da estrutura de subcategorização. O parâmetro ordem faz com que o Construtor organize os argumentos de acordo com a ordem em que aparecem na oração; desse modo, o sujeito não necessariamente será o primeiro elemento. Além desses dois parâmetros, existe também um parâmetro de quantidade, que limita o número de argumentos possíveis por estrutura de subcategorização. Essa limitação leva em consideração o fato de que os verbos têm um limite de argumentos possíveis, de modo que uma estrutura de subcategorização com, por exemplo, oito sintagmas, provavelmente está errada. Esse parâmetro pode ser modificado conforme for necessário, porém, em nosso estudo, utilizamos sempre um limite de cinco argumentos por estrutura de subcategorização[36].

Após os argumentos terem sido organizados em estruturas de subcategorização (seguindo um dos dois parâmetros disponíveis), as estruturas são armazenadas em um banco de dados. O banco de dados apresenta informações de frequência dos verbos principais extraídos, das estruturas de subcategorização vinculadas a cada verbo, dos argumentos (incluindo sua classificação sintática) e das sentenças que apresentam as estruturas de subcategorização em questão.

Filtro. Como todos os passos anteriores são automáticos, existe a possibilidade de haver ruído nos dados extraídos. Por isso, o módulo Filtro permite que se façam filtragens dos dados, de acordo com critérios de frequência. Em nossos experimentos, utilizamos um critério bem simples, apenas para limitar um pouco o tamanho do banco de dados. O critério foi a exclusão de verbos com frequência igual a 1, desse modo, essa

[36] A opção por ter um limite de cinco argumentos foi baseada na existência de verbos em português com 4 argumentos, como *comprar* e *vender*. Considerando a possibilidade de 4 argumentos, deixamos espaço para 4 argumentos mais 1 adjunto, ou para a eventual ocorrência de um verbo com 5 argumentos, algo ainda não documentado no português do Brasil.

filtragem não representou perda para a anotação, como poderá ser visto nas explicações sobre a metodologia de anotação.

5.2.1 Comentário sobre o extrator

Dos quatro módulos apresentados, o módulo Extrator talvez seja o mais importante de todos. Ele contém um conjunto de regras de extração, as quais são aplicadas às frases do *corpus* analisadas pelo *parser* PALAVRAS com árvores de dependências sintáticas. Durante a extração, com base nas informações fornecidas pelo *parser*, o sistema faz a identificação de quais verbos são auxiliares e quais são principais. Estes são utilizados, enquanto aqueles são excluídos e utilizados apenas para que possa ser reconhecido o sujeito da oração.

5.b. O cachorro foi visto por João.

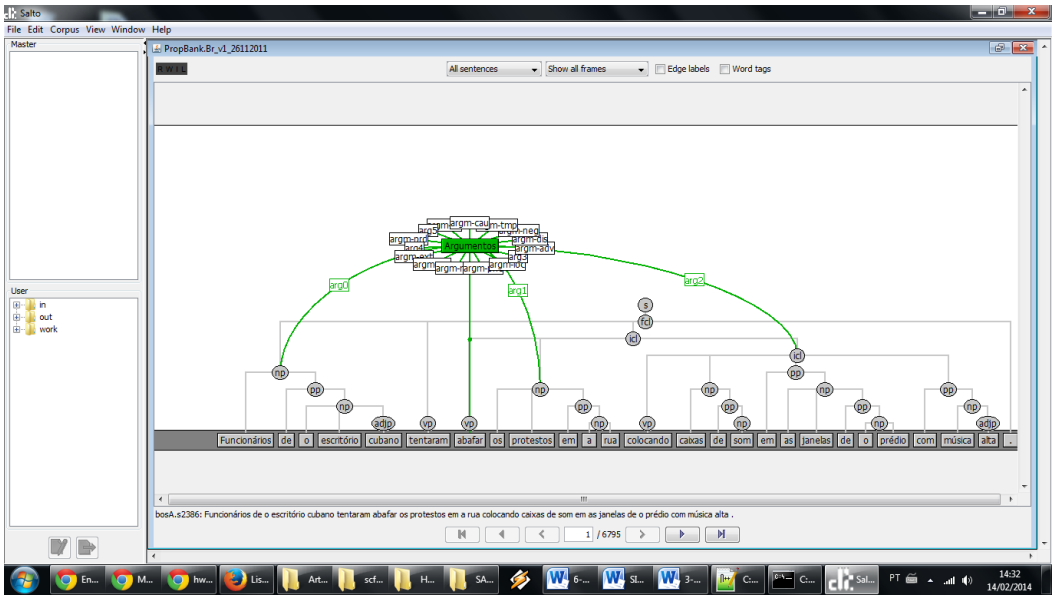
No Exemplo 5.b, o extrator reconhece ***ver*** como verbo principal. O sujeito ***o cachorro*** está ligado ao verbo auxiliar ***ser***, mas o extrator consegue recuperar essa informação e associa o sujeito ao verbo ***ver***. Desse modo, são mantidas apenas informações referentes a verbos principais.

Como mencionamos, todas as informações extraídas são identificadas por meio de regras. Assim, o extrator busca informações como, por exemplo, **@<ACC**, fornecidas pelo *parser*, as extrai e também as traduz em etiquetas mais explícitas para o anotador humano, como **OBJETO DIRETO**. Esse é um dos pontos críticos do sistema, pois, por ser baseado em regras, se as regras não forem boas, o sistema também não será bom. Como veremos ao longo deste trabalho, as regras de extração foram modificadas com o passar do tempo, de acordo com os testes realizados e seus resultados.

Apesar de termos utilizado o sistema desenvolvido inicialmente por Zanette (2010), existem outras ferramentas que poderiam ser empregadas para a anotação, como, por exemplo, a ferramenta SALTO (BURCHARDT, ERK, *et al.*, 2006). Entretanto, o sistema de anotação da ferramenta SALTO é mais complexo, deixando ao encargo do anotador a tarefa de delimitar os argumentos, como podemos ver na Figura 5.1. Por um lado, isso pode assegurar maior precisão na delimitação dos argumentos; por outro lado, aumenta a chance de erros e aumenta o trabalho dispendido na anotação. Além disso, por termos contato direto com Adriano Zanette, o desenvolvedor do

sistema que utilizamos, podíamos solicitar modificações e aprender a manusear a ferramenta com mais facilidade.

Figura 5.1 – Interface da ferramenta SALTO com exemplo retirado do PropBank.Br[37]



5.3 Interface de anotação

Os dados extraídos pelo módulo Extrator e montados pelo módulo Construtor acabam armazenados em um banco de dados em formato MySQL. Esse formato é bastante otimizado para consulta dos dados, porém, para a anotação, um banco de dados em linguagem MySQL é muito pouco intuitivo. Assim, para facilitar a anotação, criou-se uma interface de usuário que permite a visualização dos dados extraídos, com a classificação dos argumentos, de uma forma mais amigável para o anotador. Para acessar a interface, é preciso apenas ter uma ferramenta que permite a manipulação de dados em formato MySQL. Neste estudo, utilizamos para esse fim a ferramenta WampServer[38].

Como podemos ver na Figura 5.2, a interface de anotação (criada em linguagem PHP) mostra a estrutura de subcategorização (chamada de *frame*), o verbo em questão, os exemplos e os argumentos extraídos com a sua respectiva categoria sintática (com

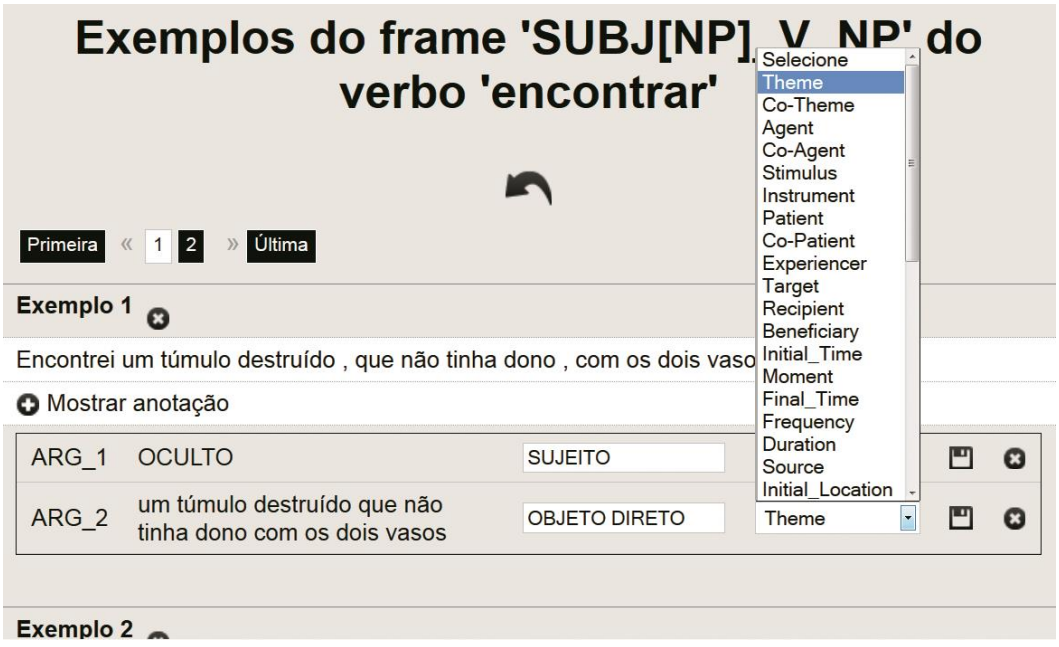
[37] Disponível para *download* em http://www.nilc.icmc.usp.br/portlex/index.php/en/downloadsingl. As linhas verdes, que indicam os argumentos, podem ser ligadas a qualquer nódulo da sentença (isto é, a qualquer círculo com etiqueta anotada pelo *parser* PALAVRAS).

[38] Disponível para *download* em http://www.wampserver.com/en/.

base na informação do *parser*). Ao anotador de papéis semânticos cabe o trabalho de criar uma lista de papéis semânticos, digitar os papéis em um arquivo de texto usando vírgula como separador e selecioná-los a partir da lista de rolagem (que pode ser vista na Figura 5.2) no momento da anotação. Com essa interface, o anotador pode se concentrar no que lhe interessa: definir a semântica dos argumentos, sem precisar delimitá-los ou procurá-los em um banco de dados. É importante ressaltar que o banco de dados está estruturado de modo a permitir apenas a seleção de um papel semântico por argumento. Sendo assim, para teorias que admitem mais de um papel (por exemplo, Gelhausen [2010]), seria necessário modificar a arquitetura do sistema.

Por fim, apesar de o extrator já deixar os dados prontos para o anotador trabalhar, a análise automática de dependências sintáticas realizada pelo *parser* PALAVRAS nem sempre é correta. Existem ruídos na análise que vão desde a simples segmentação de sentenças até a delimitação dos argumentos. Além dos possíveis ruídos decorrentes da análise automática, o extrator de estruturas de subcategorização também organiza os dados de acordo com regras, e estas nem sempre estão corretas, como já comentamos. Desse modo, existem dados que podem conter ruído no banco de dados. Como veremos na descrição da metodologia de cada um dos experimentos de anotação, grande parte desses dados ruidosos são ignorados e não são anotados.

Figura 5.2 – Amostra da interface de usuário para anotação



6 Estudos-Piloto

Agora que já apresentamos as bases teóricas e os trabalhos relacionados, além de mostrarmos os materiais básicos que usamos ao longo de todo o estudo, este capítulo é dedicado aos primeiros passos concretos em direção ao nosso objetivo principal. Nele, apresentamos os dois estudos-piloto que serviram de base para o recurso léxico que objetivamos nesta tese.

Desse modo, este capítulo será dividido em duas grandes seções: uma descrevendo brevemente o primeiro estudo-piloto (nosso primeiro teste de anotação), e outra que descreve o segundo estudo-piloto (já mais estruturado, com outro conjunto de papéis semânticos, e que serviu como base para a anotação dos dados do recurso atual).

Este capítulo se apresenta mais como parte do relato do trabalho desenvolvido durante a elaboração desta tese, de maneira que, por exemplo, os dados do estudo-piloto I são bastante negativos e serviram para refutar uma metodologia, a qual foi modificada para o estudo-piloto II e, posteriormente, veio a ser utilizada na anotação do VerbLexPor.

6.1 Estudo-piloto I

Este estudo-piloto serviu como uma primeira aproximação à anotação dos dados na prática. Nossos objetivos com ele foram os seguintes:

- testar, em um pequeno conjunto de dados, uma primeira lista de papéis semânticos (que será apresentada mais adiante), para verificar a possibilidade de utilizá-la em um conjunto maior de dados;
- observar o desempenho da ferramenta de extração de estruturas de subcategorização apresentada no Capítulo 5;
- fazer uma primeira observação da configuração das estruturas de argumentos.

Depois de apresentarmos os papéis semânticos utilizados neste estudo-piloto I, na Subseção 6.1.2, fazemos algumas considerações rápidas sobre alguns dos papéis semânticos. A Subseção 6.1.3 descreve o estado em que se encontrava a ferramenta de extração. Na Subseção 6.1.4, descrevemos a metodologia. Por fim, a Subseção 6.1.5 apresenta uma discussão dos resultados e as nossas considerações sobre as contribuições desse primeiro teste para a o andamento da tese.

6.1.1 Papéis semânticos selecionados

Para este estudo-piloto, selecionamos a lista proposta por Brumm (2008) e Gelhausen (2010), principalmente por três motivos: a lista foi desenvolvida com fundamentação na opinião de linguistas que se basearam em testes práticos de anotação; ela foi desenvolvida pensando-se em estudos multilíngues; ela é bastante extensa; e, parafraseando Perini (2008), é melhor começar com muitos papéis semânticos e depois reduzi-los do que começar com pouco e depois ter de reavaliar toda a anotação.

A lista apresenta 46 papéis semânticos ao todo, os quais são divididos em três categorias: papéis organizados em estruturas com origem e destino; papéis com dois elementos; e papéis que melhor descrevem uma situação ou contexto. Dentro da primeira categoria, temos os papéis que representam uma ação (AGENTE, PACIENTE e AÇÃO), uma experiência (EXPERIENCIADOR, EXPERIENCIADO e ESTÍMULO), um benefício (BENEFICIANTE, BENEFICIADO e BENEFÍCIO), uma posse ou troca de posse (POSSE, DONATÁRIO, RECIPIENTE e POSSUIDOR), um lugar (DIMENSÃO GEOGRÁFICA, ORIGEM, DESTINO, LOCAL e TRAJETO) ou tempo (DIMENSÃO TEMPORAL, INÍCIO, FIM, MOMENTO e FREQUÊNCIA). Na segunda categoria, encontramos nove pares de papéis semânticos: GUIA e ACOMPANHANTE, COMPARADO e MODELO, CONTRARIADO e OPOSITOR, ATOR e PAPEL, QUALIFICADO e QUALIDADE, SUBSTITUTO e SUBSTITUÍDO, TEMA e DESCRIÇÃO, TODO e PARTE, e CRIADOR e RESULTADO. Na última categoria encontram-se apenas cinco papéis: CAUSA, REQUISITO, INTENÇÃO, INSTRUMENTO e MODO.[39]

Um ponto importante dessa lista, que explica também o porquê de tantos papéis semânticos, é que ela foi desenvolvida para dar conta não só de argumentos verbais, mas também de argumentos internos de sintagmas complexos, como em "mesa de madeira", onde "de madeira" qualifica "mesa". Como este estudo aborda apenas papéis semânticos vinculados principalmente aos verbos, alguns papéis dessa lista não são utilizados.

A seguir, fazemos algumas considerações sobre o modo como os papéis semânticos apresentados nesta subseção foram anotados manualmente nas sentenças dos *corpora*.

[39] Traduzimos os nomes dos papéis semânticos do latim para o português para facilitar a compreensão. A lista completa, organizada em uma tabela com descrições básicas de características individuais, se encontra no Anexo A.

6.1.2 Anotação dos papéis semânticos

A questão da anotação é bastante complexa, pois envolve, necessariamente, um componente subjetivo, tendo em vista que a anotação é baseada no significado dos sintagmas em relação ao verbo e que esse significado não é objetivo. Existem alguns critérios bastante simplificados para a identificação de alguns papéis semânticos, tais como o AGENTE ter de ser animado ou, como propõe Cançado (2005), o AGENTE ou EXPERIENCIADOR terem a marca **desencadeador**. Contudo, tais traços, chamados de Relações Conceptuais Temáticas (RCTs) por Perini (2008)[40], não são universais e não se aplicam a todos os verbos. Algumas marcas utilizadas neste estudo-piloto, retiradas de Cançado (2005) e ampliadas por nós, se encontram no Anexo A e serviram como um guia para a anotação dos papéis semânticos; porém, as características apresentadas no Anexo A não devem ser tomadas como definitivas, mas sim como orientadoras.

6.1.3 Sistema de extração

O sistema de extração utilizado foi o mesmo que apresentamos na Seção 5.2. As regras de extração utilizadas foram as seguintes:

- As etiquetas SUBJ (SUJEITO) ou ACC (OBJETO DIRETO) detectadas na anotação do *parser* eram reconhecidas como um NP, mas o SUJEITO recebia relevância 1 e o OBJETO DIRETO, relevância 2;
- A etiqueta PIV era reconhecida como PP (OBJETO INDIRETO) de relevância 3;
- A etiqueta N era reconhecida como NP (N) de relevância 4;
- A etiqueta ADJ era reconhecida como ADJP (ADJ) de relevância 5; e
- A etiqueta ADVL era reconhecida como PP (ADJUNTO ADVERBIAL) de relevância 6, desde que não fosse um ADV (ou seja, um advérbio isolado, como, por exemplo, **previamente**, **não** etc.).

As estruturas de subcategorização eram construídas a partir da concatenação dos sintagmas, de acordo com a relevância atribuída. Exemplos de estruturas de subcategorização neste estudo-piloto são os seguintes: NP_NP (ativa), NP_PP[em] (passiva), NP (ativa), NP (passiva) etc.

---

[40] Perini (2008) se refere aos papéis semânticos como papéis temáticos, por isso o nome relações conceptuais *temáticas*.

6.1.4 Metodologia: escolha dos verbos e anotação

Para este estudo-piloto, foram selecionados quatro verbos que ocorriam em ambos os *corpora* entre os 40 verbos mais frequentes em cada corpus. Ao todo, entre os primeiros 40 verbos de cada corpus, ocorreram apenas 14 verbos em comum, sendo que os verbos **ser**, **estar**, **ter**, **apresentar** e **haver** tinham frequências muito elevadas em um ou em ambos os *corpora*, o que fugia ao escopo deste estudo-piloto (que é de testar um conjunto pequeno de dados), e o verbo **ir** teve de ser descartado porque foi reconhecido, em grande parte dos exemplos, de maneira errada pelo *parser* PALAVRAS, já que o verbo em questão era, de fato, o verbo **ser** em alguma das conjugações compartilhadas com o verbo **ir** (por exemplo, **foi**, **fomos**, **fosse** *etc*.). Dentre os oito verbos que restaram, selecionamos os quatro que apresentavam maior proximidade entre as frequências nos dois *corpora*. A Tabela 6.1 indica quais foram esses verbos e a respectiva frequência em cada um dos *corpora*.

Tabela 6.1 – Verbos Selecionados e Frequência nos *Corpora* de Cardiologia e do Diário Gaúcho

Verbo	Cardiologia	Diário Gaúcho
Encontrar	972	454
Levar	477	742
Receber	472	549
Usar	347	358

Para cada um dos verbos, foram anotados os dez primeiros exemplos de todas as estruturas de subcategorização com frequência de dez para cima. Porém, alguns casos, geralmente com frequências próximas a dez, por apresentarem exemplos ruidosos, tiveram menos de dez exemplos anotados. Por exemplo, a estrutura NP_NP_PP[de] do verbo **receber** do Diário Gaúcho teve apenas 9 exemplos anotados, apesar de sua frequência ser 12, pois 3 exemplos estavam incorretos (são comuns, por exemplo, ruídos na anotação morfossintática). Em alguns casos, os ruídos se estendiam para a estrutura de subcategorização como um todo, de modo que, para algumas estruturas, nenhum exemplo pôde ser anotado e a estrutura teve de ser descartada.

6.1.5 Discussão sobre este primeiro estudo-piloto

Começamos esta seção com o que observamos em relação ao terceiro objetivo deste estudo-piloto, que é o mais geral: observar a configuração das estruturas de subcategorização verbais em Cardiologia e no Diário Gaúcho, que concernem somente à descrição do português brasileiro nos dois gêneros textuais. Em seguida, passamos às estruturas de subcategorização em maiores detalhes para fazer considerações sobre os papéis semânticos utilizados.

Um elemento que chamou atenção na configuração das estruturas de subcategorização verbais foi a distinção entre voz ativa e passiva. Mesquita (2004) aponta que "os autores que têm o texto técnico como objeto de estudo concordam que ele apresenta as seguintes características: (...) Emprego de voz passiva". Essa mesma característica é apontada por Da Silva e Babini (2011), que tomam por base o estudo de Vidal e Cabré (2005) para o espanhol e o de Biber, Conrad e Reppen (1998) para o inglês. No breve estudo que realizamos, percebemos que a voz passiva foi mais recorrente, em termos de quantidade de estruturas de subcategorização, nos textos jornalísticos, sendo que, por exemplo, o verbo **levar** só apresentou voz passiva no Diário Gaúcho. Isso poderia ser visto como um indício de que a forma de se escrever textos especializados esteja mudando, ou, pelo menos, que talvez os textos de Cardiologia possam apresentar uma configuração diferente. Contudo, uma análise em mais larga escala é necessária para fazer qualquer afirmação mais categórica sobre o assunto; além disso, a quantidade de dados era pouca para sustentar as conclusões.

A estrutura de subcategorização mais recorrente em textos jornalísticos foi NP_NP na voz ativa, ou seja, sujeito e objeto direto, ocorrendo como mais frequente para os quatro verbos estudados. No *corpus* de Cardiologia, a estrutura de subcategorização NP_NP se apresentou como mais frequente apenas para os verbos **usar** e **receber**, sendo que, para o verbo **levar**, as ocorrências dessa estrutura de subcategorização estavam quase 90% incorretas, devido a ruídos na anotação do *parser*. Pode-se ler isso como um indício de que os textos jornalísticos usam estruturas mais simples para divulgar a informação, privilegiando uma estrutura mais direta.

Essa ideia de privilegiar uma estrutura mais direta e mais fácil de compreender pode ser vista também por meio dos papéis semânticos empregados em alguns casos. Por exemplo, enquanto nos textos jornalísticos temos, como forma mais recorrente, uma estrutura bem direta, como "(...) os homens que levaram a tevê (...)", com um AGENTE e um PACIENTE, os textos de Cardiologia, para o mesmo verbo **levar**, privilegiam

construções metafóricas cujos papéis semânticos são de CRIADOR e RESULTADO, como no exemplo "Esses achados levaram ao conceito (...)".

Existem, porém, casos como o do verbo **usar**, que apresentou basicamente as mesmas estruturas e os mesmos papéis semânticos, tendo uma variação mais forte apenas no vocabulário, como podemos ver nos exemplos "Sete pacientes usavam inibidor da enzima de conversão" do *corpus* de Cardiologia e "(...) Fábio usa um Fiat Prêmio" do Diário Gaúcho. Em ambos os casos, os papéis são de AGENTE e INSTRUMENTO, que foram papéis dominantes nas estruturas de subcategorização do verbo *usar* em ambos os *corpora*.

O verbo **usar** foi o verbo mais próximo em ambos os *corpora*, havendo bastante igualdade na anotação de papéis semânticos. O mesmo pode-se dizer do verbo **receber**, que apresentou papéis semânticos como RECIPIENTE e POSSE, BENEFÍCIO e BENEFICIADO, e EXPERIENCIADOR e EXPERIENCIADO na maioria dos casos em ambos os *corpora*. Quanto ao verbo **encontrar**, também observamos uma distribuição de papéis semânticos bastante parecida nos dois *corpora*, com predominância dos papéis AGENTE, PACIENTE e TEMA.

Esses resultados pareceriam mostrar que as diferenças nos gêneros textuais estão apenas no vocabulário e não no nível de papéis semânticos. Porém, observando-se as ocorrências do verbo **levar**, percebemos que a questão é um pouco mais complexa, pois, enquanto o *corpus* de Cardiologia privilegiou papéis como CRIADOR e RESULTADO, o *corpus* do Diário Gaúcho claramente privilegiou os papéis de AGENTE e PACIENTE; ainda que se possa dizer que existiram também configurações muito próximas, com papéis de EXPERIENCIADOR e EXPERIENCIADO, nos dois *corpora*.

No que diz respeito ao teste da aplicabilidade em larga escala da lista de papéis semânticos utilizada, o nosso primeiro objetivo específico, o que se pôde perceber é que ela realmente é muito extensa e apresenta algumas distinções que parecem não ser necessárias do ponto de vista dos papéis semânticos. O caso que mais chamou atenção (e que apresentou grandes dificuldades iniciais para a anotação manual dos papéis semânticos) foi a distinção entre os pares RECIPIENTE e POSSE, e BENEFICIADO e BENEFÍCIO. Se observarmos exemplos como os seguintes:

> *"O conceito saúde, portanto, integra o de qualidade de vida, porque* ***as pessoas*** *em bom estado de saúde não são* ***as que*** *recebem* ***bons cuidados médicos*** *(...)"*

*"O e-CYPHER (...) incluiu **pacientes que** receberam **stent** com sirolimus e foram catalogados via Internet em sua base de dados."*

Na anotação que realizamos, foi preciso fazer uma distinção entre o que era BENEFÍCIO e o que era POSSE (o que determinava, respectivamente, o BENEFICIADO e o RECIPIENTE). A partir dos exemplos acima, chegamos à conclusão de que um BENEFÍCIO é algo que não sofre qualquer alteração física ou deslocamento, nem mesmo metafórico, enquanto uma POSSE e um RECIPIENTE envolvem um deslocamento. Nos exemplos acima, "bons cuidados médicos" (BENEFÍCIO) não é algo que possa ser deslocado ou alterado fisicamente, enquanto um "stent" é algo concreto e pode ser deslocado fisicamente. Essa distinção parecia ser satisfatória, porém, ao nos depararmos com exemplos como este:

*"(...) aumento dos níveis de monóxido de carbono e hipotensão arterial em **indivíduos que** receberam **infusões destas substâncias 1**."*

percebemos que a distinção não é tão simples assim, pois uma "infusão" pode ser considerada tanto como algo concreto (como um "*stent*") quanto como um tratamento, algo abstrato (como "cuidados médicos"). O mesmo ocorre em exemplos do Diário Gaúcho, como o seguinte:

*"**O Diário Gaúcho** recebeu **a visita da Dani Bolina**, capa da Sexy deste mês e uma das gatas do "Pânico na TV", da Rede TV!."*

Nesse exemplo, também é difícil de dizer se "a visita da Dani Bolina" é algo concreto ou abstrato, principalmente no que diz respeito a deslocamento, já que claramente há um deslocamento envolvido, mesmo que "a visita", em si, não sofra deslocamento.

Essa distinção, necessária quando se separam os papéis de BENEFÍCIO e POSSE, se resume a uma questão de significado das palavras presentes na oração. Isto é, a possibilidade de deslocamento ou não, a concretude ou não dos elementos são características vinculadas ao léxico e não à estrutura de argumentos em si. Em princípio, essas características podem ser suprimidas da identificação de papéis semânticos, pois

estes devem representar a semântica do verbo e da oração, mas não necessariamente do léxico presente[41]. Como aponta Perini (2008), a estrutura de papéis semânticos não deve ser confundida com a representação conceptual temática, que seria uma representação mais elaborada dos papéis semânticos, caracterizada por um *continuum* semântico dependente dos itens lexicais empregados na oração. Os papéis semânticos se caracterizam por serem mais esquemáticos, dependentes do verbo e da estrutura sintática. Assim, tendo em vista que as estruturas sintáticas que suscitam POSSE e BENEFÍCIO são iguais em todos os contextos estudados, a sua distinção não seria necessária.

Essa mesma proposta de união de papéis semânticos pode ser feita em relação aos papéis semânticos de tempo e lugar, uma posição defendida por Perini (2008) e que encontra em nossos exemplos um reforço. Perini afirma que os elementos que distinguem LUGAR e TEMPO advêm totalmente dos itens lexicais empregados e não do verbo ou da sintaxe. Como exemplos, o autor menciona as seguintes orações (PERINI, 2008, p. 194-195):

Ele morreu em Belém.
Ele morreu em 1908.

Os alpinistas atingiram o pico.
Meu avô atingiu os noventa e cinco anos.

Nesses exemplos, os verbos são os mesmos e as estruturas de subcategorização são as mesmas, somente o que muda seriam as expressões de lugar e tempo, mas o conhecimento de que as expressões indicam lugar e tempo não está na estrutura das orações, mas sim no léxico empregado. A diferença está no conhecimento de mundo dos leitores dessas orações, que sabem que "pico" é um lugar e que "noventa e cinco anos" é uma medida temporal. Em nossos exemplos, percebemos muitos casos em que a distinção entre uma estrutura de papel semântico e outra se dava porque uma tinha o papel MOMENTO e outra o papel LOCAL. Podemos ver isso nos exemplos a seguir, com os verbos **levar** e **encontrar**:

[41] Porém, como veremos ao longo desta tese, o léxico acaba influenciando na anotação de papéis semânticos, principalmente no que diz respeito aos atributos de cada papel.

*"O evento que **em 2007** levou ao Parcão da 79 mais de 6 mil pessoas, terá a sua segunda edição das 14h às 19h no mesmo local do ano passado."*

*"José Antônio Heinzmann, pároco da Igreja Santa Rosa de Lima, **no Rubem Berta**, leva imagem de Nossa=Senhora a armazéns, campos de futebol e bares."*

*"Analisando as dificuldades intra-operatórias, **em 12 pacientes** encontramos dificuldade na canulação do óstio do seio coronariano."*

*"**Na análise de sobrevida**, não encontramos uma relação entre o grau do comprometimento cardíaco e a sobrevida, embora em muitos estudos se observe essa relação 28-37."*

Como mencionamos, a diferença entre lugar e tempo (ou LOCAL e MOMENTO, para usar os papéis semânticos de nossa lista) é uma distinção que cabe à semântica lexical, mas não a uma semântica esquemática, como é a dos papéis semânticos.

Outra consideração a ser feita diz respeito aos papéis semânticos que se apresentam em duplas (por exemplo, SUBSTITUTO e SUBSTITUÍDO, COMPARADO e MODELO, *etc*. — consulte o Anexo A para ver a lista completa), que praticamente não foram utilizados. A maioria desses papéis parece ser útil apenas se utilizássemos uma anotação com mais de um papel semântico para cada argumento. Em nosso estudo, por exemplo, utilizamos apenas os pares CRIADOR e RESULTADO, e TEMA e DESCRIÇÃO. Infelizmente, não temos dados evidenciando que os demais papéis semânticos duplos são desnecessários, porém, nos parece que eles realmente só seriam aplicáveis em casos muito restritos e que possivelmente acabariam em distinções como a de BENEFÍCIO e POSSE. Uma argumentação que já se pode fazer a esse respeito é em relação aos pares QUALIDADE e QUALIFICADO, e TEMA e DESCRIÇÃO. Considere o seguinte exemplo:

*"Atualmente esse aparelho pode ser encontrado nas unidades de atendimento, porém sua interpretação depende de **especialistas, que** muitas vezes não se encontram **presentes** no momento do exame."*

Os argumentos **que** e **presentes** foram classificados como TEMA e DESCRIÇÃO, respectivamente; no entanto, não parece haver qualquer motivo que impeça a classificação como QUALIFICADO e QUALIDADE, principalmente pelo fato de que os autores que propuseram a lista de papéis semânticos (BRUMM, 2008; GELHAUSEN, 2010) não se preocuparam em apresentar de modo mais detalhado elementos que distingam esses papéis entre si.

Assim, percebe-se que os 46 papéis semânticos podem ser reduzidos para uma lista mais condensada. No entanto, para chegarmos a tal lista, precisaríamos testar mais verbos, o que nos leva à questão da avaliação do extrator de estruturas de subcategorização.

Quanto ao desempenho, somente percebemos a questão de os objetos reflexivos não serem considerados como parte da estrutura de subcategorização, o que foi necessário modificar. As etiquetas ACC-PASS e refl até eram reconhecidas como OBJETO REFLEXIVO; contudo, não recebiam atribuição de um valor de relevância e, por isso, não eram adicionadas à estrutura de subcategorização[42], o que fazia com que os argumentos reflexivos fossem ignorados.

No restante, o sistema reconheceu muito bem os argumentos, sendo que os ruídos decorreram, em sua maioria, da anotação do *parser* PALAVRAS. Além disso, a interface de anotação é bastante simples e facilita muito o trabalho do linguista, que pode se concentrar exclusivamente no cerne do trabalho, com uma estrutura de argumentos já organizada com a anotação sintática.

O maior problema que enfrentamos foi que a extração dos resultados não pôde ser realizada automaticamente. Após ter realizado a anotação dos dados, não tínhamos nada que nos apresentasse uma informação sobre os dados que foram anotados. Aqui ainda faltava conhecimento computacional para acessar o banco de dados e extrair dele as informações requeridas para gerar análises que não fossem feitas apenas "a olho nu". Esse problema precisou ser abordado antes que pudéssemos passar para uma anotação em grande escala[43], como a que foi realizada no estudo-piloto II e na criação do próprio VerbLexPor.

---

[42] Como veremos, no segundo estudo-piloto, este erro foi corrigido e os objetos reflexivos passaram a fazer parte também das estruturas de subcategorização.

[43] Como veremos nos experimentos posteriores, esses inconvenientes foram solucionados e os resultados puderam ser analisados com maior facilidade.

Assim, a realização deste estudo-piloto apontou uma série de elementos importantes sobre os quais precisamos refletir ou que precisamos modificar antes de passarmos a uma anotação manual em grande escala. A questão mais urgente a ser abordada foi a necessidade de uma automatização da observação dos resultados a partir do banco de dados anotado. Os dados deste estudo tiveram de ser recolhidos a partir de uma observação manual, o que impediu uma visualização global dos resultados e certamente dificultou muito a observação de muitos dados. Seria muito mais complexo, por exemplo, se tivéssemos anotado vinte verbos, algo que tornaria a observação manual uma tarefa quase impossível.

Um resultado importante que observamos foi que os 46 papéis são, na verdade, muito detalhados, captando também características exclusivas do léxico presente nos argumentos. Dessa forma, percebemos que a lista de papéis semânticos pode ser reduzida. É claro que precisaríamos observar mais dados para podermos analisar com mais cuidado quais casos poderiam ser unidos e quais seriam mantidos separados.

Resumindo, o primeiro estudo-piloto realizado serviu para apontar várias informações importantes para o prosseguimento do estudo: o trabalho com o extrator de estruturas de subcategorização precisava ser retomado no que dizia respeito à apresentação dos resultados e a lista de papéis semânticos precisava de maiores testes em relação à supressão de alguns papéis.

Ao final deste estudo inicial, obtivemos importantes informações sobre como continuar o estudo e, principalmente, tivemos um primeiro contato prático com a anotação de papéis semânticos. Assim, este estudo-piloto foi importante para nortear o restante do estudo que foi realizado. Um dos principais impactos do estudo-piloto foi a percepção de que a anotação de papéis semânticos é uma tarefa que exige muito do anotador, principalmente no que diz respeito ao conhecimento linguístico.

6.2 Estudo-Piloto II

Dando sequência aos resultados do primeiro estudo-piloto, passamos a nos concentrar em uma anotação mais ampla, porém ainda com caráter de teste, que chamamos de estudo-piloto II. Neste segundo experimento, modificamos alguns detalhes na metodologia e fizemos várias alterações no extrator de estruturas de subcategorização.

Desse modo, temos várias novidades a relatar em relação ao primeiro estudo-piloto, ainda que os objetivos do experimento tenham permanecido inalterados. Por

isso, nas seções seguintes, relatamos primeiro as modificações relativas à lista de papéis semânticos e, em seguida, ao sistema de extração de estruturas de subcategorização e à metodologia. Na sequência, apresentamos os resultados obtidos no segundo estudo-piloto e os discutimos, deixando nossas considerações para o final.

6.2.1 Lista de papéis semânticos

Após o primeiro estudo-piloto, ficamos pouco impressionados com a lista de 46 papéis semânticos proposta por Brumm (2008) e Gelhausen (2010), e o fato de que decidimos por não realizar uma anotação em língua estrangeira[44] fez com que buscássemos outro tipo de lista. Além disso, apesar de a lista ter sido gerada com base na opinião de linguistas e em alguns dados concretos, a quantidade de dados testada não foi grande. Segundo Brumm (2008), foi utilizada apenas uma lista de sentenças. Assim, após várias consultas bibliográficas, optamos por utilizar uma lista de papéis descritivos e genéricos que já tivesse sido testada em uma quantidade maior de dados, o que conferiria uma maior qualidade potencial para os papéis semânticos utilizados. Por isso, após observarmos as listas empregadas em vários estudos (FrameNet, PropBank, VerbNet, entre outros), optamos por usar os papéis da VerbNet (Kipper-Schuler, 2005), seguindo a sua versão 3.2, que é a mais recente.

Como já mencionamos no Capítulo 4, já existe uma VerbNet.Br; porém, ela foi feita a partir de uma importação de dados do inglês, tomando por base o potencial interlinguístico das classes de Levin. Por ter sido um estudo pioneiro que visava a ser um primeiro passo para o estudo em português dos papéis semânticos no estilo da VerbNet, não houve um estudo linguístico mais profundo que mostrasse o quanto essa importação realmente traz dados confiáveis para o português. Desse modo, decidimos focar nossos esforços no mesmo âmbito da VerbNet.Br, utilizando uma anotação manual, que posteriormente será confrontada com os dados importados do inglês presentes na VerbNet.Br[45].

Um detalhe importante de se ressaltar é que a lista de papéis semânticos da VerbNet foi bastante modificada desde sua versão 1.0. As modificações realizadas

[44] A anotação em língua estrangeira (alemão) estava prevista no projeto original desta tese, porém, logo após o estudo-piloto I, percebemos que não seria produtivo anotarmos papéis em língua estrangeira.
[45] A comparação da nossa anotação com os dados da VerbNet.Br foi realizada após a criação do VerbLexPor léxico e é apresentada no Capítulo 11.

deveriam estar relatadas na documentação do recurso[46]; contudo, se observarmos os dados do recurso e compararmos com a documentação, é possível detectar algumas discrepâncias. Após um estudo dos papéis semânticos e dos exemplos disponíveis no recurso VerbNet, extraímos a lista de papéis semânticos efetivamente utilizada, somamos a ela os papéis potenciais presentes na documentação e, com base em nossas observações do primeiro estudo-piloto, realizamos algumas pequenas modificações. A principal modificação foi a criação do hiperônimo TARGET[47], que passou a abrigar BENEFICIARY e RECIPIENT, para os casos em que um verbo autoriza ambos. As demais modificações apenas alteraram o entendimento da hierarquia da VerbNet, mas não modificaram os papéis em si. Nas Figuras 6.1 e 6.2, podemos ver a hierarquia de papéis semânticos apresentada na documentação da VerbNet e a hierarquia que utilizamos.

Com as modificações realizadas, definiu-se uma lista com 38 papéis semânticos: THEME, CO-THEME, AGENT, CO-AGENT, STIMULUS, INSTRUMENT, PATIENT, CO-PATIENT, EXPERIENCER, TARGET, RECIPIENT, BENEFICIARY, INITIAL TIME, MOMENT, FINAL TIME, FREQUENCY, DURATION, SOURCE, INITIAL LOCATION, MATERIAL, GOAL, DESTINATION, RESULT, PRODUCT, LOCATION, TRAJECTORY, ATTRIBUTE, TOPIC, PIVOT, VALUE, EXTENT, ASSET, CAUSE, REFLEXIVE, PREDICATE, VERB, MANNER E COMPARATIVE. Pode parecer estranho o uso dos papéis semânticos em inglês, porém, por estarmos utilizando como fonte a VerbNet, acreditávamos que essa escolha simplificaria uma comparação futura do português com o inglês.

Alguns dos papéis semânticos da lista se aplicam potencialmente apenas a adjuntos, como MANNER e COMPARATIVE, outros são papéis auxiliares, como VERB e REFLEXIVE, que se aplicam, respectivamente, a argumentos que formam um significado complexo com o verbo (por exemplo, casos de verbos-suporte) e à partícula reflexiva. Informações mais detalhadas sobre a funcionalidade dos papéis semânticos utilizados podem ser encontradas no Anexo B, onde apresentamos uma tabela com a lista completa, uma descrição e alguns comentários quanto ao emprego.

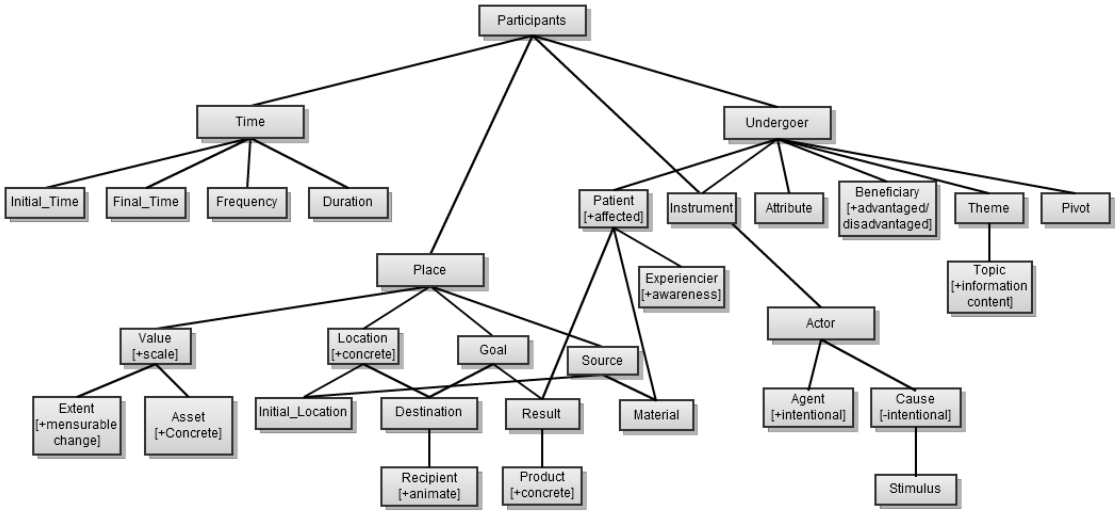
---

[46] Disponível em http://verbs.colorado.edu/~mpalmer/projects/verbnet/VerbNet3.0ReadMe.doc. Acessado em: 27/01/2015.

[47] Por termos escolhido uma lista em inglês, optamos por não traduzir os nomes dos papéis e por manter a nomenclatura toda em inglês. Assim, quando nos referirmos a papéis genéricos, utilizaremos nomes em português, como AGENTE e PACIENTE, porém, quando nos referirmos à nomenclatura empregada neste estudo-piloto, usaremos o inglês. Essa decisão por usar os nomes em inglês foi posteriormente revista, e a lista empregada no VerbLexPor, como será visto posteriormente, está em português.

É importante ressaltar aqui que, apesar de termos cogitado juntar papéis como os de tempo (TIME) e local (PLACE), como havíamos mencionado ao final do primeiro estudo-piloto, achamos por bem manter essa distinção. Ainda que local e tempo sejam marcados pelas mesmas preposições e estruturas sintáticas, a distinção entre eles para a semântica de uma oração é importante, principalmente quando levamos em conta o reconhecimento automático de significados. Assim, optamos por manter as duas categorias em vez de juntá-las ou fazer grandes modificações na hierarquia da VerbNet. As alterações que realizamos na hierarquia foram mais de cunho organizacional, tirando arestas que passavam de um grupo para outro. Desse modo, na nova hierarquia apresentada (Figura 6.2), cada papel tem apenas um papel superordenado, tornando mais clara e não ambígua a relação entre os papéis semânticos.

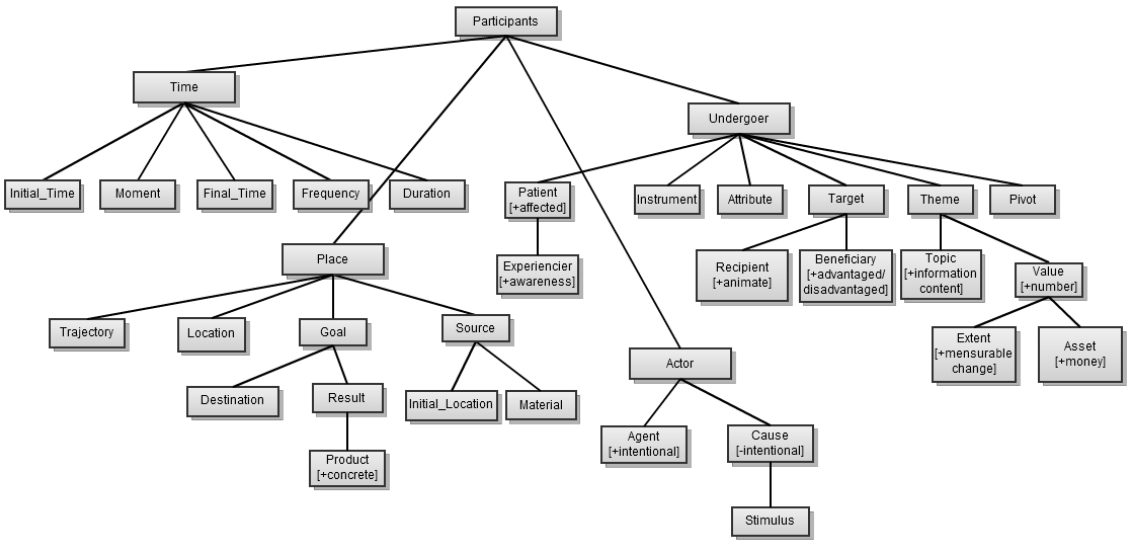
Figura 6.1 – Hierarquia de papéis semânticos utilizada na VerbNet (versão 3.2)



6.2.1 Modificações no extrator e na interface de anotação

Corrigindo os problemas que detectamos no primeiro estudo-piloto, o extrator de estruturas de subcategorização passou a apresentar as estruturas de maneira diferente. Enquanto antes as estruturas eram apresentadas como NP_NP_PP[em] ou NP_PP[para], a partir desta versão, o sujeito passou a ser explicitado na estrutura, assim como a posição do verbo.

Figura 6.2 – Hierarquia de papéis semânticos utilizada em nosso segundo estudo-piloto



Com essas modificações, as estruturas de subcategorização apresentadas passaram a ter o seguinte formato: SUBJ[NP]_V_NP_PP[em], SUBJ[NP]_V_PP[para]. Como optamos por usar sempre um sujeito, mesmo que oculto, os elementos SUBJ[NP] e V poderiam ser suprimidos, já que eles sempre estarão nas duas primeiras posições das estruturas de subcategorização; no entanto, como é possível escolher uma apresentação em que a ordem da oração define a ordem dos elementos na estrutura de subcategorização, optou-se por manter esses elementos explícitos.

Além disso, foi corrigido o fato de que os reflexivos não estavam sendo apresentados na estrutura de subcategorização. Assim, nesta versão, foi possível ver estruturas como, por exemplo, SUBJ[NP]_V_REFL_PP[em].

Uma das principais modificações ocorreu na apresentação dos resultados, pois nesta versão era possível, após a anotação, reunir os dados em uma apresentação parecida com a das estruturas de subcategorização, mostrando a categoria sintática e a anotação de papel semântico. Usando essa nova função, após a anotação, era possível, por exemplo, ver quais verbos tinham uma estrutura do tipo SUJ<AGENT>+ OBJ.DIR<THEME> e qual a sua frequência.

6.2.2 Método de anotação

Para realizar a anotação de papéis semânticos, fizemos inicialmente algumas escolhas em relação às quantidades a serem anotadas. Assim como no primeiro estudo-piloto, optamos por uma anotação amostral, almejando um teste dos papéis semânticos

apresentados pela VerbNet. Decidimos anotar, nos dois *corpora*, primeiro os 25 verbos mais frequentes do *corpus* de Cardiologia e, em seguida, também nos dois *corpora*, os 25 verbos mais frequentes do *corpus* do Diário Gaúcho, pulando os que já haviam sido anotados na primeira etapa. Assim, foram anotados 50 verbos ao todo em cada um dos *corpora*[48] — com os seguintes critérios:

- Os seguintes verbos foram excluídos: *ser*, *estar*, *ter* e *haver*.
- Foram anotadas exatamente dez sentenças de cada estrutura de subcategorização.
- Os verbos anotados tinham de estar presentes nos dois *corpora* com frequência suficiente para que pelo menos dez sentenças fossem anotadas dentro de pelo menos uma estrutura de subcategorização.

A exclusão *a priori* de quatro verbos (*ser*, *estar*, *ter* e *haver*) se deu por eles serem extremamente polissêmicos e/ou frequentes nos dois *corpora*. A anotação desses verbos com o método adotado dificilmente refletiria as suas várias facetas, além de consumir muito tempo devido à quantidade de estruturas de subcategorização existentes para cada um deles.

A escolha de dez exemplos, para cada estrutura de subcategorização, foi apenas um incremento em relação ao método usado no primeiro estudo-piloto. Com a modificação apresentada aqui, garantimos que todas as estruturas de subcategorização tivessem dez exemplos anotados. Assim, se uma estrutura tivesse 16 exemplos, mas apenas nove estivessem corretos, ela era descartada como um todo.

A presença dos verbos nos dois *corpora* foi uma exigência para a sua anotação tendo em vista o objetivo comparativo deste estudo-piloto: como queríamos comparar os resultados, achamos plausível dar prioridade para verbos presentes nos dois *corpora* em frequências anotáveis.

[48] Houve apenas uma exceção a isso. A título de curiosidade, anotamos o verbo *ir* no *corpus* do Diário Gaúcho. Assim, o Diário Gaúcho teve, na verdade, 51 verbos anotados. Esse verbo seria anotado também no *corpus* de Cardiologia, mas a sua frequência não foi suficiente.

6.2.3 Resultados e considerações sobre a anotação de papéis semânticos

Nesta subseção, expomos nossas considerações qualitativas sobre o método empregado na anotação de papéis semânticos e, em seguida, apresentamos os resultados da anotação e da comparação entre os dois corpora.

6.2.3.1 Considerações sobre a lista e o método

A lista de papéis semânticos da VerbNet se mostrou adequada na maioria dos casos, pois se aplicou bem aos argumentos dos verbos anotados. Os únicos problemas encontrados nesse sentido foram resultantes da união da lista da VerbNet com uma metodologia que não distingue entre argumentos e adjuntos. Como optamos por anotar todos os elementos que se ligassem ao verbo, considerando que a frequência seria o delimitador dos argumentos anotados, alguns dos elementos anotados, por serem de natureza adverbial, não tinham um papel semântico condizente, precisando ser anotados com papéis que se adequavam apenas parcialmente. Como veremos mais adiante, na descrição do recurso léxico gerado, esse tipo de problema foi posteriormente solucionado com a adição dos papéis semânticos específicos para adjuntos utilizados no PropBank[49].

Em geral, a anotação dos adjuntos adverbiais foi uma tarefa complexa. Observando as sentenças 6.a a 6.d a seguir, extraídas dos *corpora* anotados, temos adjuntos adverbiais com as palavras *jogo* e *estudo* (destacados em negrito) que representam parte dessa complexidade.

6.a. Eles fizeram um jogo largado e nós demos oportunidade **em um jogo** que estava em nossas mãos.
6.b. Teremos de melhorar muito em relação ao que mostramos **no primeiro jogo**, mas temos todas as condições de reverter.

[49] Isso pode parecer, em princípio, contradizer nosso posicionamento, de considerar como argumento os elementos no nível semântico, reservando à separação entre complemento e adjunto para o plano sintático. Contudo, cremos que não há problema em reconhecer que, na semântica de uma oração, nem sempre os papéis semânticos serão atribuídos pelo verbo. O que é preciso deixar claro é que não temos em mente uma separação explícita entre argumentos e adjuntos, o que seria contrário à nossa opção inicial, mas apenas pegamos emprestado os papéis semânticos usados no PropBank para complementar os papéis semânticos já existentes, de modo a deixar a lista mais robusta.

6.c. **No presente estudo**, animais adultos restritos apresentaram aumento de todos os parâmetros estereológicos analisados na aorta, sugerindo hiperplasia da túnica média.
6.d. O prognóstico utilizado para o TC6M foi demonstrado **no estudo SOLVD10**.

Poderíamos, por exemplo, anotar essas estruturas como, MOMENT, LOCATION ou mesmo INSTRUMENT, dependendo de sua situação na sentença, mas não tínhamos um papel que representasse um significado como SITUATION (stituação). Isso ocorreu porque os papéis semânticos da VerbNet foram pensados apenas para complementos, e não para adjuntos. Assim, vimos que seria necessário incluir papéis que dessem conta desses adjuntos adverbiais. Apesar desses problemas referentes aos adjuntos, a lista se mostrou adequada para a atribuição de papéis semânticos para os demais argumentos.

No que diz respeito ao método amostral escolhido, ele foi adequado para a maioria dos verbos, pois equilibra o tempo utilizado para anotar e a representatividade dos dados anotados. Porém, ficou claro que, para verbos muito polissêmicos (por exemplo, *dar*, que tem muitos usos como verbo-suporte, os quais também foram anotados), a amostragem não capta grande parte dos significados do verbo. No entanto, se aumentarmos o número de exemplos anotados a cada estrutura de subcategorização, o esforço necessário para anotar cada um dos verbos também aumentaria. Por mais que sempre exista esse problema com o método amostral (afinal, alguns dados são ignorados), durante o processo de anotação, é possível perceber quais verbos não estão representados adequadamente e, se necessário, é possível dar um tratamento especial a eles.

No que diz respeito ao extrator de estruturas de subcategorização, com a anotação de mais verbos em relação ao estudo-piloto I, percebemos que alguns elementos linguísticos das sentenças são anotados pelo *parser* PALAVRAS (BICK, 2000) de uma forma que não estava sendo levada em consideração pelo sistema. Por exemplo, agentes da passiva são anotados pelo PALAVRAS como PASS, e os objetos indiretos são anotados tanto como PIV quanto como SA; porém, o sistema estava preparado apenas para reconhecer PIVs e ADVLs. Portanto, alguns agentes da passiva acabaram não sendo reconhecidos (pois não apresentavam a marcação ADVL) e o mesmo aconteceu com os objetos indiretos marcados como SA. Para eliminar esse tipo de problema, fizemos uma análise do conjunto completo de etiquetas empregadas pelo

PALAVRAS[50] e acrescentamos ao sistema, com a respectiva descrição, as modificações necessárias para que fossem extraídos todos os argumentos julgados relevantes[51].

6.2.3.2 Exportação para XML

Uma das observações que fizemos após o primeiro estudo-piloto foi o fato de que os dados armazenados no formato MySQL eram de difícil análise, principalmente pelo fato de que os dados ficam distribuídos em diferentes tabelas dentro do banco de dados, fazendo com que a visualização do todo da informação seja mais complexa. Sendo assim, uma das medidas tomadas para mitigar esse problema de análise dos dados e também para facilitar a manipulação e a disponibilização dos dados, foi a criação de uma ferramenta que exportasse os dados para o formato XML. O formato XML, além de ser mais apropriado para a visualização do que o MySQL, também permite uma utilização mais simples por parte de outros sistemas e bancos de dados, facilitando a divulgação dos resultados no meio acadêmico.

Após uma série de discussões sobre como seria realizada a exportação dos dados, decidiu-se por abrir uma vaga para um projeto de Master I[52] que abordaria a exportação dos dados do formato MySQL para XML e importação de volta de XML para MySQL. Assim, após a seleção de candidatos, iniciou-se um trabalho de coorientação do aluno Samy Sassi, do curso de Ciências da Computação da Universidade Joseph Fourier.

Nos quatro meses da coorientação, desenvolveu-se então uma ferramenta em linguagem Python que lia o banco de dados em formato MySQL, extraia as informações necessárias e as transcrevia para o formato XML. As Figuras 6.3 e 6.4 mostram como se apresentam os dados em cada um desses formatos, apenas para dar uma ideia da diferença entre eles.

---

[50] As etiquetas com as respectivas explicações de suas funções podem ser encontradas no seguinte site: http://beta.visl.sdu.dk/visl/pt/info/portsymbol.html. Acessado em: 27/01/2015.

[51] A lista final de argumentos extraídos pelo sistema pode ser vista mais adiante, no Capítulo 8, quando expomos os materiais do recurso final.

[52] Durante esse período, estávamos realizando um estágio de doutorado-sanduíche no Laboratoire d'Informatique de Grenoble, como parte do Projeto CAMELEON (CAPES/COFECUB 707/11). Na França, um projeto de Master I é mais ou menos o equivalente a um projeto de iniciação científica no Brasil e tem uma duração de quarto meses.

Figura 6.3 – Dados apresentados em formato MySQL

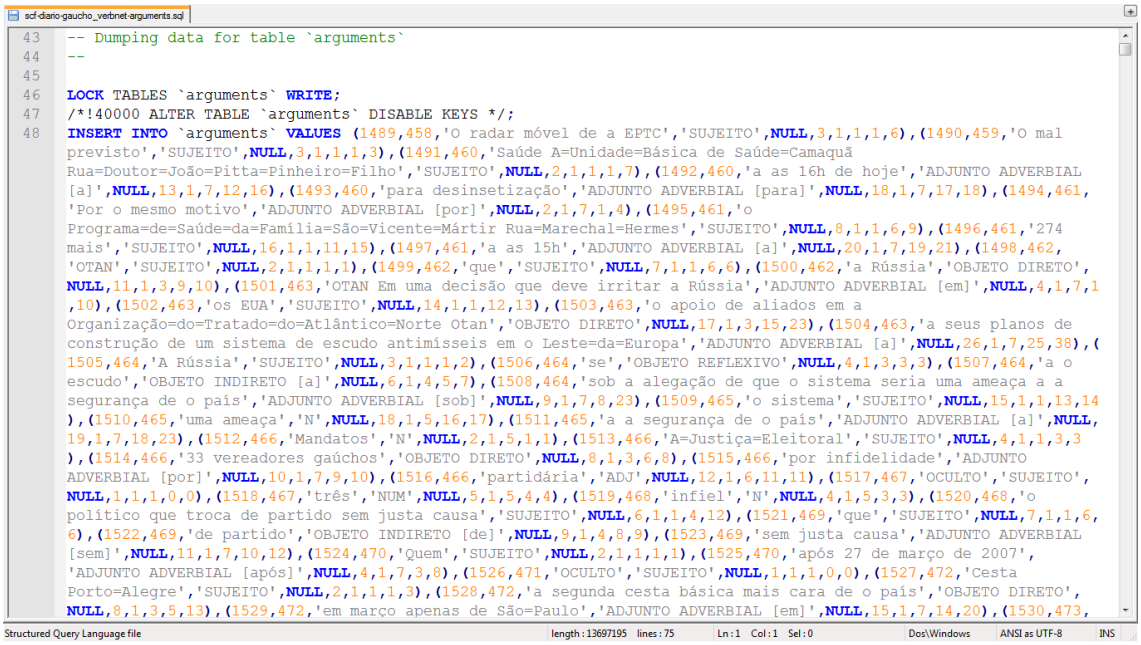
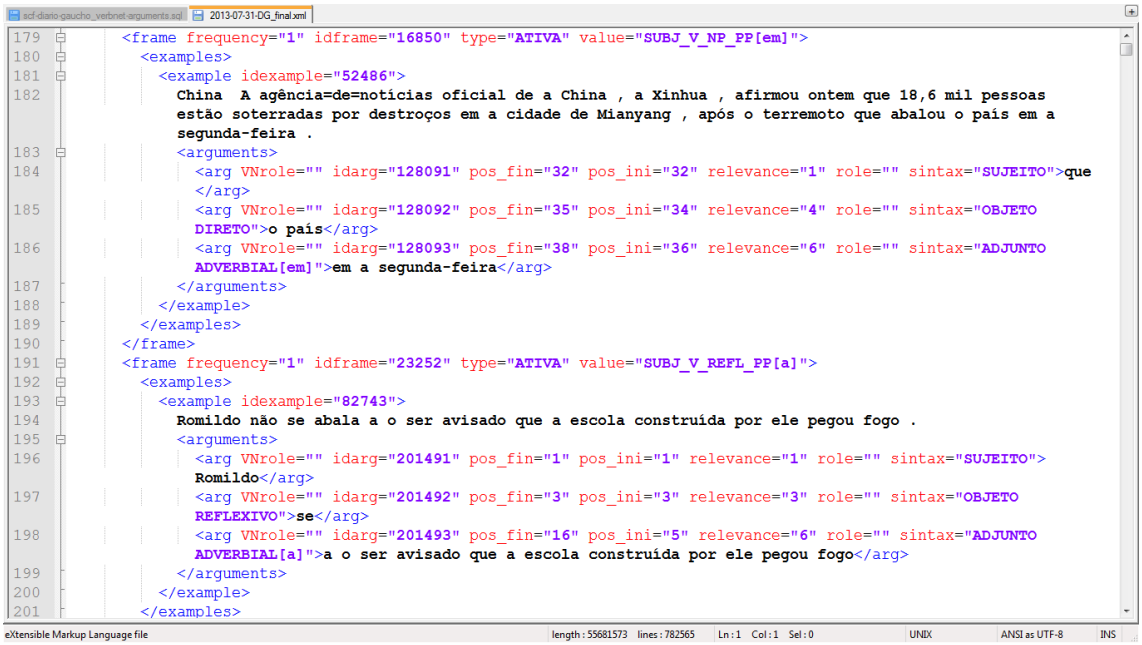


Figura 6.4 – Dados apresentados em formato XML



A conversão dos dados para o formato XML auxiliou tanto na observação dos resultados dos experimentos posteriores como na disponibilização dos dados na plataforma Jibiki (a qual apresentaremos de modo aprofundado na Seção 8.6). Essa maior facilidade de análise poderá ser vista a seguir, quando apresentamos os resultados do segundo estudo-piloto e, ainda mais adiante, quando discutirmos o agrupamento de verbos no português brasileiro e os resultados do recurso final. Assim, apesar de ter sido

uma etapa relativamente simples do estudo, a exportação dos dados para XML trouxe muitos resultados positivos e propiciou um avanço mais rápido do trabalho de análise dos dados.

6.2.3.3 Resultados da anotação e comparação entre os *corpora*

Neste estudo-piloto, realizamos a anotação de 3.400 orações (1.790 orações no corpus de Cardiologia e 1.610 no *corpus* do Diário Gaúcho). Essas orações se encontram atualmente armazenadas em um banco de dados em formato MySQL, o qual foi exportado também para XML.

No que diz respeito à diferença de frequências entre as sentenças anotadas nos *corpora*, temos exemplos bastante discrepantes. Por exemplo, o verbo *considerar*, bastante frequente em Cardiologia, com 60 sentenças anotadas, encontra no *corpus* do Diário Gaúcho uma contraparte de apenas 10 sentenças. Essas diferenças poderiam ter sido amenizadas se, durante a organização do *corpus*, tivéssemos selecionado sentenças específicas para cada verbo em vez de textos completos. No entanto, isso implicaria na construção de um novo recurso a partir do zero, o que demandaria tempo. Além disso, uma organização desse tipo poderia camuflar algumas diferenças existentes entre os dois tipos de linguagem, algo que não desejamos, tendo em vista que nosso objetivo é observar a linguagem em sua forma natural, com diferenças e semelhanças que variam desde as frequências até as estruturas.

Entre as 1790 orações do *corpus* de Cardiologia, observaram-se 304 estruturas sintático-semânticas[53] diferentes, sendo esta, que conta com apenas um argumento, a mais frequente: SUJ<Theme>; no *corpus* do Diário Gaúcho, entre as 1610 orações, encontraram-se 272 estruturas diferentes, sendo mais frequente uma estrutura com dois argumentos: SUJ<Agent>+OBJ.DIR<Theme>.

Em ambos os *corpora*, houve muitas ocorrências de estruturas sintático-semânticas com frequência 1; dentre elas, 117 estavam no *corpus* de Cardiologia e 106 no *corpus* do Diário Gaúcho. Normalmente, frequências baixas são descartadas, por não representarem informações relevantes. Em nosso caso, porém, por se tratar de uma

[53] Por **estruturas sintático-semânticas**, nos referimos às associações entre estruturas sintáticas (sujeito, objeto direto etc.) e papéis semânticos (AGENT, PATIENT etc.) em uma oração. Para simplificar a representação das estruturas sintático-semânticas, utilizaremos as seguintes abreviaturas para a sintaxe:

- SUJEITO = SUJ
- OBJETO DIRETO = OBJ.DIR
- ADJUNTO ADVERBIAL[prep.] = ADJ.ADV[prep.]

anotação manual, a baixa frequência não deve ser desconsiderada. Além disso, o fato de que existe apenas uma sentença no *corpus* até então anotada com a estrutura sintático-semântica SUJ<Theme>+ADJ.ADV [em]<Location>+ADJ.ADV [a]<Goal>[54] para o verbo *chegar* não quer dizer que haja apenas uma ocorrência de cada um dos argumentos SUJ<Theme>, ADJ.ADV[a]<Location> e ADJ.ADV[em] <Goal> para esse mesmo verbo. Durante o aprendizado de máquina de sistemas de anotação de papéis semânticos, não apenas a estrutura como um todo pode ser relevante, mas também cada um de seus elementos individuais.

Tabela 6.2 – Cinco estruturas mais frequentes no *corpus* de Cardiologia

Cardiologia		
Estrutura	**Freq.**	**Freq. %**
SUJ<Theme>	181	10,11
SUJ<Theme>+ADJ.ADV[em] <Location>	121	6,76
SUJ<Instrument>+OBJ.DIR <Theme>	102	5,70
SUJ<Agent>+OBJ.DIR<Theme>	63	3,52
SUJ<Patient>	40	2,23

Tabela 6.3 – Cinco estruturas mais frequentes no *corpus* do Diário Gaúcho

Diário Gaúcho		
Estrutura	**Freq.**	**Freq. %**
SUJ<Agent>+OBJ.DIR<Theme>	171	10,62
SUJ<Theme>	114	7,08
SUJ<Agent>	92	5,71
SUJ<Theme>+ADJ.ADV [em] <Location>	50	3,11
SUJ<Agent>+OBJ.DIR<Theme>+ADJ.ADV [em]<Location>	45	2,79

Nas Tabelas 6.2 e 6.3, podemos ver as cinco estruturas mais frequentes nos dois *corpora*. Nessas tabelas, é possível observar que, enquanto o *corpus* de Cardiologia privilegia construções passivas e intransitivas (o que explica a ocorrência de duas

[54] A sentença em questão é *Em alguns trechos, a água chegou a 1,5m de altura*.

estruturas sem objetos), o Diário Gaúcho apresenta estruturas agentivas transitivas diretas no topo, seguidas por passivas e intransitivas.

Quando observamos, no banco de dados, os verbos e sentenças que se enquadram nas estruturas mais frequentes sem objetos, percebemos que, no caso da Cardiologia, se trata, na maioria dos exemplos, de utilização de voz passiva[55], e nem tanto de intransitividade. Já no Diário Gaúcho ocorre o oposto, com uma maioria de exemplos intransitivos[56]. Isso contraria as observações realizadas em nosso primeiro estudo-piloto, quando havíamos observado uma tendência maior de apassivamento no Diário Gaúcho, o que provavelmente era um fenômeno pertinente apenas aos verbos estudados.

Tanto o Diário Gaúcho quanto o *corpus* de Cardiologia apresentam estruturas transitivas diretas em posições elevadas na lista, porém, na Cardiologia, há uma tendência para que o sujeito seja um INSTRUMENT, deixando o real agente apagado. O mesmo não se observa no Diário Gaúcho, que apresenta grande quantidade de sujeitos agentes. Esse fenômeno não é algo que se apresenta apenas entre as estruturas sintático-semânticas mais recorrentes, mas ao longo das várias estruturas existentes. A Cardiologia apresentou uma forte tendência a esconder os verdadeiros agentes, colocando em evidência os instrumentos utilizados.

Na comparação, não se pode afirmar que os *corpora* utilizem estruturas sintático-semânticas diferentes, pois quase todas as estruturas ocorrem nos dois tipos de texto. O que se percebe é mais uma tendência diferente no *corpus* especializado, sendo que o principal fator é o apagamento dos agentes. Para sustentar esse resultado com números, observamos que, dentre as 304 estruturas sintático-semânticas anotadas, apenas 31 apresentavam AGENT, enquanto no Diário Gaúcho, dentre as 272 estruturas, 121 apresentavam AGENT. Isso representa um salto de 10,19% para 44,49% entre os *corpora*.

Em termos de exemplos concretos, os sujeitos em Cardiologia tendem a ser expressões como estas (extraídas do banco de dados):

- *Estudos de o perfil lipídico*;
- *a combinação de restrição calórica com exercício físico*; e

---

[55] Alguns exemplos:
- *Foram avaliados os seguintes parâmetros:*
- *Foi observada uma distribuição igual de a população estudada em relação ao sexo.*

[56] Observou-se uma maioria de verbos como “ocorrer”, “existir”, “ficar”, “acontecer”.

- *Análises futuras de feocromocitomas com técnicas de microarray proteômica*;

enquanto o Diário Gaúcho apresenta mais sujeitos como estes:

- *o jogador*;
- *o técnico Abel=Braga*; e
- *Leona=Cavali*.

Além das expressões serem de categorias semânticas diferentes, vale aqui observar que os sujeitos em Cardiologia são muito mais extensos, o que pode ser um indicativo do gênero.

Como pode ser visto na Tabela 6.4 (Linha 1), os dados sobre a agentividade se mantêm distintos quando olhamos para o número de sentenças. A Cardiologia apresenta 198 sentenças com AGENT em 1790 (11,06%) contra 734 sentenças em um total de 1610 (45,59%) no Diário Gaúcho. Também é possível perceber que a quantidade de sentenças com INSTRUMENT (Linha 16 da Tabela 6.4) é mais de três vezes maior em Cardiologia do que no Diário Gaúcho. Outras diferenças estão no fato de que o papel PIVOT (Linha 22 da Tabela 6.4), que geralmente representa um elemento que contém outro elemento, sem participar em uma ação, ocorre quase seis vezes mais no *corpus* de Cardiologia do que no do Diário Gaúcho, e o papel GOAL (Linha 14 da Tabela 6.4), que geralmente representa um objetivo de uma ação, também é muito mais frequente naquele do que neste.

6.2.3.1 Aporte estatístico para a observação de diferenças entre as linguagens

Para avaliar se há diferenças significativas entre os dados dos corpora, recorremos à Estatística. Como nossos dados não são paramétricos e são categóricos, optamos por um teste que observa a correlação entre dois *rankings* de dados. Assim, estamos comparando aqui se os *rankings* de papéis semânticos e informações sintáticas nos dois *corpora* têm alguma correlação ou não.

Utilizando o coeficiente de correlação tau-b de Kendall[57], realizamos três experimentos com diferentes informações. No experimento 1, avaliamos a correlação

[57] O coeficiente de correlação tau-b de Kendall avalia se existe uma correlação entre os *rankings* de duas amostras. Assim, ele informa se o ranqueamento de uma amostra X é correlacionado ao ranqueamento de

entre os *rankings* dos papéis semânticos nos dois *corpora*, considerando também as informações sintáticas e a distribuição nas sentenças. Utilizamos os dados conforme estão representados nas Tabelas 6.2 e 6.3. Nesse experimento, os resultados apontaram que há uma correlação inversa entre as amostras, pois encontramos um valor de τ = -0,394 ($p < 0,001$). Assim, percebe-se que estruturas sintático-semânticas muito frequentes no corpus de Cardiologia tendem a ser pouco frequentes no corpus do Diário Gaúcho e vice-versa. Esse resultado corrobora algumas tendências observadas na análise qualitativa anterior, quando apontamos diferenças, por exemplo, no uso de AGENT e INSTRUMENT.

Tabela 6.4 – Papéis semânticos e sua frequência nos dois *corpora*

N	Papéis Semânticos	Cardiologia	Diário Gaúcho
1	AGENT	198	734
2	ATTRIBUTE	97	46
3	BENEFICIARY	113	109
4	CAUSE	120	71
5	CO-AGENT	0	16
6	COMPARATIVE	19	0
7	CO-PATIENT	19	0
8	DESTINATION	1	91
9	DURATION	38	9
10	EXPERIENCER	41	93
11	EXTENT	29	11
12	FINAL_TIME	0	11
13	FREQUENCY	2	0
14	GOAL	215	84
15	INITIAL_TIME	0	11
16	INSTRUMENT	294	91
17	LOCATION	407	274
18	MATERIAL	0	15
19	MANNER	88	30
20	MOMENT	194	202
21	PATIENT	241	212
22	PIVOT	132	23
23	RECIPIENT	0	12
24	REFLEXIVE	4	20
25	RESULT	269	257
26	SOURCE	57	2
27	STIMULUS	6	11
28	TARGET	8	51
29	THEME	1221	962
30	TOPIC	20	14
31	VALUE	12	0
32	VERB	83	44

uma amostra Y. Os valores possíveis de τ variam entre -1 e 1, sendo 0 uma indicação de que não há correlação. Os cálculos estatísticos foram realizados com a ferramenta IBM SPSS 19.

No experimento 2, observamos a correlação entre os dois *corpora* no que diz respeito a papéis semânticos associados às suas respectivas anotações sintáticas. Isto é, em vez de utilizarmos a estrutura sintático-semântica das sentenças (como fizemos no experimento 1), consideramos apenas os argumentos isolados, com suas informações sintáticas e semânticas, da forma como representamos na Tabela 6.5 (mais adiante). Com esse conjunto de dados, não houve correlação entre as duas amostras ($\tau = 0{,}031$; $p = 0{,}608$). O problema com esse resultado foi o *p* maior que 0,05. Uma das possíveis causas para isso é a pequena quantidade de dados.

Por fim, no experimento 3, consideramos apenas o *ranking* dos papéis semânticos, sem observar a anotação sintática. Os dados foram utilizados da forma como estão apresentados na Tabela 6.4. O valor de τ foi 0,521 ($p < 0{,}001$), indicando uma correlação positiva.

Desse modo, os resultados dos três experimentos mostraram que, quanto mais complexa for a informação analisada, maior é a distância entre as amostras. Quando observamos as sentenças inteiras, a correlação foi inversa; quando observamos apenas os argumentos isolados, não houve correlação; porém, quando observamos apenas a distribuição de papéis semânticos nos dois *corpora*, tivemos uma correlação positiva. É importante ressaltar que, para esses experimentos, não consideramos o verbo presente nas sentenças ou ao qual os argumentos estavam associados. Observamos apenas as informações sintáticas e de papéis semânticos de maneira isolada.

6.2.4 Considerações sobre o Estudo-Piloto II

Neste segundo estudo-piloto, já dotados de ferramentas melhores para a análise dos dados, pudemos realizar uma anotação mais ampla, que abrangeu mais sentenças e mais verbos em relação ao estudo-piloto anterior. Com essa maior quantidade de dados, pudemos também recorrer a uma breve análise estatística, que auxiliou na discussão dos resultados e permitiu uma breve observação de nossa hipótese sobre a diferença de ranqueamento dos papéis semânticos nos *corpora*.

Observamos também ainda alguns ajustes que foram feitos à lista de papéis semânticos e à ferramenta de extração de estruturas de subcategorização. Apesar dos pequenos ajustes necessários que foram detectados, cremos que este estudo-piloto foi fundamental para dar sustento à realização de uma anotação em grande escala que é o nosso experimento final.

Tabela 6.5 – Estruturas sintático-semânticas mais frequentes nos dois *corpora*

Cardiologia	**Freq.**	**Diário Gaúcho**	**Freq.**
SUJEITO<THEME>	659	SUJEITO <AGENT>	733
OBJETO DIRETO <THEME>	507	OBJETO DIRETO <THEME>	494
ADJUNTO ADVERBIAL [em] <LOCATION>	356	SUJEITO <THEME>	338
SUJEITO <INSTRUMENT>	217	ADJUNTO ADVERBIAL [em] <LOCATION>	259
SUJEITO <RESULT>	190	SUJEITO <PATIENT>	171

7 Tarefa com Múltiplos Anotadores

Após a anotação realizada no segundo estudo-piloto, desenvolvemos um experimento paralelo relacionado à anotação com múltiplos anotadores. Até então, a anotação que realizamos tem sido feita por apenas um anotador: o autor deste trabalho. No entanto, existem estudos, principalmente em língua inglesa, que afirmam conseguir altos níveis de concordância entre vários anotadores para tarefas de anotação semântica. Um exemplo disso é o estudo de Hovy, Marcus et al. (2006), que apresentou uma metodologia para se obter 90% ou mais de concordância entre anotadores. Para tal, uniram-se os *frames* do PropBank aos significados da WordNet, de modo que o anotador apontava qual era o significado do verbo, e o *frame* era automaticamente selecionado e atribuído.

Atualmente, existe um estudo (FOSSATI, GIULIANO e TONELLI, 2013) que busca levar a anotação da FrameNet para múltiplos anotadores não especialistas. Para isso, foram simplificadas as definições de cada um dos elementos do *frame* e foram conduzidos experimentos em duas etapas: a primeira etapa envolvia apenas a desambiguação do verbo, que tomou como base o trabalho de Hong e Baker (2011), ainda que pareça similar ao experimento de Hovy, Marcus et al. (2006); a segunda etapa consistia em indicar quais argumentos deveriam ser anotados com os papéis semânticos associados ao significado predefinido do verbo. Enquanto a primeira etapa obteve resultados com mais de 90% de acurácia, a segunda etapa não teve resultados tão positivos[58].

O elemento em comum nos dois trabalhos apresentados é que já existe um recurso anterior que pode ser utilizado como base. Hovy, Marcus et al. (2006) tinham o PropBank com milhares de sentenças anotadas e só buscava expandir a anotação para outros *corpora*, e Fossati, Giuliano e Tonelli (2013) tinham a FrameNet e, da mesma forma, apenas buscam expandir a anotação para outros *corpora*. O que se percebe também é que ambos se apoiam nos significados de verbos apontados pela WordNet e não são de fato uma anotação de papéis semânticos. Ambos os estudos pedem para anotadores desambiguizarem o significado de verbos com base nos significados da WordNet, e, em seguida, atribuem, de maneira automática ou semiautomática, uma

[58] Os autores apenas relatam essa impressão negativa, mas não divulgam números sobre a segunda etapa da anotação, devido ao fato de ser um trabalho em andamento.

anotação semântica com base em algum recurso já existente. Desse modo, para cada significado do verbo, já existe um *frame* de papéis semânticos predefinido, que é aplicado assim que o anotador escolhe qual o seu significado. No caso de Fossati, Giuliano e Tonelli (2013), ainda havia uma maior preocupação em verificar, na segunda etapa, se a anotação realmente fazia sentido, mas, no caso de Hovy, Marcus et al. (2006), apenas era feita uma desambiguação do significado do verbo.

Em nosso caso, não existe ainda um recurso para o português que contenha a anotação de papéis semânticos descritivos, exceto a VerbNet.Br, que foi importada de maneira semiautomática da VerbNet e ainda não foi revisada para garantir uma baixa incidência de ruído. O único recurso que temos com uma boa base de verbos é a WordNet.Br, mas aplicar apenas o passo de desambiguização dos verbos sem ter os papéis semânticos vinculados a eles não bastaria para a anotação. Desse modo, é preciso deixar claro que o ponto de partida para o experimento descrito aqui é diferente dos experimentos já realizados por outros autores.

Nossa intenção com o experimento é observar se, para a criação de um recurso com anotação de papéis semânticos, seria possível utilizar desde o princípio a anotação de múltiplos anotadores com pouco treinamento ou se é melhor utilizar apenas um anotador com bastante treinamento (que é o método que está sendo utilizado neste estudo e também no desenvolvimento da VerbNet). Aproveitamos o experimento também para observar se existe concordância na delimitação entre argumentos e adjuntos (algo que não adotamos explicitamente até então, mas que poderia ser adotado, tendo em vista que recursos como o PropBank o fazem).

7.1 Procedimento

Para o experimento, foram selecionados dez anotadores linguistas (alunos de pós-graduação em Estudos da Linguagem na UFRGS) e 25 sentenças extraídas dos *corpora* apresentados no Capítulo 5. O treinamento foi básico, consistindo apenas em uma explicação sobre a tarefa e o assunto, e no fornecimento de um manual de anotação (que apresentava a lista de papéis semânticos apresentada no Anexo B e mais algumas informações básicas sobre o procedimento da anotação).

A anotação foi realizada em papel, o que não é ideal, mas foi o método mais simples de aplicar o experimento a várias pessoas garantindo que não houvesse comunicação entre elas. A estrutura das sentenças a serem anotadas foi similar à que apresentamos na Figura 5.2 (Capítulo 5), com a ressalva de que, por ser uma anotação

em papel, não havia uma lista de rolagem para escolher as sentenças (apenas a lista de consulta no manual de anotação).

Além da anotação dos papéis semânticos, como mencionamos, também fazia parte da tarefa distinguir cada um dos elementos anotados entre argumentos e adjuntos. Para isso, foi apresentada também uma breve explicação sobre as possíveis diferenças entre argumentos e adjuntos[59]. Este é um exemplo dos dados apresentados para anotação:

O resultado de o exame para investigar vestígios de pólvora em suas mãos , para saber se ele **utilizou** arma , teve resultado negativo .
SUJ = ele _____ ( ) Arg / ( ) Adj
OD = arma _____ ( ) Arg / ( ) Adj
Comentário:

Cada uma das sentenças a ser anotada era apresentada da forma como estava no banco de dados (com a devida explicação), seguida pelos argumentos (as abreviaturas estavam descritas no manual de anotação) com um espaço para escrever o papel semântico e a opção entre argumento ou adjunto. Por fim, acrescentamos um espaço para os comentários do anotador, que poderiam ser de qualquer ordem relacionada à anotação.

7.2 Cálculo da concordância entre múltiplos anotadores

Após a anotação ter sido realizada, para observar se houve concordância entre os anotadores, utilizamos cálculos com base no coeficiente π, um dos possíveis coeficientes utilizados para a observação de concordância entre anotadores. Como, em geral, utiliza-se o coeficiente κ para essa tarefa, discutimos a seguir os motivos que nos levaram a optar por outro coeficiente.

Artstein e Poesio (2008) apresentam uma longa discussão acerca de diversos coeficientes e testes utilizados para avaliar a concordância entre anotadores. Os autores chamam atenção para o fato de que há um problema de terminologia, pois o teste desenvolvido por Fleiss (1971) acabou sendo chamado de multi-κ, apesar de tomar como base o coeficiente π e, portanto, ter um pressuposto diferente. Como existe esse

[59] Como já mencionamos anteriormente, sabemos que a distinção entre argumentos e adjuntos é um assunto bastante controverso nas teorias gramaticais, por isso, nos limitamos a mostrar que a diferença se dá em relação ao quanto determinado elemento afeta o significado do verbo, saturando-o ou não.

problema de terminologia, Artstein e Poesio (2008) propõem que se utilize κ para o teste de Cohen (1960), multi-π para o teste de Fleiss (1971) e multi-κ para o teste de Davies e Fleiss (1982). Neste estudo, seguiremos a proposta de Arstein e Poesio (2008) em relação à terminologia.

Vejamos as principais diferenças entre os coeficientes. Segundo Artstein e Poesio (2008), os testes que usam π como base partem do pressuposto de que a distribuição das etiquetas não é uniforme, mas que a distribuição entre os anotadores o é. Assim, para um dado conjunto de etiquetas, cada uma delas tem a mesma probabilidade de ser utilizada por todos os anotadores, mas algumas têm mais chance de serem utilizadas do que outras. No caso dos testes que utilizam κ como base, tanto a distribuição das etiquetas quanto a distribuição das anotações é pressuposta como não uniforme, sendo assim, todas as distribuições são consideradas independentes entre si.

Por exemplo, dado um conjunto de etiquetas AGENT, THEME e LOCATION, e três anotadores A, B e C, um teste com base em π observa a totalidade dos dados e avalia uma distribuição não uniforme para as etiquetas (por exemplo, 50% dos argumentos receberiam a etiqueta AGENT, 30% THEME e 20% LOCATION), essa mesma distribuição será aplicada a todos os anotadores: A, B e C. No caso do κ, para esse mesmo conjunto de etiquetas e anotadores, seria avaliada a distribuição das anotações para cada um dos anotadores; desse modo, teríamos, por exemplo: 40% para AGENT, 35% para THEME e 25% para LOCATION no caso do anotador A; 60% para AGENT, 20% para THEME e 20% para LOCATION no caso do anotador B; e 45% para AGENT, 45% para THEME e 10% para LOCATION no caso do anotador C. Assim, a concordância de κ leva em conta não somente a distribuição das etiquetas, mas também a anotação feita por cada um dos anotadores. Conforme apontam Artstein e Poesio (2008), na teoria, essa diferença é bastante grande, porém, na prática, ela perde um pouco a sua força, pois os coeficientes π e κ resultam em valores muito próximos, e, no caso de multi-π e multi-κ, essa diferença varia ainda menos, pois ela tende a se extinguir conforme o número de anotadores aumenta.

Como temos mais de dois anotadores, a diferença entre os coeficientes é muito pequena, mas, ainda assim, é importante que se decida por um ou outro em virtude dos pressupostos assumidos. Neste estudo, assumem-se os pressupostos de π, pois estamos avaliando a confiabilidade dos dados anotados por vários anotadores, de modo que as etiquetas devem ter uma distribuição não uniforme, mas os anotadores deveriam anotar de modo consistente e similar. Sendo assim, para verificar a concordância entre os

anotadores e também entre os pares de anotadores, empregamos, respectivamente, os testes multi-π e π. A observação da concordância entre os pares de anotadores serve principalmente para detectar *outliers* (isto é, anotadores que possivelmente não entenderam a tarefa ou que realizaram a anotação sem prestar muita atenção aos dados) e para poder dar mais confiabilidade ao multi-π. Os cálculos foram levados a cabo por meio de uma ferramenta presente no mwetoolkit (RAMISCH, VILLAVICENCIO e BOITET, 2010; 2010) que calcula vários coeficientes de concordância.

7.3 Resultados da anotação com múltiplos anotadores

Primeiramente, observamos a distinção entre argumentos e adjuntos, que consideramos ser uma tarefa mais simples (principalmente por haver apenas duas possibilidades de anotação), para observar se algum dos anotadores se caracterizava como *outlier*. Para essa observação, comparamos os anotadores em pares calculando o π entre eles.

A distinção entre argumentos e adjuntos, apesar de ser bastante controversa no caso de alguns verbos, deveria ser bastante simples na maioria dos casos. Por exemplo, em 7.a, é possível perceber que o sujeito (*O PT*) e o objeto direto (*um projeto de lei*) são argumentos, por serem necessários para que o verbo expresse seu significado completo, enquanto o adjunto adverbial (*no Congresso*) aparece apenas para acrescentar uma informação que não depende do verbo.

7.a. *O PT* ***apresentou*** *no Congresso um projeto de lei que cria contribuição social sobre fortunas.*

Como, em grande parte dos casos, a distinção é razoavelmente clara, esperávamos um alto nível de concordância nessa tarefa. Porém, não foi isso que observamos. Ao analisar os valores de π para os pares de anotadores utilizando apenas dados da distinção entre argumentos e adjuntos, percebemos que três anotadores apresentaram níveis baixos de concordância com os demais anotadores, a ponto de haver valores negativos entre eles (o que indica discordância). Uma das possíveis explicações para isso é que talvez eles não tenham compreendido a tarefa, ou simplesmente fizeram a anotação com pressa, deixando de ponderar adequadamente cada uma das instâncias a ser anotada. Dado o baixo nível de concordância entre esses

anotadores em relação aos demais, o multi-π também foi baixo, com um valor de 0,315020.

Com a retirada desses três *outliers*, o valor do coeficiente multi-π aumenta para 0,553020, mas continua abaixo dos 0,8, apontados por Neuendorf (2002, *apud* Arstein e Poesio, 2008) como mínimo necessário para que se considere que haja uma boa concordância. Assim, duas conclusões vêm imediatamente à mente: ou a tarefa não estava clara para os anotadores, ou a anotação de argumentos e adjuntos não é tão simples quanto imaginávamos.

Passemos então para a tarefa mais importante, que é a anotação de papéis semânticos. Para o cálculo do multi-π dessa tarefa, também retiramos os mesmos três *outliers*, pois assumimos que não havia por que confiar nos seus resultados em uma tarefa mais complexa, que envolve mais de trinta possíveis anotações, e não apenas duas. Assim, dentre os sete anotadores restantes, obtivemos um multi-π de 0,253407 (multi-κ = 0,256954). Esse valor é extremamente baixo, de modo que se pode dizer que praticamente não houve concordância entre os anotadores.

Observando-se as anotações individuais, percebe-se que houve alguns pontos de convergência, principalmente na atribuição dos papéis semânticos AGENT, MOMENT, LOCATION e, em alguns casos, THEME. No entanto, quando outros papéis semânticos eram requeridos, os anotadores discordaram de modo a ter, em alguns casos, uma anotação diferente para cada anotador. Em mais de um caso, em uma mesma sentença, houve total concordância em um argumento, mas discordância nos demais. Por exemplo, no caso do Exemplo 7.a, os 10 anotadores concordaram que o sujeito *O PT* desempenha a função de AGENT, já o objeto direto *um projeto de lei* teve apenas 5 anotadores concordando com o papel THEME, e o adjunto adverbial *no Congresso* contou com apenas 6 anotadores optando por LOCATION.

Outras sentenças não tiveram concordância em nenhum dos argumentos. No Exemplo 7.b, o sujeito *A versão religiosa* recebeu 4 anotações como AGENT (ainda que o sintagma indique um participante não animado e nem volitivo) e 3 como THEME (a escolha de THEME está, de certa forma, correta, pois todos os requisitos do papel são preenchidos; porém, por se tratar, nesse caso, de um verbo de estado, o objeto assume papel de THEME, e o sujeito é um PIVOT), enquanto o objeto indireto *com as mulheres Jaca ou Melancia* foi anotado como THEME por 4 anotadores e como ATTRIBUTE por 3.

7.b. *A versão religiosa não **conta** com as mulheres Jaca ou Melancia , mas todas=as velocidades estão lá , em a música* .

Apenas o Exemplo 7.c apresentou uma maior concordância entre os anotadores no que diz respeito aos dois argumentos. O sujeito *ele* foi reconhecido como AGENT pelos 10 anotadores (o que confere com o verbo de ação-processo e o sujeito animado e volitivo), enquanto o objeto direto *arma* foi anotado por 8 anotadores como INSTRUMENT.

7.c. *O resultado de o exame para investigar vestígios de pólvora em suas mãos , para saber se ele **utilizou** arma , teve resultado negativo* .

Em princípio, essa alta concordância pode parecer bom, mas acaba sendo uma prova de que é possível também concordar no erro. Neste caso específico, o argumento *arma*, anotado pela maioria como INSTRUMENT, é um caso de THEME, tendo em vista que INSTRUMENT, de acordo com as definições do manual de anotação (Anexo B), é o participante utilizado na realização de uma ação, mas não é o argumento sobre o qual a ação (evento) incide. O teste indicado pode ter sido um dos motivos que levou a uma quase unanimidade na atribuição do papel INSTRUMENT, tendo em vista que indicamos substituir o argumento por *usando x*. No caso do Exemplo 7.c, a substituição resultaria no Exemplo 7.d, o qual é incompreensível. Porém, o treinamento básico fornecido aos anotadores pode ser um dos motivos pelo qual a compreensão desse papel semântico foi errônea.

7.d. *O resultado de o exame para investigar vestígios de pólvora em suas mãos , para saber se ele **utilizou*** usando* *arma , teve resultado negativo* .

7.4 Considerações sobre a anotação com múltiplos anotadores

Existem vários motivos que podem ter levado a uma concordância baixa entre anotadores. É possível, por exemplo, que o material fornecido não tenha sido detalhado o suficiente para a realização da tarefa, ou que os anotadores não tenham entendido claramente o que deveria ser feito. Taboada e Das (2013), por exemplo, indicam que chegar a um consenso após muito treinamento é algo bem diferente de chegar a um consenso após explicar rapidamente o método a um recém-chegado num projeto. Cremos que o principal fator envolvido é a complexidade da tarefa, que requer um

treinamento muito bem-desenvolvido para que se possa chegar a níveis maiores de concordância.

Como pode ser visto no trabalho de Hovy, Marcus *et al.* (2006), a solução encontrada para se obter alto nível de concordância entre anotadores foi simplificar a tarefa ao máximo possível. Para simplificar a tarefa, no entanto, seria necessário que já tivéssemos um recurso existente, do qual pudéssemos tirar insumos para a anotação. Porém, estamos tratando aqui justamente do desenvolvimento de um recurso que ainda não existe para o português, e não da expansão do mesmo.

Algo que poderia aumentar a concordância seria uma interface de anotação mais bem-desenvolvida e mais amigável do que uma folha de papel e um manual de anotação. No entanto, não cremos que tal material conseguiria aumentar o valor da concordância (multi-π) de 0,25 para mais de 0,8 (que seria um valor aceitável para o desenvolvimento de um recurso).

Embora o processo seja complexo e necessite treinamento cuidadoso dos envolvidos, o trabalho com múltiplos anotadores geralmente é mais produtivo, acelerando o trabalho e rendendo bons resultados em relação à anotação de um único anotador. Além disso, quando se utilizam múltiplos anotadores, os dados levantados representam uma amostra da língua, em vez de apenas a descrição de um linguista. Porém, a baixa concordância averiguada neste experimento faz com que nossa tendência seja por manter a anotação com apenas um anotador, que teve um maior treinamento com o estudo de outros recursos. Essa opção de continuar com apenas um anotador, ainda que não seja ideal, é o que dispomos no momento, pois o treinamento completo de um anotador requereria recursos dos quais não dispomos para a realização desta tese. Além disso, depois que existir uma anotação, trabalhos futuros podem se aproveitar do material que existe para aperfeiçoá-lo ou levá-lo adiante.

8 Desenvolvimento do VerbLexPor

Depois de termos realizado dois estudos-piloto e ponderado sobre alguns pontos fortes e fracos em nossa metodologia, passamos à etapa final do desenvolvimento do recurso léxico com informação de papéis semânticos, que é a principal proposta deste trabalho. Primeiramente, achamos interessante dar um nome ao recurso. Sendo assim, chegamos ao nome Léxico de Verbos com Dados Sintáticos e Semânticos do Português Brasileiro, e adotamos a sigla VerbLexPor, para facilitar.

Neste capítulo, finalmente mostraremos e discutiremos em detalhes a nossa lista de papéis semânticos, mas antes discutiremos as modificações que foram realizadas no extrator de estruturas de subcategorização, que será a primeira seção deste capítulo. Em seguida, após a lista, comentamos muito rapidamente alguns aspectos da metodologia, que permaneceu praticamente inalterada em relação ao segundo estudo-piloto, e apresentamos algumas informações gerais sobre o VerbLexPor. Ao final, comparamos o VerbLexPor com dois outros recursos existentes para o português (VerbNet.Br e PropBank.Br), indicando os procedimentos realizados para possibilitar a comparação com cada um deles.

Este capítulo será apenas expositivo, de modo que reservaremos capítulos posteriores deste trabalho para discutirmos os dados. A única seção neste capítulo voltada à discussão será a que apresenta os papéis semânticos. Sem mais, passemos ao extrator.

8.1 Modificações realizadas no processo de extração

Como foi relatado no Capítulo 6, após realizarmos o segundo estudo-piloto, percebemos que as regras utilizadas pelo sistema de extração de estruturas de subcategorização não abrangiam todas as etiquetas do PALAVRAS (Bick, 2000) em relação aos elementos que nos interessavam. Assim, fizemos um estudo mais aprofundado de todas as suas etiquetas e modificamos as regras do sistema de extração de modo que todas as informações pertinentes fossem extraídas para a anotação final e fossem organizadas da maneira que achássemos mais apropriada.

Todas as regras foram sistematizadas num formato “se X, então Y”. Se não existir uma regra para determinado elemento, ele não é extraído como argumento do verbo. Basicamente, o sistema observa se um determinado elemento está presente na anotação do PALAVRAS e, se estiver, extrai a informação e atribui um novo rótulo. A

totalidade das regras é apresentada na Tabela 8.1 da seguinte maneira: a primeira coluna representa a etiqueta que é buscada entre as informações fornecidas pelo *parser* PALAVRAS; a segunda coluna indica a anotação (NP, PP etc.) mostrada na estrutura de subcategorização; a terceira coluna indica a classificação sintática atribuída ao argumento extraído, de acordo com a etiqueta encontrada na primeira coluna; por fim, a quarta coluna apresenta o índice de relevância atribuído ao argumento[60].

Tabela 8.1 – Regras utilizadas pelo extrator de estruturas de subcategorização para o desenvolvimento do recurso, apresentadas em ordem de execução

Se (etiqueta)	**Então (estrutura de subcategorização)**	**Classificação Sintática**	**Índice de Relevância**
SUBJ, ou ICL-SUBJ, ou FS-SUBJ	SUBJ	SUJEITO	1
DAT	DAT	OBJETO INDIRETO PRONOMINAL	3
ACC-PASS, ou refl	REFL	OBJETO REFLEXIVO	3
ACC	NP	OBJETO DIRETO	4
ICL-ACC, ou FS-ACC	OCL	OBJETO DIRETO ORACIONAL	4
SC e PRP, ou ICL-SC e PRP, ou FS-SC e PRP, ou OC e PRP, ou ICL-OC e PRP, ou FS-OC e PRP, ou PRED e PRP, ou ICL-PRED e PRP	PR[prep.]	PREDICATIVO[prep.]	5
SC, ou ICL-SC, ou FS-SC, ou OC, ou ICL-OC, ou FS-OC, ou PRED, ou ICL-PRED	PR	PREDICATIVO	5
PIV ou SA	PP[prep.]	OBJETO INDIRETO[prep.]	5
PASS	PP[prep.]	AGENTE DA PASSIVA[prep.]	5
ADVL, mas não ADV[61]	PP[prep.]	ADJUNTO ADVERBIAL[prep.]	6

Essa lista pode causar algum estranhamento principalmente na coluna da estrutura de subcategorização, pois usamos algumas notações que não são tradicionais, principalmente se olharmos para trabalhos como os de Preiss, Briscoe e Korhonen

[60] Lembramos que o índice de relevância é um fator adotado pelo sistema para organizar os argumentos do verbo em uma determinada ordem na estrutura de subcategorização e na interface de anotação. O número de relevância "2" (que não está presente na tabela) representa a posição do verbo.

[61] Esta regra procura por adjuntos adverbiais que não sejam apenas advérbios (ou seja, a regra deixa passar advérbios como, por exemplo, **não**, **simplesmente**, **facilmente** etc.).

(2007), e Messiant (2008), entre outros, que utilizam apenas notações como NP, PP e ADJ e ADV. Por isso, fazem-se necessárias algumas explicações.

Primeiramente, optamos por explicitar a função sintática de sujeito na estrutura de subcategorização em vez de apenas deixar o tradicional NP. Essa opção se deu para marcar a posição do sujeito, tendo em vista que a estrutura de subcategorização pode tanto apresentar uma estrutura canônica quanto a estrutura normal da oração (em que o sujeito não precisa aparecer na primeira posição). Além disso, o sujeito nem sempre precisa ser um NP (ainda que sempre possa ser substituído por um), podendo ser uma oração subjetiva. Outra notação fora do comum é o caso de DAT, que indica a presença de um pronome pessoal na posição de objeto indireto, achamos importante manter essa diferença, pois nem todos os objetos indiretos podem ser pronominalizados[62]. Optamos por não fazer esse tipo de distinção, porém, reconhecemos que é importante destacar os verbos que permitem pronominalização do objeto indireto. Por motivo parecido, optamos por explicitar quando o objeto direto é oracional, pois alguns verbos não autorizam o uso de uma oração nessa posição. A marca do reflexivo também frequentemente é deixada de fora da estrutura de subcategorização, tendo em vista que, muitas vezes, ela faz parte do próprio verbo. Aqui optamos por explicitá-la, justamente porque temos papéis semânticos para tratar dos casos em que o pronome reflexivo *se* faz parte do verbo. Por fim, distinguimos na estrutura de subcategorização os predicativos, geralmente representados como ADJP (sintagma adjetival), pois nem sempre são representados por adjetivos e, às vezes, ocorrem com preposições.

Com essa lista de regras, o único problema que restou foram os erros de anotação do PALAVRAS. Estes, de fato, permaneceram sem modificação ao longo do tempo. Alguns erros que tivemos de anotação incluem, por exemplo, a anotação automática da sigla IAM (infarto agudo do miocárdio) como verbo *ir* (na 3ª pessoa do plural do pretérito imperfeito) no *corpus* de Cardiologia, o que representou grande parte das ocorrências desse verbo no *corpus* em questão. Porém, nem todos os erros do PALAVRAS são tão gritantes; alguns são mais sutis, como, por exemplo, o reconhecimento errado de um objeto indireto como objeto direto. Numa impressão subjetiva que tivemos dos dados dos *corpora*, mesmo no caso do Diário Gaúcho, que é jornalístico, o *parser* não parece ter atingido os 96-97% apontados por Bick (2000).

[62] Isso justifica, por exemplo, a distinção feita por Bechara (1999) entre complemento relacional e objeto indireto, a qual preferimos ignorar para não termos que abrir ainda mais possibilidades de ruídos por parte do *parser*.

Agora que vimos a lista de regras do extrator, vejamos agora como ficou constituída a lista de papéis semânticos, que sofreu apenas algumas alterações após o segundo estudo-piloto.

8.2 Lista de papéis semânticos

Nossa lista de papéis semânticos não chegou a passar por modificações profundas em relação àquela utilizada no segundo estudo-piloto. Apenas introduzimos alguns papéis semânticos para tratar mais amplamente dos adjuntos adverbiais. Esses papéis semânticos adicionais foram retirados do manual de anotação do PropBank[63].

Na sequência deste texto, trataremos de cada um dos papéis semânticos de maneira individual ou em conjunto com papéis similares, apresentaremos uma definição breve e também exemplos extraídos dos *corpora*. As definições, por vezes, podem não parecer claras ou mesmo não ter uma aceitação unânime, principalmente em relação aos exemplos apresentados. É importante ressaltar que elas servem mais como norteadoras da anotação, sendo que, às vezes, é difícil determinar precisamente os limites de cada papel semântico. Existem, é claro, os exemplos prototípicos, que podem ser apresentados e se encaixam perfeitamente nas definições, mas os textos reais não são feitos só de protótipos. Por isso, é importante que se leve em consideração esta breve advertência.

Nas definições, também é possível depreender a hierarquia dos papéis semânticos[64], pois sempre iniciamos as definições por meio de um papel semântico que serve de hiperônimo. Desse modo, quando dizemos que um AGENTE é um "*ator* que realiza a ação", estamos indicando que o papel AGENTE está diretamente subordinado ao papel ATOR, o qual não faz parte da nossa anotação, pois é apenas um papel estrutural (apenas faz parte da hierarquia de papéis, mas não é usado para anotação).

Os exemplos foram retirados diretamente dos *corpora* e não foram modificados. Desse modo, é comum que contrações do tipo *pelo*, *da* etc. estejam escritas na forma descontraída: *por o*, *de a*. Além disso, os sinais de pontuação também são isolados por meio de espaços em relação à palavra da esquerda e alguns sintagmas reconhecidos pelo *parser* como unidades linguísticas são ligados por meio de sinais de igual (=). Os

---

[63] Disponível em: https://verbs.colorado.edu/propbank/EPB-Annotation-Guidelines.pdf. Acessado em: 27/02/2014.
[64] A hierarquia completa, incluindo os papéis semânticos apenas estruturais, pode ser vista no Anexo D.

elementos relevantes, como o argumento e o verbo, são destacados com negrito e sublinhado, respectivamente. Sem mais, vamos à lista:

AGENTE

Definição: ATOR que realiza a ação.

Comentário: O AGENTE pode ser reconhecido pelos traços prototípicos de desencadeador da ação, volição e animação, porém, existem casos em que a volição não existe.

Possíveis métodos/testes de detecção: Para saber se um argumento é AGENTE, não existem métodos específicos. Em geral, é preciso avaliar se o verbo expressa uma ação ou ação-processo, pois, nesse caso, na voz ativa, existe a possibilidade de o sujeito ser AGENTE. Nesse caso, é preciso atentar para os elementos de animação e desencadeamento, pois eles serão os principais definidores.

Exemplos:

Quando a saudade é bandida , **nós** a matamos em os lençóis .

A Rainha=do=Carnaval de Porto=Alegre , Emely=Ribeiro , mostra quatro fantasias criadas por Evandro .

Assim , a mortalidade em 57 pacientes que usaram NA foi de 62 % contra 82 % em aqueles **que** usaram dopamina e adrenalina , com risco relativo de 0,68 4 .

AGENTE LOCATIVO

Definição: AGENTE que ocorre quando o AGENTE *real* é referido por meio de metonímia, e o argumento é um lugar.

Comentário: Anotamos como AGENTE LOCATIVO aqueles AGENTES que poderiam, por meio de alternância, aparecer na forma de adjunto adverbial de lugar.

Possíveis métodos/testes de detecção: Colocar um AGENTE no lugar do AGENTE LOCATIVO e transferir o AGENTE LOCATIVO para uma posição de adjunto adverbial ou adnominal. Por exemplo: **Nenhum de os departamentos** resolveu o problema. (**Ninguém** *em nenhum dos departamentos* resolveu o problema.)

Exemplos:

Uma Casa **que** trabalha com leis deveria zelar por o integral cumprimento de esta , que é a lei fundamental de a vida .

Por=meio=de a sua assessoria=de=imprensa , a Secretaria=Municipal (Smed) afirma que o **município** vem trabalhando para ajustar as vagas , além=de promover ações integradas com a rede estadual , via Central=de=Matrículas .

Hoje , **a Coluna=do=Gugu** manda abraço para os leitores :

TEMA

Definição: OBJETO que não é modificado pelo evento, podendo sofrer deslocamento ou não.

Comentário: O papel TEMA talvez seja o que menos tenha informação consigo. Ele pode ser concreto ou abstrato, animado ou não, pode ser outro evento etc. Este é um dos papéis que talvez mereça um estudo mais aprofundado para delimitar os tipos de TEMA que existem.

Exemplos:

A entrada maciça de Ca2 + que ocorre em a célula durante a reperfusão se acompanha de necrose com banda de contração , caracterizada por miofilamentos hipercontraídos e importante dano mitocondrial (38-42) .

Aparece quando **o infarto** está acontecendo .

Duas garçonetes participam de um concurso em que a escolhida ganha **uma bolada em=dinheiro** .

TÓPICO

Definição: TEMA de uma conversa ou mensagem.

Comentário: Estará envolvido nos verbos de comunicação.

Exemplos:

As lesões arterioscleróticas podem ser encontradas em pacientes que não apresentam fatores de risco ou outras causas que expliquem **o aparecimento de estas lesões** .

O coronel Paulo=Roberto=Mendes , subcomandante-geral de a Brigada=Militar , reconhece o bairro como área conflagrada e admite **a falta de viaturas de o 20º BPM** .

O tenente-coronel informou ainda **que pretendia realizar operações em outras áreas de a Zona=Norte ainda em a noite de sexta-feira** .

PACIENTE

Definição: OBJETO modificado (implícita ou explicitamente) pelo evento.

Comentário: Sua marca é ser afetado pelo evento.

Exemplos:

Mesmo em as DIC crônicas , **a letalidade** atingiu 5,4 % .

A taxa de mortalidade aumenta com a progressão de a insuficiência=cardíaca ;

Aí , **o 4-4-2** virou 4-5-1 .

COAGENTE / COTEMA / COPACIENTE

Definição: AGENTE, TEMA ou PACIENTE que ocorre junto com outro AGENTE, TEMA ou PACIENTE, respectivamente.

Comentário: É o caso de verbos que admitem dois argumentos do mesmo tipo. Ocorrem às vezes na forma de adjuntos de companhia.

Possíveis métodos/testes de detecção: Inverter de posição com o AGENTE, TEMA ou PACIENTE, respectivamente, e ver se o significado da sentença permanece o mesmo.

Exemplos:

COAGENTE = Casou- se com **Fernando (Tato=Gabus)** .

COTEMA = O aumento de o risco para ocorrência de DM acompanhou **a elevação de o consumo de álcool** :

COPACIENTE = Verifica- se , conforme o esperado , que a prevalência de Ha aumenta **com a idade** .

PIVÔ

Definição: OBJETO que aparece juntamente com TEMA, mas que tem maior importância que este, diferenciando-se assim de COTEMA.

Comentário: O PIVÔ tem a mesma função de TEMA, apenas é mais importante que este devido ao foco do verbo, não podendo mudar de posição sem alterar o significado.

Possíveis métodos/testes de detecção: Inverter de posição com o TEMA e ver se o significado da sentença é modificado.

Exemplos:

Além=disso , mesmo=que alguns de estes resultados sejam verdadeiros positivos , devem representar alterações coronarianas com melhor prognóstico , visto=que **nenhum de eles** apresentou eventos coronários em o período de seguimento .

A adequada heparinização depende **de uma estratégia padronizada** , sendo importante a auditoria de a prática clínica e ações padronizadas que permitam melhorar a qualidade de o atendimento a os pacientes .

Os outros dois pacientes aguardam o mesmo tipo de correção operatória .

CAUSA

Definição: ATOR que representa o motivo de ocorrência de um evento.

Comentário: A CAUSA funciona como desencadeador da ação, mas não é animado nem tem volição.

Possíveis métodos/testes de detecção: Não existem, porém, é possível identificar se o verbo é de ação ou ação-processo e, com isso, estabelecer se seria possível haver um AGENTE. Nesse caso, se não houver nem volição, nem animação no argumento que poderia ser o AGENTE, os indícios apontam para um argumento do tipo CAUSA.

Exemplos:

Outros antioxidantes podem atuar em a eliminação de os radicais alquilperoxil , de entre eles pode- se citar os Cf , a UQH2 e o ß-caroteno 24,29,31.

Elas têm características heterogêneas , **algumas mutações** causam hipertrofia septal e **outras** causam hipertrofia apical (Arg162Trp e Gly203 Ser) .

Em a manhã de domingo , **o rompimento de uma tubulação em a esquina de as ruas Câncio=Gomes e Voluntários=da=Pátria** provocou a falta de água .

INSTRUMENTO

Definição: PROCEDIMENTO que é utilizado para realizar uma ação.

Comentário: O INSTRUMENTO muitas vezes pode assumir a posição de sujeito, para ocultar o real AGENTE.

Possíveis métodos/testes de detecção: Ocorre com verbos de ação e ação-processo. Em geral, é possível transformar o argumento em um adjunto adverbial introduzido pela preposição “com”.

Exemplos:

Estatísticas de a polícia gaúcha revelam uma mudança em o criminoso mundo de as drogas :

Já os pacientes com prótese , provavelmente já são acompanhados **por serviços mais especializados** , procurando logo o hospital terciário na=presença=de algum sinal ou sintoma .

Três pacientes que apresentaram disfunção ventricular e TVNS foram tratados também **com amiodarona** .

EXPERIENCIADOR

Definição: PACIENTE que sofre uma alteração nos sentidos ou que expressa um sentimento pessoal.

Comentário: Representa a ideia de senciência do papel PACIENTE. O participante tem de ser senciente para poder ser um EXPERIENCIADOR.

Exemplos:

Testemunhas ouviram quatro tiros e viram um Chevette perto=de o local de o crime .

Lembre que , depois de te conquistar , **eles** geralmente esquecem o caminho de=volta .

O técnico que deu a Libertadores de 1995 a o Tricolor , acha que é hora de o clube investir em o time :

ESTÍMULO

Definição: CAUSA que provoca uma reação em alguém.

Comentário: Geralmente é a causa por trás de uma experiência e, portanto, está associado ao papel de EXPERIENCIADOR. É o papel de desencadeador dos verbos de sentimento, mas não necessariamente é animado ou volitivo.

Exemplos:

Ady sofre **de glaucoma e catarata** .

Laura fica magoada **de ser a última a saber de o noivado de Bruna** .

Fico triste **por minha filha** .

RECIPIENTE

Definição: ALVO receptor de algo concreto que parte de um LUGAR INICIAL e chega até ele.

Comentário: Este papel pode se aplicar também a um interlocutor, que recebe uma mensagem.

Exemplos:

Uma=vez identificados desvios significativos , o aluno / cliente deve ser encaminhado **para um profissional especializado** , pois são casos de responsabilidade de o ortopedista e de o fisioterapeuta .

Gioconda conta **para Júlia** que Barreto doou sangue para salvar a vida de o neto .

Zidane pede **Gislaine** em casamento .

BENEFICIÁRIO

Definição: ALVO que obtém uma vantagem ou desvantagem gerada pelo evento.

Comentário: Este papel é muito parecido com o de RECIPIENTE, e a distinção entre eles é complicada em casos não prototípicos.

Exemplos:

Imagine que preparei uma arapuca **para a minha mulher** .

Em pacientes com insuficiência=cardíaca congestiva a suspensão repentina pode provocar arritmias e piora de o quadro .

As aulas são abertas **a vestibulandos interessados em aprimorar seus textos** .

ALVO

Definição: OBJETO para o qual uma ação é realizada ou que é tido como receptor de algo.

Comentário: Papel criado para servir como uma interface entre RECIPIENTE e BENEFICIÁRIO. Um ALVO deve ser animado ou poder ser interpretado como tal na oração (p.ex.: **A casa** ganhou um novo visual. — **A casa** não é um argumento animado *per se*, mas, pelo uso do verbo **ganhar**, ela recebe esse traço do uso metafórico).

Exemplos:

Zebra recebe a ajuda de amigos para concretizar o sonho de participar de uma corrida de cavalos .

Foram avaliados 106 pacientes com reperfusão (Cr) e 48 pacientes sem reperfusão (Sr) **que** receberam terapia trombolítica em a fase aguda de o infarto .

Em o total , **1.538 pacientes** receberam atorvastatina 80 mg e 1.548 , placebo .

RESULTADO

Definição: PROCEDIMENTO que passa a existir por consequência de um evento.

Exemplos:

À=medida=que os episódios de atividade reumática se manifestam , a doença reumática crônica progride e se caracteriza , principalmente , por fibrose e calcificação valvar , que causam **deformidades estruturais em as valvas** .

Uma avaliação fisioterapêutica minuciosa deve ser realizada para se estabelecer o quadro em que o paciente se encontra , para indicação de um ou outro recurso ;

Latino fez **show** em Canoas em o último dia 4 e seguiu para a noitada em o Bar=Alternativo , em Novo=Hamburgo , a o som de o Fat=Duo .

PRODUTO

Definição: RESULTADO concreto.

Comentário: Um PRODUTO é sempre concreto, representando um objeto que é produzido por uma ação. O problema deste papel é que ele tem uma grande sobreposição com o papel RESULTADO. Em nossos dados, não detectamos nenhum PRODUTO, de modo que não temos exemplos para este papel. Um exemplo prototípico de PRODUTO seria *casa* em "João construiu uma casa".

MATERIAL

Definição: FONTE que representa o que foi usado para a geração de um PRODUTO ou RESULTADO.

Comentário: Pode ser entendido como a matéria-prima.

Possíveis métodos/testes de detecção: Geralmente, pode ser detectado testando-se se o argumento responde à questão: de que é feito?

Exemplos:

Faço **de a vida** uma canção .

Sua riqueza e sua ostentação ignoram Jesus=Cristo , cientista criador de o Universo , que faz **de o embrião** sua maior criação .

Em o especial , 32 artistas apresentam os seus maiores sucessos e fazem **de a passagem de o ano** uma grande festa sertaneja .

VALOR

Definição: OBJETO que é um número.

Exemplos:

Hipocinesia e acinesia receberam **2 e 3** respectivamente .

Por a resolução de o CMN , a nova forma de cálculo permite que a TR fique , em o mínimo , **em zero** .

O gaúcho paga , no=máximo , **24 %** em impostos sobre seus rendimentos .

VARIAÇÃO

Definição: VALOR que representa uma variação positiva ou negativa mensurável.

Exemplos:

Em revisão (25) de 21 programas de exercício para pacientes com claudicação , observou- se que após o exercício físico contínuo a distância para o início de a dor aumentou em média 179 % e a média de a distância máxima tolerada aumentou **em 122 %** .

A passagem deve passar **de R$ 2 para R$ 2,10**

O índice calculado semanalmente por a Fundação=Getúlio=Vargas subiu **0,30 %** em a primeira prévia .

ATIVO

Definição: VALOR que representa dinheiro.

Exemplos:

O empréstimo deverá custar **cerca=de R$ 1,5 milhão** .

Quando Rafael ganhava **R$ 400 mensais** (R$ 100 por=pessoa em=casa) , a família recebia o benefício .

O outro , André , teria pedido **R$ 50 mil** para não revelar a história .

FONTE

Definição: PROCEDIMENTO que representa a base, referência ou ponto de partida de um evento.

Comentário: Diferencia-se de LUGAR INICIAL pelo fato de não ser um lugar, mas sim um elemento ou grupo de elementos que são tomados como inspiração, cobaias, base de comparação etc.

Exemplos:

A cintilografia cardíaca com gálio-67 foi negativa em 14 pacientes (73,7 %) e positiva em 5 pacientes (26,3 %) , **de os=quais** 4 apresentaram grau discreto de captação cardíaca (+) e apenas 1 apresentou grau moderado (+) .

É algo **de o=qual** você não poderá fugir !

Renato aproxima- se e tira Joana **de a discussão** .

FINALIDADE

Definição: PROCEDIMENTO que indica um objetivo a ser atingido.

Possíveis métodos/testes de detecção: Geralmente, responde à pergunta: para quê?

Exemplos:

O duplo-produto (DP) é considerado o melhor indicador não-invasivo **para se avaliar o trabalho de o miocárdio , durante o repouso ou esforços , sendo bastante eficiente como indicador de sobrecarga cardíaca em exercícios de=força 8,11** .

Ela vê o filho e corre **para amparar- lo** :

Procuro uma linda mulher , entre=25=e=35 anos , **para viver um grande amor com alegria , paz e fidelidade** .

RECÍPROCO

Definição: ACESSÓRIO que marca quando um pronome reflexivo indica uma ação realizada reciprocamente por dois AGENTES.

Comentário: Se o pronome reflexivo implicar em um novo papel semântico, o papel RECÍPROCO não será utilizado (consulte a Tabela 8.2, mais adiante). Este papel semântico serve como um indicador de função gramatical, assim como veremos nos casos de VERBO e SE PASSIVO, a seguir.

Possíveis métodos/testes de detecção: É preciso observar se o pronome reflexivo refere-se ao sujeito, e este é plural.

Exemplos:

Ele me ligou em o dia 28 de dezembro de 2006 , nos conhecemos e não **nos** separamos mais .

Então , já vai agilizando com a tua galera , que é certo que a gente **se** encontra lá !

Eduardo e Débora encontram- **se** em o cinema .

SE PASSIVO

Definição: ACESSÓRIO que marca quando o pronome reflexivo é utilizado como partícula apassivadora (consulte a Tabela 8.2, mais adiante).

Comentário: Este papel semântico serve como um indicador de função gramatical.

Exemplos:

Avaliar quando **se** devem realizar exames de cintilografia de perfusão de o miocárdio (CPM) , baseando- se em informações objetivas obtidas de o teste ergométrico e de a análise de os fatores clínico-epidemiológicos para doença arterial coronária (DAC) .

Realizou- **se** a análise de os coeficientes de correlação simples de as variáveis estudadas .

Gostaria de saber se eu vou arranjar emprego , amor e se os meus caminhos vão **se** abrir em 2008 .

VERBO

Definição: ACESSÓRIO que marca o uso como parte ou como portador da predicação.

Comentário: Este é um papel que indica uma função gramatical e está atrelado a verbos-suporte, em que o complemento serve como indicador do evento; a reflexivos, no caso em que estes fazem parte do verbo (consulte a Tabela 8.2, mais adiante); a sujeitos que não deveriam estar presentes (caso de verbos impessoais, em que o sujeito foi forçado automaticamente durante o processamento do texto [por exemplo, "Faz treze anos", em que foi forçado um sujeito oculto inexistente]).

Exemplos:

Esses dados dão **suporte** a a idéia de que em algumas condições de resistência a a insulina a via metabólica de insulina pode estar inibida enquanto a via de crescimento celular está preservada , resultando em a hipertrofia miocárdica induzida por a hiperinsulinemia .

A comparação de os resultados de o GC e GT obtidos em o terceiro mês de este estudo encontra- **se** em a Tabela 4 .

Por=trás de tudo que fazem , há uma filosofia de vida , um conjunto de princípios básicos de os=quais não abrem mão .

ATRIBUTO

Definição: ACESSÓRIO que serve para qualificar um PACIENTE, TEMA ou AGENTE presente na oração.

Comentário: Geralmente o atributo está relacionado a predicativos.

Exemplos:

Sabrina entra em a igreja , e Barretinho chega **sujo de lama** , gritando para que ela não se case .

Como não temos esse fármaco ideal , novas pesquisas se fazem **necessárias** .

Os anos não passam porque seus dias , ensolarados ou sombrios , permanecem **arquivados em a memória** .

TEMPO INICIAL

Definição: TEMPO que indica quando um evento se inicia.

Comentário: Aplica-se a adjuntos de tempo.

Possíveis métodos/testes de detecção: Geralmente está vinculado a adjuntos de tempo iniciados pelas preposições *de*, *desde* ou *a partir de*. Também pode ser normalmente identificado pela resposta à pergunta: quando começa?

Exemplos:

O curso , que vai **de abril** a junho , ainda tem dez vagas para os interessados .

As cenas devem ir a o ar **a=partir=de amanhã** .

Estádio lotado , presença de a torcida visitante (dez ônibus sairão de Sapucaia=do=Sul) e uma rivalidade que vem **de 2007** .

TEMPO FINAL

Definição: TEMPO que indica quando um evento termina.

Comentário: Aplica-se normalmente a adjuntos de tempo, ainda que tenha ocorrência como argumento para alguns verbos que indicam o fim de algo.

Possíveis métodos/testes de detecção: Pode ser normalmente identificado pela resposta à pergunta: quando acaba?

Exemplos:

As inscrições para o Chance vão somente **até o próximo dia 16** .

O período de licença-maternidade poderá chegar **a seis meses** em o total .

A novela termina **em o final de o mês** .

MOMENTO

Definição: TEMPO que indica quando ocorre um evento.

Comentário: Aparece apenas como adjunto.

Possíveis métodos/testes de detecção: Em geral, responde à pergunta: quando?

Exemplos:

A geração de 1941-50 apresenta taxas estimadas em a faixa etária de 30=a=39 anos **em 1980** , em a de 40 a 49 anos em 1990 , e em a de 50 a 59 anos em 2000 .

É óbvio que as receitas públicas caem **em os momentos de crise** .

Sou de touro , nasci **em 3/5/1961** .

FREQUÊNCIA

Definição: TEMPO que indica um intervalo regular em que um evento ocorre.

Comentário: Aplica-se apenas a adjuntos.

Exemplos:

Em o Brasil ocorrem aproximadamente 30 mil novos casos **por ano** de febre=reumática , de os=quais 50 % evoluem para cardite .

Em a maioria de as vezes , não mostram armas , apenas simulam um volume por=baixo=de a camisa .

Nei=Rogério=Lacerda , 47 anos , trabalha **em dias intercalados** em o Hospital=de=Clínicas de Porto=Alegre e , em as horas vagas , deixa a imaginação fluir em a fábrica de o sonho , como define sua oficina escura em os fundos de a casa .

DURAÇÃO

Definição: TEMPO que indica o período de duração de um evento.

Comentário: Este papel semântico representa a extensão de um evento no plano temporal, ao contrário dos papéis MOMENTO, TEMPO INICIAL e TEMPO FINAL, que indicam um tempo mais estático. Aplica-se apenas a adjuntos.

Possíveis métodos/testes de detecção: Tentar substituir a preposição existente por *durante* pode auxiliar na detecção deste papel, ainda que essa mesma preposição possa ser simplesmente um indicador de MOMENTO.

Exemplos:

Foram acompanhados **por 3 meses** e distribuídos aleatoriamente em 2 grupos .

Durante 6,7 anos de acompanhamento clínico , foram realizados 1.595 exames ecocardiográficos em esse grupo de pacientes .

A alíquota menor , entretanto , só vale **para os seis primeiros anos de moradia de o contribuinte estrangeiro em a Espanha** .

LUGAR INICIAL

Definição: ESPAÇO que indica de onde parte um deslocamento.

Comentário: Não necessariamente representa um espaço concreto.

Possíveis métodos/testes de detecção: Geralmente pode ser detectado pela pergunta: de onde?

Exemplos:

Em um erro em a 30ª volta , o carro de Felipe saiu **de a pista** e atolou em a areia .

Os milaneses deixaram **o estádio** com a certeza de que valeu a pena esperar seis meses desde a sua contratação .

Vai **de o funk** a a música religiosa , passando por a música tradicionalista e por a música negra norte-americana .

DESTINO

Definição: ESPAÇO que indica para onde algo se desloca.

Comentário: Não necessariamente representa um espaço concreto.

Possíveis métodos/testes de detecção: Geralmente pode ser detectado pela pergunta: para onde? Porém, também é preciso observar se há deslocamento de um TEMA no evento, pois, nesse caso, podemos ter um DESTINO que responde apenas à pergunta: onde? Isso é o que ocorre em um dos exemplos a seguir, com o verbo *colocar* e o argumento *em o móvel*.

Exemplos:

O interesse em o atendimento pré-hospitalar de o infarto agudo de o miocárdio foi desenvolvido em a década de 60 , em decorrência de o grande número de óbitos observado antes que o paciente chegasse **a o hospital** (4) .

Um de os agressores jogou um sofá velho sobre a vítima , colocou álcool **em o móvel** e ateou fogo .

Ele tem nove anos e passou **para a quarta série** .

LUGAR

Definição: ESPAÇO que indica onde um evento ocorre.

Comentário: Não necessariamente representa um espaço concreto. Aplica-se apenas a adjuntos.

Possíveis métodos/testes de detecção: Em geral, responde à pergunta: onde? Porém, não se aplica se for o ponto final do deslocamento de um TEMA.

Exemplos:

Observou- se em apenas um caso , reversão espontânea a o ritmo sinusal , fato este que ocorre em até 50 % de os casos de FA **em humanos** .

A chuva abriu um valo **em a Estrada=Ricardo=Vieira=de=Barcellos** , que liga Itapuã a Viamão , impossibilitando a passagem de os veículos .

A quarta fuga aconteceu em 2007 , novamente **em a Pej** .

TRAJETÓRIA

Definição: ESPAÇO que indica o intervalo espacial entre um ponto e outro ao longo do qual algo se desloca.

Comentário: Não necessariamente representa um espaço concreto. É um papel bastante difícil de detectar, pois nem sempre vem introduzido de preposições como *ao longo de*, que seriam prototípicas para este papel.

Exemplos:

Está em liberdade provisória enquanto o processo segue **seu curso** .

A porto-alegrense Aldair=Jurema=Brazeiro=Gruski , que nunca havia participado de a procissão , foi uma de as fiéis que subiram **o morro** .

O caminhão de o lixo não sobe mais **a rua** e temos dificuldade de locomoção em a altura de o número 400 .

MODO

Definição: ADJUNTO que indica o modo como um evento ocorre.

Comentário: Aplica-se apenas a adjuntos. Em alguns casos, parece-se muito com o papel ATRIBUTO, porém se aplica a adjuntos adverbiais e não a predicados.

Exemplos:

As variáveis categóricas foram apresentadas **em número e porcentual** .

Resumo De a NotíciaSusepe vai pedir a a Justiça que o assaltante passe a cumprir a pena **em regime fechado** , quando for recapturado .

O segundo é a costela , que vem **em baixa** .

COMPANHIA

Definição: ADJUNTO que indica com quem um evento ocorre.

Comentário: Aplica-se apenas a adjuntos. Distingue-se do papel COAGENTE por não permitir a reformulação da frase com AGENTE e COAGENTE juntos na função de sujeito.

Exemplos:

Jogou mal **contra o Jaciara** .

Universitária vive **com a lembrança de seu antigo namorado que desapareceu de a escola sem qualquer explicação** .

Médico=É=Ferido=A=Tiros Em o RioAté a noite de ontem , era grave o estado de saúde de o médico ortopedista Lídio=Toledo=de=Araújo=Filho , 35 anos , filho de o também ortopedista Lídio=Toledo , que trabalhou **com a Seleção=Brasileira de futebol** .

COMPARAÇÃO

Definição: ADJUNTO que indica uma comparação entre dois objetos.

Comentário: Aplica-se apenas a adjuntos.

Exemplos:

Os grupos A e B não apresentaram diferença significante **em=relação=a a medicação** (tab .

A a semelhança de a esclerose sistêmica , também o lúpus=eritematoso sistêmico pode apresentar significativo número de pacientes com sinais de hipertensão arterial pulmonar quando se submete uma população de pacientes a a ecocardiografia com Doppler .

Em contraste com outras doenças cardíacas que vêm apresentando declínio em as últimas décadas , a incidência de a insuficiência=cardíaca está aumentando .

SITUAÇÃO

Definição: ADJUNTO que indica uma situação, não é nem um Lugar nem um Momento.

Comentário: Aplica-se apenas a adjuntos.

Exemplos:

Em a verdade , a disfunção ventricular esquerda , diferentemente de o que acontece **em as demais valvopatias** , não é a principal determinante de a sobrevida de os pacientes com estenose mitral corrigida .

Estes fatores associados podem , **em a volta a a rotina** , trazer certo desconforto , e assim é preciso que sejam retomadas as rotinas de hábitos saudáveis .

Em uma de essas , ela disse que seria capaz de " andar com um pinto no=meio=de as pernas " , referindo- se a os micos que eles pagam **em o Big=Boss** .

DIREÇÃO

Definição: ADJUNTO que indica uma direção.

Comentário: Aplica-se apenas a adjuntos. Este papel aponta direções espaciais como *para cima*, *para baixo* etc.

Exemplos:

Um olhava **para o chão** e o outro cochilava .

Durante a perseguição , houve troca de tiros e os assaltantes chegaram a jogar granadas **em a direção de os policiais** .

O buraco está vindo **em a direção de a minha casa** .

DISCURSO

Definição: ADJUNTO que indica conexões textuais e discursivas.

Comentário: Aplica-se apenas a adjuntos que funcionam como conectores.

Exemplos:

Em conclusão , não encontramos diferença estatística significativa entre as variáveis analisadas por o Eco e RMC em ambos os grupos clínicos .

Em esse sentido , Stefanelli recentemente publicou os achados de o seguimento de 38 jovens :

Em paralelo , uma outra via de agressão se torna importante quando os mesmos mediadores acarretam a down-regulation de receptores alfa 2-adrenérgicos e a bloqueio a o influxo de cálcio transmembrana , com diminuição de a resposta de as catecolaminas .

EXEMPLO

Definição: ADJUNTO que indica uma exemplificação.

Comentário: Aplica-se apenas a adjuntos. Os três exemplos que arrolamos a seguir são os únicos anotados em nossos *corpora*.

Exemplos:

Eletrocardiogramas adicionais eram realizados quando o paciente apresentasse sintomas , **como palpitações e / ou suspeição de arritmias** .

No=entanto , o eco apresenta grandes desafios , **como problemas com a janela acústica e a sua grande dependência de o examinador** .

por=último , a administração de outros agentes antiinflamatórios , **como os inibidores de a COX-2** , não apresentam benefícios cardiovasculares e o conhecimento sobre seus potenciais efeitos adversos ainda é limitado .

ADJUNTO CAUSAL

Definição: ADJUNTO que indica o motivo de um evento.

Comentário: Distingue-se do papel CAUSA por aplicar-se apenas a adjuntos.

Exemplos:

Esse maior risco aconteceu **em decorrência de o aumento de a incidência de acidente=vascular=cerebral isquêmico e de o desenvolvimento ou piora de a insuficiência=cardíaca** .

Desnutrição e hipoproteinemia podem ocorrer em a evolução de a insuficiência=cardíaca **por diversos motivos** , (14) entre os=quais salientaremos apenas dois :

Ele será julgado **por homicídio simples** .

ADVÉRBIO

Definição: ADJUNTO aplicável a adjuntos que não se enquadram na definição de nenhum dos outros papéis semânticos.

Comentário: Aplica-se apenas a adjuntos.

Exemplos:

O resultado foi que o treinamento de moderada intensidade , **em o cômpito geral** , provocou mais benefícios que o treinamento de alta intensidade .

Para piorar , o esgoto corre a=céu=aberto .

Para Tcheco , o reerguimento de a equipe não depende de um único jogador .

Alguns papéis merecem algumas observações além das que apresentamos na descrição da lista. Por exemplo, no português, existe uma série de funções que podem ser atribuídas à partícula *se*. Assim, buscamos referência para que pudéssemos ter uma anotação padronizada para suas diversas funções. O esquema a que chegamos pode ser visto na Tabela 8.2. As informações dessa tabela foram retiradas do tutorial do PropBank.Br[65], apenas acrescentamos os papéis semânticos que utilizamos para cada caso.

Nesse quesito, em algumas sentenças, o pronome *se* com função apassivadora foi reconhecido pelo PALAVRAS como sujeito da oração. Nesses casos em particular, optamos por anotar como se fosse um caso de sujeito indeterminado, e não com o papel SE PASSIVO. Tomamos essa opção em conformidade com as observações do pronome *se* como um índice de indeterminação de sujeito, independentemente da regência, conforme apontam Bechara (1999) e Cunha e Cintra (1985).

Tabela 8.2 – Uso do pronome **se**[66]

Anotação do PropBank.Br	**Descrição**	**Exemplo**	**Papel Semântico**
SE-REF-OD	Pronome reflexivo (objeto direto)	Ele se feriu	Papel normal (por exemplo: PACIENTE/TEMA/EXPERIENCIADOR)
SE-REF-OI	Pronome reflexivo (objeto indireto)	Ele se deu um presente	Papel normal (por exemplo: RECIPIENTE/BENEFICIÁRIO)
SE-REC	Pronome recíproco	Eles se encontraram	RECÍPROCO
SE-PAS	Partícula apassivadora	Vendem-se casas	SE_PASSIVO
SE-IND	Partícula de indeterminação de sujeito	Concordou-se com tudo.	Papel normal (por exemplo: AGENTE/CAUSA/ESTÍMULO)
SE-EXP	Partícula expletiva	Acabou-se a festa.	VERBO
SE-VPR	Partícula integrada ao verbo	Apaixonou-se e arrependeu-se.	VERBO

Outro papel que acrescentamos e que não é comum foi o papel semântico AGENTE LOCATIVO. Esse papel foi criado para dar conta dos casos em que os sujeitos são, na verdade, um adjunto adverbial de lugar que foi promovido à posição de sujeito por meio de metonímia. Esse papel pode, em princípio, ser categorizado junto com o

[65] Disponível em: http://www.nilc.icmc.usp.br/portlex/images/arquivos/propbank-br/propbank.br%20tutorial.pdf. Acessado em: 27/02/2014.

[66] Esta tabela foi extraída do tutorial do PropBank.Br. Os papéis semânticos correspondentes foram acrescentados.

papel AGENTE, mas preferimos distinguir entre os dois porque essa distinção oferece mais informação de uma forma simples de detectar.

Em relação ao que tínhamos no segundo estudo-piloto, as modificações ficam por parte da inserção de alguns papéis semânticos que serviram para dar conta de casos que ainda não estavam previstos. Também decidimos por usar uma anotação em português, já que, se necessário, a conversão dos nomes dos papéis semânticos do português para outra língua (ou mesmo outra codificação) pode ser feita sem dificuldade agora que os dados se encontram também em formato XML.

8.3 Metodologia

A metodologia do desenvolvimento do recurso permaneceu praticamente inalterada em relação à que apresentamos na descrição do segundo estudo-piloto (Seção 6.2.3). A única modificação que realizamos foi privilegiar a anotação dos dados do *corpus* do Diário Gaúcho. Desse modo, iniciamos a anotação seguindo a ordem de frequência dos verbos do Diário Gaúcho e somente anotávamos os dados também no *corpus* de Cardiologia se o verbo em questão tivesse frequência suficiente para ser anotado. Os demais procedimentos continuaram inalterados: não anotamos os verbos *ser*, *estar*, *ter* e *haver*; e não anotamos estruturas de subcategorização que não apresentavam pelo menos dez exemplos corretos.

8.4 Dados do VerbLexPor

Nesta seção, apresentaremos dados quantitativos do recurso, como as frequências de cada papel semântico nos dois *corpora*, a frequência de combinações entre sintaxe e semântica etc. Os dados serão mostrados principalmente por meio de diferentes tabelas, para facilitar a sua visualização.

Antes de passar aos dados mais detalhados, cabe informarmos alguns dados básicos do recurso como ele se encontra hoje. No que diz respeito ao *corpus* do Diário Gaúcho, o recurso conta com 191 verbos anotados, totalizando 5.301 sentenças e 11.089 argumentos. Já no *corpus* de Cardiologia anotamos 77 verbos (sendo 76 deles também anotados no Diário Gaúcho), resultando em 1.931 sentenças e 4.192 argumentos. Além dessas sentenças que têm anotação de papéis semânticos, existem milhares de outras sentenças no *corpus* que estão anotadas com as funções sintáticas dos diferentes argumentos. Desse modo, ainda que o recurso não esteja completamente anotado com papéis semânticos, as demais sentenças presentes no recurso não deixam de apresentar

informações sintáticas que foram extraídas com base na anotação do *parser* PALAVRAS.

Passando então aos dados de papéis semânticos propriamente ditos, na Tabela 8.3, podemos observar como ficaram distribuídos os papéis semânticos nos *corpora*. Apresentamos nela cada um dos papéis semânticos por ordem decrescente de frequência total, discriminando as ocorrências no Diário Gaúcho e em Cardiologia, juntamente com as representações percentuais. Lembramos que o papel semântico PRODUTO não teve ocorrência nas amostras que anotamos em nossos *corpora* e, por isso, temos apenas 45 papéis semânticos na Tabela 8.3, em vez dos 46 papéis que seriam esperados.

Tabela 8.3 – Papéis semânticos utilizados e sua frequência nos *corpora*

#	Papel Semântico	Freq. DG	DG %	Freq. Cardio	Cardio %	Freq. Total	Total %
1	TEMA	3.015	27,19%	1.416	33,78%	4.431	29,00%
2	AGENTE	2.540	22,91%	254	6,06%	2.794	18,28%
3	LUGAR	540	4,87%	143	3,41%	683	4,47%
4	RESULTADO	363	3,27%	289	6,89%	652	4,27%
5	PACIENTE	497	4,48%	145	3,46%	642	4,20%
6	EXPERIENCIADOR	591	5,33%	47	1,12%	638	4,18%
7	PIVÔ	345	3,11%	282	6,73%	627	4,10%
8	VERBO	407	3,67%	184	4,39%	591	3,87%
9	TÓPICO	453	4,09%	68	1,62%	521	3,41%
10	CAUSA	191	1,72%	222	5,30%	413	2,70%
11	MOMENTO	306	2,76%	87	2,08%	393	2,57%
12	FINALIDADE	257	2,32%	130	3,10%	387	2,53%
13	INSTRUMENTO	152	1,37%	208	4,96%	360	2,36%
14	SITUAÇÃO	176	1,59%	162	3,86%	338	2,21%
15	ATRIBUTO	194	1,75%	136	3,24%	330	2,16%
16	DESTINO	187	1,69%	8	0,19%	195	1,28%
17	RECIPIENTE	169	1,52%	13	0,31%	182	1,19%
18	BENEFICIÁRIO	110	0,99%	68	1,62%	178	1,16%
19	MODO	83	0,75%	77	1,84%	160	1,05%
20	COTEMA	41	0,37%	48	1,15%	89	0,58%
21	AGENTE LOCATIVO	72	0,65%	3	0,07%	75	0,49%
22	ALVO	37	0,33%	30	0,72%	67	0,44%
23	ADJ. CAUSAL	20	0,18%	45	1,07%	65	0,43%
24	ESTÍMULO	60	0,54%	0	0,00%	60	0,39%
25	FONTE	20	0,18%	35	0,83%	55	0,36%
26	LUGAR INICIAL	47	0,42%	2	0,05%	49	0,32%
27	ATIVO	40	0,36%	0	0,00%	40	0,26%
28	COAGENTE	39	0,35%	0	0,00%	39	0,26%
29	DURAÇÃO	10	0,09%	24	0,57%	34	0,22%
30	RECIPROCO	24	0,22%	0	0,00%	24	0,16%

#	Papel Semântico	Freq. DG	DG %	Freq. Cardio	Cardio %	Freq. Total	Total %
31	SE PASSIVO	1	0,01%	20	0,48%	21	0,14%
32	COMPARAÇÃO	3	0,03%	18	0,43%	21	0,14%
33	COMPANHIA	16	0,14%	3	0,07%	19	0,12%
34	TEMPO FINAL	17	0,15%	0	0,00%	17	0,11%
35	TEMPO INICIAL	14	0,13%	0	0,00%	14	0,09%
36	VARIAÇÃO	11	0,10%	1	0,02%	12	0,08%
37	MATERIAL	11	0,10%	0	0,00%	11	0,07%
38	COPACIENTE	0	0,00%	9	0,21%	9	0,06%
39	DISCURSO	4	0,04%	5	0,12%	9	0,06%
40	DIREÇÃO	8	0,07%	0	0,00%	8	0,05%
41	ADVERBIO	5	0,05%	3	0,07%	8	0,05%
42	VALOR	5	0,05%	1	0,02%	6	0,04%
43	TRAJETÓRIA	6	0,05%	0	0,00%	6	0,04%
44	FREQUÊNCIA	2	0,02%	3	0,07%	5	0,03%
45	EXEMPLO	0	0,00%	3	0,07%	3	0,02%
Total		11.089	100,00%	4.192	100,00%	15.281	100,00%

Na Tabela 8.3, podemos ver que o papel semântico tema tem a maior frequência em ambos os corpora. Isso corrobora nossa percepção de que o papel TEMA poderia ser subdividido em outras categorias, de maneira a ser semanticamente mais preciso. A título de especulação, seria possível, por exemplo, distinguir um TEMA que indicasse movimento. No entanto, para conseguir realizar uma subdivisão desse papel semântico tão predominante, é preciso estudar os dados que temos com bastante cuidado, observando os verbos envolvidos e outros traços relativos ao argumento em questão. Esse tipo de estudo requer tempo e, infelizmente, não pôde ser realizado no escopo deste trabalho, que está mais preocupado com o desenvolvimento de um recurso, mas é um estudo que pretendemos realizar no futuro.

Tabela 8.4 – Estruturas sintático-semânticas no Diário Gaúcho (amostra)

Estrutura	Freq	Freq. %
SUJEITO<agente>	2.511	22,64%
OBJETO DIRETO<tema>	1.343	12,11%
SUJEITO<tema>	1.010	9,11%
SUJEITO<experienciador>	584	5,27%
ADJUNTO ADVERBIAL[em]<lugar>	426	3,84%
SUJEITO<paciente>	351	3,17%
OBJ DIR ORACIONAL<tema>	344	3,10%
SUJEITO<pivo>	324	2,92%
ADJUNTO ADVERBIAL[em]<momento>	263	2,37%
OBJ DIR ORACIONAL<topico>	255	2,30%

Tabela 8.5 – Estruturas sintático-semânticas em Cardiologia (amostra)

Estrutura	Freq	Freq. %
SUJEITO<tema>	684	16,32%
OBJETO DIRETO<tema>	480	11,45%
SUJEITO<pivo>	272	6,49%
SUJEITO<agente>	236	5,63%
SUJEITO<causa>	197	4,70%
OBJ DIR ORACIONAL<tema>	190	4,53%
ADJUNTO ADVERBIAL[em]<lugar>	136	3,24%
ADJUNTO ADVERBIAL[em]<situacao>	132	3,15%
SUJEITO<resultado>	130	3,10%
SUJEITO<instrumento>	120	2,86%

Dando prosseguimento à apresentação dos dados, mostramos, na Tabela 8.4 e 8.5, as estruturas sintático-semânticas mais frequentes nos dois corpora. Por estrutura sintático-semântica, entendemos a associação entre um papel semântico e uma função sintática. Diferentemente da Tabela 8.3, essas tabelas não são exaustivas, tendo em vista que existem dezenas de estruturas sintático-semânticas nos corpora.

Ainda que esta não seja a seção que reservamos para a discussão dos dados, podemos observar claramente que os sujeitos em Cardiologia e nos textos do Diário Gaúcho são bastante diferentes. É claro que é preciso levar em conta também a diferença dos verbos anotados (tendo em vista que apenas 76 deles são compartilhados), por isso, pedimos paciência ao leitor, pois isso será devidamente levado a cabo no Capítulo 10, quando apresentaremos uma discussão mais aprofundada dos resultados que obtivemos em relação às nossas hipóteses e questões de pesquisa.

Tabela 8.6 – Sentenças sintático-semânticas no Diário Gaúcho (amostra)

Sentença sintático-semântica	Freq.	Freq. %
SUJEITO<agente> + OBJETO DIRETO<tema>	441	8,3%
SUJEITO<agente>	362	6,8%
SUJEITO<tema>	259	4,9%
SUJEITO<agente> + OBJ DIR ORACIONAL<topico>	175	3,3%
SUJEITO<experienciador> + OBJETO DIRETO<tema>	134	2,5%
SUJEITO<experienciador> + OBJ DIR ORACIONAL<tema>	129	2,4%
SUJEITO<pivo> + OBJETO DIRETO<tema>	121	2,3%
SUJEITO<agente> + OBJETO DIRETO<topico>	98	1,8%
SUJEITO<paciente>	91	1,7%
SUJEITO<agente> + OBJ DIR ORACIONAL<tema>	89	1,7%

Tabela 8.7 – Sentenças sintático-semânticas em Cardiologia (amostra)

Sentença sintático-semântica	Freq.	Freq. %
SUJEITO<tema>	150	7,77%
SUJEITO<pivo> + OBJETO DIRETO<tema>	109	5,64%
SUJEITO<agente> + OBJETO DIRETO<tema>	64	3,31%
SUJEITO<tema> + ADJUNTO ADVERBIAL[em]<lugar>	49	2,54%
SUJEITO<instrumento> + OBJ DIR ORACIONAL<tema>	44	2,28%
SUJEITO<tema> + OBJETO REFLEXIVO<verbo> + PREDICATIVO<atributo>	40	2,07%
SUJEITO<agente> + OBJ DIR ORACIONAL<tema>	40	2,07%
SUJEITO<experienciador> + OBJ DIR ORACIONAL<tema>	40	2,07%
SUJEITO<instrumento> + OBJETO DIRETO<tema>	39	2,02%
SUJEITO<causa> + OBJETO DIRETO<tema>	36	1,86%

Ampliando um pouco mais às associações de dados, podemos observar quais foram as estruturas sintático-semânticas mais comuns nas sentenças. Para tal, basta observarmos como ocorre cada estrutura sintático-semântica associada às demais estruturas sintático-semânticas presentes na mesma sentença. Dessa forma, por falta de um nome melhor, podemos dizer que temos uma espécie de **sentença sintático-semântica**. É esse tipo de informação que mostramos nas Tabelas 8.6 e 8.7.

Novamente, as listas têm de ser amostrais, pois temos literalmente centenas de ocorrências nos *corpora*. Essas tabelas são as que revelam mais informações sobre a individualidade das sentenças vinculadas aos verbos. Esse tipo de informação foi o que utilizamos para fazer parte dos nossos experimentos com agrupamentos de verbos, sobre os quais comentaremos no Capítulo 9. Porém, antes de passarmos ao agrupamento de verbos, reservamos mais uma seção deste capítulo para observar como nosso recurso se compara com outros dois recursos existentes no Brasil: o PropBank.Br (DURAN e ALUÍSIO, 2011; 2012) e a VerbNet.Br (SCARTON, 2013).

8.5 Comparação com outros recursos

Nesta seção, procuramos mostrar, com dados quantitativos, como o VerbLexPor se compara a outros recursos já existentes que tratam de papéis semânticos. Como os três recursos em questão (VerbNet.Br, PropBank.Br e VerbLexPor) são diferentes entre si, seja por opções teóricas, seja por questões de detalhes na implementação, as comparações não puderam ser realizadas diretamente, sem modificações. Desse modo,

ao apresentarmos as comparações, também relatamos as modificações que tiveram de ser realizadas para que elas fossem possíveis.

8.5.1 VerbLexPor vs. PropBank.Br

O PropBank.Br (DURAN e ALUÍSIO, 2011; 2012) já foi apresentado muito brevemente na Seção 4.3, porém, aqui trataremos de alguns detalhes mais aprofundados. O PropBank.Br se parece com o VerbLexPor no sentido de que ambos partem de um conjunto de sentenças e têm anotações semânticas feitas com base na anotação sintática realizada por um *parser*. A comparação que desejamos realizar diz respeito à porcentagem de sentenças que ambos os recursos têm anotadas de maneira similar. Desse modo, fizemos uma série de alterações em nossos dados para permitir essa comparação, mas antes vamos a alguns dados do recurso em foco.

O PropBank.Br conta com 5.537 instâncias anotadas, partindo de um total de 3.164 sentenças (algumas sentenças foram reproduzidas, de acordo com a quantidade de verbos principais presentes). O número de instâncias anotadas é parecido com o nosso (no caso do *corpus* do Diário Gaúcho), porém o PropBank.Br tem muito mais verbos anotados, totalizando 992 verbos diferentes. Isso dá uma média de 5,58 sentenças anotadas por verbo, enquanto o VerbLexPor conta com uma média de 27,75 sentenças por verbo no *corpus* do Diário Gaúcho e 25,08 sentenças por verbo no caso dos artigos de Cardiologia. Isso indica que nosso recurso tem muito mais redundância na anotação do que o PropBank.Br, que privilegiou um maior número de verbos. Assim, para a comparação, precisamos usar apenas os verbos que estão presentes nos dois recursos, caso contrário, teríamos uma disparidade que não refletiria a realidade que queremos comparar.

Outro elemento que temos de diferente é o tipo de papel semântico usado para a anotação. Como informamos na Seção 4.3 desta tese, o PropBank.Br usa papéis semânticos numerados (A0-A5), enquanto nosso recurso usa papéis semânticos descritivos. Por isso, antes de realizarmos uma comparação entre os recursos, foi necessário converter os papéis semânticos descritivos para papéis numerados. Nesse mesmo quesito dos papéis semânticos, houve também um problema no que diz respeito aos papéis semânticos usados para adjuntos. Ainda que existentes em ambos os recursos de maneira mais ou menos parecida, os adjuntos adverbiais ocorrem de maneira mais ou menos aleatória nas amostras textuais. Como o PropBank possui um baixo índice de redundância (apenas 5,58 sentenças em média por verbo), uma tentativa de comparar as

anotações semânticas de adjuntos adverbiais seria provavelmente frustrada e apenas levaria a ruídos. Desse modo, optamos por excluir os papéis semânticos próprios para adjuntos da comparação.

Restava-nos então olhar como fazer a tradução dos papéis semânticos descritivos para numerados. No manual de anotação do PropBank.Br, temos claramente definidos os papéis A0 e A1, como podemos ver a seguir:

Papel Semântico – PropBank.Br	Papel Semântico Descritivo
arg0	Agente ou causador
arg1	Paciente, experienciador ou tema

Infelizmente, para os demais papéis semânticos (A2 – A5), a definição é *ad hoc* e depende diretamente do verbo em questão. Diante desse problema, optamos por não considerarmos os demais papéis semânticos para a comparação. A lista completa dos papéis traduzidos é a seguinte:

Papel Semântico Descritivo	Papel Semântico Numerado
AGENTE	A0
COAGENTE	A0
CAUSA	A0
INSTRUMENTO	A0
ESTÍMULO	A0
EXPERIENCIADOR	A1
PACIENTE	A1
COPACIENTE	A1
TEMA	A1
COTEMA	A1
TÓPICO	A1
PIVÔ	A1

Assim, de nossos 46 papéis semânticos, pudemos realizar uma comparação com apenas 12 deles (reduzidos para apenas dois: A0 e A1). Ainda assim, cremos que, mesmo partindo de um princípio de 50% de similaridade, é melhor uma comparação reduzida do que nenhuma comparação, pois isso já oferece algum indício de como os recursos são compatíveis entre si.

A comparação foi realizada por meio de pares de verbo e papel semântico. Assim, primeiro observávamos se existia, por exemplo, o verbo *fazer* associado ao papel semântico AGENTE (*fazer*+AGENTE) em um dos recursos. Em seguida, procurávamos por esse mesmo par no outro recurso.

Apesar de todos os cortes nos papéis semânticos que tivemos de fazer, tivemos uma intersecção de verbos bastante alta nos dois recursos: 183 verbos do *corpus* do Diário Gaúcho e todos os 77 verbos do *corpus* de Cardiologia estavam presentes no PropBank.Br, o que mostra a grande abrangência desse recurso.

Começando pelo *corpus* do Diário Gaúcho, na comparação dos pares de verbo e papel semântico para os 183 verbos, tivemos um total de 363 pares, enquanto o PropBank.Br, para esses mesmos 183 verbos, apresentou 348 pares. A intersecção entre os dois recursos, considerando apenas o *corpus* do Diário Gaúcho, foi de 306 pares. Os resultados, em termos de precisão, abrangência e medida f, se encontram a seguir:

Precisão	Abrangência	Medida f
84,30	87,93	86,08

Realizando o mesmo procedimento para o *corpus* de Cardiologia em relação ao PropBank.Br, para os 77 verbos, tivemos um total de 132 pares no *corpus* de Cardiologia, e um total de 144 pares no PropBank.Br. A intersecção foi de 119 pares. Os resultados, em termos de precisão, abrangência e medida f, se encontram a seguir:

Precisão	Abrangência	Medida f
90,15	82,64	86,23

Esses resultados parecem indicar que a metodologia adotada pelo PropBank.Br seja melhor em termos de custo benefício, já que, com menor esforço (média de 5,58 sentenças por verbo, em oposição às 27,75 em média que usamos). Tal percepção, no entanto, é precoce, pois tivemos que usar apenas alguns dos papéis semânticos para fazer a comparação, o que reduziu em muito as chances de vermos possíveis diferenças.

De qualquer modo, os papéis de causa/agentividade (A0) e paciente/tema (A1) foram, de fato, os mais frequentes em nosso *corpus* do Diário Gaúcho, e um dos pontos que mais chamou atenção no *corpus* de Cardiologia: o fato de ter muitos INSTRUMENTOS

como sujeito acabou sendo ignorado na comparação, pois não há um papel semântico predefinido para o caso de INSTRUMENTO no PropBank.Br.

Depois dessa comparação mais básica com o PropBank.Br, passamos à comparação do VerbLexPor com um recurso que usa de fato papéis semânticos descritivos.

8.5.2 VerbLexPor vs. VerbNet.Br

A comparação que realizamos com a VerbNet.Br (SCARTON, 2013) foi feita de modo similar à que mostramos em relação ao PropBank.Br. A diferença é que a VerbNet.Br, por ter uma estrutura mais parecida com a VerbNet (KIPPER-SCHULER, 2005), permitiu uma comparação mais direta dos papéis semânticos, sem tantas modificações. Diferentemente do PropBank.Br, que foi anotado manualmente, a VerbNet.Br foi importada de maneira semiautomática a partir da VerbNet do inglês. Para tal, foram usadas as associações existentes entre a VerbNet e a WordNet, e entre a WordNet e a WordNet.Br. Desse modo, quando havia um *synset* na WordNet.Br que fosse sinônimo ou quase sinônimo de um *synset* na WordNet do inglês, usavam-se as associações entre os recursos para importar a anotação da VerbNet relativa ao *synset* em questão para a VerbNet.Br. Assim, a VerbNet.Br conseguiu importar muita informação de maneira semiautomática e construir um recurso bastante robusto que foi o primeiro no Brasil nos moldes da VerbNet.

A VerbNet.Br conta com um acervo de 5.368 verbos[67]. Os dados disponibilizados até então (em formato CSV ou SQL) dão conta desses verbos associados aos papéis semânticos cabíveis, além de outras informações pertinentes à VerbNet.Br. Dessa forma, estão disponíveis 22.359 pares compostos por verbos associados a papéis semânticos.

Ainda assim, existem alguns problemas na criação semiautomática de recursos semânticos. A VerbNet.Br teve como base as características interlinguísticas dos papéis semânticos da VerbNet. O problema é que os verbos do português nem sempre se comportam como os verbos do inglês. Assim, a chance de haver ruído no recurso final é grande, como mostramos mais adiante com alguns poucos exemplos. Desse modo, o

[67] Deve-se levar em consideração que, por exemplo, o verbo *abençoar* e sua forma reflexiva, *abençoar-se*, são considerados separadamente, por uma questão de metodologia.

teste de comparação que realizamos aqui serve também para, de certa forma, validar o conteúdo importado do inglês para a VerbNet.Br.

Antes de passarmos à comparação, descrevemos a seguir algumas modificações que tiveram de ser feitas nos papéis semânticos. Apesar de usarmos a mesma base da VerbNet, acrescentamos papéis específicos para adjuntos e também modificamos um pouco outros papéis semânticos. Além disso, uma questão bastante crítica é que, em nossa tese, usamos a versão 3.2 da VerbNet como base para a lista de papéis semânticos, a qual surgiu apenas em 2013, quando a VerbNet.Br já estava em vias de conclusão.

As diferenças entre as listas de papéis semânticos da versão 3.1 e da versão 3.2 da VerbNet são salientes, então precisamos olhar nos dois manuais para encontrar as traduções devidas. Desse modo, todos os papéis semânticos foram traduzidos da VerbNet.Br para o formato que adotamos, em português, de acordo com a lista a seguir (observe que os papéis estão sem acentos e símbolos especiais, pois estão assim também no banco de dados):

Papel VerbNet.Br	Papel VerbLexPor
Actor	agente
Actor1	agente
Actor2	coagente
Agent	agente
Asset	ativo
Attribute	atributo
Beneficiary	alvo
Cause	causa
Destination	destino
Experiencer	experienciador
Extent	variacao
Instrument	instrumento
Material	material
Patient	paciente
Patient1	paciente
Patient2	copaciente

Papel VerbNet.Br	Papel VerbLexPor
Product	produto
Recipient	alvo
Source	fonte
Stimulus	estimulo
Theme	tema
Theme1	tema
Theme2	cotema
Topic	topico
Value	valor

Além dessa tradução, que envolveu passar tanto da versão 3.1 da VerbNet para a versão 3.2 como adaptar para o português, também tivemos de simplificar dois papéis semânticos da lista do VerbLexPor, tendo em vista que eles são ramificações de papéis da VerbNet:

Papel VerbLexPor	Papel Simplificado
agente_locativo	agente
recipiente	alvo
beneficiario	alvo

O caso do papel ALVO, como pode ser visto nos dois quadros, é um pouco mais complicado. Tendo em vista que ele foi criado como um hiperônimo de RECIPIENTE e BENEFICIÁRIO, levamos em consideração várias possibilidades de tradução dos papéis semânticos, desde traduzirmos de ALVO para BENEFICIÁRIO até traduzirmos RECIPIENTE e BENEFICIÁRIO (e os respectivos correspondentes em inglês) para ALVO. A diferença entre todas essas possibilidades foi muito pequena (não passando de 2 pares), mas optamos por mostrar a que teve mais resultados compatíveis, de modo que apresentaremos em detalhes quantitativos apenas a que está ilustrada no quadro acima.

Como pode ser visto nos quadros, assim como no caso da comparação com o PropBank.Br, não pudemos utilizar os papéis que podem ser usados para adjuntos, pois a VerbNet.Br, assim como a VerbNet, está organizada de modo que apenas alguns adjuntos são considerados. Assim, foram excluídos os papéis semânticos a seguir:

DURAÇÃO, ADJ. CAUSAL, ADVÉRBIO, COMPANHIA, COMPARAÇÃO, DIREÇÃO, DISCURSO, FINALIDADE, FREQUÊNCIA, LUGAR INICIAL, MODO, PIVÔ, SITUAÇÃO, TEMPO FINAL, TEMPO INICIAL, TRAJETÓRIA, EXEMPLO, LUGAR e MOMENTO. Isso também inclui os papéis semânticos LOCATION e TIME da VerbNet (os equivalentes de lugar e tempo).

Por fim, alguns papéis semânticos não puderam ser usados simplesmente por não haver equivalentes. Os papéis semânticos RECIPROCO, SE PASSIVO e VERBO representam fenômenos que não existem em inglês, ou não foram levados em consideração no recurso, de modo que também não estão presentes na VerbNet.Br. Além disso, o papel semântico PREDICATE, que existe na versão 3.1 da VerbNet, foi excluído da versão 3.2, de modo que ele não está na nossa lista e não tem uma correspondência na versão 3.2, não podendo ser levado em consideração.

Depois dessas alterações, sempre necessárias quando se comparam recursos com bases teóricas diferentes, passamos aos resultados da comparação. A metodologia foi exatamente a mesma usada para a comparação com o PropBank.Br (Seção 8.5.1).

Começando pelo *corpus* do Diário Gaúcho, tivemos 166 verbos em comum com a VerbNet. O Diário Gaúcho teve então 582 pares válidos, contra 865 da VerbNet.Br. Na intersecção entre os dois, tivemos 395 pares em comum. Desse modo, os resultados de precisão, abrangência e medida f ficaram da seguinte maneira:

Precisão	Abrangência	Medida f
67,86%	45,66%	54,59%

Prosseguindo para o *corpus* de Cardiologia, tivemos 69 verbos compatíveis. A VerbNet.Br apresentou 359 pares elegíveis, contra 207 do *corpus* de Cardiologia. Na intersecção, obtivemos 132 pares. Desse modo, os resultados de precisão, abrangência e medida f ficaram da seguinte maneira:

Precisão	Abrangência	Medida f
63,77%	36,77%	46,64%

Como pudemos ver, os resultados da comparação entre a VerbNet.Br e o VerbLexPor apontam para apenas cerca de 50% de similaridade. Por um lado, um dos fatores que pode influenciar nessa baixa similaridade é o fato de que a VerbNet.Br conta com muitos casos de polissemia, de modo que um verbo pode estar em muitas classes

diferentes. Essa polissemia é algo que nossos dados amostrais talvez não reflitam. Por outro lado, a discrepância pode ter se originado pela importação dos dados do inglês para o português, pois tal importação requer uma compatibilidade talvez inexistente entre os dados nas duas línguas.

É preciso observar que existem alguns exemplos bastante curiosos de ruídos na VerbNet.Br, tais como o caso do verbo *correr* e *rasgar* estarem no mesmo grupo (classe *escape 51.1*) e compartilharem os mesmos papéis semânticos, inclusive com uma indicação de que o verbo *rasgar* seja intransitivo com o papel THEME. Existem casos em que *rasgar* e *correr* realmente podem assumir um significado próximo, mas *rasgar* sempre vai requerer um objeto direto para expressar a ideia de *percorrer*.

Podemos também observar o verbo *abaixar*, marcado com os papel semântico ATTRIBUTE em sua forma intransitiva, mesmo que o papel ATTRIBUTE se refira, em geral, a predicativos de acordo com a descrição da VerbNet.

A falta de exemplos na VerbNet.Br também é, de certo modo, um empecilho para o seu uso linguístico, pois faz com que alguns dos dados apresentados sejam difíceis de compreender. Temos, por exemplo, o verbo *capitular* como membro do grupo *declare 29.4*. Sem um exemplo de uso ou mesmo uma indicação da sintaxe do verbo, é difícil de associar *capitular* a verbos como *coroar* ou *denominar*. Esses são apenas alguns dos exemplos que encontramos em uma análise não extensiva dos dados[68].

8.5.3 Resumo das Comparações

Atualmente, o PropBank.Br (DURAN e ALUÍSIO, 2011; 2012) e a VerbNet.Br (SCARTON, 2013) são os dois recursos disponíveis no Brasil que podem ser comparados, de uma forma ou de outra, com o VerbLexPor. Existem também os recursos no estilo da FrameNet (BAKER, FILLMORE e LOWE, 1998), como discutimos no Capítulo 4, porém, o tipo de papel semântico usado nesses recursos são muito específicos em relação ao contexto, de modo que não teríamos como fazer uma comparação.

A quantidade de modificações requeridas infelizmente reduziu em muito o potencial do resultado da comparação com o PropBank.Br. Como as teorias por trás dos

[68] Esses dados estão disponíveis para consulta no site da VerbNet.Br: http://143.107.183.175:21380/verbnetbr/index.html

dois recursos são muito distintas, os dados que restaram para a comparação foram muito restritos. Dessa forma, os resultados dessa comparação devem ser observados com cautela.

No que diz respeito à comparação com a VerbNet.Br, o procedimento de comparação foi bem mais simples, mas a similaridade foi menor. Como a VerbNet.Br foi construída semiautomaticamente por meio de associações entre outros recursos, fica difícil de saber se ela pode ser considerada como um padrão-ouro, ou se nosso recurso, construído manualmente, seria uma melhor referência. Como vimos em alguns breves exemplos, algumas informações na VerbNet.Br, por não ter um contexto de uso, ou por terem sido importadas semiautomaticamente, contêm ruídos.

8.6 Disponibilização do VerbLexPor

Nesta seção, apresentamos mais um passo que realizamos ao final do segundo estudo-piloto e que aplicamos também para o VerbLexPor. Nosso recurso é composto por um banco de dados com sentenças anotadas com papéis semânticos, mas, até o final do segundo estudo-piloto, ainda não havíamos posto esses dados à disposição de quem estivesse interessado. Por isso, iniciamos um trabalho que visou a importar os dados do VerbLexPor para uma plataforma de livre acesso na Internet, em colaboração com um dos desenvolvedores da plataforma Jibiki (MANGEOT, 2006), o Prof. Mathieu Mangeot. Neste capítulo, apresentamos os procedimentos realizados para a disponibilização *on-line* de nossos dados, assim como a própria plataforma Jibiki. Também indicamos outros meios disponíveis para *download* do VerbLexPor.

8.6.1 A plataforma Jibiki

A plataforma Jibiki[69] faz parte do projeto Papillon (BOITET, MANGEOT e SÉRASSET, 2002), cujo principal objetivo é desenvolver bancos de dados lexicais multilíngues e disponibilizá-los gratuitamente *on-line*. O projeto Papillon conta atualmente com contribuições de dez línguas, entre elas o português. O principal recurso do Papillon, o Dicionário Papillon, foi desenvolvido a partir de dicionários monolíngues em oito línguas, de modo que suas entradas estão vinculadas por meio de uma interlíngua. Dessa forma, por exemplo, as entradas do dicionário vietnamita estão vinculadas às suas correspondentes nas demais línguas do dicionário.

[69] http://jibiki.univ-savoie.fr/jibiki/Home.po.

Nesse âmbito, a plataforma Jibiki é uma ferramenta que permite a consulta *on-line* dos dados gerados no projeto Papillon, facilitando a divulgação das informações. A plataforma apresenta diversas possibilidades de busca e pode ser facilmente atualizada, apresentando até mesmo uma interface própria para edição dos dados. A importação é feita a partir de um arquivo XML, que é indexado às estruturas presentes na plataforma, de modo que os dados são importados e disponibilizados para consulta. A Figura 8.1 mostra a página inicial do projeto Papillon via plataforma Jibiki.

Figura 8.1 – Plataforma Jibiki. Página inicial.

8.6.1.1 Importação dos dados

Como já possuíamos uma versão de nossos dados em formato XML, a importação foi bastante facilitada, mas ainda requereu bastante trabalho. Como a plataforma Jibiki é destinada principalmente a dados de linguagem comum, optamos por fazer *upload* apenas dos dados do *corpus* do Diário Gaúcho.

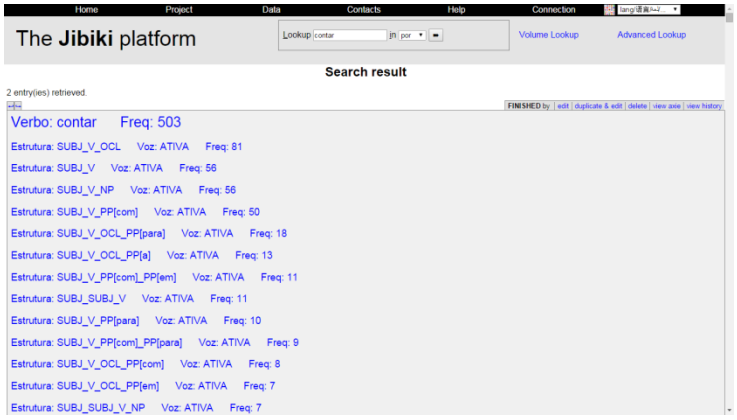
Em primeiro lugar, por se tratar de uma plataforma dedicada a dicionários de tamanho (em *bytes*) não muito grande, precisamos reduzir a quantidade de informações que havia no XML. Desse modo, foi necessário remover os dados sobre o verbo *ser*, pois ele tem uma grande quantidade de informações, porém não foi anotado semanticamente por nós, de modo que nos pareceu ser uma perda aceitável, ainda que se trate do verbo mais frequente no *corpus*.

Em seguida, precisamos realizar a indexação do formato de dados do arquivo XML para o formato da plataforma Jibiki. Nesse processo, os nodos e as demais informações presentes na estrutura XML são vinculados às categorias presentes na estrutura da plataforma. Assim, foi preciso indicar quais nodos correspondiam às entradas, quais informações indicavam a sintaxe, a semântica etc.

Por fim, a importação foi realizada com sucesso. Restava então apenas modificar a interface de visualização. Ela foi programada em linguagem XSLT, a qual ainda não dominávamos. Assim iniciou-se um estudo dessa linguagem e vários testes para modificar a interface e a deixar da forma como queríamos.

Acessando o *site* da plataforma Jibiki e selecionando a língua portuguesa, é possível consultar os dados deste estudo. Pode-se também observar que a interface de consulta dos dados em português é bem diferente daquela das demais línguas, tendo em vista que o português apresenta dados bem diferenciados. Nas Figuras 8.2 e 8.3, apresentamos imagens de nossos dados como estão disponíveis atualmente na plataforma Jibiki. Essas figuras apresentam o estado atual em que se encontra a interface de consulta. O banco de dados atual se encontra com os dados atuais do VerbLexPor.

Figura 8.2 – Plataforma Jibiki. Resultados do verbo **contar** nos dados de língua portuguesa. Informações de estruturas de subcategorização, voz e frequência.



Para acessar os dados, basta acessar o *site*, selecionar a língua portuguesa e digitar um verbo; se ele estiver presente no *corpus*, a consulta retornará uma lista com todas as estruturas de subcategorização do verbo consultado. Clicando com o mouse em cada uma das estruturas (ou simplesmente puxando a barra de rolagem para baixo), é

possível consultar as sentenças do *corpus* que correspondem a cada uma estruturas de subcategorização, conforme está ilustrado na Figura 8.3.

Figura 8.3 – Plataforma Jibiki. Resultados do verbo **fazer** nos dados de língua portuguesa. Informações de exemplos da estrutura de subcategorização, sintaxe e papéis semânticos.

8.6.2 Projeto CAMELEON

Tendo em vista nossa participação no Projeto CAMELEON (CAPES/COFECUB 707/11), tivemos a possibilidade de disponibilizar os dados para *download* no site do projeto[70], juntamente com algumas informações básicas sobre o estudo. Nesse site, disponibilizamos os dois arquivos XML (um de cada *corpus*) e todas as tabelas dos dois *corpora* no formato SQL, de maneira que também os dados do *corpus* de Cardiologia podem ser baixados.

O *download* é gratuito e pode ser feito por qualquer pessoa que tenha acesso à Internet. Desse modo, o trabalho realizado já está disponível para qualquer interessado que queira pesquisar as informações sintáticas e semânticas do VerbLexPor.

8.6.3 Considerações sobre a disponibilização do VerbLexPor

Com a disponibilização dos dados do VerbLexPor das duas formas apresentadas acima, garantimos que os dados possam ser consultados por quem quer que esteja interessado. Por um lado, se o interesse for uma consulta básica a verbos específicos, a plataforma Jibiki tem uma interface mais amigável para a análise. Por outro lado, se o

[70] Site: http://cameleon.imag.fr/xwiki/bin/view/Main/Semantic%20role%20labels%20corpus%20-%20Brazilian%20Portuguese. Acessado em: 06/02/2015.

interesse for uma análise contrastiva ou um estudo que requeira alguma manipulação dos dados, os formatos XML ou SQL contêm as informações necessárias para tal fim.

Esse passo da disponibilização dos dados garantiu um dos nossos objetivos, que era permitir que outros pesquisadores usassem nossos dados da forma como bem entendessem. Por constituírem um recurso léxico, os dados do VerbLexPor só encontram sua real utilidade se forem empregados ou integrados a outros recursos e pesquisas. Sendo assim, a disponibilização dos dados sempre foi uma de nossas preocupações desde o início desta tese, e cremos que os dois meios encontrados fazem jus às nossas intenções.

8.7 Fechamento do capítulo

Neste capítulo, discutimos a lista de papéis semânticos usada e apresentamos exemplos retirados do *corpus* para todos os casos em que foi possível encontrá-los; a metodologia foi apresentada muito rapidamente, tendo em vista que ela não foi muito modificada em relação à que foi detalhada no Capítulo 6; mostramos dados quantitativos e comparativos do VerbLexPor; e, por fim, indicamos onde os dados do VerbLexPor podem ser encontrados e baixados. Com os dados de que dispomos, já poderíamos começar a análise e discussão dos dados, retomando nossas questões de pesquisa e hipóteses. Contudo, antes de darmos esse passo, passamos para um outro experimento que desenvolvemos a partir dos dados que foram levantados: o agrupamento de verbos. Esse será o assunto do nosso próximo capítulo.

9 Agrupamentos de Verbos

Os experimentos que apresentamos agora estão vinculado aos resultados observados em nosso segundo estudo-piloto e aos dados do VerbLexPor. Neste capítulo, descrevemos dois experimentos de agrupamento de verbos que visaram a reproduzir de forma automática ou semiautomática a tarefa desenvolvida por Levin (1993). Os experimentos de agrupamento de verbos, como se apresentam aqui, podem ser vistos como uma consequência da anotação com papéis semânticos e como uma das possíveis aplicações do recurso que desenvolvemos.

O motivo de termos realizado esses experimentos se explica pelo fato de que, se houvesse grupos de verbos semanticamente próximos já delimitados para o português, a anotação de papéis semânticos poderia ser feita com base nos grupos, e não com base em cada um dos verbos. Por exemplo, como apresentamos no Capítulo 4, a VerbNet (KIPPER-SCHULER, 2005) utilizou o sistema de classes desenvolvido por Levin para aplicar os papéis semânticos anotados a milhares de verbos, apesar de terem sido efetivamente anotadas 272 classes. Assim, como ilustração, a anotação realizada para a classe 77 (que reúne alguns conceitos prototípicos do verbo *accept*), que consiste em apontar a existência de um AGENTE e um TEMA, serviu para todos os oito verbos presentes na classe 77: *accept*, *understand*, *encourage*, *discourage*, *disprefer*, *reject*, *repent*, *rue*.

Tendo em mente esse uso de classes verbais como facilitador da anotação e como multiplicador de resultados, o objetivo destes experimentos foi agrupar verbos que fossem semanticamente próximos com base em informações sintáticas e semânticas. Também analisamos essas informações individualmente, para que pudéssemos observar a contribuição dos dados sintáticos e semânticos para o agrupamento. Realizamos, como mencionado, dois experimentos: a) um com base nos dados do segundo estudo-piloto; e b) um com base nos dados atuais do VerbLexPor.

A maior parte das informações foi extraída automaticamente de nosso banco de dados; porém, como veremos na metodologia, alguns dados do Experimento I foram levantados manualmente durante a anotação realizada no segundo estudo-piloto, enquanto os dados do Experimento II foram levantados automaticamente. Desse modo, este capítulo estará dividido em duas partes, cada uma relatando um dos experimentos, os quais têm o mesmo objetivo: criar um agrupamento dos verbos semanticamente similares. Em ambos os estudos, também queríamos observar se a anotação semântica

auxilia na tarefa, de modo que temos resultados que consideram os papéis semânticos (além de outras informações) e resultados que só levam em consideração a sintaxe. Em cada uma das seções, relatamos a metodologia e os resultados dos experimentos, chamando atenção, desde já, que os resultados do Experimento II foram muito superiores aos resultados do Experimento I, o qual serviu quase como um estudo-piloto para o agrupamento de verbos. Ao final do capítulo, fazemos uma retomada do capítulo e uma consideração geral sobre os resultados obtidos.

9.1 Experimento I

Neste experimento, usamos uma metodologia manual e outra automática para agrupar os dados que tínhamos ao final do segundo estudo-piloto (Capítulo 6, Seção 2). O trabalho foi desenvolvido dentro do Projeto CAMELEON, em parceria com o Prof. Dr. Carlos Ramisch. Nossa hipótese de pesquisa, com esse experimento, era que os dados levantados manualmente (as alternâncias sintáticas), que veremos mais adiante, teriam resultados melhores para o agrupamento de verbos. Essa hipótese se baseia no trabalho de Levin (1993), que usou alternâncias sintáticas em sua classificação. Também levantamos a hipótese de que os papéis semânticos melhorariam o desempenho do agrupamento, tendo em vista que eles fornecem informações semânticas sobre os verbos em questão.

9.1.1 Metodologia

Após o segundo estudo-piloto, tínhamos disponíveis em nosso banco de dados informações sintáticas e semânticas. Pensando no trabalho de Levin (1993), ficavam faltando apenas informações sobre as alternâncias sintáticas[71] permitidas pelos verbos. Assim, primeiramente consultou-se a literatura para encontrar alternâncias sintáticas que são comuns em português[72]. Essa busca partiu das alternâncias descritas por Levin (1993), limitando-se apenas às que podem ocorrer em português, e, em seguida, passou a uma fase de acréscimo de outras alternâncias discutidas na literatura sobre a língua

[71] Alternâncias sintáticas são as diferentes estruturas sintáticas que um verbo admite. Por exemplo, o verbo "comer" pode ser encontrado tanto na forma transitiva direta (*Pedro comeu uma maça.*) quanto na forma intransitiva (*Pedro já comeu.*). Essas diferentes possibilidades são chamadas de alternâncias sintáticas (ou diátese, ou alternâncias de diátese).

[72] Como mencionamos na Seção 2.2, as estruturas de subcategorização podem ser empregadas como indicadoras das alternâncias sintáticas, porém, preferimos utilizar também um método semiautomático, que será explicado mais adiante, para reconhecer as alternâncias sintáticas possíveis para cada verbo.

portuguesa (CANÇADO, 1996; CHAGAS DE SOUZA, 1999; 2001; ÁVILA, 2006; CIRÍACO, 2007; MORAES, 2008; AMARAL, 2010). Ao todo, chegamos a dezoito alternâncias sintáticas possíveis, as quais estão listadas no Anexo C, com exemplos. Essas alternâncias são discutidas de modo resumido em Levin (1993) e Scarton (2013).

A partir da lista de alternâncias gerada, desenvolvemos um sistema em Python que permitia fazer automaticamente as conversões necessárias para cada sentença. Assim, para cada verbo, as alternâncias foram geradas a partir de uma sentença-modelo[73] e alimentadas automaticamente ao buscador do Google para que se obtivesse a sua frequência. A geração das alternâncias funcionava da seguinte maneira: a partir de uma sentença real encontrada nos *corpora* estudados, simplificávamos sua estrutura e gerávamos uma sentença-exemplo, como a que apresentamos no Exemplo 9.a, a seguir:

9.a. ele avaliou o resultado com o estudo no hospital

Como podemos ver, a sentença não possui pontuação ou letras maiúsculas, se apresenta na forma ativa e possui sempre dois adjuntos adverbiais no final: um representando um possível instrumento e outro representando um lugar. Essa sentença-exemplo era então automaticamente processada e convertida para as dezoito alternâncias que encontramos. Por exemplo, ela era convertida para o Exemplo 9.b, que representa a alternância passiva (e ignora os adjuntos adverbiais):

9.b. o resultado foi avaliado por ele

Cada uma das alternâncias geradas era enviada automaticamente para o buscador do Google[74], o qual retornava o número de ocorrências de cada uma delas. Os resultados eram armazenados em um arquivo CSV e, em seguida, era feita uma validação manual em relação às alternâncias sintáticas possíveis para cada verbo. Nesse procedimento semiautomático, puderam ser utilizados apenas verbos transitivos diretos,

[73] Um exemplo de sentença-modelo é o seguinte: *Pedro confirmou a história com seu depoimento na delegacia.*

[74] A sentença era enviada entre aspas e com três formas possíveis para o verbo (presente, pretérito perfeito e pretérito imperfeito).

principalmente pelo fato de que os intransitivos, transitivos indiretos e pronominais não são bem descritos na bibliografia no que diz respeito às alternâncias possíveis[75].

Após a validação das alternâncias, juntamos todos os dados que tínhamos disponíveis sobre os verbos, ou seja, as alternâncias sintáticas, as estruturas de subcategorização, as informações de classificação sintática (sujeito, objeto direto etc.) e a anotação de papéis semânticos disponíveis no banco de dados. Tendo essas informações, utilizamos o algoritmo de Lin (1998) para verificar a similaridade entre os verbos. Esse algoritmo calcula a similaridade de elementos de acordo com seus atributos e, por vezes, aponta características que podem escapar ao olho humano, como, por exemplo, a similaridade de comportamento sintático-semântico entre antônimos.

Os atributos utilizados neste experimento variaram de acordo com os dados que tínhamos à disposição; por isso, organizamos os dados em diferentes grupos de teste, de acordo com os atributos utilizados, conforme exemplificamos na Tabela 9.1.

Tabela 9.1 – Exemplos das quatro categorias de atributos para o agrupamento[76]

Método de Agrupamento	Verbo	Atributo 1	Atributo 2	Atributo 3
1	Apresentar	Alternância passiva	N/A	N/A
2	Apresentar	SUBJ[NP]_V_NP	10	N/A
3	Apresentar	SUJEITO	AGENT	23
4	Apresentar	SUBJ[NP]_V_NP	AGENT + THEME	5

A Tabela 9.1 apresenta apenas exemplos possíveis da disposição dos dados. Ela indica que, para este experimento, os dados foram divididos em quatro grupos, que correspondem aos seguintes atributos: Método 1 – apenas alternâncias sintáticas manualmente verificadas; Método 2 – estruturas de subcategorização e frequência; Método 3 – classificação sintática, papel semântico e frequência; e Método 4 – estrutura

[75] Chegamos a trabalhar também com verbos intransitivos, identificando possíveis alternâncias; porém, os resultados ainda estão muito incipientes para se fazer uma discussão sobre o assunto.

[76] Os atributos representados variam bastante, podendo representar tipos de alternâncias sintáticas, estruturas de subcategorização, classificações sintáticas, papéis semânticos ou frequência.

de subcategorização da sentença, estrutura de papéis semânticos da sentença e frequência. No caso dos Métodos 2 a 4, por serem dados extraídos diretamente do banco de dados, havia ainda uma distinção relativa aos *corpora* utilizados, de modo que, para cada um desses grupos, houve uma subdivisão entre os dados extraídos do *corpus* de Cardiologia e do Diário Gaúcho.

Para avaliar a acurácia dos resultados, utilizamos como padrão-ouro os dados do Thesaurus eletrônico para o Português do Brasil (TeP), versão 2.0, desenvolvido por Maziero, Pardo, Di Felippo e Dias da Silva (2008). No TeP 2.0, existem milhares de *synsets* organizados conforme os pressupostos da WordNet (FELLBAUM, 1998). Um ponto importante que se deve ressaltar é que a medida proposta por Lin (1998) é de similaridade, e não de sinonímia, dessa forma, os resultados em relação ao padrão-ouro devem ser visto com certa cautela, pois não são exatamente os mesmos critérios utilizados. Algo que é claramente complicado nesse caso é que a definição de similaridade está vinculada a um comportamento em relação a determinados atributos. Esse assunto será tratado com mais detalhes na seção a seguir, juntamente com os resultados.

9.1.2 Resultados e discussão

Com os resultados do cálculo de Lin (1998) aplicado aos grupos representados na Tabela 9.1, foi possível observar quais tipos de dados foram mais precisos em relação ao padrão-ouro. Para tal, também foi necessário estabelecer pontos de corte em relação aos dados. O cálculo de Lin resulta em valores que variam de 0 a 1[77]. Testamos, então, três pontos de corte diferentes (>0,0; ≥0,3; e ≥0,7)[78]. Os resultados dos três pontos de corte com a média da acurácia são apresentados na Tabela 9.2.

Como se pode ver na Tabela 9.2, a acurácia foi baixa. Um dos motivos que pode ter gerado esses resultados baixo é o padrão-ouro utilizado. Como mencionamos, o TeP 2.0 não apresenta exatamente o mesmo tipo de relação e também não abrange todas as relações possíveis da língua. Existem pares de verbos identificados como similares que não são contemplados pelo TeP, mas que são, de fato, similares. Por exemplo, em uma

[77] Quanto mais próximo de 1 for o resultado, maior a similaridade entre os verbos.

[78] Até onde sabemos, não existem pontos de corte pré-estabelecidos pela literatura; por isso, estabelecemos esses pontos de modo experimental, sem uma base prévia que desse suporte a eles. Como os cálculos eram realizados automaticamente, poderíamos ter usado quaisquer e quantos pontos de corte quiséssemos; porém, cremos que esses três representem bem a extratificação dos resultados que poderiam ser obtidos.

avaliação manual, percebe-se que a relação entre pares de verbos como *demonstrar* / *apresentar*, depreendida automaticamente em todas as categorias do ponto de corte >0,0 no *corpus* do Diário Gaúcho, tem uma similaridade inclusive de sinonímia[79] que não é refletida no TeP 2.0. Também temos pares como *aumentar* / *melhorar* (sinonímia), *identificar* / *conhecer* (hiperonímia), *oferecer* / *revelar* (sinonímia), *variar* / *reduzir* (hiperonímia), *permitir* / *confirmar* (sinonímia). Todos esses exemplos, que, em uma observação humana, podem ser considerados como similares, não aparecem como tal no TeP, seja por não ser o foco do recurso lexical (casos de hiperonímia) ou por simplesmente não ser abrangido por ele (casos de sinonímia). Em nosso caso, como temos poucos verbos sob análise, esse tipo de não correspondência pesou bastante no cálculo da acurácia.

Tabela 9.2 – Médias da acurácia dos resultados em relação aos três pontos de corte de acordo com o *corpus* e o método de agrupamento[80]

	>0,0	**≥0,3**	**≥0,7**
***Corpus* – Método de Agrupamento**	**Acurácia média (%)**	**Acurácia média (%)**	**Acurácia média (%)**
Independente de *corpus* – 1	13,16	<u>15,31</u>	<u>16,15</u>
Cardiologia – 2	12,42	11,02	2,13
Diário Gaúcho – 2	13,91	13,23	7,20
Cardiologia – 3	12,24	10,78	3,87
Diário Gaúcho – 3	15,38	9,90	9,59
Cardiologia – 4	11,52	9,48	7,14
Diário Gaúcho – 4	<u>17,73</u>	7,38	3,94

Isso não quer dizer que o TeP não é um recurso confiável, nem quer dizer que nossos resultados foram baixos porque o padrão-ouro não é apropriado[81], apenas

[79] Para observar a possibilidade das relações, utilizamos o mesmo conceito da WordNet (FELLBAUM, 1998) que é a substituição em contexto, o qual também foi empregado no TeP para as questões de sinonímia.

[80] Os valores sublinhados indicam os resultados mais altos para cada um dos pontos de corte.

[81] Recentemente, foi publicado um padrão-ouro que talvez fosse mais apropriado, por ter sido construído com os mesmos princípios de Levin (1993). Porém, esse padrão-ouro, apresentado por Scarton (2013),

ressaltamos que os resultados devem ser relativizados. Contudo, é preciso reconhecer que os resultados foram ruins; afinal, a quantidade de informação disponível para cada um dos verbos é bastante grande, advinda não apenas de uma classificação sintático-semântica, mas também de uma classificação manual de alternâncias. Quanto à diferença entre os *corpora*, já era esperado que os dados do *corpus* de Cardiologia fossem menos precisos que os do *corpus* do Diário Gaúcho, tendo em vista que os dados do TeP fazem referência à linguagem comum e não à especializada.

No caso das alternâncias (Método 1), por já ter sido uma opção testada e defendida por Levin (1993) para o inglês, esperávamos que os resultados fossem mais precisos. O que se viu, porém, foi que os resultados se mostraram menos precisos do que o uso de estruturas de subcategorização e estruturas de papéis semânticos (Método 4) no caso do ponto de corte >0,0. No entanto, a categoria das alternâncias foi a única que resultou em um aumento da acurácia juntamente com o aumento do ponto de corte. Com esses resultados, nossa hipótese de que as alternâncias sintáticas teriam resultados melhores para o agrupamento foi comprovada nos pontos de corte >0,3 e >0,7, mas não no ponto de corte >0,0.

Nossa segunda hipótese, de que os papéis semânticos melhorariam a classificação na porção dos dados levantados automaticamente (Métodos 3 e 4 em relação ao Método 2), não foi confirmada. No ponto de corte >0,3, a acurácia dos Métodos 3 e 4 foi menor do que a do Método 2. Nos demais pontos de corte, os resultados oscilaram, de modo que não é possível afirmar que os papéis semânticos melhoraram o agrupamento.

Um resultado positivo (ainda que relacionado apenas indiretamente ao experimento) foi a constatação de que a anotação de papéis semânticos se torna mais consistente após a análise das possíveis alternâncias sintáticas. Com elas, ficam mais claras, por exemplo, as alterações de posições entre sujeitos agentes e sujeitos oblíquos (casos de AGENTE e AGENTE LOCATIVO), o uso de INSTRUMENTOS na posição de sujeito etc. Isso facilita a anotação de papéis semânticos e dá mais consistência ao trabalho do anotador.

Depois que observamos os resultados do agrupamento, bastante inferiores aos de outros trabalhos, como, por exemplo, Scarton (2013), chegamos à conclusão de que

---

por ser amostral e abranger apenas algumas classes, infelizmente contém apenas 15 dos verbos com os quais trabalhamos. Desse modo, teríamos que descartar a maior parte de nossos verbos para poder utilizá-lo, o que não nos pareceu ser uma boa prática.

precisaríamos modificar o método para melhorar os resultados. O uso da medida de similaridade de Lin (1998) não rendeu os resultados positivos que esperávamos. Além disso, como tínhamos poucos dados, ficou difícil tirar conclusões generalizantes. Portanto, decidimos partir para um novo experimento, com uma nova metodologia.

9.2 Experimento II

Para realizar um novo experimento, esperamos até que tivéssemos os dados completos do VerbLexPor, tendo em vista que os dados do segundo estudo-piloto haviam sido poucos para as análises estatísticas. No caso do VerbLexPor, ainda não temos uma grande quantidade de dados, mas, por exemplo, no caso do Diário Gaúcho, temos quase quatro vezes mais verbos anotados, de modo que a confiabilidade dos resultados é maior. Ainda assim, como veremos mais adiante, optamos por um grau de confiança maior nos testes estatísticos. De posse de mais dados, optamos por retomar nossa hipótese do experimento anterior: os papéis semânticos melhoram o agrupamento de verbos.

O trabalho neste segundo experimento foi realizado em parceria com o doutorando Rodrigo de Sousa Wilkens, do Programa de Pós-Graduação em Computação da UFRGS, orientando da Profª. Drª. Aline Villavicencio.

9.2.1 Metodologia

Neste segundo experimento com agrupamento de verbos, optamos por usar um procedimento totalmente automático, exceto pela criação do padrão-ouro. Explicaremos nesta seção, passo a passo, a metodologia utilizada, começando pelos dados, passando pelas ferramentas, até chegarmos à comparação com o padrão-ouro.

Os dados que utilizamos foram provenientes do banco de dados do VerbLexPor em sua versão SQL (que contém também informações da anotação do *parser* PALAVRAS). Optamos por utilizar essa versão para podermos aproveitar as etiquetas semânticas que o PALAVRAS (BICK, 2000) anota para cada palavra. Desse modo, tínhamos informações sintáticas, na forma das estruturas de subcategorização, e informações semânticas, na forma dos papéis semânticos que anotamos e de etiquetas semânticas que o PALAVRAS usa. Não utilizamos informações de frequência, como, por exemplo, frequência da estrutura de subcategorização, por entender que essa informação, por se basear no *corpus* como um todo, poderia influenciar negativamente nos resultados.

Cada instância analisada se baseava em uma sentença do banco de dados. Desse modo, cada instância era composta pelo verbo da sentença em questão e por informações sintáticas e semânticas atribuídas a cada um dos argumentos desse mesmo verbo: sintaxe de estrutura, em formato de estrutura de subcategorização (SUBJ, NP, PP etc.); sintaxe, em formato de classificação (SUJEITO, OBJETO DIRETO, OBJETO INDIRETO etc.); sintaxe, na forma de voz ativa ou passiva; semântica, por meio dos papéis semânticos; e semântica, por meio das etiquetas do PALAVRAS. Cada informação dessas foi considerada como um grupo de atributos vinculados à instância. A seguir, apresentamos um exemplo simplificado de instância do verbo *encontrar*:

- encontrar
    - Arg1 = SUBJ – SUJEITO – H – agente
    - Arg2 = NP – OBJETO_DIRETO – co – tema

Essa instância representa os valores presentes no arquivo ARFF que foi usado para gerar os resultados. Cada instância, na realidade, é um vetor e possui centenas de atributos, distribuídos em variáveis binárias (sintaxe e semântica do PALAVRAS – cada variável representa uma das possíveis classificações, como SUJEITO, OBJETO DIRETO, state, tool etc.), variáveis não binárias (número do argumento, número da sentença etc.) e variáveis categóricas (verbos e papéis semânticos). Como cada argumento foi considerado de modo independente dos outros argumentos do verbo, cada um deles apresentava seus próprios atributos de sintaxe e semântica.

Com base nessas informações de sintaxe e semântica de cada argumento presente em cada instância, as instâncias de cada um dos verbos anotados no banco de dados foi agrupada por meio do algoritmo k-Means. O agrupamento automático foi levado a cabo por meio da ferramenta WEKA 3 (HALL, FRANK, *et al.*, 2009), sendo que foram testadas várias formas de agrupamento, de acordo com os atributos considerados. Desse modo, por meio de seleção dos atributos de cada instância, realizamos quatro agrupamentos diferentes:

- Agrupamento 1: Sintaxe + Papéis Semânticos + Semântica do PALAVRAS
- Agrupamento 2: Sintaxe + Papéis Semânticos
- Agrupamento 3: Apenas Sintaxe

- Agrupamento 4: Apenas Papéis Semânticos

Com base nesses dados, podíamos observar qual era a contribuição de cada um dos grupos de atributos para o agrupamento. O agrupamento foi realizado com base em dados de cada um dos *corpora* (Cardiologia e Diário Gaúcho) de modo independente.

Por fim, os resultados do agrupamento realizado pelo WEKA foram avaliados em relação a um padrão-ouro desenvolvido especialmente para esta tarefa de agrupamento, de modo que foram utilizados dois padrões-ouro: um para os resultados do *corpus* de Cardiologia e outro para os resultados do Diário Gaúcho. Os padrões-ouro foram criados com base nos dados do TeP 2.0 (MAZIERO, PARDO, *et al.*, 2008) e do padrão-ouro utilizado por Scarton (2013). O procedimento de criação do padrão-ouro foi semiautomático e seguiu estes passos:

1. Partindo do TeP 2.0, foram usados apenas os grupos que tinham os verbos que tínhamos anotado em cada um dos *corpora*.
2. Desses grupos selecionados, foram excluídos os verbos que não tínhamos anotado.
3. Os passos 1 e 2 foram aplicados ao padrão-ouro utilizado por Scarton (2013).
4. Foram removidos todos os grupos que fossem iguais.
5. Foram removidos grupos que estivessem contidos dentro de outros grupos.
6. Foram removidos todos os grupos que continham apenas um verbo.
7. Lemos cada um dos grupos restantes e removemos manualmente os grupos ou verbos individuais que apresentavam algum problema (como, por exemplo, significados arcaicos ou não sinonímicos).
8. Separamos uma lista de verbos dos nossos dados que não estavam em nenhum dos grupos que sobraram.
9. Analisamos grupo por grupo para ver se algum dos verbos que não estavam em nenhum deles não se aplicava ou poderia formar um novo grupo.
10. Inserimos grupos com apenas um verbo para cada verbo que não estivesse contemplado nos demais grupos.

Com esse procedimento, cremos termos chegado a dois padrões-ouro justos para os dados que testamos, pois eles contêm apenas os dados que usamos e foram avaliados manualmente para garantir que não houvesse problemas. A comparação entre os resultados e os padrões-ouro se deram por meio do cálculo do índice de similaridade de Jaccard, sendo que, a partir da comparação de todos os grupos gerados com todos os

grupos do padrão-ouro, selecionamos o que apresentava o índice máximo de similaridade.

É importante ressaltar que a quantidade predefinida de grupos utilizada no WEKA para o algoritmo k-Means foi baseada no padrão-ouro criado. Assim, por exemplo, para gerar os resultados do *corpus* de Cardiologia, fornecemos ao algoritmo o número de 60 grupos, que foi o número existente em nosso padrão-ouro.

Definidos esses procedimentos, podemos passar aos resultados.

9.2.2 Resultados do agrupamento

Os resultados numéricos do agrupamento estão resumidos na Tabela 9.3, a seguir. A primeira informação que salta aos olhos nessa tabela é que o método de agrupamento 1 (com sintaxe, papéis semânticos e semântica do PALAVRAS) e o método de agrupamento 2 (com sintaxe e papéis semânticos) geraram resultados exatamente iguais. A segunda informação é que o melhor resultado, em ambos os *corpora*, está no método 4, que utiliza apenas os atributos de papéis semânticos.

Tabela 9.3 – Resultado do agrupamento de verbos de acordo com o método de agrupamento e o *corpus*

Método	Cardiologia	DG
Agrupamento 1	47,63%	37,87%
Agrupamento 2	47,63%	37,87%
Agrupamento 3	49,92%	43,45%
Agrupamento 4	52,30%	43,79%

Se a diferença entre os métodos 2, 3 e 4 fosse realmente significativa, então teríamos a informação de que os dados semânticos e sintáticos juntos não contribuem para agrupar as palavras em grupos semânticos coesos, mas o uso de apenas papéis semânticos ou apenas sintaxe contribui.

Por um lado, entendemos que, para o agrupamento de verbos, uma classificação de papéis semânticos não é estritamente necessária, já que basta termos uma boa anotação sintática para atingirmos resultados satisfatórios que posteriormente poderão ser anotados semanticamente. Por outro lado, os resultados sugerem que os papéis semânticos são relativamente independentes da sintaxe no que diz respeito ao significado dos verbos agrupados. Tendo em vista que a análise de papéis semânticos é bastante influenciada pela sintaxe dos verbos em questão, decidimos observar se as

diferenças apresentadas eram significativas e, para tal, realizamos uma validação cruzada de 10 iterações. Como nosso conjunto de dados é bastante restrito, principalmente por provir de uma anotação manual de apenas um anotador, usamos um intervalo de confiança de 99% para avaliar a significância da diferença. Usando esses parâmetros, em ambos os *corpora*, a diferença entre os métodos 2, 3 e 4 nos três métodos não foi significativa. Isso quer dizer que os três métodos empregados são estatisticamente iguais num intervalo de confiança de 99%. Isso mostra que tanto a sintaxe quanto a semântica têm uma contribuição similar para o agrupamento de verbos, o que ressalta o vínculo existente entre esses dois fatores na distinção do significado de um verbo.

Observando os resultados do agrupamento como um todo, tivemos resultados em torno de 50% de similaridade com o padrão-ouro no *corpus* de Cardiologia e em torno de 40% no *corpus* do Diário Gaúcho. Uma questão que resulta desses valores é a seguinte: porque o *corpus* de linguagem especializada apresenta dados de agrupamento mais similares ao padrão-ouro do que o *corpus* de linguagem não especializada? Uma possibilidade é que, como os padrões-ouro foram criados especificamente com base no número de verbos de cada uma das amostras, a quantidade de dados pode ter influenciado nesse cálculo, já que o número de verbos anotados no *corpus* de Cardiologia é menos da metade do *corpus* do Diário Gaúcho.

Além do cálculo de similaridade de Jaccard, também realizamos um cálculo de precisão, abrangência e medida f, conforme apresentamos na Tabela 9.4. Esse cálculo tomou por base os grupos que apresentaram máxima similaridade, considerando os verbos presentes no grupo gerado pelo experimento e os verbos presentes no grupo eleito como mais similar no padrão-ouro. Excluímos o método de agrupamento 1 por ele ter rendido resultados exatamente iguais ao método 2, porém usando mais dados (o que indica que esses dados extras não foram aproveitados no agrupamento).

Infelizmente, apesar de termos realizado os cálculos de precisão, abrangência e medida f, nossos resultados ainda não podem ser comparados com os do estudo de Scarton (2013), simplesmente pelo fato de que nosso padrão-ouro foi construído de maneira diferente, o que impede, por exemplo, o cálculo de acurácia ponderada de classe, pois o resultado, com base em nossos dados, seria sempre 100%, já que todos os verbos do padrão-ouro estão necessariamente no agrupamento. Isso desequilibraria demais os resultados e provavelmente não refletiria a realidade.

Tabela 9.4 – Precisão, abrangência e medida f
para cada um dos métodos de agrupamento utilizados

Corpus	Método	Precisão	Abrangência	Medida f
Cardiologia	Agrupamento 2	34,04%	68,97%	45,58%
	Agrupamento 3	36,27%	64,91%	46,54%
	Agrupamento 4	36,79%	61,74%	46,10%
Diário Gaúcho	Agrupamento 2	26,07%	58,57%	36,08%
	Agrupamento 3	28,83%	53,47%	37,46%
	Agrupamento 4	30,28%	52,92%	38,52%

9.3 Considerações sobre os agrupamentos

Nosso primeiro experimento de agrupamento obteve resultados muito baixos, mas rendeu uma metodologia mais interessante e consistente para a anotação de papéis semânticos, que é o principal interesse desta tese. No segundo experimento, os resultados foram melhores, ainda que estejam longe de serem adequados para um agrupamento confiável e consistente de verbos.

Nossos resultados gerais ainda estão, em sua maioria, abaixo dos 50% de similaridade e também na medida f. Por isso, ainda é preciso investir maiores esforços nessa tarefa de agrupamentos para podermos melhorar os resultados. Contudo, como o agrupamento de verbos não é o escopo desta tese, que tinha por objetivo apenas observar as informações utilizadas para realizá-lo, encerramos por aqui a parte dedicada a essa tarefa.

10 Análise e Discussão dos Dados do VerbLexPor

Finalmente chegou o momento de esmiuçarmos os dados que obtivemos com a anotação dos dois *corpora*, seja em sua natureza contrastiva (entre si ou com outros recursos), seja em sua própria constituição. Optamos por misturar neste capítulo tanto a análise quanto a discussão dos dados, para não distanciarmos tanto uma etapa da outra, tendo em vista a importância de ambas. Em alguns momentos, reproduziremos também dados que já foram apresentados anteriormente, principalmente os do Capítulo 8, que serviram para descrever o recurso.

Neste capítulo, analisaremos primeiramente os dados dos dois *corpora*, primeiramente separados e, depois, em contraste. Em seguida, retomamos nossas questões de pesquisa e hipóteses à luz de nossos resultados. Por fim, fazemos uma breve consideração sobre os dados antes de passarmos ao capítulo final desta tese.

10.1 Análise dos dados

Nesta seção, colocamos os dados do VerbLexPor sob a lupa para vermos o que eles representam para a descrição do português. Começamos pelos dados do Diário Gaúcho, passamos para os dados de Cardiologia e, por fim, comparamos ambos.

10.1.1 Diário Gaúcho

Observando os dados do *corpus* do Diário Gaúcho, a primeira característica que nos salta aos olhos é a simplicidade da estrutura semântica dos textos. A grande maioria das sentenças tem uma estrutura que envolve AGENTE ou TEMA. Esses dois papéis semânticos são responsáveis por mais de 47% das anotações em todo o *corpus* (conforme pode ser visto na Tabela 8.3).

Isso pode apontar para algumas conclusões não necessariamente complementares: a) os dois papéis precisam ser mais bem estudados, tendo em vista que sua prevalência também é um fator que pode não fornecer muita informação semântica distintiva sobre os verbos; b) os verbos estudados se concentravam no espectro de ação-processo, pendendo para a agentividade, e no espectro estativo, com grande presença de sentenças com TEMAS sem AGENTES; e, por isso, c) os verbos mais frequentes do português são de ação-processo e estativos.

No Capítulo 8, apresentamos a Tabela 8.6, em que tínhamos as sentenças sintático-semânticas. A seguir, na Tabela 10.1, reproduzimos os dados com uma

pequena diferença: retiramos dela os papéis semânticos específicos para adjuntos adverbiais. Desse modo, temos apenas as ocorrências de papéis semânticos de complementos.

Tabela 10.1 – Sentenças sintático-semânticas do *corpus* do Diário Gaúcho, desconsiderando os papéis semânticos de adjuntos (amostra)

Sentença	Freq.	%
SUJEITO<agente> + OBJETO DIRETO<tema>	663	12,51%
SUJEITO<tema>	629	11,87%
SUJEITO<agente>	569	10,73%
SUJEITO<agente> + OBJ DIR ORACIONAL<topico>	184	3,47%
SUJEITO<experienciador> + OBJETO DIRETO<tema>	161	3,04%
SUJEITO<pivo> + OBJETO DIRETO<tema>	159	3,00%
SUJEITO<paciente>	151	2,85%
SUJEITO<experienciador> + OBJ DIR ORACIONAL<tema>	129	2,43%
SUJEITO<agente> + OBJETO DIRETO<topico>	111	2,09%
SUJEITO<agente> + OBJ DIR ORACIONAL<tema>	109	2,06%

Na Tabela 10.1, fica ainda mais visível a preponderância dos papéis AGENTE e TEMA em relação aos demais, sendo que mais de 35% das sentenças anotadas no Diário Gaúcho têm exclusivamente esses dois papéis semânticos como obrigatórios. O maior destaque, como nos lembramos da Tabela 8.3, fica para o papel de TEMA, responsável por mais de 27% (3.015) dos argumentos anotados (considerando a anotação de adjuntos). Depois desses dois papéis predominantes, encontramos, na sequência, os papéis de EXPERIENCIADOR, PACIENTE e TÓPICO, todos com mais ou menos a mesma porção dos argumentos (entre 4 e 5% cada).

Saindo um pouco do campo da semântica e observando apenas aspectos sintáticos, temos um reino supremo da voz ativa, sendo ela responsável por mais de 93% das sentenças anotadas. A estrutura de subcategorização mais frequente foi SUBJ_V_NP (mais de 22% das ocorrências), e as construções transitivas diretas e intransitivas foram as que ficaram no topo, sendo que apenas entre as três mais frequentes (SUBJ_V_NP, SUBJ_V e SUBJ_V_OCL, respectivamente) já temos um domínio de mais de 46% das estruturas de subcategorização. Se somarmos essas três a outras formas básicas, como reflexivas ou copulativas, o índice atinge os 55%. Isso mostra uma maioria de orações diretas e sem uso de sintagmas preposicionados (sejam adjuntos adverbiais ou complementos indiretos).

Todos esses dados fazem menção a uma linguagem direta e que propicia uma maior facilidade de associação entre os verbos e seus argumentos. O fato de que se emprega muito mais a voz ativa em vez da passiva indica uma explicitação maior dos agentes presentes na linguagem.

Agora que discutimos mais alguns resultados do *corpus* do Diário Gaúcho, vamos ver alguns dados do *corpus* de Cardiologia.

10.1.2 Cardiologia

Nos dados anotados do *corpus* de Cardiologia, o papel semântico predominante, sem um segundo colocado próximo, é TEMA, responsável por mais de 33% dos argumentos anotados. O segundo colocado, com quase 7% de ocorrência, é RESULTADO, seguido de perto por PIVÔ e, em seguida, com pouco mais de 6%, AGENTE. A partir desses dados, podemos concluir que o foco na Cardiologia se desloca bastante dos agentes para os objetos envolvidos na área especializada. Os agentes, ainda que explícitos em alguns casos, recuam para o segundo plano.

Isso fica ainda mais claro quando olhamos para a Tabela 10.2, com dados amostrais de sentenças sintático-semânticas da Cardiologia sem considerar papéis semânticos de adjuntos. Nela percebemos que as formas agentivas aparecem apenas nas posições 3 e 10, sendo responsáveis por muito poucas das sentenças anotadas no *corpus*: apenas pouco mais de 13% delas. Já o papel TEMA ocorre em mais de 73% das sentenças, seguido por RESULTADO, que ocorre em pouco mais de 15%, e por PIVÔ, com quase 15%. Isso representa praticamente um monopólio dos objetos e da inatividade.

Passando para uma análise sintática, a voz ativa foi bastante superior à voz passiva, mas a proporção foi de 74,57% contra 25,43%, respectivamente. Olhando para as estruturas de subcategorização, a voz passiva já aparece na terceira estrutura mais frequente (SUBJ_V) e ocorre em 4 das dez primeiras estruturas. As estruturas com sintagmas preposicionados (adjuntos adverbiais e complementos indiretos) foram maioria, estando em mais de 51% das estruturas. Isso indica uma tendência a orações ampliadas, com acréscimo de informações não essenciais ao significado da oração.

Tabela 10.2 – Sentenças sintático-semânticas do *corpus* de Cardiologia, desconsiderando os papéis semânticos de adjuntos (amostra)

Sentença	Freq.	%
SUJEITO<tema>	411	21,38 %
SUJEITO<pivo> + OBJETO DIRETO<tema>	182	9,47%
SUJEITO<agente> + OBJETO DIRETO<tema>	93	4,84%
SUJEITO<resultado>	64	3,33%
SUJEITO<tema> + PREDICATIVO<atributo>	60	3,12%
SUJEITO<instrumento> + OBJETO DIRETO<tema>	51	2,65%
SUJEITO<tema> + OBJETO REFLEXIVO<verbo> + PREDICATIVO<atributo>	50	2,60%
SUJEITO<causa> + OBJETO DIRETO<tema>	46	2,39%
SUJEITO<instrumento> + OBJ DIR ORACIONAL<tema>	45	2,34%
SUJEITO<agente> + OBJ DIR ORACIONAL<tema>	44	2,29%

10.1.3 Contraste entre Diário Gaúcho e Cardiologia

Após termos visto algumas informações individuais dos *corpora*, passamos agora a uma comparação entre os dados de ambos. Nossa intenção aqui é observar as semelhanças e/ou diferenças entre as estruturas sintáticas e semânticas entre a linguagem comum (representada pelo *corpus* do Diário Gaúcho) e uma linguagem especializada (representada pelo *corpus* de Cardiologia).

Nosso procedimento será muito parecido com o que já mostramos nas Seções 6.2.4.3 e 6.2.4.4. Começaremos pela análise estatística e, em seguida, passamos a considerações qualitativas sobre os dados em contraste.

10.1.3.1 Análise estatística

Para a análise estatística, usamos a mesma metodologia apresentada no Capítulo 6: a partir de listas de dados com frequências nos dois *corpora*, aplicamos o teste de correlação tau-b de Kendall com a ferramenta IBM SPSS 19. Conforme explicamos anteriormente, esse teste observa se há correlação entre os *rankings* de duas amostras. Desse modo, podemos ver se os *corpora* apresentam as informações num *ranking* de frequência parecido.

Diferentemente do segundo estudo-piloto, onde tínhamos apenas verbos iguais nos dois *corpora*, agora temos 191 verbos no Diário Gaúcho (DG) e apenas 77 no *corpus* de Cardiologia, sendo que 76 desses verbos são iguais nos dois *corpora*. Sendo

assim, optamos por fazer uma avaliação usando o teste de correlação tanto nos dados totais quanto apenas nos dados relativos a verbos que foram anotados nos dois *corpora*. As diferenças, como veremos, não foram grandes, sendo que os valores que mostrarmos entre parênteses correspondem aos dados totais, sem seleção de verbos.

Para facilitar a compreensão, reproduziremos nesta seção partes de várias tabelas que já foram vistas ao longo desta tese. Essas reproduções parciais de tabelas servirão apenas para orientar o leitor e facilitar o entendimento dos testes realizados, mas não acrescentarão informações novas que não tenham sido vistas em outros capítulos. As informações novas ficarão por parte dos testes realizados, que indicarão os graus de correlação entre os dados.

Começaremos então observando o nível mais abstrato de informação que temos: apenas o *ranking* de papéis semânticos. Para realizar o teste tau-b de Kendall, todos os dados correspondentes foram organizados conforme apresentamos na amostra a seguir (a totalidade dos dados pode ser vista na Tabela 8.3):

Papel	DG	Cardio
TEMA	3015	1416
AGENTE	2540	254
EXPERIENCIADOR	591	47
LUGAR	540	143
PACIENTE	497	145
etc.		

Com base apenas nessa estrutura mais abstrata formada apenas por papéis semânticos, o valor do tau-b (τ_b) foi 0,51, com $p < 0,01$ ($\tau_b = 0,48$; $p < 0,01$). Esse valor corresponde a uma correlação positiva. No entanto, essa é a única configuração em que temos correlação. Como veremos a seguir, todas as demais configurações resultaram em valores muito próximos de zero.

Quando diminuímos a abstração e acrescentamos um vínculo entre a sintaxe e a semântica (como nas Tabelas 8.4 e 8.5), partindo de dados como o que apresentamos a seguir, os valores de τ_b passam a 0,16, com $p < 0,01$ ($\tau_b = 0,13$; $p < 0,01$).

Sintaxe+Papel Semântico	DG	Cardio
SUJEITO<agente>	2511	236
OBJETO DIRETO<tema>	1343	480
SUJEITO<tema>	1010	684
SUJEITO<experienciador>	584	46
ADJUNTO ADVERBIAL[em]<lugar>	426	136
Etc.		

Se chegarmos ao nível da sentença sintático-semântica (como nas Tabelas 8.6 e 8.7), com a associação dos papéis semânticos e da sintaxe de todos os argumentos em torno do verbo, a correlação passa a ser negativa, com τ_b = -0,27 e p < 0,01 (τ_b = -0,28; p < 0,01).

Sintaxe+Papel	DG	Cardio
SUJEITO<agente> + OBJETO DIRETO<tema>	663	93
SUJEITO<tema>	629	411
SUJEITO<agente>	569	5
SUJEITO<agente> + OBJ DIR ORACIONAL<topico>	184	27
SUJEITO<experienciador> + OBJETO DIRETO<tema>	161	5
Etc.		

Como existem também papéis semânticos atribuídos somente a adjuntos que podem influenciar esta categoria, optamos por excluir das sentenças sintático-semânticas os papéis específicos de adjuntos (como DISCURSO, SITUAÇÃO, MOMENTO etc.), assim como apresentamos nas Tabelas 10.1 e 10.2. Com essa remoção, a correlação ficou muito próxima de zero, mas com valor negativo: τ_b = -0,08 e p = 0,061 (τ_b = -0,09; p = 0,013). Podemos observar que, quando retiramos esses papéis semânticos específicos para adjuntos, nossos dados para os verbos que foram anotados nos dois *corpora* acabam gerando valores apenas marginalmente significativos para um intervalo de confiança de 95% (p = 0,061), enquanto os dados que não levam em conta os verbos selecionados permanecem significativos (p = 0,013).

Dessa forma, o que essas estatísticas indicam é a mesma tendência indicada em nosso segundo estudo-piloto: quanto maior a complexidade dos dados e menor a abstração, menor é a correlação existente entre os dados. Isso mostra que as anotações do *corpus* de Cardiologia não têm dependência em relação às anotações do *corpus* do Diário Gaúcho e que, portanto, são diferentes entre si, exceto quando se observa apenas os papéis semânticos empregados. O problema com isso é que os papéis semânticos dificilmente podem ser compreendidos sem seu vínculo com a sintaxe, que, nesta

análise estatística, serve também como uma portadora da informação do verbo, já que uma análise de correlação verbo por verbo não apresenta dados suficientes para uma avaliação significativa dos dados.

Agora que observamos as diferenças e semelhanças através de uma breve análise estatística, passamos a observar os dados de um ponto de vista qualitativo.

10.1.3.2 Análise Qualitativa

Para esta análise qualitativa contrastiva, optamos por observar apenas os dados dos 76 verbos anotados nos dois *corpora*, tendo em vista que já apresentamos dados relativos aos *corpora* individuais no Capítulo 8. Esta análise servirá como um complemento a esse capítulo, pois algumas das observações feitas aqui em contraste já foram indicadas quando realizamos a descrição do recurso. Começaremos vendo diferenças mais amplas no campo dos papéis semânticos em geral e então passaremos a explicitar diferenças mais específicas de verbos individuais.

Tabela 10.3 – Papéis semânticos relativos apenas aos 76 verbos anotados em comum nos dois *corpora* (sem os papéis semânticos específicos para adjuntos)

Papel	DG	%	Cardio	%
TEMA	1.474	29,37%	1.396	37,61%
AGENTE	979	19,51%	254	6,84%
LUGAR	328	6,54%	143	3,85%
PIVÔ	275	5,48%	282	7,60%
RESULTADO	268	5,34%	289	7,79%
VERBO	229	4,56%	184	4,96%
EXPERIENCIADOR	205	4,08%	47	1,27%
TÓPICO	198	3,95%	68	1,83%
PACIENTE	180	3,59%	145	3,91%
FINALIDADE	150	2,99%	130	3,50%
CAUSA	134	2,67%	202	5,44%
ATRIBUTO	120	2,39%	136	3,66%
INSTRUMENTO	89	1,77%	208	5,60%
DESTINO	87	1,73%	8	0,22%
RECIPIENTE	75	1,49%	13	0,35%
BENEFICIÁRIO	48	0,96%	58	1,56%
AGENTE LOCATIVO	39	0,78%	3	0,08%
ALVO	23	0,46%	30	0,81%
LUGAR INICIAL	10	0,20%	2	0,05%
COTEMA	8	0,16%	48	1,29%
FONTE	5	0,10%	35	0,94%
VALOR	4	0,08%	1	0,03%

Papel	DG	%	Cardio	%
VARIAÇÃO	3	0,06%	1	0,03%
ESTÍMULO	41	0,82%	0	0,00%
RECÍPROCO	24	0,48%	0	0,00%
MATERIAL	11	0,22%	0	0,00%
ATIVO	7	0,14%	0	0,00%
COAGENTE	5	0,10%	0	0,00%
SE PASSIVO	0	0,00%	20	0,54%
COPACIENTE	0	0,00%	9	0,24%

Uma das diferenças que já era possível perceber a partir da descrição presente no Capítulo 8 é que os papéis semânticos de AGENTE e INSTRUMENTO têm uma predominância diferente nos dois *corpora*. Porém, observando a Tabela 10.3, vemos que não são apenas eles os protagonistas de diferenças.

Observando a Tabela 10.3, vemos que a diferença de ocorrência de AGENTES é quase três vezes maior (em porcentagem) no *corpus* do Diário Gaúcho em relação ao de Cardiologia, enquanto o inverso é verdadeiro para os papéis de INSTRUMENTO e CAUSA (este apenas o dobro no *corpus* de Cardiologia). Isso reforça a observação inicial de que o *corpus* de Cardiologia tende a suprimir de certa forma os AGENTES. Porém, por estarmos agora mostrando um contraste que envolve os mesmos verbos nos dois *corpora*, essa supressão dos AGENTES em apenas um *corpus* também mostra que isso não é um fator ligado aos verbos em questão, mas sim ao gênero textual envolvido[82].

Se observarmos em que posição ocorrem esses papéis semânticos, observamos que a distribuição do papel INSTRUMENTO ocorre mais predominantemente na posição de sujeito, e não de adjunto adverbial (Tabela 10.4). Isso indica que o papel INSTRUMENTO, frequentemente considerado um papel de adjuntos, na realidade, aparece mais frequentemente na posição de sujeito, devido a casos de alternância sintática.

Tabela 10.4 – Função sintática do papel semântico INSTRUMENTO nos *corpora*

Sintaxe + Papel	DG	Cardio
SUJEITO<instrumento>	66	120
AGENTE DA PASSIVA[por]<instrumento>	14	16
ADJUNTO ADVERBIAL[em]<instrumento>	6	31
ADJUNTO ADVERBIAL[com]<instrumento>	1	12
ADJUNTO ADVERBIAL[por=meio=de]<instrumento>	0	10
OBJETO INDIRETO[com]<instrumento>	0	7

[82] Consulte Finatto, Eichler e Del Pino (2003) para uma observação semelhante na área da Química.

Sintaxe + Papel	DG	Cardio
ADJUNTO ADVERBIAL[por]<instrumento>	0	7
OBJETO INDIRETO[por]<instrumento>	0	5
ADJUNTO ADVERBIAL[de]<instrumento>	1	0
OBJETO INDIRETO[em]<instrumento>	1	0

Outros papéis semânticos que chamaram atenção pela diferença entre os dois *corpora* foram os papéis vinculados à percepção, como é o caso de EXPERIENCIADOR e ESTÍMULO (sendo que este último só ocorreu no *corpus* do Diário Gaúcho). Isso indica que o gênero textual artigo de Cardiologia tende a não demonstrar percepções próprias dos envolvidos, mas sim se expressar de maneira objetiva, evitando verbos de percepção. Seria possível pensar que essa diferença se deu simplesmente porque não anotamos verbos suficientes de percepção, mas a verdade é que anotamos verbos como *amar*, *ver*, *ouvir*, *acreditar*, *lembrar* etc. Eles apenas não ocorreram em quantidade suficiente no *corpus* de Cardiologia (exceto pelos verbos *acreditar* e *lembrar*, que foram anotados nos dois *corpora*).

Seguindo essa mesma linha dos verbos de percepção, temos também uma grande diferença entre os dois *corpora* no que diz respeito aos verbos de (inter)locução, geralmente vinculados a um papel de TÓPICO como objeto direto ou indireto. O *corpus* do Diário Gaúcho apresentou mais que o dobro (em porcentagem) de ocorrência do papel TÓPICO. O interessante a observar nesse caso é que os verbos em questão foram também bastante diferentes. No *corpus* de Cardiologia, temos apenas quatro verbos anotados com o papel TÓPICO: afirmar, dizer, explicar e registrar; enquanto, no *corpus* do Diário Gaúcho, temos onze verbos: admitir, confirmar, contar, dizer, exigir, explicar, fazer, lembrar, reconhecer, registrar e revelar; lembramos que todos esses doze verbos mencionados foram anotados nos dois *corpora*, alguns deles simplesmente não apresentavam o argumento que seria TÓPICO, enquanto outros, como o verbo *fazer*, indicam um caso claro de verbo-suporte.

Esses casos de verbo-suporte também foram um elemento que nos chamou atenção. Alguns estudos de Terminologia Textual, principalmente de língua alemã, indicam que, nas linguagens especializadas, os substantivos deverbais têm predominância sobre os verbos, esvaziando o significado destes (HOFFMANN, 1998; 1998; WEINRICH, 2005, p. 988). Contudo, o que percebemos em nosso estudo foi que, em ambos os *corpora*, a ocorrência de verbos-suporte foi parecida, o que é indicado também pelo fato de o papel VERBO ter uma porcentagem de ocorrência próxima nos

dois *corpora*. Olhando para cada caso do papel semântico VERBO nos dois *corpora*, percebemos que os casos em que há verbo-suporte são muito mais frequentes no Diário Gaúcho, com 150 casos, do que no *corpus* de Cardiologia, que apresenta apenas 68 casos desse tipo. Isso é o oposto do que propõem Hoffmann (HOFFMANN, 1998; 1998) e Weinrich (2005), mas corrobora os dados levantados em estudo anterior (ZILIO, 2009), indicando que a ocorrência de verbos-suporte pode não ser um traço tão marcante quanto se assume em linguagens especializadas ou, ao menos, não na Cardiologia. Por um lado, é possível que essa diferença em nossas averiguações se dê pelo fato de que os três estudos mencionados são de língua alemã, enquanto, de nossos estudos, apenas o de 2009 envolvia a língua alemã. Por outro lado, isso pode ser um indício de uma mudança nas linguagens especializadas nos últimos anos, afinal, os estudos de língua alemã mencionados anteriormente, ainda que tenham data recente, foram realizados nos anos 80 ou antes.

Essas foram as considerações que julgamos mais importantes no que diz respeito aos aspectos semânticos do VerbLexPor no contraste das duas linguagens abordadas. Agora passamos a olhar para aspectos sintáticos. Por não se tratar de dados da anotação de papéis semânticos, usamos todos os dados dos *corpora* para averiguar as informações sintáticas; também consideramos apenas dados percentuais, tendo em vista que, como mostramos no Capítulo 5, os *corpora* têm tamanhos diferentes.

O primeiro aspecto que observamos foi a questão da voz empregada no discurso. Hoffmann (1998) chama atenção para a importância acentuada da voz passiva no discurso especializado. Desse modo, observamos se o VerbLexPor corroborava essa hipótese. Considerando todas as estruturas de subcategorização do VerbLexPor, o *corpus* de Cardiologia contou com 22.540 estruturas, contra 23.779 do Diário Gaúcho. Esses números são bastante próximos, com menos de 5% de diferença, mas preferimos considerar apenas os dados percentuais para garantir uma medição mais justa. Olhando para a voz ativa, vemos que, no *corpus* de Cardiologia, ela é responsável por 79,17% das estruturas de subcategorização do *corpus*. Já no *corpus* do Diário Gaúcho, essa dominação sobe para 93,29%. Desse modo, a voz passiva acaba sendo três vezes mais frequente no *corpus* de Cardiologia, sendo responsável por 20,83% das estruturas, contra apenas 6,71% no *corpus* do Diário Gaúcho.

Tabela 10.5 – As cinco estruturas de subcategorização mais frequentes em ambos os *corpora*

Sintaxe	DG	%	Cardio	%
SUBJ_V_NP+ATIVA	4.404	18,52%	3.509	15,57%
SUBJ_V_OCL+ATIVA	2.690	11,31%	1.136	5,04%
SUBJ_V+ATIVA	2.031	8,54%	862	3,82%
SUBJ_V_NP_PP[em]+ATIVA	924	3,89%	761	3,38%
SUBJ_V_PP[em]+ATIVA	821	3,45%	666	2,95%

Com isso, fica clara a importância elevada da voz passiva no *corpus* de linguagem especializada, em oposição à sua presença pálida no *corpus* de linguagem comum. Ainda assim as cinco estruturas de subcategorização mais frequentes em ambos os *corpora* são as mesmas, são na voz ativa e seguem a mesma ordem, como podemos ver na Tabela 10.5. Apenas nas estruturas seguintes é que os *corpora* se distinguem e que surgem as duas primeiras estruturas de subcategorização na voz passiva no *corpus* de Cardiologia.

Apesar de ter menos voz passiva, o *corpus* do Diário Gaúcho apresenta muito mais incidência de orações objetivas diretas, sendo que 15,89% das estruturas de subcategorização contêm esse tipo de estrutura sintática. No *corpus* de Cardiologia, esse tipo de estrutura aparece em apenas 7,98% das estruturas de subcategorização. Considerando somente as estruturas que contêm algum tipo de objeto direto (oracional ou não), a incidência de objetos diretos oracionais é de 15,99% no *corpus* de Cardiologia e de 27,74% no *corpus* do Diário Gaúcho. Sendo assim, estruturas oracionais mais complexas ocorrem com mais frequência no *corpus* de linguagem menos especializada. Essa informação vai contra nossas expectativas, pois esperávamos que a linguagem comum optasse pela forma mais simples, sem a adição de orações dentro de orações.

Essas foram algumas das informações que nos chamaram a atenção do ponto de vista qualitativo. Agora que terminamos a parte de análise dos dados do *corpus* de maneira contrastiva, vamos observar como essas informações respondem nossas questões de pesquisa e como ficam as nossas hipóteses.

10.2 Questões de pesquisa e hipóteses

Tendo já analisado algumas das diversas possibilidades que o VerbLexPor apresenta, vamos ver agora como essas análises nos ajudam a responder a nossas questões de pesquisa e a confirmar ou refutar nossas hipóteses. Optamos por deixar esta parte separada da análise feita nas seções anteriores para dar a ela o devido destaque, de modo que nossas questões e hipóteses não ficassem espalhadas ao longo do texto, mas sim concentradas em uma seção específica.

Retomaremos aqui primeiro cada uma das questões com suas respectivas respostas e, em seguida, cada uma das hipóteses com as informações que coletamos para refutá-las ou confirma-las. Começamos então pela primeira questão de pesquisa que levantamos lá no início, na Seção 1.4:

- Como se caracterizam as estruturas argumentais de verbos do português brasileiro em textos de jornalismo popular?

A resposta a esta pergunta pode ser encontrada no Capítulo 8, quando descrevemos o VerbLexPor. Partindo da observação do *corpus* do Diário Gaúcho, que é nosso representante de linguagem não especializada, podemos observar quais são as estruturas mais e menos frequentes. É preciso ter em mente aqui que não analisamos todos os verbos existentes, mas sim uma amostra de menos de 10% dos verbos totais do Diário Gaúcho. No entanto, essa amostra dos 191 verbos mais frequentes é responsável por mais de 51% das sentenças no Diário Gaúcho. Desse modo, temos uma boa amostra para observar as estruturas argumentais de verbos.

Observando as sentenças anotadas no *corpus*, mais de 47% delas envolve o argumento sujeito<AGENTE> e mais de 25%, o objeto direto<TEMA>. Esses são os dois argumentos que se apresentam mais frequentemente nas orações. Lembramos que as frequências apresentadas são relativas ao número total de sentenças anotadas no *corpus*, ou seja, 5.301. Desse modo, ainda que quase todas elas tenham um sujeito, nem todas têm objetos diretos, e menos ainda têm objetos indiretos.

No que diz respeito apenas aos tipos de sujeito, se unirmos AGENTES (47,37%), TEMAS (19,05%) e EXPERIENCIADORES (11,02%), temos mais de 77% dos sujeitos de todas as sentenças. Se acrescentarmos o papel semântico PIVÔ como sujeito a esse cálculo, então passamos dos 84%. Olhando para os objetos diretos (oracionais ou não), temos o papel TEMA (31,82%) como principal participante, seguido de longe por TÓPICO

(7,47%) e RESULTADO (3,72%). Os demais participantes objetos diretos se distribuem entre diversos papéis semânticos. Ao observarmos apenas os objetos indiretos, podemos indicar TEMA (4,76% - para comparação, esse valor é mais de 27% de todos os objetos indiretos) como papel predominante, seguido por DESTINO (1,97%) e RECIPIENTE (1,84%).

Se passarmos a observar as estruturas argumentais de orações inteiras, então temos sujeito<AGENTE> + objeto direto <TEMA> como a estrutura mais frequente, com 8,3% (como pode ser visto na Tabela 8.6). Se somarmos a essa estrutura a ocorrência de objeto direto oracional<TEMA>, então a porcentagem sobe para 10%. Ressaltamos aqui o fato de que os papéis semânticos específicos para adjuntos foram excluídos dessa contagem, tendo em vista que eles não são parte da estrutura argumental *per se*. Seguindo a tabela, temos duas estruturas de orações intransitivas na sequência, com sujeito<AGENTE> (6,8%) e sujeito<TEMA> (4,9%). Só então se quebra a hegemonia de AGENTES e TEMAS, com a introdução de uma estrutura com sujeito <AGENTE> + objeto direto oracional <TÓPICO> (3,3%), se somarmos a isso a ocorrência de objeto direto não oracional, então essa mesma estrutura passa a 5,1%, superando a oração com apenas sujeito <TEMA>. Na sequência, entram duas estruturas oracionais com sujeito<EXPERIENCIADOR>, uma com objeto direto<TEMA> (2,5%), e a outra com objeto direto oracional<TEMA> (2,4%). Apenas com essas estruturas, já temos 31,7% de todas elas. Os outros 68,3% são representados por 662 estruturas diferentes.

Agora que mostramos as estruturas mais frequentes, que representam uma grande parte das estruturas argumentais em nosso *corpus*, passamos à nossa segunda questão de pesquisa:

- Se existirem, quais são as diferenças que marcam textos especializados em relação a textos não especializados no que diz respeito às estruturas sintáticas e semânticas?

Para responder a essa questão, recorremos aos dados que acabamos de apresentar neste capítulo, quando comparamos os dados dos dois *corpora*. Começando pelas estruturas sintáticas, vimos que um dos fatores que diferenciaram os dois *corpora* foi a incidência de voz passiva, que teve uma predominância maior no *corpus* de Cardiologia. No que diz respeito a estruturas sintático-semânticas, vimos que a ocorrência de sujeito<AGENTE> foi muito maior no *corpus* do Diário Gaúcho, e que o

corpus de Cardiologia apresenta uma tendência a suprimir os agentes das orações e substituí-los por instrumentos ou pela própria voz passiva.

As maiores semelhanças ficaram por parte das estruturas de subcategorização, sendo que as cinco primeiras nos dois *corpora*, responsáveis por uma grande porcentagem do total, são exatamente as mesmas e seguem a mesma ordem. Também tivemos o papel semântico TEMA como mais frequente nos dois *corpora*.

Temos também as informações referentes ao ranqueamento dos itens sintáticos e semânticos, que apontam para semelhanças e diferenças entre os dois gêneros estudados, mas preferimos deixar para apontar esses resultados na discussão da segunda hipótese, que virá mais adiante. Por ora, vamos à primeira hipótese:

- ***Diferentes gêneros textuais podem compartilhar um conjunto de papéis semânticos descritivos genéricos.***

Esta hipótese foi confirmada pela própria existência do recurso que aqui apresentamos. O teste desta hipótese foi um dos motivos pelo qual realizamos dois estudos-piloto antes de iniciarmos a anotação que deu origem ao VerbLexPor, e também foi o motivo pelo qual realizamos algumas modificações na lista de papéis semânticos ao longo desta tese. Em última instância, a lista de papéis semânticos que utilizamos não apresentou dificuldades para ser empregada tanto na Cardiologia quanto nos textos do Diário Gaúcho, de modo que ambos os gêneros textuais envolvidos utilizaram o mesmo conjunto de papéis semânticos descritivos sem qualquer problema de compatibilidade.

Se houve algo que talvez tenha deixado a desejar, isso se refere a alguns papéis, como TEMA e AGENTE, serem muito genéricos e não passarem uma semântica muito precisa das orações anotadas. É claro que existiram casos que não foram prototípicos e que podem, à primeira vista, parecer um encaixe forçado a uma categoria; porém, se olharmos para as descrições dos papéis semânticos que utilizamos e observarmos os exemplos existentes no *corpus*, é mais provável que, em caso de discórdia por parte do observador, a primeira intenção seja de trocar o papel semântico empregado por outro papel semântico existente na lista, e não por um papel semântico que não está na lista.

Passemos então à segunda hipótese:

- ***O que define a especificidade dos domínios nos* corpora *estudados é o* ranking *dos papéis semânticos.***

Essa hipótese foi confirmada. É possível discordar disso com base no fato de que a correlação entre as listas de frequência dos papéis semânticos isolados foi positiva, com $\tau > 0,5$, o que representa uma correlação média. No entanto, ao longo do estudo, percebemos que os papéis semânticos estão necessariamente vinculados à sintaxe e aos verbos ou outros elementos presentes nas orações e sentenças. Dessa forma, fica claro que o ranqueamento dos papéis semânticos foi marcadamente diferente nos dois domínios estudados, sendo que bastou vincular os papéis semânticos a suas categorias sintáticas para termos praticamente uma inexistência de correlação entre as listas, chegando a uma correlação negativa quando comparadas as orações como um todo.

Pudemos observar também o ranqueamento do ponto de vista qualitativo, com destaque para alguns elementos específicos como a distribuição dos papéis semânticos AGENTE e INSTRUMENTO no domínio da Cardiologia e do Diário Gaúcho. Assim, ainda que a correlação entre os dois no que diz respeito aos papéis semânticos isolados tenha sido média, existem alguns papéis semânticos que demonstram um comportamento distinto, o que pode ser visto como relevante para os Estudos Terminológicos no que diz respeito à caracterização dos domínios.

Passamos agora para breves considerações sobre este capítulo antes de passarmos para as nossas considerações finais sobre esta tese.

10.3 Considerações

Neste capítulo, organizamos a discussão dos resultados em duas partes: uma para apresentar e discutir uma análise do VerbLexPor, e outra para mostrar brevemente como os resultados apresentados serviram para responder às nossas questões de pesquisa e nortear a confirmação de nossas hipóteses. Na parte de análise, optamos por separar os dados específicos do *corpus* do Diário Gaúcho dos dados de Cardiologia, e observamos apenas alguns aspectos importantes em cada um dos *corpora*. Observamos que o Diário Gaúcho tem um uso predominante de sentenças mais simples, com presença de AGENTE e sem extensões preposicionais entre as estruturas sintáticas mais frequentes. Na Cardiologia, vimos que 51% das orações têm sintagmas preposicionados e que os papéis predominantes são TEMA e, bem menos frequentes, RESULTADO e PIVÔ.

Na análise contrastiva que realizamos entre os *corpora*, ficou clara a maior importância da voz passiva no *corpus* de Cardiologia, e confirmamos a ascensão dos INSTRUMENTOS à posição de sujeito. Nas medidas estatísticas observadas, vimos que, quanto mais específicos foram os dados, maior a distância entre os dois *corpora*.

Assim, neste capítulo, discutimos alguns dos resultados que mais nos chamaram a atenção no VerbLexPor. Os dados disponíveis se dispõem a uma série de outros possíveis estudos que aqui não foram contemplados, mas que poderão ser realizados futuramente por nós ou por outros pesquisadores interessados. Essas e outras questões serão abordadas no capítulo seguinte, que encerra esta tese.

11 Considerações Finais

Ao longo desta tese, realizamos uma série de experimentos diferentes, sejam eles relacionados diretamente ao recurso que desenvolvemos com o VerbLexPor, ou aos procedimentos que fizeram parte do seu desenvolvimento. Tínhamos, no início do trabalho, dois objetivos, que eram:

> *desenvolver um recurso léxico com informações sobre papéis semânticos para o português.*

e

> *realizar uma comparação entre as sentenças e verbos nos gêneros textuais especializado e não especializado.*

Agora que chegamos ao final desta tese, podemos olhar para o que realizamos e ver que nossos dois objetivos foram cumpridos, mas também percebemos que, ao realizar esses objetivos, novas informações surgiram e novas possibilidades de estudo foram se abrindo. Neste capítulo final, nossa intenção é retomar os pontos principais do trabalho que foi realizado, mas também apontar alguns dos caminhos relacionados que ainda precisam ser trilhados para podermos seguir avançando no conhecimento e na pesquisa de verbos e papéis semânticos.

Em relação ao primeiro objetivo, está bem claro que já temos um recurso léxico desenvolvido, com informações de papéis semânticos e dividido em dois gêneros textuais distintos: textos de jornalismo popular e artigos científicos de Cardiologia. Anotamos ao todo 192 verbos, sendo 191 no *corpus* do Diário Gaúcho e 76 no *corpus* de Cardiologia (destes, 75 em comum com o Diário Gaúcho). Os 191 verbos do Diário Gaúcho foram os mais frequentes do *corpus* (exceto por quatro verbos que foram excluídos, conforme explicado na metodologia). A anotação foi amostral, mas ainda assim resultou em 5.301 sentenças anotadas no *corpus* do Diário Gaúcho e 1.931 sentenças no *corpus* de Cardiologia, totalizando mais de quinze mil argumentos anotados nos dois *corpora*. Esse material todo se encontra atualmente disponível para *download* em dois formatos (XML ou SQL) no site do Projeto CAMELEON e o *corpus* do Diário Gaúcho está também disponível para consulta *online* na Plataforma Jibiki.

Com essas facilidades, mesmo que o interessado não tenha grandes habilidades computacionais, ainda é possível analisar *online* as anotações realizadas.

Mesmo com todos esses dados disponíveis, ainda temos muita coisa por fazer. Estamos em vias de desenvolver um anotador automático de papéis semânticos, partindo de iniciativas como a de Alva-Manchego (2013), para facilitar a anotação de outros *corpora* ou mesmo para a expansão do próprio VerbLexPor. O ponto em que chegamos é o fim desta tese, mas não é o ponto final do recurso, que ainda pode e deve ser expandido para outros gêneros textuais e para abranger mais verbos.

Uma das principais críticas que o trabalho tem recebido é justamente pelo fato de a anotação ter sido feita por apenas um linguista, já que a confiabilidade dos dados anotados pode ser questionada, tendo em vista que não há uma medida que confirme que mais de uma pessoa teria anotado daquela maneira. Essa crítica é válida e, infelizmente, na atual circunstância em que nos encontramos, não temos nem mesmo como realizar um teste para avaliar a anotação, tendo em vista que os testes de comparação realizados no Capítulo 8 submeteram muitos dados a alterações, pois ainda não temos, de fato, um padrão-ouro ou outra anotação do mesmo gênero que possa ser considerada como um termo de comparação. No entanto, como ficou claro pelos estudos-piloto realizados e apresentados no Capítulo 6, nossa experiência com anotação de papéis semânticos não é casual. Houve um período extenso de estudo de teorias e listas de papéis semânticos antes de chegarmos onde estamos, e cremos que o resultado obtido está à altura do que se espera de um recurso anotado com papéis semânticos descritivos.

Ainda assim, também temos perspectivas de avançar na anotação e, se possível, contar com mais anotadores no futuro, esse foi o motivo pelo qual um dos experimentos que realizamos foi a anotação de papéis semânticos por múltiplos anotadores (Capítulo 7). Pensamos nesse experimento simples porque, apesar de confiarmos na qualidade de nosso trabalho, somente um trabalho com mais anotadores permite verificar estatisticamente essa mesma qualidade. No Capítulo 7, vimos que nossos resultados de concordância entre os anotadores foram baixos, mas também obtivemos dados importantes para a realização de um trabalho com mais de um anotador. Agora temos uma noção melhor do que está envolvido num trabalho como esse, principalmente no que diz respeito ao treinamento dos anotadores antes de realizar a tarefa, que é bastante complexa. Também aprendemos com esse estudo que a interface de anotação precisa ser mais amigável e, agora que temos uma quantidade razoável de dados já anotados e que

estamos em vias de desenvolver um anotador de papéis semânticos, temos maiores perspectivas para auxiliar a anotação de dados com mais anotadores. Falta-nos ainda um projeto exclusivo dedicado à anotação, mas isso é uma questão de tempo e planejamento.

Voltando aos nossos objetivos, e passando ao segundo objetivo, cremos que a quantidade de informações que levantamos nesta tese em relação aos gêneros textuais estudados satisfaz esse objetivo. Muitas das informações observadas neste estudo, principalmente no que diz respeito ao nosso *corpus* especializado, nunca foram avaliadas, pois não há, em nosso conhecimento, outro *corpus* de Cardiologia de língua portuguesa anotado com papéis semânticos. Uma das informações que nos chamou a atenção, por exemplo, foi a presença de muito mais sujeitos com papel semântico INSTRUMENTO do que no *corpus* de linguagem comum. Outros estudos, incluindo o do próprio Swales (1990), já apontaram que existe uma tendência à impessoalidade nos textos técnicos e científicos. Porém, até então, o que se havia verificado era uma ocorrência mais acentuada de voz passiva (assim como observamos) e de sujeitos não humanos, mas ainda não se havia apontado que tipo de sujeitos eram esses. Com o estudo que realizamos, é possível observar qual papel semântico foi atribuído a vários sujeitos, inclusive aqueles que, de fato, são sujeitos com papel de AGENTE.

Mas nossa análise não se reteve apenas a ver o que o *corpus* de Cardiologia apresentava de diferente em relação ao *corpus* do Diário Gaúcho. Observamos que o texto presente no Diário Gaúcho tende a apresentar estruturas semânticas simples em suas sentenças, sendo que mais de 50% dos argumentos foram anotados apenas com AGENTE ou TEMA. Isso pode ser uma marca do próprio gênero textual, que busca passar informações de maneira direta e sem complicações. As predominâncias de AGENTE e TEMA também indicam que o uso de verbos de ação-processo e de verbos estativos é maior em relação ao uso de verbos de processo, como já havia anunciado Borba (1990) em seu dicionário. Na Cardiologia, como mencionamos, os papéis semânticos AGENTE e EXPERIENCIADOR perdem sua importância e, em seu lugar, crescem os papéis TEMA, RESULTADO, INSTRUMENTO, PIVÔ e CAUSA.

No que diz respeito ao papel TEMA, ressaltamos mais de uma vez que ele parece ser pouco expressivo semanticamente. Em nossa lista, temos incorporadas algumas divisões desse papel semântico, como, por exemplo, PACIENTE (um TEMA que é afetado) e TÓPICO (o TEMA de uma interlocução). Porém, um estudo mais aprofundado dos argumentos anotados com esse papel podem revelar a existência de outras informações

relevantes para uma maior delimitação desse papel semântico. Com certeza, agora que temos um recurso anotado, esse tipo de estudo se tornou bem mais fácil, ainda que se trate de mais de dois mil argumentos só no *corpus* do Diário Gaúcho.

Voltando o olhar para nossas hipóteses, as duas foram confirmadas, demonstrando que gêneros textuais diferentes podem compartilhar um mesmo conjunto de papéis semânticos descritivos e que o que os diferencia nesse quesito é o ranqueamento desses papéis em sua associação com a sintaxe. A comprovação dessas hipóteses tem várias implicações importantes para o conhecimento linguístico. Ao mostrarmos que apenas uma lista de papéis semânticos pode ser compartilhada por gêneros textuais diferentes, mostramos também que temos uma lista robusta que pode ser utilizada na anotação de outros gêneros textuais. Além disso, quando observamos que os papéis semânticos no *corpus* especializado tiveram um comportamento diferente, contribuímos para um novo ponto de vista na distinção dos gêneros textuais e apresentamos uma nova informação importante para a Terminologia.

Por fim, realizamos dois experimentos relativos ao agrupamento de verbos, na esperança de que fosse possível encontrar um método confiável de agrupar verbos semanticamente próximos com base nos dados de que dispomos. Nossos resultados não foram muito promissores, sendo que eles ficaram, em sua maioria, abaixo de 50% no índice de similaridade de Jaccard. Porém, uma informação ficou clara a partir dos resultados: tanto a sintaxe quanto a semântica parecem contribuir igualmente para o agrupamento de verbos. Isso mostrou que ambas influenciam no significado dos verbos, e que ambas são igualmente confiáveis para a realização de agrupamentos. Desse modo, é possível expandir o agrupamento de verbos para englobar também aqueles que ainda não receberam anotações semânticas, pois o uso exclusivo da sintaxe se mostrou como um parâmetro confiável.

Esses experimentos de agrupamento tiveram, além de seu resultado para o próprio objetivo de agrupar verbos, também uma resultado colateral de mostrar um dos inúmeros possíveis empregos do VerbLexPor para o PLN. Assim, além de termos apresentado dados que enriquecem o conhecimento linguístico sobre o português através principalmente de informações sobre gêneros textuais, também aproveitamos para trabalhar uma das várias possibilidade de aplicação para o PLN, ressaltando mais uma vez o caráter interdisciplinar desta tese.

Além desse caráter interdisciplinar, tivemos também um trabalho de colaboração internacional que apresentamos nesta tese. Primeiramente, tivemos a colaboração com o

Prof. Dr. Carlos Ramisch para o primeiro teste de agrupamento de verbos e para os testes de ranqueamento dos papéis semânticos. E, juntamente com esses trabalhos, também tivemos a transposição do banco de dados do *corpus* do Diário Gaúcho para a plataforma Jibiki, desenvolvida pelo Prof. Dr. Mathieu Mangeot (2006), que trabalhou diretamente conosco nessa empreitada durante o estágio de doutorado-sanduíche que realizamos no Laboratoire d'Informatique de Grenoble. Com essas colaborações, avançamos em nosso estudo e disponibilizamos alguns dados do VerbLexPor para consulta *online*, facilitando a consulta para os interessados que desejam apenas verificar rapidamente algumas informações, sem precisar percorrer um arquivo XML ou um banco de dados em MySQL.

Assim, ao longo do trabalho colaborativo com diferentes colaboradores, em diferentes institutos e instituições de pesquisa, os resultados interdisciplinares e de colaboração internacional que esta tese apresentou serviram para fortalecer laços entre equipes de pesquisa que se esforçam para o desenvolvimento de um conhecimento linguístico em comum, cada uma com seus diferentes pontos de vista e aplicações. Dessa forma, o VerbLexPor está disponível para que novos estudos possam ser realizados sobre essa base de conhecimento que apresentamos nesta tese e que colocamos à disposição da comunidade acadêmica.

Bibliografia

ALVA-MANCHEGO, F. E. **Anotação automática semissupervisionada de papéis semânticos para o português do Brasil**. USP. São Carlos, p. 137. 2013.

AMARAL, L. **Os Verbos de Modo de Movimento no Português Brasileiro**. Belo Horizonte: UFMG, 2010. Trabalho de Conclusão de Curso.

ARTSTEIN, R.; POESIO, M. Inter-Coder Agreement for Computational Linguistics. **Computational Linguistics**, 34, n. 4, 2008. 555-596.

ÁVILA, M. C. **Propriedades semânticas e alternâncias sintáticas do verbo:** um exercício exploratório de delimitação do significado. Araraquara: UNESP, 2006. Dissertação de Mestrado.

BAKER, C. F.; FILLMORE, C. J.; LOWE, J. B. **The Berkeley FrameNet project**. COLING-ACL '98: Proceedings of the Conference. Montreal, Canadá: [s.n.]. 1998. p. 86-90.

BAKHTIN, M. **Estética da criação verbal**. Tradução de Maria Ermantina Galvão G. Pereira. São Paulo: Martins Fontes, 1997.

BEAUGRANDE, R.-A. D.; DRESSLER, W. Introduction to Text Linguistics. **Site do Prof. Beaugrande**, Berlim, 2002. Disponivel em: <http://beaugrande.com/introduction_to_text_linguistics.htm>. Acesso em: 26 fev. 2015.

BECHARA, E. **Moderna gramática portuguesa**. 37ª. ed. Rio de Janeiro: Lucerna, 1999.

BERBER SARDINHA, T. **Lingüística de Corpus**. Barueri: Manole, 2004.

BERTOLDI, A.; CHISHMAN, R. L. O. **Desafios para a anotacão semântica de textos jurídicos:** limites no uso da FrameNet e rotas alternativas. Anais do X Encontro de Linguística de Corpus. Belo Horizonte, MG: Faculdade de Letras da UFMG. 2012. p. 103-121.

BEVILACQUA, C. R. **Unidades fraseológicas especializadas eventivas:** descripción y reglas de formación en el ámbito de la energía solar. Barcelona: IULA/UPF, 2004. Tese de doutorado. Orientador: Maria Teresa Cabré Castellví.

BIBER, D.; CONRAD, S.; REPPEN, R. **Corpus Linguistics:** Investigating language structure and use. Cambridge: CUP, 1998.

BICK, E. **The Parsing System PALAVRAS:** automatic grammatical analysis of Portuguese in a constraint grammar framework. Aarhus: Aarhus University Press, 2000.

BOITET, C.; MANGEOT, M.; SÉRASSET, G. **The Papillon project:** cooperatively building a multilingual lexical data-base to derive open source dictionaries & lexicons. Proceedings of On NLP and XML (NLPXML 2002), COLING Workshop. Taipei, Taiwan: [s.n.]. 2002. p. 9-15.

BORBA, F. D. S. **Dicionário Gramatical de Verbos do Português Contemporâneo do Brasil**. São Paulo: UNESP, 1990.

BORBA, F. D. S. **Dicionário de Usos do Português do Brasil**. São Paulo: Ática, 2002.

BOUQUET, S. **Introdução à leitura de Saussure**. São Paulo: Cultrix, 1997.

BRANCO, A. et al. **The Portuguese Language in the Digital Era / A língua portuguesa na era digital**. Heidelberg, Nova Iorque: Springer, 2012.

BRUMM, T. **Erstellung eines Systems thematischer Rollen mit Hilfe einer multiplen Fallstudie**. [S.l.]: [s.n.], 2008. 103 p. TCC. Orientador: Tom Gelhausen.

BURCHARDT, A. et al. **SALTO - A Versatile Multi-Level Annotation Tool**. Proceedings of LREC 2006. [S.l.]: [s.n.]. 2006.

CAMPO, A. A.; ARAQUE, I. R. Corpus Pattern Analysis in determining specialised uses of verbal lexical units. **Terminàlia 7**, Barcelona, 2013. 26-33.

CANÇADO, M. Verbos Psicológicos: Análise Descritiva dos Dados do Português Brasileiro. **Revista de Estudos da Linguagem**, 4, n. 1, 1996. 89-114.

CANÇADO, M. Posições Argumentais e Propriedades Semânticas. **DELTA**, São Paulo, n. 21, 2005. 23-56.

CANÇADO, M. Argumentos: Complementos e Adjuntos. **Revista Alfa**, São Paulo, 53, n. 1, 2009. 35-59.

CANÇADO, M. Verbal Alternations in Brazilian Portuguese: a Lexical Semantic Approach. **Studies in Hispanic and Lusophone Linguistics**, 3, n. 1, 2010. 77-111.

CANÇADO, M.; GODOY, L.; AMARAL, L. **The construction of a catalog of Brazilian Portuguese verbs**. Proceedings of the Workshop on Recent Developments and Applications of Lexical-Semantic Resources (LexSem 2012), in conjunction with KONVENS 2012. Viena, Itália: [s.n.]. 2012. p. 438-445.

CANÇADO, M.; GODOY, L.; AMARAL, L. **Catálogo de verbos do português brasileiro:** classificação verbal Segundo a decomposição de predicados: volume 1: verbos de mudança. Belo Horizonte: Editora UFMG, 2013.

CHAGAS DE SOUZA, P. **A Alternância Causativa no Português do Brasil:** Defaults num Léxico Gerativo. São Paulo: Universidade de São Paulo, 1999. Tese de Doutorado em Linguística. Faculdade de Filosofia, Letras e Ciências Humanas, Universidade de São Paulo.

CHAGAS DE SOUZA, P. **Notas Sobre a Construção Adversativa**. Anais do 4º Encontro do Círculo de Estudos Linguísticos do Sul (CELSUL). Curitiba, PR: [s.n.]. 2001.

CHISHMAN, R. L. O.; SPADER, D.; PADILHA, J. G. Kicktionary_Br: um relato sobre a anotação semântica de um corpus voltado ao domínio do futebol. **Revista Veredas**, 17, 2013. 101-116.

CIRÍACO, L. S. **A alternância causativo/ergativa no PB:** restrições e propriedades semânticas. Belo Horizonte: [s.n.], 2007. Dissertação (Mestrado em Linguística). Faculdade de Letras, Universidade Federal de Minas Gerais.

COHEN, J. A coefficient of agreement for nominal scales. **Educational and Psychological Measurement**, 20, 1960. 37-46.

CUNHA, C. F. D.; CINTRA, L. F. L. **Nova gramática do Português Contemporâneo**. Rio de Janeiro: Nova Fronteira, 1985.

DA SILVA, E. B.; BABINI, M. A preparação de material terminológico em língua inglesa por meio de ferramentas linguístico-computacionais. **Trabalhos em Linguística Aplicada**, 50, n. 1, jan.-jul. 2011.

DAVIES, M.; FLEISS, J. L. Measuring agreement for multinomial data. **Biometrics**, 38, n. 4, 1982. 1047–1051.

DIAS-DA-SILVA, B. C. **A face tecnológica dos estudos da linguagem:** o processamento automático das línguas naturais. [S.l.]: Unesp, 1996. Tese de doutorado. Orientador: Telmo Correia Arrais.

DIAS-DA-SILVA, B. C. O estudo Lingüístico-Computacional da Linguagem. **Letras de Hoje**, Porto Alegre, 41, n. 2, 2006. 103-138.

DIAS-DA-SILVA, B. C... I. **A construção da base da wordnet.br:** conquistas e desafios. Proceedings of the Third Workshop in Information and Human Language Technology (TIL 2005), in conjunction with XXV Congresso da Sociedade Brasileira de Computação. São Leopoldo, RS: [s.n.]. 2005. p. 2238–2247.

DIAS-DA-SILVA, B. C.; FELIPPO, A. D.; NUNES, M. D. G. V. **The automatic mapping of Princeton WordNet lexicalconceptual relations onto the Brazilian Portuguese WordNet database**. Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC 2008). Marrakech, Morocco: [s.n.]. 2008. p. 1535-1541.

DOWTY, D. Thematic Proto-Roles and Argument Selection. **Language**, 67, n. 3, Set. 1991. 547-619.

DURAN, M. S.; ALUÍSIO, S. M. **Propbank-Br:** a Brazilian Portuguese corpus annotated with semantic role labels. Proceedings of the 8th Symposium in Information and Human Language Technology. Cuiabá, MT: [s.n.]. 2011.

DURAN, M. S.; ALUÍSIO, S. M. **Propbank-Br:** a Brazilian treebank annotated with semantic role labels. Proceedings of the LREC 2012. Istambul, Turquia: [s.n.]. 2012.

FELLBAUM, C. **WordNet:** An electronic lexical database. Cambridge, Massachusetts: MIT Press, 1998.

FENG, M.; SUN, W.; NEY, H. **Semantic cohesion model for phrase-based SMT**. Proceedings of COLING 2012. Mumbai, India: [s.n.]. 2012. p. 867–878.

FERNANDES, F. **Dicionário de verbos e regimes**. 4ª. ed. Porto Alegre: Ed. Globo, 1963.

FILLMORE, C. J. **The case for case**. Proceedings of the Texas Symposium on Language Universals. [S.l.]: [s.n.]. 1967.

FINATTO, M. J. B. et al. **Características do jornalismo popular:** avaliação da inteligibilidade e auxílio à descrição do gênero. VIII Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana, 2011, Cuiabá - MT. Anais do STIL 2011. Cuiabá: Sociedade Brasileira de Computação. 2011. p. 30-39.

FINATTO, M. J. B.; EICHLER, M. L.; DEL PINO, J. C. Sujeitos e agentes de poder e dever em textos sobre equilíbrio químico: aspectos lingüístico-terminológicos e aspectos conceituais da enunciação científica e o ensino-aprendizagem de Química. **Revista Organon**, Porto Alegre, 32-33, n. 16, 2003. 83-104.

FLEISS, J. L. Measuring nominal scale agreement among many raters. **Psychological Bulletin**, 76, n. 5, 1971. 378–382.

FOSSATI, M.; GIULIANO, C.; TONELLI, S. **Outsourcing FrameNet to the Crowd**. Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics. Sofia, Bulgária: [s.n.]. 2013. p. 742–747.

FRANCHI, C. Teoria da adjunção: predicação e relações temáticas. **Revista Estudos da Linguagem**, 11, n. 2, 2003. 155-176.

FRANCHI, C.; CANÇADO, M. Teoria generalizada dos papéis temáticos. **Revista Estudos da Linguagem**, 11, n. 2, 2003. 83-123.

FRANCIS, N.; KUCERA, H. **Brown Corpus**. Providence: Brown University, 1964.

GELHAUSEN, T. **Modellextraktion aus natürlichen Sprachen:** Eine Methode zur systematischen Erstellung von Domänenmodellen. Karlsruhe: KIT Scientific Publishing, 2010. Dissertation, Karlsruher Institut für Technologie.

GILDEA, D.; JURAFSKY, D. Automatic Semantic Role Labeling. **Computer Linguistics**, Cambridge, 28, n. 3, 2002. 245-288.

GRUBER, J. S. **Studies in Lexical Relations**. MIT. [S.l.]. 1965. Orientador: Edward S. Klima.

HALL, M. et al. The WEKA Data Mining Software: An Update. In: _____ **SIGKDD Explorations**. 1. ed. Hamilton: University of Waikato, v. 11, 2009.

HARRIS, Z. S. The structure of science information. **Journal of Biomedical Informatics**, 35, 2002. 215–221.

HOFFMANN, L. Grundbegriffe der Fachsprachenlinguistik. In: _____ **Germanistisches Jahrbuch für Nordeuropa. Deutsche Fachsprachen in Forschung und Lehre**. Helsinki, Estocolmo: [s.n.], v. VII, 1988. p. 9-16.

HOFFMANN, L. Fachsprachen als Subsprachen. **Fachsprachen: ein internationales Handbuch zur Fachsprachenforschung und Terminologiewissenschaft**, Berlim, Nova Iorque, 1, 1998.

HOFFMANN, L. Syntaktische und morphologische Eigenschaften von Fachsprachen. **Fachsprachen: ein internationales Handbuch zur Fachsprachenforschung und Terminologiewissenschaft**, Berlim, Nova Iorque, I, 1998.

HONG, J.; BAKER, C. F. **How good is the crowd at "real" wsd?** Proceedings of the Fifth Law Workshop (LAW V). Portland, Oregon: [s.n.]. 2011. p. 30-37.

HOVY, E. et al. **OntoNotes:** The 90% solution. Proceedings of the Human Language Technology Conference of the North American Chapter of the ACL. Nova Iorque: [s.n.]. 2006. p. 57–60.

IENCO, D.; VILLATA, S.; BOSCO, C. **Automatic extraction of subcategorization frames for Italian**. Proceedings of the LREC 2008. [S.l.]: [s.n.]. 2008.

JACKENDOFF, R. S. **Semantic Structures**. Cambridge: MIT Press, v. 18, 1990. Current Studies in Linguistic Series.

JACKENDOFF, R. S. Conceptual Semantics. In: MAIENBORN, C.; HEUSINGER, K. V.; PORTNER, P. **Semantics:** An International Handbook of Natural Language Meaning. [S.l.]: De Gruyter Mouton, v. 1, 2011. p. 688-709.

JONES, B. et al. **Semantics-based machine translation with hyperedge replacement grammars**. Proceedings of COLING 2012. Mumbai, India: [s.n.]. 2012. p. 1359–1376.

JURAFSKY, D.; MARTIN, J. H. **Speech and Language Processing:** An Introduction to Natural Language Processing, Speech Recognition, and Computational Linguistics. Upper Saddle River, NJ: Prentice-Hall, 2000.

KASPER, S. **A comparison of 'thematic role' theories**. Philipps-Universität Marburg. [S.l.]. 2008. Dissertação de mestrado.

KIPPER-SCHULER, K. **VerbNet: a broad-coverage, comprehensive verb lexicon**. University of Pennsylvania. [S.l.]. 2005. Tese de doutorado. Orientador: Martha S. Palmer.

KONG, F.; ZHOU, G. **Exploring local and global semantic information for event pronoun resolution**. Proceedings of COLING 2012. Mumbai, India: [s.n.]. 2012. p. 1475–1488.

LEVIN, B. **English Verb Classes and Alternations:** A Preliminary Investigation. Chicago: University of Chicago Press, 1993.

LEVIN, B.; RAPPAPORT-HOVAV, M. **Argument Realization**. Cambridge, Nova Iorque, Melbourne, Madri, Cape Town, Singapure, São Paulo: Cambridge University Press, 2005.

LIMA, B. D. A. F. D. **Valência dos verbos de vitória e derrota em português**. Belo Horizonte: UFMG, 2007. Dissertação de Mestrado.

LIMA, V. L. S. D.; NUNES, M. D. G. V.; VIEIRA, R. **Desafios do Processamento de Línguas Naturais**. Anais do XXVII Congresso da SBC. [S.l.]: [s.n.]. 2007. p. 2202-2216.

LIN, D. **Automatic retrieval and clustering of similar words**. Proceedings of the 17th International Conference on Computational Linguistics, Association for Computational Linguistics. Morristown, NJ: [s.n.]. 1998. p. 768-774.

LOPER, E.; YI, S.-T.; PALMER, M. **Combining Lexical Resources:** Mapping Between PropBank and VerbNet. Proceedings of the 7th International Workshop on Computational Semantics. Tilburg, Holanda: [s.n.]. 2007.

LORENTE, M. **Verbos y fraseología en los discursos de especialidad**. XI Jornadas de Lingüística: homenaje al profesor José Luis Guijarro Morales. Cádiz: Universidad de Cádiz - Servicio de Publicaciones. 2009. p. 55-84.

LUFT, C. P. **Dicionário de regência verbal**. São Paulo: Ática, 1996.

MACIEL, A. M. B. **Para o reconhecimento da especificidade do termo jurídico**. UFRGS. Porto Alegre. 2001. Tese de doutorado. Orientador: Maria da Graça Krieger.

MANGEOT, M. **Dictionary Building with the Jibiki Platform**. Proceedings of EURALEX 2006, Software Demonstration. Torino: [s.n.]. 2006.

MANNING, C. D. **Automatic aquisition of a large subcategorization dictionary from corpora**. ACL '93 Proceedings of the 31st annual meeting on Association for Computational Linguistics. [S.l.]: [s.n.]. 1993. p. 235-242.

MARCUSCHI, L. A. Gêneros textuais: definição e funcionalidade. In: DIONISIO, Â. P.; MACHADO, A. R.; BEZERRA, M. A. **Gêneros textuais e ensino**. Rio de Janeiro: Lucerna, 2002. p. 19-36.

MAZIERO, E. G. et al. **A Base de Dados Lexical e a Interface Web do TeP 2.0 - Thesaurus Eletronico para o Portugues do Brasil**. VI TIL. [S.l.]: [s.n.]. 2008. p. 390–392.

MESQUITA, E. M. D. C. Algumas considerações sobre os textos técnico e jornalístico. **Linguagem: estudos e pesquisas, catalão**, 4-5, 2004.

MESSIANT, C. **A subcategorization acquisition system for French verbs**. Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Techonologies. Columbus, Ohio: [s.n.]. 2008. p. 55-60.

MESSIANT, C.; KORHONEN, A.; POIBEAU, T. **LexSchem:** A Large Subcategorization Lexicon for French Verbs. Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC). Marrakech, Marrocos: [s.n.]. 2008.

MORAES, H. R. **Aspectos sintaticamente relevantes do significado lexical: estudo dos verbos de movimento**. Universidade Estadual Paulista. Araraquara. 2008. Tese de doutorado.

NEVES, M. H. D. M. **Gramática de Usos do Português**. São Paulo: Unesp, 2000.

NEVES, M. H. D. M. Le poids de la notion tesniérienne de centralité du verbe dans les analyses linguistiques. **Synergies Brésil**, 13, 2013. 35-47.

NUNES, M. D. G. V. O Processamento de Línguas Naturais: Para quê e para quem? **Notas Didáticas do ICMC**, São Carlos, 2008.

OTHERO, G. D. Á. Lingüística Computacional: uma breve introdução. **Letras de Hoje**, Porto Alegre, 41, n. 2, 2006. 341-351.

OTHERO, G. D. Á.; MENUZZI, S. D. M. **Lingüística Computacional:** teoria e prática. São Paulo: Parábola Editorial, 2005.

PALMER, M. **Semlink:** Linking PropBank, VerbNet and FrameNet. Proceedings of the Generative Lexicon Conference. Pisa, Itália: [s.n.]. 2009.

PALMER, M.; GILDEA, D.; KINGSBURY, P. The Proposition Bank: A Corpus Annotated with Semantic Roles. **Computational Linguistics Journal**, 31, n. 1, 2005.

PERINI, M. A. **Estudos de Gramática Descritiva:** as valências verbais. São Paulo: Parábola Editorial, 2008.

PICHT, H. **Fachsprachliche Phraseologie – die terminologische Funktion von Verben**. Terminology and Knowledge Engineering. Proceedings of the International Congress on Terminology and Knowledge Engineering. Frankfurt a.M.: INDEKS Verlag. 1987. p. 21-34.

POSSAMAI, V.; LEIPNITZ, L. **Os estudos de gênero e a tradução:** uma relação proveitosa demonstrada por meio da abordagem da tradução de artigos científicos. Anais do 4º SIGET. Simpósio Internacional de Estudos de Gêneros Textuais. Tubarão: UNISUL. 2007. p. 2016-2027.

PREISS, J.; BRISCOE, T.; KORHONEN, A. **A System for Large-scale Acquisition of Verbal, Nominal and Adjectival Subcategorization Frames from Corpora**. Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics. Praga, República Tcheca: [s.n.]. 2007.

RAMISCH, C.; VILLAVICENCIO, A.; BOITET, C. **mwetoolkit:** a Framework for Multiword Expression Identification. Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC 2010). Valetta, Malta: [s.n.]. 2010.

RAMISCH, C.; VILLAVICENCIO, A.; BOITET, C. **Web-based and combined language models:** a case study on noun compound identification. Proceedings of the 23rd International Conference on Computational Linguistics (COLING 2010). Pequim: [s.n.]. 2010.

ROSA, J. L. G. **Fundamentos da Inteligência Artificial**. 1ª. ed. São Paulo: LTC, 2011.

SALOMÃO, M. FrameNet Brasil: um trabalho em progresso. **Calidoscópio**, 7, n. 3, 2009. 171-182.

SANTOS, D.; CARDOSO, N. **Reconhecimento de entidades mencionadas em português: Documentação e actas do HAREM, a primeira avaliação conjunta na área**. Linguateca. [S.l.]. 2007.

SAUSSURE, F. D. **Curso de Lingüística Geral**. 27ª. ed. São Paulo: Cultrix, 2006. Organizado por Charles Bally, Albert Sechehaye, com a colaboração de Albert Riedlinger. Tradução de Antônio Chelini, José Paulo Paes, Izidoro Blikstein. 1ª edição original em francês de 1916.

SCARTON, C. **VerbNet.Br: construção semiautomática de um léxico verbal online e independente de domínio para o português do Brasil**. NILC/USP. [S.l.]. 2013. Dissertação de mestrado. Orientador: Sandra Maria Aluísio.

SCHULTE IM WALDE, S. **A Subcategorisation Lexicon for German Verbs induced from a Lexicalised PCFG**. Proceedings of the 3rd Conference on Language Resources and Evaluation. Las Palmas de Gran Canaria, Espanha: [s.n.]. 2002. p. 1351–1357.

SCOTT, M. Oxford Wordsmith Tools, version 4.0, 2007. Disponivel em: <http://www.lexically.net/downloads/version4/wordsmith.pdf>. Acesso em: 08 mar. 2015.

STUBBS, M. **Text and Corpus Analysis:** Computer Assisted Studies of Language and Culture. [S.l.]: Wiley, 1996.

SUN, L.; KORHONEN, A. **Improving verb clustering with automatically acquired selectional preferences**. Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing (EMNLP 2009). Singapura: [s.n.]. 2009. p. 638-647.

SWALES, J. M. **Genre analysis:** English in academic and research settings. Cambridge: Cambridge University Press, 1990.

TABOADA, M.; DAS, D. Annotation upon Annotation: Adding Signalling Information to a Corpus of Discourse Relations. **Dialogue and Discourse**, 4, n. 2, 2013. 249-281.

VIDAL, V.; CABRÉ, M. T. **Estrategias para la ensenanza de combinaciones léxicas metaforicas en un curso de lenguas para fines específicos**. Lingüística aplicada en la sociedad de la información y la comunicación. Palma de Mallorca: Universitat de les Illes Balears. 2005. p. 187-195.

VIEIRA, R.; LIMA, V. L. S. D. **Lingüística computacional:** princípios e aplicações. Anais do Congresso da Sociedade Brasileira de Computação. Fortaleza: SBC. 2001. p. 47-88.

WEINRICH, H. **Textgrammatik der deutschen Sprache**. 3ª. ed. Hildesheim, Zurique, Nova Iorque: Georg Olms Verlag, 2005.

YOSHIKAWA, K. et al. **Sentence compression with semantic role constraints**. Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics. Jeju Island, Korea: [s.n.]. 2012. p. 349–353.

ZANETTE, A. **Aquisição de Subcategorization Frames para Verbos da Língua Portuguesa**. UFRGS. [S.l.]. 2010. Projeto de Diplomação. Orientadora: Aline Villavicencio.

ZANETTE, A.; SCARTON, C.; ZILIO, L. **Automatic extraction of subcategorization frames from corpora:** an approach to Portuguese. Proceedings of PROPOR 2012 - Demonstration Session. Coimbra, Portugal: [s.n.]. 2012.

ZAPIRAN, B.; AGIRRE, E.; MÀRQUEZ, L. **Robustness and Generalization of Role Sets:** PropBank vs. VerbNet. Proceedings of the ACL-08: HLT. Association for Computational Linguistics. Columbus, Ohio: [s.n.]. 2008. p. 2008.

ZILIO, L. **Colocações especializadas e 'Komposita' : um estudo constrastivo alemão-português na área de cardiologia**. UFRGS. Porto Alegre. 2009. Dissertação de Mestrado. Orientador: Maria José Bocorny Finatto.

ZILIO, L. TERMO E VALOR LINGUÍSTICO: UMA ABORDAGEM ENSAÍSTICA. **CADERNOS DO IL**, Porto Alegre, 42, 2011.

ZILIO, L. Colocações Especializadas em Alemão e Português na Área de Cardiologia. **Tradterm**, São Paulo, 20, dezembro 2012. 146-177.

ZILIO, L.; ZANETTE, A.; SCARTON, C. **Extração automática de estruturas de subcategorização a partir de corpora em português**. Anais do ELC 2012, XI Encontro de Linguística de Corpus. São Carlos. SP: [s.n.]. 2012.

ZILIO, L.; ZANETTE, A.; SCARTON, C. Automatic extraction of subcategorization frames from portuguese corpora. In: ALUISIO, S. M.; TAGNIN, S. E. O.; (EDS.) **New Languages Technologies and Linguistic Research:** a Two-Way Road. Cambridge: Cambridge Scholars Publishing, 2014. p. 78-96.

Anexo A

Nas tabelas deste anexo, adotamos as seguintes abreviaturas para facilitar a descrição dos papéis utilizados:

C = Controle = pode parar a ação (um teste possível é usar a locução "dicidiu não mais" ou "decidiu parar de").
D = Desencadeador = vinculado ao agente, mas também pode estar no Experienciador.
A = Afetado = mudança de um estado A para um estado B (a mudança pode ser de posse, lugar, estado mental ou físico *etc*.).
E = Estativo = não sofre alteração em relação à ação, ao processo ou ao estado em questão – não pode estar junto com afetado ou desencadeador.
F = elemento Físico = indica algo que é concreto, em oposição a abstrato.
M = processo Mental = algo que indica premeditação ou um envolvimento mental.

Tabela de Papéis Semânticos Utilizados no Estudo-Piloto I e Características Adicionais

Temas com origem e destino			
Papel	**Tradução**	**Descrição**	**Características**
actus	ação	ação que é realizada por alguém ou algo	ocorre com verbo suporte e alguns outros verbos
agens	agente	pessoa ou coisa que realiza a ação	D + (C)
patiens	paciente	pessoa ou coisa afetada por uma ação	A + (F)
notio	experienciado	impressão, sensação, conceito, imagem, ideia ou experiência que é sentida por alguém ou algo	
stimulus	estímulo	pessoa ou coisa que gera alguma sensação em outra pessoa ou coisa	D
experior	experienciador	aquele que sente algo	A + M
favor	benefício	vantagem ou desvantagem de uma pessoa ou coisa	E
fautor	beneficiante	pessoa ou coisa que gera um benefício ou um malefício para outra pessoa ou coisa	D
beneficiens	beneficiado	pessoa ou coisa que recebe um benefício ou malefício	A
habitum	posse	pessoa ou coisa que é possuída (mesmo que temporariamente), recebida ou dada a outra	(A) + F

		pessoa ou coisa	
donor	donatário	pessoa ou coisa que dá algo ou alguém	D ou A
recipient	recipiente	pessoa ou coisa que recebe algo ou alguém	D ou A
possessor	possuidor	pessoa ou coisa que possui algo ou alguém	E
locus dimensio	dimensão geográfica	medida de um lugar	E
locus origo	local de origem	ponto de origem de algo ou alguém	E
locus destinatio	local de destino	ponto de destino de algo ou alguém	E
locus positio	local	posição geográfica de algo ou alguém	E
limes	trajeto	caminho percorrido	E
tempus dimensio	dimensão temporal	medida do tempo	E
tempus origo	início	início	E
tempus destinatio	fim	fim	E
tempus positio	momento	algum ponto no tempo (ou uma situação)	E
frequens	frequência	frequência ou várias ocorrências de uma ação	E

Temas com dois elementos			
Papel	**Tradução**	**Descrição**	**Características**
dux	guia	pessoa ou coisa que é acompanhada	verbo de estado
comes	acompanhante	pessoa ou coisa que acompanha algo ou alguém	verbo de estado
compariens	comparado	pessoa ou coisa comparada	verbo de estado
comparand	modelo	pessoa ou coisa à qual se compara algo ou alguém	verbo de estado
contrariens	contrariado	pessoa ou coisa que tem um adversário	verbo de estado
contrarius	opositor	adversário de algo ou alguém	verbo de estado
figens	ator	pessoa ou coisa que desempenha um papel	verbo de estado
fictum	papel	papel desempenhado por alguém ou algo	verbo de estado
qualifitiens	qualificado	pessoa ou coisa sendo qualificada	verbo de estado
qualitas	qualidade	qualidade de algo ou alguém	verbo de estado
substituens	substituto	pessoa ou algo que substitui ou representa algo ou alguém	verbo de estado
substitutus	substituído	pessoa ou coisa substituída ou representada por algo ou alguém	verbo de estado
thematiens	tema	elemento cujo conteúdo ou assunto é descrito	verbo de estado
thema	descrição	conteúdo ou assunto de uma observação	verbo de estado

omnium	todo	pessoa ou coisa que é composta por partes	verbo de estado
pars	parte	parte de um todo	verbo de estado
creator	criador	pessoa ou coisa que cria (gera ou serve de estímulo) ou destrói algo ou alguém	verbo de estado
opus	resultado	elemento criado ou destruído por alguém ou algo	verbo de estado

Papéis que descrevem melhor uma ação ou situação			
Papel	**Tradução**	**Descrição**	**Características**
causa	causa	relação que representa a causa de uma ação ou que não consegue impedir uma ação	E; substituir por "por isso", "porque"
sumptio	requisito	requisito de uma ação ou pressuposto sob o qual uma ação ocorre	E; substituir por "é necessário que"
intentio	intenção	motivo de uma ação	E + M; substituir por "ele/ela quis"
intrumentum	instrumento	Instrumento, parâmetro, medida com/sem o(a) qual uma ação é executada	E; substituir por "por meio de X", "através de X", "usando X"
modus	modo	maneira como uma ação é executada	E; substituir por "assim", "dessa forma"

Anexo B

Tabela de Papéis Semânticos Utilizados no Estudo-Piloto II com Descrições e Comentários

Papel	Categoria/Papel Superordenado*	Descrição	Comentário ou Teste
Initial_Time	Time	Tempo em que uma ação se inicia.	Quando começa?
Moment	Time	Tempo em que ocorre uma ação. Pode ser também usado para o caso de uma determinada situação ou condição (p.ex.: em qualquer idade).	Quando? Em que situação/momento?
Final_Time	Time	Tempo em que uma ação termina.	Quando acaba?
Frequency	Time	Intervalo regular em que uma ação ocorre.	De quanto em quanto tempo?
Duration	Time	Período de duração de uma ação.	Durante que período?
Source	Place	Lugar (físico ou metafórico) de onde algo é retirado ou do qual algo se desloca.	
Initial_Location	Source	Lugar físico (pode ser fictício) de onde parte um deslocamento.	De onde?
Material	Source	Lugar metafórico que serve de ponto de partida para a geração de um produto ou resultado.	De quê (é feito)?
Goal	Place	Lugar (físico ou metafórico) para onde algo se desloca ou ponto final de um processo - pode ser entendido como um objetivo, uma finalidade.	Para quê?
Destination	Goal	Lugar físico (pode ser fictício) para onde algo se desloca.	Aonde?
Result	Goal	Lugar metafórico que é o ponto final de um processo.	
Product	Result	Resultado concreto.	
Location	Place	Lugar (físico ou metafórico, real ou fictício) onde uma ação ocorre.	Onde?
Trajectory	Place	Intervalo espacial entre um ponto e outro ao longo do qual algo se desloca.	
Agent	Actor	Aquilo/aquele que realiza a ação.	

Co-Agent	Actor	O mesmo que Agent. É usado apenas quando há dois agentes participando de uma ação, sendo que ambos podem trocar de lugar na frase sem alteração de significado.	Não se aplica em caso de sujeito composto.
Stimulus	Actor	Aquilo que provoca uma reação em alguém. Fonte de uma experiência.	
Instrument	Undergoer	Aquilo que é utilizado para realizar uma ação.	Pode-se testar com o verbo "usar" (Ex: Fez isso usando uma faca). O Instrument pode ser reconhecido testando se é possível utilizar a preposição "em" e a voz passiva (Ex: Isso foi visto na análise multivariada), além da preposição "com".
Atribute	Undergoer	Adjetivo ou sintagma que serve para qualificar um paciente, tema ou agente presente na oração.	
Target	Undergoer	Elemento para o qual uma ação é realizada ou que é tido como receptor de algo. Papel criado para servir como uma interface entre Recipient e Beneficiary. Uma Target deve ser animada ou poder ser interpretada como tal na oração (p.ex.: **A casa** ganhou um novo visual. - **A casa** não é um argumento animado *per se*, mas, pelo uso do verbo **ganhar**, ela ganha esse traço - uso metafórico.).	
Recipient	Target	Target receptora de algo concreto que parte de uma fonte e chega até ela.	
Beneficiary	Target	Target que experiencia uma vantagem ou desvantagem gerada pela ação.	
Theme	Undergoer	Elemento presente na ação que não é modificado por ela, podendo sofrer deslocamento ou não.	Theme é o papel mais frequente, podendo estar no lugar de sujeito,

			objeto ou mesmo de complementos preposicionados.
Co-Theme	Undergoer	O mesmo que Theme. É usado apenas quando há dois temas participando de um evento, sendo que ambos podem trocar de lugar na frase sem alteração de significado.	
Topic	Theme	Theme de uma conversa.	Sempre que se tratar de um assunto ou de uma mensagem, se usa Topic em vez de Theme. Verbos relacionados ao diálogo.
Patient	Undergoer	Elemento modificado (implícita ou explicitamente) pela ação ou pelo processo. A sua marca é ser afetado.	
Co-Patient	Undergoer	Mesmo que Co-Theme, porém, aplicado ao caso de Patient. É usado apenas quando há dois pacientes participando de um evento, sendo que ambos podem trocar de lugar na frase sem alteração de significado.	
Experiencer	Patient	Patient que sofre uma alteração psicológica.	

Pivot	Undergoer	Elemento que aparece juntamente com Theme, mas que tem maior importância que este, diferenciando-se assim de Co-Theme. O Pivot tem a mesma função de Theme, apenas é mais importante que este devido ao foco do verbo, não podendo mudar de posição sem alterar o significado.	Usado em verbos que apresentam um estado/característica (físico(a) ou mental) ou uma propriedade/posse. Não há ação ou processo envolvidos. O Pivot, assim como o Theme, não sofre alteração e permanece como está durante o evento.
Value	Undergoer	Qualquer número expresso.	
Extent	Value	Value que representa uma variação positiva ou negativa mensurável.	
Asset	Value	Value que representa dinheiro.	
Cause	Actor	Aquilo que representa a causa de um evento.	Por quê?
Reflexive	Nenhum	Marca quando um pronome reflexivo refere-se estritamente ao sujeito, sem incorrer em um novo papel semântico. Se o pronome reflexivo implicar em um novo papel semântico, este papel não será utilizado.	
Verb	Nenhum	Utilizado com verbos suporte, em que o objeto serve como indicador do evento.	
Manner	Nenhum	Representa o modo como algo foi realizado.	Como?
Comparative	Nenhum	Representa uma comparação entre dois objetos ou especifica um exemplo.	

*A categoria é um indicador da posição do papel semântico dentro da hierarquia de papéis.

Anexo C

Neste anexo, listamos as dezoito alternâncias[83] que foram verificadas para a realização do estudo de agrupamento de verbos apresentado no Capítulo 9.

Alternância ativa-passiva
NP^1 V NP^2 - João comeu a maçã.
NP^2 V (PP^1) – A maçã foi comida (por João).

Alternância ativa-passiva adjetiva
NP^1 V NP^2 - Joana preocupava a mãe.
NP^2 V (PP^1) - A mãe ficava preocupada (com Joana).

NP^1 V NP^2 - Joana preocupava a mãe.
NP^2 V (PP^1) - A mãe estava preocupada (com Joana).

Inversão locativa ou alternância pós-verbal (permite que o sujeito seja colocado após o verbo e que a posição do sujeito seja ocupada por um sintagma preposicionado)
NP V PP - João vive na cidade grande.
PP V NP – Na cidade grande vive João.

Alternância locativa
NP^1 V NP^2 PP^3 - Eles besuntaram manteiga no pão.
NP^1 V NP^3 PP^2 - Eles besuntaram o pão com manteiga.

Alternância não causativa
NP^1 V NP^2 - A água encheu o tanque.
NP^2 V PP^1 - O tanque encheu de/com água.

Alternância dativa – não ocorre em português
NP^1 V NP^2 PP^3 - John sold a car to Bill.
NP^1 V NP^3 NP^2 - John sold Bill a car.

Alternância de sujeito oblíquo
NP^1 V NP^2 PP^3 - Eu sequei as roupas no sol.
NP^3 V NP^2 - O sol secou as roupas.

Alternância conativa
NP^1 V NP^2 - Carla tocou João.
NP^1 V PP^2 - Carla tocou em João.

Alternância de alçamento de parte do corpo
NP^1 V NP^2 - João cortou o braço de Ana.
NP^1 V NP^3 PP^2 - João cortou Ana no braço.

[83] Os números sobrescritos nas estruturas de subcategorização servem para facilitar a compreensão quanto ao tipo de deslocamento e/ou modificação que aconteceu com cada um dos sintagmas.

Alternância reflexiva
NP^1 V NP^2 - Eu apresentei uma solução.
NP^2 V REFL – Uma solução se apresentou.

Alternância ergativa
NP^1 V NP^2 - Joana preocupou a mãe.
NP^2 V REFL PP^1 - A mãe se preocupou com Joana.

Alternância medial ou causativa-ergativa
NP^1 V NP^2 - Eu quebrei o vaso.
NP^2 V – O vaso (se) quebrou.

Alternância causativa I (transitiva-intransitiva, transitiva-reflexiva – as causativas podem ter um agente ou uma causa no lugar do sujeito)
NP^1 V NP^2 - João tocou a campainha. (agente)
NP^2 V – A campainha tocou.

NP^1 V NP^2 - João quebrou o prato. (agente)
NP^2 V (REFL) – O prato (se) quebrou.

NP^1 V NP^2 PP^3 - O vento aproximou o barco da praia. (causa)
NP^2 V PP^3 - O barco se aproximou da praia. (o reflexivo é obrigatório)

NP^1 V NP^2 - O frio aumentou a incidência de doenças. (causa)
NP^2 V - A incidência de doenças aumentou. (o reflexivo é impossível)

Alternância causativa II (causativa por que o sujeito é uma CAUSE, segundo Cançado [1996])
NP^1 V NP^2 PP^3 - Joana preocupa a mãe com sua arrogância.
NP^3 V NP^2 - A arrogância de Joana preocupa a mãe.

Alternância adversativa
NP^1 V NP^2 - João arrebentou o carro.
NP^1 V PP^2 - João arrebentou com o carro.

Alternância com oração causativa encabeçada
NP^1 V NP^2 - João teme o cachorro.
NP V NP^1 V NP^2 - O amigo faz João temer o cachorro.

Agente reflexivo não-intencional
NP V REFL – Paula se cortou. (O agente não teve a intenção de se cortar...)

Verbos que ocorrem somente de uma forma
Choveu.
Há tapetes.

Anexo D

Hierarquia dos papéis semânticos utilizados no VerbLexPor