Proceedings of Symposium in Information and Human Language Technology. Natal, RN, Brazil, November 4–7, 2015.

VerbLexPor: um recurso léxico com anotação de papéis semânticos para o português

Leonardo Zilio[1], **Maria José B. Finatto**[2], **Aline Villavicencio**[1]

[1]Instituto de Informática – Universidade Federal do Rio Grande do Sul (UFRGS)

[2]Instituto de Letras – Universidade Federal do Rio Grande do Sul (UFRGS)

{ziliotradutor,mariafinatto}@gmail.com, avillavicencio@inf.ufrgs.br

Abstract. *Semantic role labeling offers vital information for both Linguistics and Natural Language Processing tasks. In this article, we present a lexical resource for Portuguese annotated with semantic roles: VerbLexPor. The resource is a database with verbs and sentences extracted from both a domain specific corpus and a non-specialized generic one. Annotation was manually carried out by a linguist using VerbNet-like semantic roles. The resource has more than 6 thousand annotated sentences and 15 thousand annotated arguments, and is available for download as XML or SQL files. The paper also describes a comparative analysis between the two corpora, showing that the distribution of semantic roles in a general domain is different from that in specific domain.*

Resumo. *A anotação de papéis semânticos oferece informações importantes tanto para tarefas da Linguística quanto do Processamento da Linguagem Natural. Neste artigo, apresentamos um recurso léxico com anotação de papéis semânticos para o português: o VerbLexPor. O recurso é um banco de dados organizado a partir de verbos e sentenças extraídos de dois corpora: um especializado e outro não especializado. A anotação foi feita manualmente por um linguista com papéis semânticos descritivos. O recurso conta com mais de 6 mil instâncias e 15 mil argumentos anotados e se encontra disponível para download nos formatos XML e SQL. Este artigo também descreve uma análise comparativa entre os dois corpora, mostrando que a distribuição de papéis semânticos na linguagem não especializada é diferente da linguagem especializada.*

1. Introdução

Muitos dos avanços recentes na Linguistica Computacional (LC), Processamento de Linguagem Natural (PLN) e áreas afins se devem à disponibilização de recursos lexicais e ontológicos para a comunidade, tais como o WordNet [Fellbaum 1998] e a FrameNet [Baker et al. 1998]. Em particular, recursos léxicos com informações de papéis semânticos de verbos representam uma contribuição interdisciplinar para essas áreas. Na Linguística, esse tipo de recurso subsidia a descrição da língua em foco, tendo em vista que representa um catálogo estruturado de seus verbos com as respectivas informações sintáticas e semânticas. No PLN, esse tipo de recurso pode ser empregado para a análise semântica de sentenças, o reconhecimento automático de significado e outras tarefas associadas. Temos, por exemplo, trabalhos que usam informação semântica para resolução de anáforas [Kong and Zhou 2012], sumarização automática [Yoshikawa et al. 2012],

tradução automática [Feng et al. 2012, Jones et al. 2012] etc. Para o português do Brasil, há três recursos relativamente similares que contemplam verbos e argumentos: o PropBank.Br [Duran et al. 2011, Duran and Aluísio 2012], a VerbNet.Br [Scarton 2013] e a FrameNet Brasil [Salomão 2009].

Neste artigo, apresentamos um recurso léxico diferenciado com informações de papéis semânticos, o VerbLexPor, que foi extraído de dois *corpora*: um de domínio específico com linguagem especializada (artigos de Cardiologia) e outro genérico com linguagem não especializada (textos do jornal Diário Gaúcho). O recurso foi anotado por um linguista com papéis semânticos descritivos no estilo VerbNet [Schuler 2005]. Uma análise comparativa entre os papéis semânticos utilizados em cada um indica um uso diferenciado de papéis como AGENTE, INSTRUMENTO, CAUSA etc.

Na Seção 2, apresentamos trabalhos desenvolvidos para o português que apresentam anotação de papéis semânticos. A Seção 3 apresenta os materiais e o método utilizados. A Seção 4 apresenta os resultados, descrevendo o recurso. A conclusão e discussão de trabalhos futuros são apresentados na Seção 5.

2. Trabalhos relacionados

Nesta seção, apresentamos alguns recursos com anotação de papéis semânticos. Descrevemos recursos baseados na FrameNet [Baker et al. 1998], o PropBank.Br [Duran et al. 2011, Duran and Aluísio 2012] e a VerbNet.Br [Scarton 2013], que são os recursos que mais se assemelham ao VerbLexPor. Ao final, discutimos brevemente as semelhanças e diferenças entre eles.

2.1. Anotações no estilo FrameNet

A FrameNet [Baker et al. 1998] adota papéis semânticos bem específicos e os anota em relação ao domínio e ao contexto. Por exemplo, os papéis semânticos do frame DECISÃO (Copa do Mundo) podem ser VENCEDOR, PERDEDOR, TORNEIO e FINAL. Essa abordagem 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, se estiverem no mesmo cenário comunicativo.

No Brasil, a FrameNet Brasil [Salomão 2009] utiliza essa mesma abordagem. Existem também anotações de *frames* de alguns domínios específicos, como, por exemplo, o Kicktionary_Br [Chishman et al. 2013], que trabalha com textos sobre o futebol, e a anotação de textos jurídicos [Bertoldi and Chishman 2012].

2.2. PropBank.Br

O projeto PropBank.Br [Duran et al. 2011, Duran and Aluísio 2012] utiliza papéis semânticos numerados e contém 5.537 instâncias anotadas com ARG0 a ARG5, além de ter papéis específicos para adjuntos, como, por exemplo, ARG-TMP (para adjuntos adverbiais de tempo). No total, foram anotadas 3.164 sentenças (algumas sentenças foram replicadas, de acordo com a quantidade de verbos principais presentes) e 992 verbos diferentes[1].

[1]Dados verificados diretamente na versão 1.0 em formato CONLL, disponível em: `http://143.107.183.175:21380/portlex/index.php/en/downloadsingl`.

2.3. VerbNet.Br

A VerbNet.Br [Scarton 2013] se propôs a transpor as anotações do inglês para o português aproveitando-se das conexões que existem entre a VerbNet [Schuler 2005], a WordNet [Fellbaum 1998] e a WordNet.Br [Dias-da Silva 2005, Dias-da Silva et al. 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.

A VerbNet.Br conta com um acervo de 5.368 verbos (considerando-se diferentes os casos de verbo pronominal; por exemplo, *apresentar* e *apresentar-se* são considerados como dois verbos)[2]. Os dados disponibilizados dão conta desses verbos associados aos papéis semânticos importados da VerbNet.

2.4. A inter-relação dos recursos

As diferenças entre as anotações no estilo VerbNet, PropBank e FrameNet estão 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 de todas, com seis papéis numerados (ARG0 a ARG5) que se aplicam a qualquer contexto, configurando-se como protopapéis.

No que diz respeito à estrutura, a FrameNet apresenta *corpora* anotados, ou seja, a anotação ocorre no texto corrido; o PropBank extrai sentenças de *corpora* e as anota; e a VerbNet apresenta uma estrutura mais dicionarística, em que o verbo (ou classe de verbos) é apresentado juntamente com suas anotações semânticas e sentenças-exemplo. Nesse sentido, a VerbNet.Br se afastou um pouco de sua original, pois as sentenças-exemplo foram extraídas diretamente de *corpus*.

3. Materiais e Método

Nesta seção, apresentamos os *corpora* utilizados, a ferramenta de anotação, a lista de papéis semânticos e, por fim, a metodologia.

3.1. Corpora

Como queríamos comparar textos especializados e não especializados, foram utilizados dois *corpora*. Para representar os textos especializados, selecionamos um *corpus* composto por artigos científicos da área da Cardiologia compilado por Zilio [Zilio 2009, Zilio 2012]. Para representar os textos não especializados, selecionamos o *corpus* de textos do jornal popular Diário Gaúcho, compilado pelo projeto PorPopular[3]. Na Tabela 1, podemos ver a constituição dos *corpora* em relação ao número de palavras.

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

[2]Dados verificados diretamente na versão 1.0 em formato SQL, disponível em: `http://143.107.183.175:21380/portlex/images/arquivos/verbnetbr/verbnetbr.zip`.

[3]`http://www.ufrgs.br/textecc/porlexbras/porpopular/index.php`.

Table 1. Tamanho dos corpora

Corpus	Nº de palavras
Cardiologia	1.605.250
Diário Gaúcho	1.049.487

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 hábito de leitura para acompanhar um texto mais técnico ou científico.

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).

Ambos os corpora foram analisados automaticamente pelo parser PALAVRAS [Bick 2000] com árvores de dependências sintáticas. Nessa anotação de dependências, o *corpus* apresenta uma associação entre os elementos sintáticos das sentenças.

3.2. Extrator de Estruturas de Subcategorização

Neste estudo, usamos um extrator de estruturas de subcategorização [Zanette 2010, Zilio et al. 2014] para preparar os dados para a anotação. As estruturas de subcategorização podem ser compreendidas como uma forma simplificada da estrutura sintática. Essas estruturas são utilizadas pelo extrator de estruturas de subcategorização para organizar conjuntos de sentenças numa mesma categoria, de acordo com sua esrtutura sintática. O sistema é dividido em quatro módulos: Leitor, Extrator, Construtor e Filtro.

O módulo **Leitor** lê e reconhece cada uma das sentenças do corpus, e a entrega para o módulo extrator, ele permite que a entrada seja de vários formatos (TXT, XML etc.).

Para cada verbo conjugado reconhecido em cada uma das sentenças, o módulo **Extrator** gera tantas cópias da sentença quantos forem os verbos conjugados e extrai as dependências de cada um, tentando classificá-las em termos de estrutura de subcategorização, de acordo com o tipo de argumento[4], que pode ser, por exemplo:

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

Na Tabela 2, apresentamos todas as regras de extração que foram utilizadas pelo sistema.

Este módulo também reconhece se o verbo conjugado é auxiliar ou modal de acordo com a anotação do *parser* e busca automaticamente o verbo principal da oração,

[4]Essas sentenças duplicadas, classificadas por verbos e estrutura de subcategorização formam nossas instâncias de anotação, de modo que temos, em cada instância, um verbo principal e suas dependências.

Table 2. 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[38]	PP[prep.]	ADJUNTO ADVERBIAL[prep.]	6

o qual é passado para o próximo módulo. Além disso, o sujeito é considerado um argumento obrigatório pelo Extrator: na ausência de um sujeito explícito, o módulo assume um sujeito oculto. Isso garante que não haja estruturas de subcategorização diferentes para um mesmo verbo devido à explicitação de sujeito.

O módulo Extrator também reconhece a classificação sintática de cada sintagma, com base nas informações do parser, e a utiliza para atribuir um valor de relevância para cada sintagma (por exemplo: 1 para sujeito, 3 para objeto direto etc.). 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, distinguindo, assim, estruturas de subcategorização que seriam iguais, exceto pelo tipo de voz.

O módulo **Construtor** recebe as informações do Extrator, monta a estrutura de subcategorização com base nos valores de relevância e organiza as informações em um banco de dados. O banco de dados apresenta informações de frequência dos verbos principais, das estruturas de subcategorização, das sentenças e dos argumentos (incluindo sua classificação sintática).

O módulo **Filtro** permite que os dados sejam filtrados pela frequência. O critério que utilizamos foi a exclusão de verbos com frequência igual a 1.

3.3. Lista de Papéis Semânticos

Nossa lista de papéis semânticos é resultado de uma série de experimentos prévios de anotação, nos quais testávamos uma lista e analisávamos a anotação gerada com vistas a aprimorar a lista. No VerbLexPor, usamos principalmente os papéis semânticos da VerbNet 3.2, mas acrescentamos papéis semânticos específicos para adjuntos, os quais foram retirados do PropBank. Além disso, criamos alguns poucos papéis semânticos que achamos úteis para determinados tipos de argumento específicos do português (por exemplo, a partícula/pronome *se*, que possui diversas funções) ou para argumentos que

não haviam sido considerados na VerbNet (por exemplo, casos de verbo suporte, em que o papel de predicador e atribuidor de papel semântico está com o objeto direto ou indireto do verbo principal).

A lista completa é composta por 46 papéis semânticos. Alguns deles são papéis auxiliares, como, por exemplo, o papel *verbo*, que é usado para marcar casos de verbo-suporte, em que o objeto direto (ou indireto) é o real atribuidor de papéis, e casos em que a partícula *se* faz parte do verbo e não é um argumento reflexivo.

Por questões de espaço, não apresentaremos aqui cada um dos papéis utilizados, porém, uma explicação detalhada e com exemplos de cada um deles pode ser encontrada na Seção 8.2 e no Anexo D em Zilio [Zilio 2015]. Na Tabela 4, mais adiante, mostramos uma lista dos papéis semânticos mais utilizados com a respectiva frequência nos dois *corpora*.

3.4. Método

Com os materiais apresentados nas seções anteriores, o processo de desenvolvimento do recurso seguiu os seguintes passos:

- Organização e anotação dos *corpora* com o *parser* PALAVRAS;
- Processamento dos *corpora* com o extrator de estruturas de subcategorização para montagem do banco de dados;
- Seleção de verbos e orações para a anotação; e
- Anotação dos argumentos das orações selecionadas.

No que diz respeito à seleção de dados para a anotação, fizemos algumas escolhas em relação às quantidades a serem anotadas. Optamos por uma anotação amostral, anotando os verbos do Diário Gaúcho, seguindo a ordem de frequência e anotando os mesmos verbos, sempre que possível, também no *corpus* de Cardiologia. Assim, a anotação foi feita nos dois *corpora*, conforme os seguintes critérios:

- Estes verbos foram excluídos da anotação: ser, estar, ter e haver;
- Para todos os verbos selecionados, foram anotadas exatamente dez sentenças de cada uma das estruturas de subcategorização do verbo.

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.

Com essa metodologia, garantimos que todas as estruturas de subcategorização tivessem dez exemplos anotados. Assim, se uma estrutura tivesse 16 exemplos, mas apenas nove estivessem corretos (por exemplo, as demais apresentavam erros de *parser*), ela era descartada como um todo.

A anotação de papéis semânticos propriamente dita foi realizada através de uma interface de anotação em PHP que apresentava os dados do banco de uma maneira estruturada de acordo com os seguintes níveis:

- Verbos

- Estruturas de subcategorização
- Sentenças

Os dois primeiros níveis (verbos e estruturas de subcategorização) são organizacionais, e estavam estruturados de acordo com uma ordem crescente de frequência. Assim, a partir da lista em ordem de frequência dos verbos, era possível selecionar um verbo e, no segundo nível, ver todas as estruturas de subcategorização do verbo em questão. Ao selecionar uma estrutura de subcategorização nesse segundo nível, tínhamos então acesso às sentenças, organizadas por ordem de ocorrência no *corpus*, cada uma com seus respectivos argumentos devidamente destacados, como podemos ver na Figura 1.

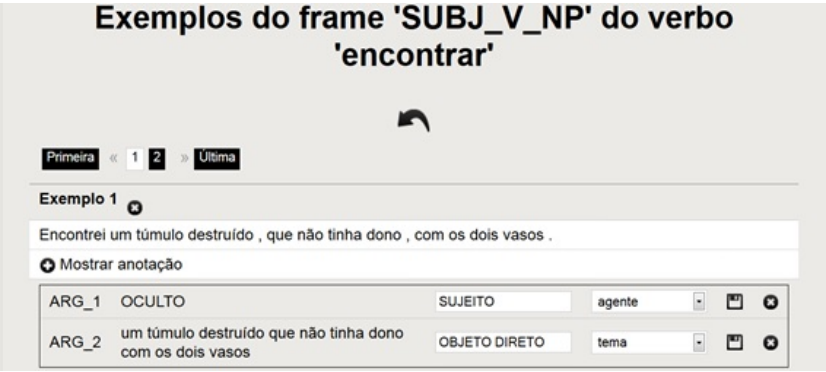

Figure 1. Interface de anotação dos dados

4. Resultados

Nesta seção, apresentamos dados quantitativos do VerbLexPor, mostrando o que o recurso disponibiliza para os usuários. Na Tabela 3, podemos ver os dados básicos do recurso, com o número de instâncias e de argumentos anotados.

Table 3. Dados básicos do VerbLexPor

	DG	Cardiologia
Verbos	191	77
Orações	5.301	1.931
Argumentos	11.089	4.192

Além das mais de seis mil sentenças que têm anotação de papéis semânticos, existem milhares de outras sentenças nos *corpora* que estão anotadas com as funções sintáticas dos diferentes argumentos, de acordo com a classificação do extrator de estruturas de subcategorização. Desse modo, ainda que o recurso não esteja completamente anotado com papéis semânticos, as demais sentenças presentes no banco de dados do recurso apresentam informações sintáticas que foram extraídas com base na anotação do parser PALAVRAS.

Na Tabela 4, podemos observar que, exceto pelo papel semântico TEMA, que é o mais frequente em ambos os *corpora*, os papéis são empregados de maneira bastante

Table 4. Papéis semânticos mais frequentes nos dois *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%

distinta nos dois *corpora*. No Diário Gaúcho, temos uma predominância de AGENTES, enquanto no *corpus* de Cardiologia, os papéis que assumem posições mais frequentes são RESULTADO, PIVÔ, CAUSA E INSTRUMENTO, que têm frequências similares ao papel AGENTE.

Um destaque cabe ao papel INSTRUMENTO, que, em muitos casos, entra na posição do AGENTE no *corpus* de Cardiologia. Podemos ver um exemplo disso nas seguintes sentenças (os INSTRUMENTOS estão em negrito):

- Outro aspecto controverso refere-se ao fato de que **a administração de digitais nas primeiras horas após infarto agudo do miocárdio** poderia aumentar a prevalência de arritmias.
- **Os estudos experimentais** confirmam essa suspeita.
- **A chamada histerese AV** procura permitir que a ativação ventricular se faça espontaneamente pelo sistema de condução cardíaco, por meio de prolongamento automático do intervalo AV do marcapasso.

Também observamos que o *corpus* de Cardiologia apresentou baixa ocorrência do papel semântico EXPERIENCIADOR, que é um dos mais frequentes no Diário Gaúcho.

Em seguida, analisamos informações sintáticas e semânticas de sentença, como as que apresentamos a seguir, nos dois *corpora*:

- SUJEITO<agente> + OBJETO DIRETO<tema>
- SUJEITO<experienciador> + OBJETO DIRETO<tema>
- SUJEITO<tema> + OBJETO REFLEXIVO<verbo> + PREDICATIVO<atributo>

Com essas informações sintáticas e semânticas, realizamos um teste de correlação usando o coeficiente de correlação tau-b de Kendall para observar se a anotação nos dois *corpora* era semelhante. Nesse teste, desconsideramos os papéis de adjuntos[5] e utilizamos

[5]Optamos por retirar da correlação os papéis de adjuntos, pois eles não são atribuídos pelos verbos,

apenas os verbos que foram anotados nos dois *corpora*. O resultado foi τ_b = -0,09 (p = 0,013), o que indica que não há correlação entre as anotações nos dois *corpora*. Isso aponta para um uso diferente dos papéis semânticos em gêneros textuais distintos.

5. Considerações finais

O recurso léxico desenvolvido apresenta uma riqueza de informações semânticas para ser analisada. Em relação aos demais recursos similares existentes para o português, nosso recurso se diferencia por ser um híbrido da VerbNet e do PropBank. As sentenças estão anotadas com papéis semânticos similares aos da VerbNet, porém, a anotação é feita em cima de sentenças extraídas de *corpora*. O recurso com mais de 6 mil instâncias e 15 mil argumentos anotados se encontra disponível para download nos formatos XML e SQL[6].

As anotações em textos especializados e não especializados foram diferentes, com baixa correlação entre as sentenças anotadas e com algumas diferenças entre papéis semânticos específicos, como, por exemplo, os papéis AGENTE e INSTRUMENTO.

6. Agradecimentos

Parte dos resultados apresentados neste trabalho foram obtidos no projeto *Simplificação Textual de Expressões Complexas*, patrocinado pela Samsung Eletrônica da Amazônia Ltda. através da lei número 8.248/91. Também agradecemos ao CNPq (processos 142356/2011-5 e 312184/2012-3) e à CAPES (processo 12537/12-8).

References

Baker, C. F., Fillmore, C. J., and Lowe, J. B. (1998). The berkeley framenet project. In *Proceedings of the 17th international conference on Computational linguistics-Volume 1*, pages 86–90. Association for Computational Linguistics.

Bertoldi, A. and Chishman, R. L. (2012). Desafios para a anotação semântica de textos jurídicos: limites no uso da framenet e rotas alternativas. In *Anais do X Encontro de Linguística de Corpus*, pages 103–121.

Bick, E. (2000). *The parsing system" Palavras": Automatic grammatical analysis of Portuguese in a constraint grammar framework*. Aarhus Universitetsforlag.

Chishman, R., Souza, D., and Padilha, J. (2013). Kicktionary_br: Um relato sobre a anotação semântica de um corpus voltado ao domínio do futebol.[kicktionary_br: A report on the semantic annotation of a corpus covering the domain of soccer].

Dias-da Silva, B. 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*, pages 2238–2247.

Dias-da Silva, B. C., Di Felippo, A., and Nunes, M. d. G. V. (2008). The automatic mapping of princeton wordnet lexical-conceptual relations onto the brazilian portuguese wordnet database. In *LREC*, volume 6, pages 335–342.

então podem aparecer, teoricamente, com qualquer verbo em qualquer sentença, o que desiquilibraria os resultados da correlação na comparação entre os verbos.

[6]O download pode ser feito no site: `http://cameleon.imag.fr/xwiki/bin/view/Main/Semantic\%20role\%20labels\%20corpus\%20-\%20Brazilian\%20Portuguese`.

Duran, M. S. and Aluísio, S. M. (2012). Propbank-br: a brazilian treebank annotated with semantic role labels. In *LREC*, pages 1862–1867.

Duran, M. S., Aluísio, S. M., et al. (2011). Propbank-br: a brazilian portuguese corpus annotated with semantic role labels. In *Proceedings of the 8th Symposium in Information and Human Language Technology, Cuiabá/MT, Brazil*.

Fellbaum, C. (1998). *WordNet*. Wiley Online Library.

Feng, M., Sun, W., and Ney, H. (2012). Semantic cohesion model for phrase-based smt. In *COLING*, pages 867–878.

Finatto, M. J. B., Scarton, C. E., Rocha, A., and Aluísio, S. (2011). Características do jornalismo popular: avaliação da inteligibilidade e auxílio à descrição do gênero. In *Proceedings of the 8th Brazilian Symposium in Information and Human Language Technology*.

Jones, B., Andreas, J., Bauer, D., Hermann, K. M., and Knight, K. (2012). Semantics-based machine translation with hyperedge replacement grammars. In *COLING*, pages 1359–1376.

Kong, F. and Zhou, G. (2012). Exploring local and global semantic information for event pronoun resolution. In *COLING*, pages 1475–1488. Citeseer.

Salomão, M. M. M. (2009). Framenet brasil: um trabalho em progresso. *Calidoscópio*, 7(3):171–182.

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

Schuler, K. K. (2005). Verbnet: A broad-coverage, comprehensive verb lexicon.

Yoshikawa, K., Hirao, T., Iida, R., and Okumura, M. (2012). Sentence compression with semantic role constraints. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers-Volume 2*, pages 349–353. Association for Computational Linguistics.

Zanette, A. (2010). Aquisição de subcategorization frames para verbos da língua portuguesa.

Zilio, L. (2009). Colocações especializadas e 'komposita': um estudo constrastivo alemão-português na área de cardiologia.

Zilio, L. (2012). Colocações especializadas em alemão e português na área de cardiologia. *Tradterm*, 20:146–177.

Zilio, L. (2015). *VerbLexPor: um recurso léxico com anotação de papéis semânticos para o português. UFRGS*. PhD thesis, Tese de doutorado orientada por Maria José Bocorny Finatto e Aline Villavicencio.

Zilio, L., Zanette, A., and Scarton, C. (2014). Automatic extraction of subcategorization frames from corpora. In *New Languages Technologies and Linguistic Research: a Two-Way Road*. Cambridge Scholars Publishing.