VerbLexPor: A Lexical Resource with Semantic Roles for Portuguese

Leonardo Zilio[1], **Maria José Bocorny Finatto**[2], **Aline Villavicencio**[1]
[1]Institute of Informatics, Federal University of Rio Grande do Sul (Brazil),
[2]Institute of Linguistics, Federal University of Rio Grande do Sul (Brazil),
Avenida Bento Gonçalves, 9500 - Agronomia, Porto Alegre - RS, Brazil, 91509-900
lzilio@inf.ufrgs.br, mfinatto@terra.com.br, avillavicencio@inf.ufrgs.br

Abstract

This paper presents a lexical resource developed for Portuguese. The resource contains sentences annotated with semantic roles. The sentences were extracted from two domains: Cardiology research papers and newspaper articles. Both corpora were analyzed with the PALAVRAS parser and subsequently processed with a subcategorization frames extractor, so that each sentence that contained at least one main verb was stored in a database together with its syntactic organization. The annotation was manually carried out by a linguist using an annotation interface. Both the annotated and non-annotated data were exported to an XML format, which is readily available for download. The reason behind exporting non-annotated data is that there is syntactic information collected from the parser annotation in the non-annotated data, and this could be useful for other researchers. The sentences from both corpora were annotated separately, so that it is possible to access sentences either from the Cardiology or from the newspaper corpus. The full resource presents more than seven thousand semantically annotated sentences, containing 192 different verbs and more than 15 thousand individual arguments and adjuncts.

Keywords: Semantic Role Labeling, Lexical Resource, Corpus

1. Introduction

Lexical resources with semantic role information are of great use in NLP applications, for example, in anaphora resolution (Kong and Zhou, 2012), automatic summarization (Yoshikawa et al., 2012) and automatic translation (Feng et al., 2012; Jones et al., 2012). Semantic roles are labels that describe the semantic function of an argument in relation to a determined word class (usually, a verb). For instance, in Sentence A, we have a simple clause, in which the subject (*Roger*) carried out the action (represented by the verb), while the direct object (*the vase*) is being modified by the event. Hence, we have the semantic roles of AGENT and PATIENT. On the other hand, in Sentence B, although we still have a subject (*the vase*), it is no longer the performer of the action, but the one who suffers a change of state (from being in one piece to being in more than one piece), so it plays the role of PATIENT.

- Sentence A: [Roger] broke [the vase].
- Sentence B: [The vase] broke.

In this simple example, we can see that semantic roles are closely related to syntax, but also depends on the semantic of the word it relates to (in this case, the verb).

For Portuguese, there are only a few initiatives that were dedicated to developing resource with this type of semantic information: PropBank.Br (Duran et al., 2011; Duran and Aluísio, 2012), VerbNet.Br (Scarton, 2013) and FrameNet Brasil (Salomão, 2009) are the most prominent, and all of them were directly derived from their English counterparts, as their names clearly point out.

VerbLexPor[1] is a lexical resource with semantic role annotation for Portuguese that portrays the semantic function of Portuguese verb arguments, while also presenting the syntactic structure of each clause. The resource we describe here is linguistically motivated, having arguments manually annotated by a linguist, but it has also a Computational Linguistic apparatus in its methodological base. Its is also the first ready-to-use resource with semantic role labeling for Portuguese that uses VerbNet-like roles and is manually annotated.

This paper is organized in the following way: in Section 2. we describe related work; in Section 3. we present the corpora and tools that were used for developing verbLexPor; in Section 4. we show the main attributes and information which can be found in VerbLexPor; and Section 5. point out the achievements and future developments of this research study.

2. Related Work

The FrameNet project (Baker et al., 1998) uses semantic roles that are specific to a domain or frame. For instance, in the semantic structure of the frame **judgment** there are four core semantic roles (COGNIZER, EVALUEE, EXPRESSOR, and REASON), but the roles are completely different when another frame, such as **perception_active**, is taken into account, even if the verb in question is the same. For Portuguese, FrameNet Brasil (Salomão, 2009) uses this approach and has already annotated various frames for generic and domain-specific vocabulary.

The PropBank project (Palmer et al., 2005) builds on an idea similar to FrameNet, but instead of using descriptive roles (such as SOURCE, GOAL, and THEME) it uses numbered roles (such as ARG0, ARG1, and ARG2). For Portuguese, the PropBank.Br project (Duran et al., 2011; Du-

[1] VerbLexPor is readily available for download in XML and SQL formats: `http://cameleon.imag.fr/xwiki/bin/view/Main/Semantic\%20role\%20labels\%20corpus\%20-\%20Brazilian\%20Portuguese`. It is also possible to look up for syntactic and semantic information on verbs of the resource accessing the Jibiki platform(Mangeot-Nagata, 2006): `http://jibiki.univ-savoie.fr/jibiki/Home.po`.

ran and Aluísio, 2012) has annotated 5.537 instances[2] with ARG0 to ARG5. In addition to these numbered roles, PropBank (as does FrameNet) also has roles for adjuncts (such as ARG-TMP, for adjuncts that express time).

The VerbNet project (Schuler, 2005) uses descriptive roles, such as the ones in FrameNet, but instead of having different roles for each frame, it has a single set of roles that applies to all verbs and arguments. For Portuguese, the VerbNet.Br project (Scarton, 2013) semi-automatically translated the VerbNet structure to Portuguese by means of the inter-connections between VerbNet, WordNet (Fellbaum, 1998) and WordNet.Br (Dias-da Silva, 2005; Dias-da Silva et al., 2008). Applying this method, for aligned synsets in WordNet and WordNet.Br, the roles could be directly imported from English into Portuguese.

3. Methodology

3.1. Corpora

VerbLexPor involves two different domains: a generic one and a specialized one; to reflect this we used two corpora. For the specialized domain, we used a corpus composed of Cardiology papers (Zilio, 2009; Zilio, 2012). For the generic domain, we used a corpus of newspaper articles extracted from Diário Gaúcho, a newspaper that has people with lower reading skills as target audience (Finatto et al., 2011); this corpus was compiled within the project PorPopular[3]. Table 3.1. shows types and tokens of the raw data in the corpora.

Corpus	Types	Tokens
Diário Gaúcho	42K	1M
Cardiology	33K	1.4M

Table 1: Statistics of the Corpora

Both corpora were parsed with dependency trees from the PALAVRAS parser (Bick, 2000), so that the syntactic information was explicit in the parsed sentences. Both corpora were then processed with a subcategorization frames extractor (Zanette, 2010; Zilio et al., 2014), which is described in the next section.

3.2. Subcategorization Frames Extractor

To organize the data from the parsed corpora into a database, we used a subcategorization frame extractor developed by Zanette (2010) and Zilio et al. (2014). Subcategorization frames can be seen as a simplified form of syntactic structure in which only the types of phrases are relevant (such as noun phrase and prepositional phrase).For instance, Sentence C would be rendered as NP_NP_PP (two noun phrases followed by a prepositional phrase).

- Sentence C: [Leila] gave [the book] [to Paul].

The subcategorization frames extractor system is divided into four modules: Reader, Extractor, Builder, and Filter.

The **Reader** module reads and recognizes each sentence in the corpus, delivering it to the Extractor module. It allows for multiple input formats (TXT, XML etc.).

For each conjugated verb in each of the sentences, the **Extractor** module generates that many copies of the sentence and extracts the dependencies of each conjugated verb together with its own copy of the original sentence. After the extraction, it classifies each dependency as a type of phrase, provided it fits the extraction rules.

This module is also responsible for recognizing if the conjugated verb is an auxiliary or modal verb, in case of which it searches for the main verb of the clause according to the information from the parser. It is also important to note that, in our extraction, the subject is considered obligatory, so that the nonexistence of an explicit subject will cause the extractor to fill the subject role with an occult subject, so that two clauses are not considered to have different SCFs based only in the existence or not of an explicit subject.

Based on the syntactic classification, the Extractor module also attributes a relevance index for each phrase, so that a canonical order can be attributed by the Builder module. Lastly, the Extractor module classifies the clause in regard to active or passive voice.

The **Builder** module receives the information from the Extractor module and puts everything together, creating the SCFs and organizing them, together with information about sentence, frequency, verbs etc., in a database.

The **Filter** module allows data to be filtered by frequency. In this module, we used a frequency criterion of 1 for excluding verbs with less than 2 occurrences.

3.3. Semantic Role List

Our semantic role list is a result from various previous annotation experiments (Zilio et al., 2013; Zilio et al., 2014), in which we tested different lists and then analyzed the results to see which one was best for Portuguese. For VerbLexPor, we used a list of semantic roles that is a mix between the roles used in VerbNet (Schuler, 2005) and in PropBank.Br (Duran et al., 2011; Duran and Aluísio, 2012), so that we cover not only arguments, but also adjuncts[4].

During the annotation process, we noticed that we had to create some roles for special cases that are relevant for Portuguese, such as the case of the particle ”se”, which can have many different functions, and specially the case of support verb constructions which are absent in VerbNet. For these cases, we created the roles VERB (for the particle ”se”, when it is part of the verb, and for support verb constructions) and SE_PASSIV (for cases in which the particle ”se” indicates the use of passive voice). Those are organizational and auxiliary roles, but not necessarily a ”semantic role” in its strict sense.

[2] Data from version 1.0 in CONLL format, available at the following website: http://143.107.183.175:21380/portlex/index.php/en/downloadsingl.

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

[4] We normally refer to both arguments and adjuncts as arguments for the sake of simplicity, but there is a difference between them in the sense saturation of the verb. Adjuncts are elements that are not semantically required by the verb, while arguments fulfill a role that is directly ruled by the verb. There are, though, many borderline cases, not only in Portuguese, in which it is difficult to determine whether a phrase is an argument or and adjunct.

The complete list is composed of 46 semantic roles. As said above, some of them are used as auxiliary roles, but most of them are roles such as AGENT, PATIENT, THEME, EXPERIENCER, STIMULUS, etc. All roles were classified in a hierarchical structure, having PARTICIPANT as topmost role, and branching into seven distinct major categories: actor, object, process, time, space, space, time, adjunct, and accessory.

We will not describe here each of the semantic roles, since the complete list with full explanation can be found in (Zilio, 2015), but here are the description of the roles mentioned above:

- **AGENT**: ACTOR that performs the action.
- **THEME**: OBJECT that is not modified by the event; may suffer displacement.
- **PATIENT**: OBJECT that is (implicitly or explicitly) modified by the event.
- **PIVOT**: OBJECT that appears together with THEME, but that has higher semantic stress than the latter in the clause.
- **EXPERIENCER**: PATIENT that perceives a modification or event by means of one of the sensory organs or that expresses a personal feeling.
- **STIMULUS**: CAUSE that forces a sensory or emotional reaction in someone.

As can be seen in the examples above, all definitions follow the principle of genus proximum and differentia specifica, so that a clear hierarchical three can be formed.

3.4. Annotation Method

These were the broad steps we followed to develop the resource:

- Extraction of SCFs;
- Selection of verbs and clauses to be annotated; and
- Annotation

We used the SCF extractor mentioned in Section 3.2. to extract SCF patterns from the corpora and store verb, sentence, and argument data in a database (Figure 1) for each corpus. The selection of the data for annotation was done by order of the most frequent verbs in the Diário Gaúcho corpus (DG). We annotated the verb first in the DG, and then proceeded to annotate the same verb on the Cardiology corpus. The annotation in both corpora followed these criteria:

- These verbs were excluded from the annotation process: *ser (to be), estar (to be), ter (to have, to be),* and *haver (to have, to be)*;
- For all the annotated verbs, exactly ten sentences were annotated for each SCF that contained ten or more sentences.

The a priori exclusion of four verbs (*ser, estar, ter,* and *haver*) was due to their extreme polysemy and/or frequency in both corpora. The sampling method used for the annotation would most certainly not reflect most of their facets, and it would consume a lot of time that could be dedicated to a greater number of verbs.

The annotation instance was thus composed of a sentence, but only one main verb of each sentence was annotated at a time, so that sentences with multiple clauses were replicated in the database, and each clause was annotated separately. It is important to point out that there are only main verbs in the database[5]. An annotation instance can be seen in the interface (Figure 2), which displayed the data from the database in a more clean and easy-to-understand way. The interface sorted the information according to frequency in the following levels:

1. Verbs
2. Subcategorization frames
3. Sentences

Both the first and second levels were used only for organization purposes. Beginning with the frequency order of verbs, one can select a verb and see all the SCFs related to it. Selecting a SCF would take the annotator to the third level, in which all sentences regarding that SCF and verb are shown. The sentences are organized in order of appearance in the corpus, each with their respective arguments and adjuncts. The annotation was made through a combo box near each of the arguments, as can be seen in Figure 2.

4. Results

Table 4. displays basic data from the resource, such as number of sentences and arguments annotated. Apart from the more than seven thousand instances (i.e. clause plus arguments) annotated with semantic role labels, there are thousands of other sentences annotated with syntactic functions for each of the arguments (as a step of the subcategorization frame extraction process).

	Diário Gaúcho	**Cardiology**
Verbs	191	77
Instances	5,301	1,931
Arguments	11,089	4,192

Table 2: Semantically annotated data in VerbLexPor in both Diário Gaúcho and Cardiology corpora

In Table 4., we present the fifteen most frequent semantic role labels in both corpora: Diário Gaúcho (DG) and Cardiology. As we can see, THEME is the most frequent role in both corpora, but then the semantic roles change, showing that domain has influence in semantic role distribution.

Using a sample with the verbs that were annotated in both corpora (76 verbs), we tested the correlation index between syntactic and semantic information in Cardiology

[5] We will better explain the process of extraction of SCFs and organization of the database in the final version of the paper.

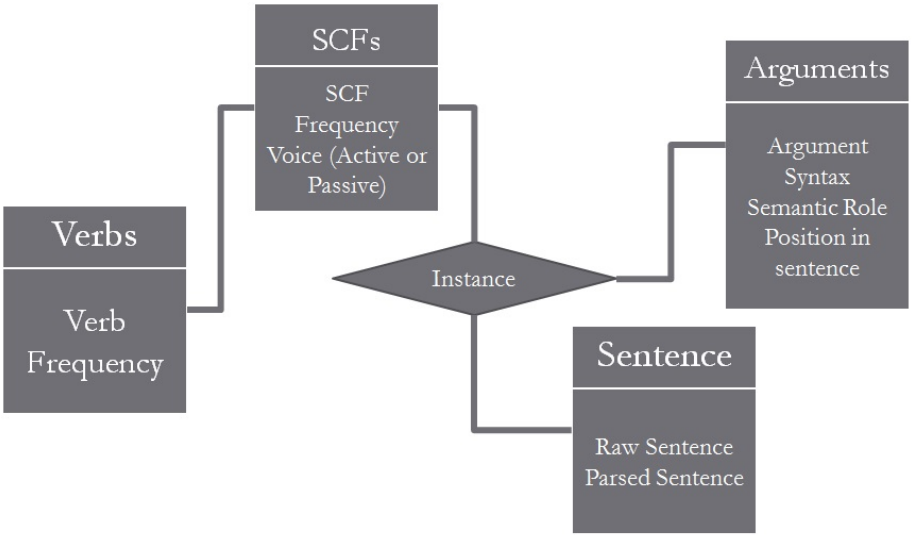
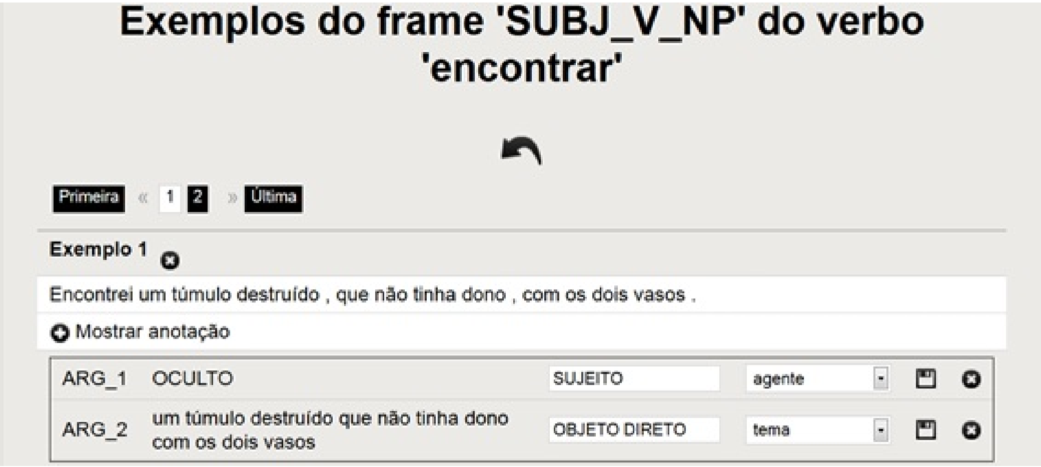


Figure 1: Database Structure



Exemplos do frame 'SUBJ_V_NP' do verbo 'encontrar'

Exemplo 1

Encontrei um túmulo destruído , que não tinha dono , com os dois vasos .

Mostrar anotação

ARG_1	OCULTO	SUJEITO	agente
ARG_2	um túmulo destruído que não tinha dono com os dois vasos	OBJETO DIRETO	tema

Figure 2: Annotation interface

and newspaper texts. The selected correlation index was Kendall's τ_b, due to the non-parametric data and the accounting for ties. The type of data we used for the correlation test was the frequency of syntactic and semantic structures on both corpora, here are some examples of these structures:

- SUBJECT<agent> + DIRECT OBJECT<theme>
- SUBJECT<experiencer> + DIRECT OBJECT<theme>
- SUBJECT<theme> + REFLEXIVE OBJECT<verb> + PREDICATIVE<attribute>

For the test to be more accurate, we did not consider adjuncts[6], so that there would be no difference in the corpora only due to the presence of random adjuncts. The result was a correlation of τ_b = -0,09 (p = 0,013), which indicates no correlation at all between the corpora, showing the influence of genre distinction in the semantic role level.

Table 4. already showed us that, while the newspaper texts have a clear preference (higher frequency) for AGENT after THEME, the Cardiology corpus has a more varied distribution, with several semantic roles almost tying for second (RESULT, AGENT, and PIVOT, closely followed by CAUSE and INSTRUMENT). The incidence of EXPERIENCER and TOPIC in both corpora is also an interesting and clear difference, showing that human feelings and interactions carries much less importance in the Cardiology corpus in relation to the newspaper corpus.

With the structure used to build VerbLexPor, it contains both the sentence oriented annotation of PropBank.Br (Duran et al., 2011; Duran and Aluísio, 2012), where each annotation has an example sentence attached to it, and the descriptive roles that are present in VerbNet.Br (Scarton, 2013). The fact that the frames were manually annotated by a linguist also confers more precision for the data, even if at a cost of coverage.

5. Final Remarks

This abstract presented VerbLexPor, a resource for Portuguese that contains information on the syntax and se-

[6] To achieve that, we removed syntactic structures annotated with those semantic roles that were specific for adjuncts.

Semantic Role	DG Freq.	% DG	Cardiology Frequency	% Cardiology	Total Freq.	% Total
THEME	3.015	27.19%	1.416	33.78%	4.431	29.00%
AGENT	2.540	22.91%	254	6.06%	2.794	18.28%
PLACE	540	4.87%	143	3.41%	683	4.47%
RESULT	363	3.27%	289	6.89%	652	4.27%
PATIENT	497	4.48%	145	3.46%	642	4.20%
EXPERIENCER	591	5.33%	47	1.12%	638	4.18%
PIVOT	345	3.11%	282	6.73%	627	4.10%
VERB	407	3.67%	184	4.39%	591	3.87%
TOPIC	453	4.09%	68	1.62%	521	3.41%
CAUSE	191	1.72%	222	5.30%	413	2.70%
MOMENT	306	2.76%	87	2.08%	393	2.57%
GOAL	257	2.32%	130	3.10%	387	2.53%
INSTRUMENT	152	1.37%	208	4.96%	360	2.36%
SITUATION	176	1.59%	162	3.86%	338	2.21%
ATTRIBUTE	194	1.75%	136	3.24%	330	2.16%

Table 3: Most frequent semantic roles in both corpora

mantic of sentences. The resource was manually annotated with descriptive semantic roles, such as AGENT, THEME, PACIENT, PLACE, EXPERIENCER, etc. The annotated resource comprehends more than seven thousand annotated instances, which correspond to more than 15 thousand annotated arguments for 192 verbs in both Cardiology and newspaper contexts. The resource is readily available for download in XML and SQL formats, and is also available for direct search in the Jibiki platform.

6. Acknowledgements

This research was partially developed in the context of the project *Text Simplification of Complex Expressions*, sponsored by Samsung Eletrônica da Amazônia Ltda., in the terms of the law n. 8.248/91. We would also like to thank the Brazilian funding agencies CNPq (process 142356/2011-5) and CAPES (process 12537/12-8) and the project RITA (STIC-AmSud/CAPES 14STIC-10).

7. Bibliographical 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.

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

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.

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.

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.

Mangeot-Nagata, M. (2006). Dictionary building with the jibiki platform. In *Atti del XII Congresso Internazionale di Lessicografia: Torino, 6-9 settembre 2006*, pages 185–188.

Palmer, M., Gildea, D., and Kingsbury, P. (2005). The proposition bank: An annotated corpus of semantic roles. *Computational linguistics*, 31(1):71–106.

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*. Ph.D. 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., Ramisch, C., and Finatto, M. J. B. (2013). Desenvolvimento de um recurso léxico com papéis semânticos para o português. *Linguamática*, 5(2):23–41.

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.

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*. Ph.D. thesis, Tese de doutorado orientada por Maria José Bocorny Finatto e Aline Villavicencio.