Desenvolvimento de um recurso léxico com papéis semânticos para o português

Developing a lexical resource annotated with semantic roles for Portuguese

Leonardo Zilio
Universidade Federal do Rio Grande do Sul
ziliotradutor@gmail.com

Carlos Ramisch
Laboratoire d'Informatique Fondamentale de Marseille
carlos.ramisch@lif.univ-mrs.fr

Maria José Bocorny Finatto
Universidade Federal do Rio Grande do Sul
mariafinatto@gmail.com

Resumo

Os objetivos deste estudo são os seguintes: apresentar uma metodologia para desenvolver um recurso léxico com informações semânticas; comparar papéis semânticos de verbos em linguagem especializada e não especializada; e observar a anotação de papéis semânticos por vários anotadores.

Foram desenvolvidos dois experimentos relacionados à anotação de papéis semânticos em português: comparação de um *corpus* de linguagem especializada com um *corpus* de linguagem não especializada; e teste da concordância entre diversos anotadores na atribuição de papéis semânticos.

Quanto aos resultados, observaram-se diferenças qualitativas entre os *corpora* estudados, sendo o apagamento de agentes um traço marcante do *corpus* especializado. A não concordância averiguada entre vários anotadores indica que a tarefa é complexa, requerendo mais treinamento ou uma maior simplificação da tarefa, o que não parece ser possível no atual estágio de desenvolvimento.

Palavras chave

Linguística Computacional, Processamento de Linguagem Natural, anotação de papéis semânticos, recursos léxicos

Abstract

The objectives of this study are as follows: to present a methodology for the development of a lexical resource with semantic information; to compare semantic roles in specialized and non-specialized language; and to observe the semantic role labeling (SRL) made by a group of annotators.

Two experiments revolving around SRL in Portuguese were developed: a comparison between data in specialized and non-specialized language corpora; and an annotation evaluation for verifying the agreement among multiple annotators for the task of SRL.

As for results, a qualitative difference between the corpora was observed, and the most prominent feature was the omission of agents in specialized texts. There was little agreement among annotators, which points toward the necessity of more training, or a simplification of the task, which does not seem to be possible at this stage of development.

Keywords

Computational Linguistics, Natural Language Processing, semantic role labeling, lexical resources

1 Introdução

A área de Processamento de Linguagem Natural (PLN) tem por objetivo facilitar a interação entre o computador e as pessoas, de modo que essa interação seja o mais natural possível, por meio do uso de línguas naturais. Nesse âmbito, a tecnologia da linguagem se concretiza como um grande desenvolvimento na história do ser humano, sendo comparada por Branco et al. (2012) "com a invenção da imprensa por Gutenberg". Tendo isso em vista, é importante desenvolver um esforço colaborativo entre várias áreas do conhecimento, incluindo a Ciência da Computação e a Linguística.

Nesse contexto, a Linguística pode se constituir numa fonte de conhecimento e recursos que, somados ao trabalho do PLN, contribuem para a interação homem-máquina, seja para a redação de um texto, seja para a interpretação de um comando de voz etc. Ao lado dos estudos de léxico, morfologia, sintaxe e texto, a semântica também tem uma função a desempenhar, pois em seu âmbito se encontra o estudo dos significados. Existem vários tipos de abordagens semânticas, desde as que observam o léxico e o seu valor na língua (por vezes, sem observar os contextos de uso de uma palavra), até as que tentam reconhecer os significados nos textos ou na interação com o mundo. A abordagem neste artigo parte principalmente da sintaxe e do léxico, e enfoca o significado de verbos em termos de seus papéis semânticos. Para isso, observamos o léxico em contexto e levamos em consideração a sintaxe em

torno dos verbos. Discutiremos os papéis semânticos de forma mais detalhada ao longo do artigo, porém, cabe aqui apresentar um breve exemplo. Observe-se a seguinte sentença:

1. [O homem] bateu [no cachorro].

Na sentença 1, 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 *no cachorro* tem o papel de PACIENTE (ou ARG1), isto é, ele é o participante no evento afetado pela ação. Assim, a semântica dos papéis se configura como uma abstração do significado da oração.

Apesar de o português ser atualmente a quinta língua mais utilizada na Internet[1], a quantidade de recursos semânticos disponíveis para o seu processamento automático ainda é pequena. Estamos distantes de outras línguas que recebem mais investimento no desenvolvimento de recursos e ferramentas para o processamento da linguagem, como é o caso, por exemplo, do inglês (Branco et al., 2012).

Neste artigo, trata-se da criação de um recurso semântico para o português brasileiro que possa ser utilizado para o processamento semântico de verbos do português. Assim, os objetivos deste artigo são:

- Apresentar os métodos para o desenvolvimento de um recurso léxico com informações semânticas[2].
- Comparar papéis semânticos de verbos em linguagem especializada com aqueles dos mesmos verbos em linguagem não especializada.
- Observar a concordância entre anotadores em uma tarefa de anotação de papéis semânticos.

Este artigo está dividido da seguinte maneira: na Seção 2, apresentamos brevemente alguns conceitos e trabalhos relacionados a este estudo; na Seção 3, detalhamos e discutimos o método utilizado para o desenvolvimento de um recurso léxico com informações sobre papéis semânticos; na Seção 4, apresentamos os resultados da anotação de papéis semânticos e realizamos a comparação entre linguagem comum e especializada à luz dos papéis semânticos; na Seção 5, relatamos e discutimos um experimento realizado com múltiplos anotadores; por fim, na Seção 6, expomos nossas considerações finais.

2 Conceitos e trabalhos relacionados

Nesta seção, procuramos apresentar brevemente a base teórica deste artigo. Na Seção 2.1, discorremos sobre papéis semânticos e estruturas de subcategorização; na Seção 2.2, discutimos trabalhos como o de Levin (1993) e recursos como a VerbNet, o PropBank e a FrameNet.

2.1 Papéis semânticos e estruturas de subcategorização

2.1.1 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 se comentou rapidamente na Seção 1, os papéis semânticos representam uma forma abstrata de semântica: "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, porém, elas não são nem tão específicas quanto uma semântica lexical (por exemplo, acepções em dicionários), nem tão abstratas quanto uma semântica puramente sintática (por exemplo, a utilização de categorias sintáticas como sujeito e objeto direto como indícios de diferenciação semântica). 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.

Para exemplificar o que são os papéis semânticos, tomemos como exemplo as sentenças a seguir[3]:

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

[1] Dados de 2010, retirados a 10 de setembro de 2013 do site http://www.internetworldstats.com/stats7.htm.

[2] Por *recurso léxico com informações semânticas*, estamos nos referindo a um banco de dados que contenha sentenças do português anotadas com papéis semânticos.

[3] Os exemplos são inventados. Não provêm dos corpora envolvidos no estudo. Opta-se aqui por usar 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.

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 do verbo permite. Os elementos a que chamamos atenção aqui, porém, não são os sintáticos, mas sim os semânticos. Em 2a, *João* está executando uma ação, 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 2b, 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*; assim, ela permanece como PACIENTE (ou ARG1), porque a ação está sendo executada por um elemento não divulgado na sentença. Na sentença 2c, 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.

A partir desse exemplo, pode-se perceber que os elementos sintáticos (sujeitos, objetos etc.) nem sempre têm uma semântica óbvia. Desse modo, discriminar os papéis desempenhados pelos elementos sintáticos em diversos contextos pode ajudar no processamento automático de textos. Por exemplo, 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 verbos de compra das quais Google seja o sujeito, pois seriam ignoradas frases como esta: *[Android Inc.] foi comprada [pela Google] [em 2005]*.

Nos exemplos fornecidos até aqui, apresentamos duas possibilidades de anotar os papéis semânticos: a forma descritiva (AGENTE, PACIENTE, INSTRUMENTO etc.) ou 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, Kingsbury e Gildea (2005) ao desenvolverem o PropBank. Esses trabalhos serão discutidos na Seção 2.2.

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); Cançado, Godoy e Amaral (2012).

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 com bastante propriedade por Levin e Rappaport-Hovav (2005). 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 frases, retiradas de Kasper (2008), deveriam ser interpretadas:

3a. The cardinal loaded bottles on the wagon.
(O cardeal colocou garrafas na carroça.)
3b. The cardinal loaded the wagon with bottles.
(O cardeal carregou a carroça com garrafas.)[4]

A interpretação, conforme indicada por Jackendoff (1990), é de que em 3a as garrafas não preenchem a carroça, enquanto em 3b a carroça está completamente cheia. Porém, Fillmore (1968, *apud* 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 no primeiro exemplo, 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*:

4a. O estudo encontrou a doença em 15 pacientes.
4b. O estudo encontrou 15 pacientes com a doença.

[4] 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.

Assim como nos exemplos 3a e 3b do inglês, a estrutura sintática das sentenças 4a e 4b apresentam diferenças claras devido ao emprego de diferentes preposições; porém, as duas sentenças podem ser consideradas paráfrases. Por um lado, as duas sentenças podem indicar que os pesquisadores encontraram a doença nos pacientes. Desse modo, o objeto encontrado, nas duas sentenças, é a doença, pois ela está sendo procurada, e não os pacientes (os pesquisadores sabem onde os pacientes estão). Por outro lado, a sentença 4b 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. Partindo apenas dessas frases escritas, uma pessoa pode interpretar o sentido das duas formas. Desse modo, há uma ambiguidade que só pode ser desfeita pela observação direta da realidade. Como a atribuição de papéis semânticos não entra no domínio da Pragmática, torna-se inviável esse tipo de atribuição.

2.1.2 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), 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), "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 morfologia básica. Por exemplo, na seguinte sentença:

5. João viu Maria.

A classificação sintática da sentença-exemplo 5 seria: *João* = sujeito; *viu* = verbo/predicado; *Maria* = objeto direto. Porém, na classificação de estrutura de subcategorização, essa mesma sentença teria a seguinte análise: *João* = NP (do inglês, *nominal phrase*) [5] 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 do 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 (pode se 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 na Seção 3. 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 (2012).

2.2 Trabalhos relacionados

Começamos esta seção com o trabalho de Levin (1993), para depois prosseguirmos com a VerbNet, o PropBank, a FrameNet e a WordNet.

O trabalho de Levin (1993) é importante não só para o inglês, a língua utilizada, mas para a Linguística como um todo, pois mostrou que é possível agrupar verbos semanticamente próximos a partir de suas estruturas sintáticas. Apesar de

[5] Em nosso estudo, privilegiamos o uso das siglas em inglês, por ser a forma utilizada durante o trabalho de anotação.

haver várias críticas ao trabalho desenvolvido[6], 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 chegar à Tabela 1.

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

Tabela 1: Comportamento dos verbos *break*, *cut*, *hit* e *touch*.

Assim, percebe-se 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 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 como o de Levin (1993)[7], porém, Cançado, Godoy e Amaral (2012) já apresentaram um projeto que intenta levar a cabo essa empreitada.

Partindo das classes de Levin (1993), Kipper-Schuler (2005) desenvolveu a VerbNet. O recurso apresentado na VerbNet contém as classes de Levin associadas aos papéis semânticos que podem ser apresentados pelos verbos de cada uma das classes. No estágio atual da VerbNet (versão 3.2), foram utilizados efetivamente 30 papéis semânticos, partindo de uma lista com 36 papéis.

Por partir das classes de Levin, a anotação de apenas 191 classes (na versão 1.0) já dava cobertura para 4.173 verbos. Atualmente, com o acréscimo de outras classes de verbos, já existe anotação para cerca de 5.800 verbos, divididos em 272 classes.

Para o português, além do nosso trabalho, sabemos da existência 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, a WordNet e a WordNet.Br. Desse modo, para as classes sinônimas entre a WordNet (Fellbaum, 1998) e a WordNet.Br (Dias-da-Silva, 2005; Dias-da-Silva, Di Felippo e Nunes, 2008), os papéis foram importados diretamente do inglês para os verbos em português. Desse modo, já existe uma VerbNet.Br, porém, ela foi construída de modo semiautomático, podendo conter erros, e apresenta apenas aquelas classes que são 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 entre as WordNets do inglês e do português. O trabalho aqui apresentado parte do português e se baseia em uma anotação manual dos dados por um linguista. Assim, apesar de nosso estudo ser menos abrangente, ele apresenta uma menor propensão a erros.

Como mencionamos, o trabalho de Scarton (2013) usou como base os alinhamentos entre a WordNet de Princeton (Fellbaum, 1998) e a WordNet.Br (Dias-da-Silva, 2005; Dias-da-Silva, Di Felippo e Nunes, 2008). As WordNets são recursos que apresentam *synsets* (conjuntos de sinônimos) e as relações entre eles. As relações podem ser de hiperonímia, antonímia, holonímia etc. Além disso, existem definições formais para os possíveis significados de cada um dos *synsets*. Por tomarem *synsets* como base, e não palavras soltas, as relações construídas entre apenas dois *synsets*

[6] 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.

[7] 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).

cobrem várias palavras, o que amplia muito a abrangência desse recurso.

Voltando para 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, Kingsbury e Gildea, 2005), que apresenta sentenças de um *corpus* anotadas com papéis numerados.

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 muito 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) foi responsável por realizar a vinculação dos papéis semânticos da VerbNet às sentenças do PropBank, de modo que as sentenças estão atualmente anotadas também com papéis 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[8] e contém mais de 6 mil instâncias anotadas.

Por fim, existe ainda outro tipo de anotação de papéis semânticos, bastante difundida, que toma como base os cenários comunicativos, chamados de *frames*[9]. É assim que se estrutura a FrameNet (Baker, Fillmore e Lowe, 1998), um projeto que tem por objetivo anotar os papéis semânticos de cada elemento 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[10]. A FrameNet Brasil (Salomão, 2009) utiliza essa mesma abordagem.

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 *frame*. Os papéis da VerbNet são menos específicos, tentando apresentar uma descrição abstrata da semântica que pode ser aplicada para qualquer contexto. Já o PropBank apresenta a solução mais abstrata, pois apenas cinco papéis (ARG0 a ARG4) se aplicam a qualquer contexto, configurando-se como protopapéis.

3 Metodologia de construção do recurso

O cerne deste estudo é o desenvolvimento de um recurso léxico para a língua portuguesa que contenha informações de papéis semânticos. Para efeitos de comparação, esse recurso pode ser entendido como uma mistura entre a VerbNet e o PropBank. Ele contém sentenças extraídas de *corpora* como base para a anotação, assim como o PropBank, porém usa papéis semânticos descritivos, como a VerbNet.

Além disso, queremos comparar os verbos presentes em textos especializados com aqueles presentes em textos não especializados, de modo que, como será visto, utilizamos dois *corpora* de maneira contrastiva.

Para desenvolver esse recurso e a comparação entre os dois tipos de texto, realizamos um estudo-piloto com uma amostra de cinquenta verbos. Nesta seção, serão apresentados os materiais e os métodos empregados nesse estudo-piloto.

3.1 Materiais

3.1.1 *Corpora*

Neste trabalho, foram utilizados dois *corpora*: um composto por textos especializados e outro composto por textos não especializados. Para representar os textos especializados, selecionamos o *corpus* composto por artigos científicos da área da Cardiologia compilado por Zilio (2009). Para representar os textos não especializados, selecionamos o *corpus* de textos do jornal popular Diário Gaúcho, compilado no âmbito do projeto

[8] Disponível no site (acessado em 15/10/2013): http://www.nilc.icmc.usp.br/portlex/index.php/en/projects/propbankbringl.

[9] É importante deixar claro que a palavra *frame* é bastante polissêmica. Neste artigo, trataremos 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). Procuraremos deixar claro pelo contexto qual é o tipo de *frame* a que nos referimos.

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

PorPopular [11]. Na Tabela 2, podemos ver a constituição dos *corpora* em relação ao número de palavras. Ambos os *corpora* foram analisados automaticamente pelo *parser* PALAVRAS (Bick, 2000) com árvores de dependências sintáticas.

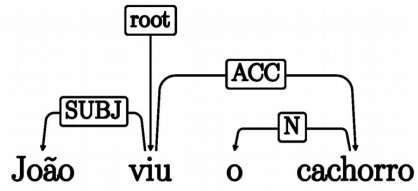
Corpus	Nº de palavras[12]
Cardiologia	1.605.250
Diário Gaúcho	1.049.487

Tabela 2: Tamanho dos *corpora*

Nessa anotação de dependências, o *corpus* anotado apresenta uma hierarquia de ligações entre os elementos sintáticos das sentenças. Um exemplo disso pode ser visto na seguinte sentença analisada com o *parser* PALAVRAS:

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 dessa sentença, se observarmos os valores após a cerquilha (#), é possível ver quais elementos estão ligados diretamente aos verbos e, com isso, extrair os argumentos. Isso porque 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 então utilizada por um extrator de estruturas de subcategorização, que reconhece automaticamente os argumentos dos verbos e os organiza em um banco de dados.

[11] 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 neste artigo porque nosso *corpus* não compreende a totalidade de textos.

[12] Os números de palavras foram observados com a ferramenta WordSmith Tools, versão 4.0 (Scott, 2004).

3.1.2 Extrator de estruturas de subcategorização

O extrator de estruturas de subcategorização (Zanette, Scarton e Zilio, 2012; Zilio, Zanette e Scarton, 2012) é um *software* que realiza a preparação dos dados para a anotação. O extrator contém um conjunto de regras de extração, que 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. Por exemplo, na sentença:

O cachorro foi visto por João.

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 o associa ao verbo *ver*. Desse modo, são mantidas apenas informações referentes a verbos principais.

Todas as informações extraídas são identificadas por meio de regras. Assim, o extrator busca informações como *@<ACC*, fornecidas pelo *parser*, as extrai e também as traduz em etiquetas mais explícitas para o anotador humano, como *OBJETO DIRETO*.

Após extrair as informações, os dados são armazenados em um banco de dados em formato MySQL. Para facilitar a anotação, existe uma interface de usuário que permite a visualização dos dados extraídos, com a classificação dos argumentos em uma ordem predefinida, assim como a anotação de papéis semânticos. Tal interface pode ser vista na Figura 1.

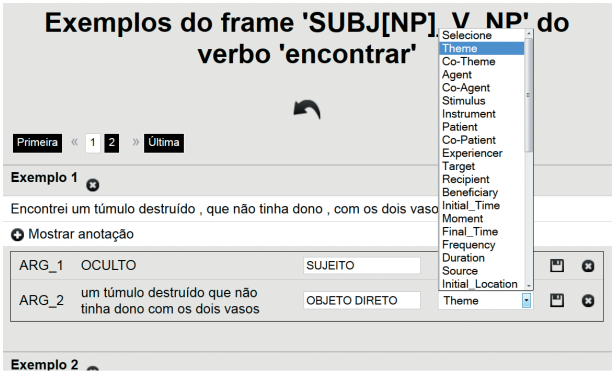


Figura 1: Interface de usuário para anotação

Essa interface mostra a estrutura de subcategorização (chamada de *frame*), o verbo em questão, os exemplos e os argumentos extraídos com a sua respectiva categoria sintática. Essas

informações são extraídas por meio de regras diretamente da anotação já presente nos *corpora*.

Ao anotador humano de papéis semânticos cabe o trabalho de criar uma lista de papéis semânticos, digitá-los em um arquivo de texto separados por vírgulas e selecioná-los a partir da lista de rolagem (que pode ser vista na Figura 1) no momento da anotação. Com essa interface de anotação, o anotador pode se concentrar no que lhe interessa: definir a semântica dos argumentos, sem precisar delimitá-los.

É importante ressaltar que, na estrutura atual, o banco de dados permite 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.

Outra característica importante da extração automática é que ela não distingue argumentos de adjuntos. 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". Assim, em nossa anotação, não faremos uma distinção *strito sensu* entre adjuntos e argumentos; contudo, existem papéis que são potencialmente atribuídos apenas a adjuntos. Para contornar o problema de distinção entre argumentos e adjuntos, utilizamos a frequência como delimitador. Desse modo, se um elemento ocorre dez vezes ou mais junto ao verbo, ele é anotado, recebendo um *status* de argumento. Não sustentamos que todos os elementos anotados de acordo com esse princípio sejam realmente argumentos, porém, a sua existência tem uma influência sobre o significado de uma sentença, e esse significado deve ser reconhecido para se obter um bom desempenho na interpretação de um texto.

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 erros de análise que vão desde a simples segmentação de sentenças até a delimitação dos argumentos. Além dos possíveis erros decorrentes da análise automática, o extrator de estruturas de subcategorização também organiza os dados de acordo com regras, e estas nem sempre são corretas. Desse modo, existem dados que podem conter ruído no banco de dados. Como veremos mais adiante, na Seção 3.2, esses dados errados são ignorados e não são anotados.

3.1.2 Lista de papéis semânticos

Outro elemento importante para a anotação de papéis semânticos em *corpora* é a definição de uma lista de papéis. Como vimos anteriormente, as listas podem se estender desde alguns poucos papéis (como no caso do PropBank) até dezenas de papéis (como no caso da VerbNet e da FrameNet), dependendo da abordagem escolhida.

Neste estudo, optamos por uma lista de papéis descritivos e genéricos, nos moldes da VerbNet. Optamos por esse tipo de lista tendo em vista que já existe um PropBank.Br que utiliza uma lista de protopapéis e também uma FrameNet.Br que usa papéis específicos para o contexto. A VerbNet.Br, apesar de existir, foi feita a partir de uma importação de dados do inglês, tomando por base o potencial interlinguístico das classes de Levin. No entanto, não houve um estudo linguístico mais profundo que mostrasse o quanto essa importação realmente seria aplicável ao português. Desse modo, decidimos focar nossos esforços no mesmo âmbito da VerbNet.Br, porém partindo de uma anotação manual, que posteriormente poderá ser confrontada com os dados importados do inglês presentes na VerbNet.Br.

Em trabalho anterior (Zilio, Zanette e Scarton, 2012), apresentamos um breve estudo com uma lista de 46 papéis semânticos proposta por Brumm (2008). Ao final do estudo, percebemos que os papéis não eram adequados ao nosso tipo de abordagem, por serem muito específicos em alguns casos, e muito genéricos em outros. Não havia um equilíbrio nos papéis, provavelmente pelo fato de que eles ainda não haviam sido testados em dados reais. A falta de dados anotados com os papéis propostos por Brumm também faz com que não haja exemplos concretos que possam ser observados para melhor compreender a funcionalidade de cada um dos papéis. Por isso, neste estudo, optamos por utilizar uma lista já testada, revisada e com exemplos que podem ser livremente acessados na internet: a lista da VerbNet (Kipper, 2005) em sua versão 3.2. Essa lista, além de permitir posteriormente uma comparação direta com os dados da VerbNet.Br, também nos pareceu ser a melhor para suprir as falhas que observamos na lista de Brumm (2008).

Após um estudo dos papéis semânticos e dos exemplos disponíveis na VerbNet, realizamos algumas pequenas modificações na lista. A principal modificação foi a criação do hiperônimo TARGET[13], que passou a abrigar BENEFICIARY e

[13] 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,

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.

No total, 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, acreditamos que essa escolha simplificará uma comparação futura do português com o inglês.

Alguns desses papéis 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.

Para manter o artigo sucinto, não explicaremos aqui as funcionalidades específicas de cada um dos papéis. Essas informações podem ser obtidas na documentação da VerbNet[14].

3.2 Método de anotação

Para realizar a anotação de papéis semânticos nas sentenças dos *corpora*, fizemos inicialmente algumas escolhas em relação às quantidades a serem anotadas. Neste estudo, optamos por uma anotação amostral, almejando um teste dos papéis semânticos apresentados pela VerbNet. Optamos por 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, excluindo os que já haviam sido anotados na primeira etapa. Foram anotados, dessa forma, 50 verbos ao todo em cada um dos *corpora*[15] — 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 em Zilio, Zanette e Scarton (2012). Com a modificação apresentada aqui, garantimos que todas as estruturas de subcategorização tenham dez exemplos anotados. Se uma estrutura tem 16 exemplos, mas apenas nove estão corretos, ela é 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. Não nos adiantava anotar verbos em apenas um dos *corpora*, pois não seria possível comparar os resultados.

Neste estudo-piloto, a anotação foi desenvolvida por apenas um anotador linguista treinado[16], o qual teve acesso a um manual de anotação com a descrição dos papéis semânticos e de alguns exemplos retirados ou da VerbNet ou da anotação realizada em um estudo anterior. Foi realizado também um experimento com múltiplos anotadores, o qual será relatado mais adiante, na Seção 5, porém, para este estudo-piloto, usamos apenas um anotador.

4 Resultados e considerações sobre a anotação de papéis semânticos

Nesta seção, 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.

porém, quando nos referirmos à nomenclatura empregada no estudo, usaremos o inglês.

[14] Disponível no site: http://verbs.colorado.edu/~mpalmer/projects/verbnet/downloads.html.

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

[16] O anotador mencionado é o primeiro autor deste artigo.

4.1 Considerações sobre 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 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 entre argumento e adjunto, alguns elementos anotados, por serem de natureza adverbial, não tinham um papel semântico condizente, precisando ser anotados com papéis que se adequavam apenas parcialmente. Esse tipo de problema pode ser solucionado se adicionarmos os papéis semânticos específicos para adjuntos que são utilizados no PropBank.

Em geral, a anotação de adjuntos adverbiais foi uma tarefa complexa. Observando as sentenças 6a a 6d a seguir, extraídas dos *corpora*, temos adjuntos adverbiais que contêm *jogo* e *estudo* (destacados em negrito). Esses adjuntos adverbiais apresentam uma dificuldade para a atribuição de um papel semântico. 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. Isso ocorreu porque os papéis semânticos da VerbNet foram pensados apenas para argumentos e não para adjuntos. Assim, cremos que seja necessária fazer uma separação entre papéis para argumentos e papéis para adjuntos, como fez o projeto PropBank. Apesar de alguns problemas no que diz respeito aos adjuntos adverbiais, os demais argumentos sempre tinham algum papel semântico na lista que se adequava.

6a. Eles fizeram um jogo largado e nós demos oportunidade ***em um jogo*** *que estava em nossas mãos.*
6b. Teremos de melhorar muito em relação ao que mostramos ***no primeiro jogo****, mas temos todas as condições de reverter.*
6c. ***No presente estudo****, animais adultos restritos apresentaram aumento de todos os parâmetros estereológicos analisados na aorta, sugerindo hiperplasia da túnica média.*
6d. O prognóstico utilizado para o TC6M foi demonstrado ***no estudo SOLVD10****.*

No que diz respeito ao método amostral escolhido, ele se mostrou adequado para a maioria dos verbos, pois representa bom equilíbrio entre tempo utilizado para anotar e representatividade dos dados anotados. Porém, ficou claro que, para verbos muito polissêmicos (por exemplo, *dar*), 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, talvez tornando impossível uma anotação de muitos verbos em um tempo aceitável, tendo em vista que temos apenas um anotador. Por mais que sempre exista um problema com o método amostral (afinal, alguns dados são ignorados), depois que a anotação é feita, é possível perceber quais verbos não estão representados adequadamente e, se necessário, é possível dar um tratamento especial a eles.

A ferramenta utilizada para a extração e anotação dos dados desenvolvida por Zanette (2010) é bastante versátil e pode ser adaptada às necessidades do anotador. Por exemplo, para alterar a lista de papéis semânticos, basta modificar um arquivo de texto. Entretanto, com a anotação de mais verbos em relação ao estudo anterior (Zilio, Zanette e Scarton, 2012), 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 garantir que não haverá mais esse tipo de problema, fizemos uma análise do conjunto completo de etiquetas empregadas pelo PALAVRAS[17] e acrescentamos ao sistema, com a respectiva descrição, as modificações necessárias.

Apesar de termos utilizado a ferramenta desenvolvida por Zanette (2010), existem outras ferramentas que poderiam ser empregadas para a anotação, como a ferramenta SALTO (Burchardt et al., 2006). Entretanto, o sistema de anotação dessa ferramenta é muito mais complexo, deixando ao encargo do anotador a tarefa de delimitar os argumentos. Por um lado, isso garante uma maior precisão na delimitação dos argumentos; por outro lado, aumenta a chance de erros e aumenta o trabalho dispendido na anotação.

[17] As etiquetas com as respectivas explicações de suas funções em http://beta.visl.sdu.dk/visl/pt/info/portsymbol.html.

4.2 Resultados da anotação e comparação entre os *corpora*

Seguindo os métodos apontados na Seção 3.2, 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[18]. Com os dados em formato XML, o compartilhamento dos dados se torna mais simples, pois o formato XML é mais acessível do que o formato MySQL. No atual estágio, a distribuição do recurso anotado requer apenas o compartilhamento de um arquivo específico para cada *corpus*[19].

No que diz respeito à diferença de frequências entre os *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 na organização do *corpus* através da seleção de sentenças específicas em vez de textos completos. Porém, 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, o que não seria bom para este estudo; afinal, almejamos 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[20] 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 serem informações atribuídas manualmente, dificilmente a baixa frequência pode ser desconsiderada sem um bom motivo. 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>[21] para o verbo *chegar* não quer dizer que haja apenas uma ocorrência de cada uma das associações sintático-semânticas SUJ<Theme>, ADJ.ADV[a]<Loca-tion> 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.

Nas Tabelas 3 e 4, 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 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[22], e nem tanto de intransitividade. Já no Diário Gaúcho ocorre o oposto, com uma maioria de exemplos intransitivos[23]. Isso contraria nossas observações em estudo anterior[24], quando havíamos observado

[18] A ferramenta de exportação foi desenvolvida por Samy Sassi sob orientação de Leonardo Zilio, Carlos Ramisch e Mathieu Mangeot.

[19] Os arquivos encontram-se disponíveis para *download* gratuito em http://cameleon.imag.fr/xwiki/bin/view/Main/Resources.

[20] 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.]

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

[22] Alguns exemplos:

- *Foram avaliados os seguintes parâmetros:*
- *Foi observada uma distribuição igual de a população estudada em relação ao sexo.*
- *Não foram demonstradas evidências consistentes de o papel de níveis circulantes de MIF, IL-6 e sCD40L como marcadores de SCA.*

[23] Observou-se uma maioria de verbos como "ocorrer", "existir", "ficar", "acontecer".

[24] Para aquele estudo, foram selecionados quatro verbos com frequências próximas nos dois *corpora*: *levar*, *encontrar*, *usar* e *receber* (Zilio, Zanette e Scarton, 2012). Os resultados sobre o apassivamento, porém, não foram divulgados. Na época, os autores consideraram que seria muito precipitado publicar resultados de um fenômeno tão amplo levando em consideração apenas quatro verbos. Mesmo agora, com nossos 50 verbos, por mais que sejam os mais frequentes presentes nos dois corpora, talvez não tenhamos resultados representativos o suficiente para uma conclusão, devido à

uma tendência maior de apassivamento no Diário Gaúcho, o que provavelmente era um fenômeno pertinente apenas aos verbos estudados.

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 3: Cinco estruturas mais frequentes no *corpus* de Cardiologia

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

Tabela 4: Cinco estruturas mais frequentes no *corpus* do Diário Gaúcho

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 um agente, enquanto no Diário Gaúcho, dentre as 272 estruturas, 121 apresentavam agente. 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
- *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*.

Como pode ser visto na Tabela 6 (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) 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), 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), que geralmente representa um objetivo de uma ação, também é muito mais frequente naquele do que neste.

Utilizando o coeficiente de correlação tau-b de Kendall [25], realizamos três experimentos com diferentes informações.

No experimento 1, avaliamos a correlação 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 3 e 4. Nesse experimento, os resultados apontaram que há uma correlação inversa entre as amostras, pois encontramos um valor de $\tau = -0{,}394$ ($p < 0{,}001$). 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

amplitude do fenômeno. Assim, nossos resultados devem ser observados com cautela.

[25] 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 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.

resultado corrobora algumas tendências observadas na análise qualitativa anterior.

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 5. Com esse conjunto de dados, não houve correlação entre as duas amostras ($\tau = 0{,}031$; $p = 0{,}608$).

Por fim, no experimento 3, consideramos apenas o *ranking* dos papéis semânticos, sem observar a anotação sintática. Os dados foram utilizados exatamente da forma como estão apresentados na Tabela 6. 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. 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.

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

Tabela 5: Dados sintático-semânticos dos dois *corpora*[26]

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

Tabela 6: Papéis semânticos e sua frequência nos dois *corpora*

5 Experimento com múltiplos anotadores

Além da anotação, também desenvolvemos um experimento paralelo relacionado ao tema. O experimento foi uma tentativa de passar a anotação a múltiplos anotadores.

Existem estudos que já observaram a tarefa de anotação com mais de um anotador. Hovy et al. (2006) apresentaram uma solução 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 tal, 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, bastante similar ao experimento de Hovy et al. (2006), porém, que tomou como base o trabalho de Hong e Baker (2011); a segunda etapa era 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 (apesar de o único verbo apresentado ter ficado em 81,9%), a segunda etapa não teve resultados tão positivos.

O elemento em comum nos dois trabalhos apresentados é que já existe um recurso anterior

[26] Essa tabela apresenta apenas os dados mais frequentes, a título de exemplo; porém, para o cálculo do *p*, utilizamos a tabela completa, que contém mais de 140 linhas.

que pode ser utilizado como base. Hovy 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) têm a FrameNet e, da mesma forma, apenas buscam expandir a anotação para outros *corpora*.

Em nosso caso, não existe ainda um recurso para o português que contenha a anotação de papéis semânticos descritivos. 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 mais útil utilizar a anotação de múltiplos anotadores com pouco treinamento ou de apenas um com muito treinamento (que é o método que está sendo utilizado).

5.1 Procedimento

Para o experimento, foram selecionados dez anotadores linguistas (alunos de pós-graduação em Linguística da UFRGS) e 25 sentenças extraídas dos *corpora* apresentados na Seção 3.1.1. O treinamento foi básico, consistindo apenas em uma explicação sobre a tarefa e o assunto, e no fornecimento de um manual de anotação.

No manual de anotação, cada um dos papéis semânticos que poderiam ser utilizados foi apresentado ao lado de uma descrição, como pode ser visto neste exemplo:

LOCATION	Lugar (físico ou metafórico, real ou fictício) onde uma ação ocorre.

A estrutura das sentenças a serem anotadas foi similar à que apresentamos na Figura 1, com a ressalva de que, por ser uma anotação em papel, não havia uma lista de rolagem para escolher as sentenças (apenas uma lista para consulta no manual de anotação).

Além da anotação dos papéis semânticos, também fazia parte da tarefa a distinção de cada um dos elementos anotados entre argumentos e adjuntos. Para tal, foi apresentada também uma breve explicação sobre a diferença entre argumentos e adjuntos[27].

Eis aqui 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, seguida pelos argumentos (as abreviaturas estavam explicitadas 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.

5.2 Cálculo de concordância

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. Em geral, utiliza-se o coeficiente κ para essa tarefa, por isso, 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 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 das etiquetas 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.

[27] 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.

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 muito 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, pessoas que não entenderam a tarefa ou que realizaram a anotação sem prestar muita atenção aos dados) e 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, 2010a; 2010b) que calcula vários coeficientes de concordância.

5.3 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, na sentença 7, a seguir, 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. *O PT* ***apresentou*** *no Congresso um projeto de lei que cria contribuição social sobre fortunas.*

Por isso, 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 argumento e adjunto, 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 (multi-κ = 0,320770).

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*, afinal, se eles não haviam compreendido (ou haviam feito às pressas) a tarefa de distinguir entre argumento e adjunto, cremos que não havia por que confiar nos seus resultados em uma tarefa muito 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 do papel AGENT, MOMENT, LOCATION e, em alguns casos, THEME. No entanto, quando outros papéis 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 da sentença 7, acima, os 10 anotadores concordaram que o sujeito *O PT* desempenha a função de AGENT, no entanto, 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. Por exemplo, a sentença 8 não teve concordância em nenhum dos argumentos. O sujeito *A versão religiosa* recebeu 4 anotações como AGENT e 3 como THEME, enquanto o objeto indireto *com as mulheres Jaca ou Melancia* foi anotado como THEME por 4 anotadores e como ATTRIBUTE por 3.

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

Apenas a sentença 9 apresentou uma maior concordância entre os anotadores no que diz respeito aos dois argumentos.

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

O sujeito *ele* foi reconhecido como AGENT pelos 10 anotadores, enquanto o objeto direto *arma* foi anotado por 8 anotadores como INSTRUMENT.

Existem vários motivos que podem ter levado a uma concordância tão baixa. 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. Porém, 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 et al. (2006), a solução encontrada para se obter alto nível de concordância foi simplificar a tarefa o 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.

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, como a VerbNet, o PropBank e o PropBank.Br, e com anotações-teste antes de iniciar o trabalho.

6 Conclusões

Realizamos a anotação de uma quantidade amostral de verbos em dois *corpora* com base na metodologia proposta. Desse modo, temos um recurso lexical com informação sobre papéis semânticos disponível em um formato amostral. Após a anotação, foi possível observar semelhanças e diferenças nos papéis semânticos atribuídos para verbos em textos especializados e não especializados.

Quanto à metodologia de anotação, a lista de papéis escolhida foi suficiente para realizar a anotação dos argumentos dos verbos. Porém, a anotação de elementos que podem ser considerados adjuntos se mostrou complexa. Assim, cremos ser necessária uma modificação para incluir papéis específicos de adjuntos, nos mesmos moldes do PropBank.

A opção por uma anotação de algumas sentenças por verbo, e não da totalidade de sentenças, é válida, pois a maioria dos verbos não apresentou uma grande polissemia. Assim, cremos ser prudente manter a metodologia empregada e, se necessário, dar um tratamento especial aos verbos mais polissêmicos. Por exemplo, para os casos de verbos como *dar*, *fazer* e *ir*, realmente seria necessário um olhar mais cuidadoso, pois a quantidade de significados desses verbos é muito grande, e a anotação amostral não dá conta de suas várias facetas.

A ferramenta utilizada para a extração e anotação dos dados apresentou versatilidade e pretendemos continuar com o seu uso. Além disso,

com a possibilidade de exportar os dados para o formato XML, que é mais amigável, a disponibilização dos dados e seu compartilhamento se tornam mais simples.

No que diz respeito à comparação entre linguagem comum e linguagem especializada, os dados estatísticos mostram que, conforme aumenta a complexidade dos dados (anotação de papéis semânticos -> de argumentos -> de sentenças), aumenta também a distância entre as duas amostras.

Essa diferença entre os dados dos *corpora* pode ser visualizada qualitativamente no que diz respeito a alguns elementos específicos, tais como a utilização de INSTRUMENTS na posição de sujeitos, com um apagamento do agente, que é uma marca dos textos da Cardiologia. Além disso, papéis como PIVOT e GOAL também foram mais frequentes na Cardiologia do que no Diário Gaúcho.

A avaliação da concordância entre vários anotadores permitiu que observássemos o quão complexa é a tarefa de atribuição papéis semânticos e de distinção entre argumentos e adjuntos. Os índices multi-π encontrados entre anotadores linguistas ficaram abaixo dos limites apontados na literatura como mínimos para a existência de concordância. A literatura apresenta estudos com múltiplos anotadores com resultados superiores (Hovy et al., 2006; Fossati, Giuliano e Tonelli, 2013), porém, esses estudos se baseiam em recursos já existentes, o que permite simplificar as decisões do anotador. Em nosso caso, seria possível simplificar a anotação (por exemplo, reduzir a lista de papéis semânticos), mas isso modificaria muito o recurso final a ser gerado. Assim, acreditamos que o trabalho de anotação deve ser realizado inicialmente por apenas um anotador treinado. Posteriormente, de posse de um recurso já desenvolvido, será possível simplificar a anotação tomando por base as informações do recurso existente para realizar um novo teste com múltiplos anotadores.

Agradecimentos

Agradecemos ao CNPq e à CAPES, pelo financiamento e também ao Projeto CAMELEON (CAPES-Cofecub 707/11) pelo apoio e oportunidade de intercâmbio no exterior.

Referências

Afonso, Susana, Eckhard Bick, Renato Haber e Diana Santos. 2001. Floresta sintá(c)tica: um treebank para o português. In: *Actas do XVII Encontro da Associação Portuguesa de Linguística*, APL, Lisboa, Outubro de 2001.

Artstein, Ron e Massimo Poesio. 2008. Inter-Coder Agreement for Computational Linguistics. In: *Computational Linguistics*, 34(4):555-596. ACL

Baker, Collin F., Charles J. Fillmore e John B. Lowe. 1998. The Berkeley FrameNet project. In: *COLING-ACL '98*: Proceedings of the Conference. Montreal, Canada 1998, pp. 86-90.

Bick, Eckhardt. 2000. *The Parsing System* PALAVRAS*: Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework*. Aarhus: Aarhus University Press. http://beta.visl.sdu.dk/~eckhard/pdf/PLP20-amilo.ps.pdf

Branco, António *et al*. 2012. *The Portuguese Language in the Digital Era* / *A língua portuguesa na era digital*. Heidelberg, Nova Iorque: Springer.

Brumm, T. 2008. *Erstellung eines Systems thematischer Rollen mit Hilfe einer multiplen Fallstudie*. Trabalho de conclusão de curso. http://www.ipd.kit.edu/Tichy/uploads/arbeiten/135/StudienarbeitBrumm.pdf

Burchardt, Ajoscha, Katrin Erk, Anette Frank, Andrea Kowalski e Sebastian Pado. 2006. SALTO - A Versatile Multi-Level Annotation Tool. In: *Proceedings of LREC 2006*.

Cançado, Márcia. 2009. Argumentos: Complementos e Adjuntos. In: *Revista Alfa*, São Paulo, 53 (1): 35-59.

Cançado, Márcia. 2010. Verbal Alternations in Brazilian Portuguese: a Lexical Semantic Approach. In: *Studies in Hispanic and Lusophone Linguistics*, 3 (1), p. 77-111.

Cançado, Márcia, Luisa Godoy e Luana Amaral. 2012. The construction of a catalog of Brazilian Portuguese verbs. In: *Proceedings of the Workshop on Recent Developments and Applications of Lexical-Semantic Resources (LexSem 2012), in conjunction with KONVENS 2012*. Viena, Itália, pp. 438-445.

Cohen, J. 1960. A coefficient of agreement for nominal scales. In: *Educational and Psychological Measurement*, 20, p. 37-46.

Cohen, J. 1968. Weighted kappa: nominal scale agreement with provision for scaled disagreement or partial credit. In: *Psychological Bulletin*, 70, p. 213-220.

Davies, Mark e Joseph L. Fleiss. 1982. Measuring agreement for multinomial data. In: *Biometrics*, 38(4), p. 1047–1051.

Dias-Da-Silva, Bento C. 2005. A construção da base da wordnet.br: conquistas e desafios. In: *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, Brasil, pp. 2238–2247.

Dias-Da-Silva, Bento C., Ariani Di Felippo e Maria das Graças Volpe Nunes. 2008. The automatic mapping of Princeton WordNet lexicalconceptual relations onto the Brazilian Portuguese WordNet database. In *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC 2008),* Marrakech, Morocco, pp. 1535-1541.

Dowty, David. 1991. Thematic Proto-Roles and Argument Selection. In: *Language*, Vol. 67, No. 3. (Sep., 1991), pp. 547-619.

Duran, Magali Sanches e Sandra Maria Aluísio. 2011. Propbank-Br: a Brazilian Portuguese *corpus* annotated with semantic role labels. In: *Proceedings of the 8th Symposium in Information and Human Language Technology*, October 24-26, Cuiabá/MT, Brazil.

Duran, Magali Sanches e Sandra Maria Aluísio. 2012. Propbank-Br: a Brazilian treebank annotated with semantic role labels. In: *Proceedings of the LREC 2012*, May 21-27, Istanbul, Turquia.

Fellbaum, C. (1998) *WordNet*: An electronic lexical database. MIT Press. Cambridge, Massachusetts.

Fillmore, Charles J. 1967. The case for case. In: Bach, Emmon e Robert Harms (Eds.). *Proceedings of the Texas Symposium on Language Universals*, April 13-15, 1967.

Fleiss, J. L. 1971. Measuring nominal scale agreement among many raters. In: *Psychological Bulletin*, v. 76, n. 5, p. 378–382.

Fossati, Marco, Claudio Giuliano e Sara Tonelli. 2013. Outsourcing FrameNet to the Crowd. In: *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, p. 742–747, Sofia, Bulgaria.

Franchi, Carlos e Márcia Cançado. 2003. Teoria generalizada dos papéis temáticos. *Revista Estudos da Linguagem*, v. 11, n. 2.

Gelhausen, T. 2010. *Modellextraktion aus natürlichen Sprachen*: Eine Methode zur systematischen Erstellung von Domänenmodellen. Karlsruhe: KIT Scientific Publishing. Tese de doutorado, Karlsruher Institut für Technologie.

Gildea, Daniel e Martin Jurafsky. 2002. Automatic Semantic Role Labeling. In: *Computer Linguistics*, 28(3), p. 245-288. Cambridge: MIT Press.

Gruber, J.S. 1965. *Studies in Lexical Relations*. MIT. Tese de doutorado. Orientador: Edward S. Klima.

Hong, Jisup e Collin F Baker. 2011. How good is the crowd at "real" wsd? In: *Proceedings of the Fifth Law Workshop (LAW V)*, p. 30-37, Portland, Oregon.

Hovy, Eduard, Mitchell Marcus, Martha Palmer, Lance Ramshaw e Ralph Weischedel. (2006) OntoNotes: The 90% solution. In: *Proceedings of the Human Language Technology Conference of the North American Chapter of the ACL*, New York, p. 57–60.

Ienco, Dino, Serena Villata e Cristina Bosco. 2008. Automatic extraction of subcategorization frames for Italian. In: *Proceedings of the LREC 2008*.
http://www.di.unito.it/~ienco/ienco_LREC08.pdf

Jackendoff, R.S. (1990) *Semantic Structures*. Current Studies in Linguistic Series, v. 18. Cambridge: MIT Press.

Kasper, Simon. 2008. *A comparison of 'thematic role' theories*. Philipps-Universität Marburg. Dissertação de mestrado.

Kipper-Schuler, K. 2005. *VerbNet*: a broad-coverage, comprehensive verb lexicon. University of Pennsylvania. Tese de doutorado orientada por Martha S. Palmer.

Levin, Beth. 1993. *English Verb Classes and Alternations: A Preliminary Investigation*. Chicago: University of Chicago Press.

Levin, Beth e Malka Rappaport-Hovav. 2005. *Argument Realization*. Cambridge, Nova Iorque, Melbourne, Madri, Cape Town, Singapure, São Paulo: Cambridge University Press.

Lima, Bruno de A. F. de. 2007. Valência dos verbos de vitória e derrota em português. Dissertação de Mestrado. Belo Horizonte: UFMG.

Lin, Dekang. 1998. Automatic retrieval and clustering of similar words. In: *Proceedings of the 17th International Conference on Computational Linguistics*, Association for Computational Linguistics, Morristown, NJ, EUA, pp. 768-774.

Loper, Edward, Szu-ting Yi e Martha Palmer. 2007. Combining lexical resources: Mapping between propbank and verbnet. In: *Proceedings of the 7th International Workshop on Computational Linguistics*.

Manning, Christopher D. 1993. Automatic aquisition of a large subcategorization dictionary from corpora. In: *ACL '93 Proceedings of the 31st annual meeting on Association for Computational Linguistics*, p. 235-242.

Maziero, Erick G., Tiago A. S. Pardo, Ariani Di Felippo e Bento C. Dias-Da-Silva. 2008. A Base de Dados Lexical e a Interface Web do TeP 2.0 - Thesaurus Eletronico para o Portugues do Brasil. In: *VI TIL*, pp. 390–392.

Messiant, Cédric. (2008) A subcategorization acquisition system for French verbs. In: *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Techonologies*, Columbus, Ohio, 55-60.

Messiant, Cédric, Anna Korhonen e Thierry Poibeau. 2008. LexSchem: A Large Subcategorization Lexicon for French Verbs. In: *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC)*. Marrakech, Marrocos. http://www.lrec-conf.org/proceedings/lrec2008/pdf/142_paper.pdf

Muniz, M. C. M. 2003. *Léxicos Computacionais: Desafios na Construção de um Léxico de Português Brasileiro*. Monografia de Qualificação. Instituto de Ciências Matemáticas de São Carlos, USP. 50p.

Muniz, M. C. M. 2004. *A construção de recursos lingüístico-computacionais para o português do Brasil: o projeto de Unitex-PB*. Dissertação de Mestrado. Instituto de Ciências Matemáticas de São Carlos, USP. 72p.

Palmer, Martha, Daniel Gildea e Paul Kingsbury. 2005. "The Proposition Bank: A *Corpus* Annotated with Semantic Roles", *Computational Linguistics Journal*, 31:1.

Perini, Mário Alberto. 2008. *Estudos de Gramática Descritiva*: as valências verbais. São Paulo: Parábola Editorial.

Preiss, Judita, Ted Briscoe e Anna Korhonen. 2007. A System for Large-scale Acquisition of Verbal, Nominal and Adjectival Subcategorization Frames from *Corpora*. In: *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics*, Praga, República Tcheca, 2007. Disponível em: http://www.cl.cam.ac.uk/~alk23/acl-07.pdf.

Ramisch, Carlos, Aline Villavicencio e Christian Boitet. 2010a. mwetoolkit: a Framework for Multiword Expression Identification. In: *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC 2010)*, Valetta, Malta, Maio de 2010.

Ramisch, Carlos, Aline Villavicencio e Christian Boitet. 2010b. Web-based and combined language models: a case study on noun compound identification. In: *Proceedings of the 23rd International Conference on Computational Linguistics (COLING 2010)*, Beijing, China, Agosto de 2010.

Salomão, Margarida. 2009. FrameNet Brasil: um trabalho em progresso. *Calidoscópio* 7(3), 171-182.

Scarton, Carolina. 2013. *VerbNet.Br*: construção semiautomática de um léxico verbal online e independente de domínio para o português do Brasil. NILC/USP. Dissertação de mestrado orientada por Sandra Maria Aluísio.

Schulte im Walde, Sabine. 2002. A Subcategorisation Lexicon for German Verbs induced from a Lexicalised PCFG. In: *Proceedings of the 3rd Conference on Language Resources and Evaluation*, v. IV, Las Palmas de Gran Canaria, Espanha, p. 1351–1357. Disponível em: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.16.8846&rep=rep1&type=pdf

Scott, M. 2004. *Wordsmith Tools version 4*. Oxford: Oxford University Press.

Scott, W. A. 1955. Reliability of content analysis: The case of nominal scale coding. In: *Public Opinion Quarterly*, 19(3), p. 321–325.

Zanette, Adriano. 2010. *Aquisição de* Subcategorization Frames *para Verbos da Língua Portuguesa*. Projeto de Diplomação. UFRGS. Orientadora: Aline Villavicencio.

Zanette, Adriano, Carolina Scarton e Leonardo Zilio. 2012. Automatic extraction of subcategorization frames from *corpora*: an approach to Portuguese. In: *Proceedings of PROPOR 2012* - Demonstration Session. Coimbra, Portugal.

Zapiran, B., E. Agirre e L. Màrquez. 2008. Robustness and Generalization of Role Sets: PropBank vs. VerbNet. In: *Proceedings of the ACL-08: HLT*. Association for Computational Linguistics, Columbus, Ohio, June, 2008.

Zilio, Leonardo. 2009. *Colocações especializadas e* Komposita*: um estudo contrastivo alemão-português na área de Cardiologia*. Dissertação de Mestrado. Orientadora: Maria José Bocorny Finatto. Disponível em: http://www.lume.ufrgs.br/bitstream/handle/10183/16877/000706196.pdf?sequence=1

Zilio, Leonardo, Adriano Zanette e Carolina Scarton. 2012. Extração automática de estruturas de subcategorização a partir de *corpora* em português, in: *Anais do ELC 2012*, XI Encontro de Linguística de *Corpus*, São Carlos - SP.