Towards a cross-linguistic VerbNet-style lexicon for Brazilian Portuguese

Carolina Scarton, Sandra Aluísio

Center of Computational Linguistics (NILC), University of São Paulo
Av. Trabalhador São-Carlense, 400. 13560-970 São Carlos/SP, Brazil
carol@icmc.usp.br, sandra@icmc.usp.br

Abstract

This paper presents preliminary results of the Brazilian Portuguese Verbnet (VerbNet.Br). This resource is being built by using other existing Computational Lexical Resources via a semi-automatic method. We identified, automatically, 5688 verbs as candidate members of VerbNet.Br, which are distributed in 257 classes inherited from VerbNet. These preliminary results give us some directions of future work and, since the results were automatically generated, a manual revision of the complete resource is highly desirable.

1. Introduction

The task of building Computational Lexical Resources (CLRs) and making them publicly available is one of the most important tasks of Natural Language Processing (NLP) area. CLRs are used in many other applications in NLP, such as automatic summarization, machine translation and opinion mining. Specially, CLRs that treat the syntactic and semantic behaviour of verbs are very important to the tasks of information retrieval (Croch and King, 2005), semantic parser building (Shi and Mihalcea, 2005), semantic role labeling (Swier and Stevenson, 2004), word sense disambiguation (Girju et al., 2005), and many others. The reason for this is that verbs contain information about sentence roles, such as the argument position, that could be provided by knowing the verb.

The English language has a tradition in building CLRs. The most widely known are WordNet (Fellbaum, 1998), PropBank and its frame files (Palmer et al., 2005), FrameNet (Baker et al., 2005) and VerbNet (Kipper, 2005). All of these resources have information about verbs, but in a different way: WordNet contains deep semantic relations of verbs, such as synonym and hyperonym; PropBank has information about verbs and their arguments with semantic role annotation; FrameNet groups verbs according to the scenario in which these verbs appear; and VerbNet groups verbs according to their syntactic and semantic behaviours. VerbNet-style follows Levin's hypothesis (Levin, 1993), in which verbs that share the same syntactic behaviour also share some semantic components. As an example (from Levin (1993)), let's observe verbs *to spray* and *to load* (sentences 1 and 2).

1. Sharon *sprayed* water on the plants / Sharon *sprayed* the plants with water

2. The farmer *loaded* apples into the cart / The farmer *loaded* the cart with apples

It is possible to see that the verb *to spray* in 1 and *to load* in 2 share the same syntactic behaviour (the objects changed places) and the semantic of these verbs is related to putting and covering something. This alternation of arguments is called diathesis alternation. In this example, it is also possible to see that the semantic of Levin's verb classes is superficial: we can not say that the verb *to spray* is a synonym of the verb *to load*. To fulfill this gap, VerbNet has mappings to WordNet, which has deeper semantic relations.

Brazilian Portuguese language lacks CLRs. There are some initiatives like WordNet.Br (Dias da Silva et al., 2008), that is based on and aligned to WordNet. This resource is the most complete for Brazilian Portuguese language. However, only the verb database is in an advanced stage (it is finished, but without manual validation), currently consisting of 5,860 verbs in 3,713 *synsets*. Other initiatives are PropBank.Br (Duran and Aluisio, 2011), FrameNet.Br (Salomao, 2009) and FrameCorp (Bertoldi and Chishman, 2009). The first one is based on PropBank and the second and third are based on FrameNet.

However, none of these resources tackles the syntactic/semantic interface of the verbs. Therefore, we proposed VerbNet.Br (Scarton, 2011), which is a VerbNet for Brazilian Portuguese language, directly aligned to VerbNet. This is why we started our work from a manual step, which involved manual translation of diathesis alternations of VerbNet from English into Portuguese (see more details in Section 3.1).

Whereas CLRs inspired on WordNet, PropBank and FrameNet have been built by using manual approaches based on corpora, several approaches to build verbnets for other languages employed completely automatic methods, focusing on machine learning. Studies like Joanis and Stevenson (2003), Sun et al. (2008), Sun et al. (2009), Kipper (2005), Merlo and Stevenson (2001) and Sun and Korhonen (2011) for English language, Merlo et al. (2002) for Italian language, Schulte im Walde (2006) for German language, Ferrer (2004) for Spanish language and Sun et al. (2010) for French language focuse on machine learning. Most of these researches used information of frames subcategorization as features for machine learning methods. Subcategorization frames provides information about the syntactic realization of verbs as well as diathesis alternations.

To build VerbNet.Br, we are considering the hypothesis that Levin's verb classes have a cross-linguistic potential - this hypothesis was enunciated by Jackendoff (1990) and verified by Merlo et al. (2002) for Italian, Sun et al. (2010) for French and Kipper (2005) for Portuguese. Using that, we proposed a semi-automatic method to build the VerbNet.Br by using the alignments between WordNet.Br and WordNet and the mappings between VerbNet and WordNet. We also have the hypothesis that this semi-automatic method will present better results (results with more precision) than the completely automatic methods.

In this paper we present the current state of VerbNet.Br project by showing a complete run in the method we have chosen and some preliminary results. In section 2, we present a literature review of CLRs and the relation of these and VerbNet.Br. We also present in this section the relation of VerbNet.Br and some completely automatic methods. In section 3, we present the method to build VerbNet.Br. In section 4, we present preliminary results of VerbNet.Br, using as examples the classes ”Equip-13.4.2”, ”Remove-10.1” and ”Banish-10.2” inherited automatically from VerbNet. Finally, in section 5, we present some conclusions and future work.

2. Literature review

Since VerbNet.Br has been built by using VerbNet, WordNet and WordNet.Br, our literature review is focused on these three resources. Moreover, we also present some completely automatic approaches that are related to our research.

2.1. WordNet

WordNet (Fellbaum, 1998) is the most used CLR. The main semantic relation of this kind of CLR is synonymy - *synsets* are based in this relation. Because of this, Wordnet is composed by four classes: nouns, adjectives, adverbs and verbs (words from different syntactic classes are not synonyms). The verb database contains 11,306 verbs and 13,508 *synsets*.

By using WordNet, wordnets to other languages has been built. MultiWordNet (Bentivogli et al., 2002) and EuroWordNet (Vossen, 2004) are large projects that aim to build wordnets to many other languages such as Italian, Spanish, German, French and Portuguese. WordNet.Br is also based on WordNet.

2.2. WordNet.Br

The Brazilian Portuguese wordnet (called WordNet.Br) (Dias da Silva et al., 2008) is based on WordNet and aligned to it. This CLR is the most complete for Brazilian Portuguese language and has the verb database finished but still under validation. WordNet.Br used the following method:

- A linguist selected a verb in Portuguese;
- Then, he/she searched in a Portuguese-English dictionary for the verb in English that best fitted in the sense in Portuguese;
- After that, he/she searched in WordNet for the *synset* that best fitted in the sense;
- Finally, the linguist decided what kind of relation the *synsets* had. The options were: EQ_SYNONYM (perfect synonym), EQ_NEAR_SYNONYM (imperfect synonym), EQ_HAS_HYPONYM (hyponymy relation) and EQ_HAS_HYPERNYM (hypernymy relation). These relations were defined by Vossen (2004) in the EuroWordNet project.

Figure 1 (from Felippo and Dias da Silva (2007)) shows an example of a *synset* of WordNet aligned to a *synset* of WordNet.Br by using the EQ_SYNONYM alignment.

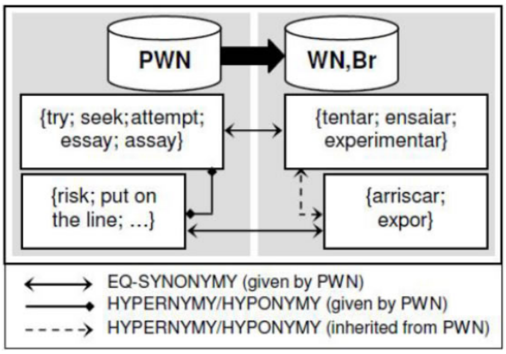

PWN
WN,Br
{try; seek;attempt; essay; assay}
{tentar; ensaiar; experimentar}
{risk; put on the line; ...}
{arriscar; expor}
EQ-SYNONYMY (given by PWN)
HYPERNYMY/HYPONYMY (given by PWN)
HYPERNYMY/HYPONYMY (inherited from PWN)

Figure 1: Example of a *synset* alignment between WordNet and WordNet.Br (Felippo and Dias da Silva, 2007)

As you can see in Figure 1, the other semantic relations, like hypernymy, can be inherited by WordNet.Br from WordNet. This is possible because of the alignment between the *synsets*.

2.3. VerbNet

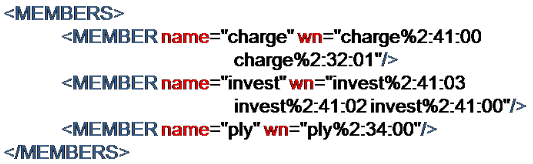
VerbNet (Kipper, 2005) has syntactic and semantic information about English verbs. It is based on Levin's hypothesis of verb classes. This CLR has mappings to PropBank, FrameNet and WordNet.

Verb classes have a group of members, thematic roles, selective restrictions, syntactic frames and semantic predicates. Table 1 shows the structure of ”Equip-13.4.2”, which is a class of VerbNet.

<table>
<tr><td colspan="3">Equip-13.4.2</td></tr>
<tr><td colspan="3"><b>Thematic roles and Selectional restrictions:</b> Agent [+animate — +organization], Theme and Recipient [+animate — +organization]</td></tr>
<tr><td colspan="3"><b>Members:</b> charge, invest, ply, arm, equip, rearm, redress, regale, reward, saddle, treat, armor, burden, compensate, encumber, overburden, weight</td></tr>
<tr><td colspan="3"><b>Frames:</b></td></tr>
<tr><td>NP V NP PP</td><td>Brown equipped Jones with a camera.</td><td>Agent V Recipient with Theme</td></tr>
<tr><td><b>Semantic Predicates</b></td><td colspan="2">(1) has_possession(start(E), Agent, Theme); (2) has_possession(end(E), Recipient, Theme); (3) transfer(during(E), Theme); (4) cause(Agent, E)</td></tr>
</table>

Table 1: The structure of ”Equip-13.4.2” class of VerbNet

Each member could be mapped to one or more *synsets* of WordNet, as we can see in Figure 2. The mappings are represented by ”wn” tags.



```
<MEMBERS>
        <MEMBER name="charge" wn="charge%2:41:00
                                    charge%2:32:01"/>
        <MEMBER name="invest" wn="invest%2:41:03
                                    invest%2:41:02 invest%2:41:00"/>
        <MEMBER name="ply" wn="ply%2:34:00"/>
</MEMBERS>
```

Figure 2: Example of the mappings between VerbNet and WordNet

2.4. Automatic methods

Some studies grouped verbs by using machine learning methods in large corpora. Although the method proposed here is semi-automatic and based on other resources, we also used some techniques of these studies and we intend to compare the results of our method with the results of a machine learning method.

For the English language, studies of Joanis and Stevenson (2003), Merlo and Stevenson (2001), Kipper (2005), Sun et al. (2008) and Sun et al. (2009) presented methods to group verbs automatically. Especially, Kipper (2005) made experiments with machine learning to improve the VerbNet. Sun et al. (2008), Sun et al. (2009) and Joanis and Stevenson (2003) considered the Levin's taxonomy to put verbs into classes.

For other languages, we can cite Sun et al. (2010) (French), Ferrer (2004) (Spanish), Merlo et al. (2002) (Italian) and Schulte im Walde (2006) (German). Specifically, Sun et al. (2010) used a gold standard to compare with the machine learning results. The building of this gold standard was quite similar to our method to build VerbNet.Br. Besides that, Sun et al. (2010), Merlo et al. (2002) and Schulte im Walde (2006) also considered the Levin's taxonomy.

Most of these researches used subcategorization frames as features for machine learning. In our approach, we use subcategorization frames too, but in a different way (see Section 3). However, we also intend to evaluate the results of our semi-automatic method, comparing them with the results of a completely automatic method that will use machine learning with subcategorization frames as features.

3. Building VerbNet.Br

Although Scarton (2011) reported the method developed to build the VerbNet.Br, such paper is available only in Portuguese and, for this reason, we decided to quickly describe it here. The proposed method is composed by four stages (Sections 3.1, 3.2, 3.3 and 3.4, respectively, present the four stages). We based our experiments on version 3.0 of VerbNet and we only considered the classes defined by Levin (1993) without the subclasses and extensions proposed by Kipper (2005).

3.1. Stage 1: Manual translation of diathesis alternations of VerbNet from English into Portuguese

The Stage 1 (under development) is the direct translation of diathesis alternations from English into Portuguese, manually. For example, Table 1 presents only one diathesis alternation for the class "Equip-13.4.2": "NP V NP with NP", that means, a noun phrase followed by a verb, followed by a noun phrase, followed by the preposition "with", followed by a noun phrase. This alternation can be directly translated into Portuguese:"NP V NP *com* NP". To do that, we just replaced the preposition "with" in English for the preposition *com* in Portuguese. In this step, we only consider the alternations that can be directly translated. If an alternation doesn't occur in Portuguese or if it occurs in a different way, it is not translated.

We decided to translate only the alternations that fits perfectly into Portuguese because of two reasons. The first one is that we did not have specialized people to do this task. The task is being developed by a native speaker of Portuguese, who does not have linguistic expertise. The second one is that we intend to identify the similarity between English and Portuguese diathesis alternations and find out how many diathesis alternations are shared by both languages. Besides that, we intend firstly to establish the perfect alignments and, after, deal with the other cases. As future work, we intend to extend VerbNet.Br with alternations that were not directly translated and with alternations that appear in Portuguese, but not in English, such as phrases without subject.

3.2. Stage 2: Automatic search of diathesis alternations of Brazilian Portuguese verbs in corpus

The Stage 2 (finished) is the search for diathesis alternations of verbs in Portuguese in corpus. This step was carried out by using the subcategorization frames extractor tool developed by Zanette (2010). This tool, based on Messiant (2008) developed for the French language, uses a corpus, tagged by PALAVRAS parser (Bick, 2005), to identify the syntactic behaviour of verbs. In other words, the search was for patterns like "NP V NP", "NP V com NP", etc (Zanette et al., 2012).

The Lácio-ref (Aluísio et al., 2004), a Brazilian Portuguese corpus from Lácio-Web project, was used in this stage. This corpus has, approximately, 9 million words and it is divided into five genres: scientific, informative, law, literary, and instructional. We identified 8,926 verbs and 196,130 frames. However, these numbers also contain incorrect verbs and incorrect frames that will be discarded by using a threshold frequency.

For example, the verbs of class "Equip-13.4.2" should present in the corpus the pattern "NP V NP *com* NP" as defined in the Stage 1.

3.3. Stage 3: Automatic generation of candidate members of VerbNet.Br by using other CLRs

The Stage 3 (finished) was the generation of candidate members for classes of VerbNet.Br, by using the mappings between VerbNet and WordNet and the alignments between WordNet and WordNet.Br. Figure 3 shows how this stage was developed: for each class in VerbNet, we searched firstly the *synsets* of WordNet mapped to each verb member, then we searched for the *synsets* of WordNet.Br and thus the members of these Portuguese *synsets* were defined as the candidate members. We defined 4,063 verbs as candidate members in 207 classes.

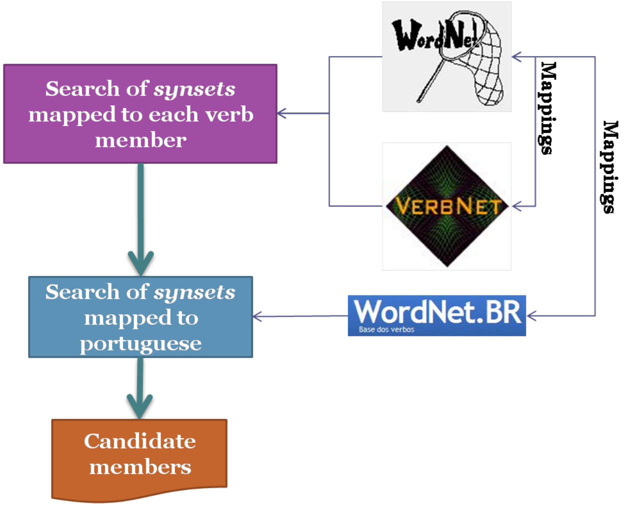

Search of *synsets* mapped to each verb member
WordNet
Mappings
Mappings
VERBNET
Search of *synsets* mapped to portuguese
WordNet.BR
Base dos verbos
Candidate members

Figure 3: Candidate Members definition

For the class ”Equip-13.4.2” we identified 38 candidate members, such as *dotar* (to gift) and *armar* (to arm).

3.4. Stage 4: Selection of members of VerbNet.Br CLRs

Finally, the Stage 4 (future work) will use all the others together. Figure 4 shows an illustration of how this stage will work.

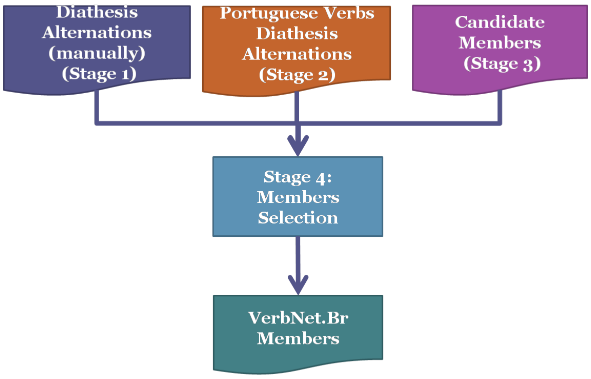

Diathesis Alternations (manually) (Stage 1)
Portuguese Verbs Diathesis Alternations (Stage 2)
Candidate Members (Stage 3)
Stage 4: Members Selection
VerbNet.Br Members

Figure 4: Stage 4: putting all stages together

As may be seen in Figure 4, this stage will use all the other stages to select the members of VerbNet.Br. For each candidate member, we will compare the diathesis alternations identified in the Stage 2 with the diathesis alternations translated in the Stage 1. If the candidate member presents in the corpus (Stage 2) a certain frequency of the diathesis alternations defined in Stage 1, it will be selected, if not, it will be discarded. Some results of this stage, from a pilot test, will be presented in the next section.

4. Experiments

This section contains the preliminary results of VerbNet.Br. Since the Stages 2 and 3 are already done, we carried out an experiment with three classes taken from the Stage 1. The classes selected were ”Equip-13.4.2”, which is shown in Table 1, ”Remove-10.1”, shown in Table 2, and ”Banish-10.2”, shown in Table 3.

Remove-10.1		
Thematic roles and Selectional restrictions: Agent [+int_control — +organization], Theme and Source [+location]		
Members: abolish, abstract, cull, deduct, delete, depose, disgorge, dislodge, disengage, draw, eject, eliminate, eradicate, excise, excommunicate, expel, extirpate, extract, extrude, lop, omit, ostracize, partition, pry, reap, retract, roust, separate, shoo, subtract, uproot, winkle, wrench, withdraw, oust, discharge, dismiss, evict, remove, sever		
Frames:		
NP V NP	Doug removed the smudges.	Agent V Theme
Semantic Predicates	(1) cause(Agent, E) (2) location(start(E), Theme, ?Source) (3) not(location(end(E), Theme, ?Source))	
NP V NP PP.source	Doug removed the smudges from the tabletop.	Agent V Theme +src Source
Semantic Predicates	(1) cause(Agent, E); (2) location(start(E), Theme, Source); (3) not(location(end(E), Theme, Source))	

Table 2: The structure of ”Remove-10.1” class of VerbNet

Section 4.1 presents materials and methods. Section 4.2 contains the preliminary results for the three classes cited above.

4.1. Materials and methods

Since the Stages 2 and 3 are stored in a MySQL database, it was easy to recover the data and to compare it. The Stage 1 is being developed in XML files, making automatic information recovery easy too.

The subcategorization frames identified in Stage 2 needed to be filtered out mainly because of some parsing errors like adjuncts tagged as arguments. Therefore, the Maximum Likelihood Estimate (MLE), used in previous work (Ferrer, 2004), was applied in this phase. The MLE is the ratio of the frequency of a verb frame to the whole frequency of the verb. We considered a threshold of 0,05 (the same adopted by Ferrer (2004)).

We also needed to decide how many diathesis alternations we would consider to select a candidate member. For these preliminary experiments, the rate of 60% was our choice, although we will also test other values. This was important because some diathesis alternations defined in the Stage 1 did not occur in the corpus (the alternation could be easily and correctly generated, but they were never used by native speakers). The rate of 60% was chosen empirically. As future work, we intend to vary this rate (50%, 70%, etc) and to evaluate the impact of this rate in the overall precision and recall.

Banish-10.2		
Thematic roles and Selectional restrictions: Agent [+animate — +organization], Theme[+animate], Source [+location] and Destination [+location — -region]		
Members: banish, deport, evacuate, expel, extradite, recall, remove, shanghai		
Frames:		
NP V NP	The king banished the general.	Agent V Theme
Semantic Predicates	(1) cause(Agent, E); (2) location(start(E), Theme, ?Source); (3) location(end(E), Theme, ?Destination)	
NP V NP PP.source	The king banished the general from the army.	Agent V Theme +src Source
Semantic Predicates	(1) cause(Agent, E); (2) location(start(E), Theme, Source); (3) not(location(end(E), Theme, Source))	
NP V NP PP.destination	The king deported the general to the isle.	Agent V Theme to Destination
Semantic Predicates	(1) cause(Agent, E); (2) location(start(E), Theme, ?Source); (3) location(end(E), Theme, Destination)	

Table 3: The structure of ”Banish-10.2” class of VerbNet

4.1.1. Preliminary Results

In this section we present some preliminary results of VerbNet.Br, by using the classes ”Equip-13.4.2”, ”Banish-10.2” and ”Remove-10.1”.

Equip-13.4.2

The class ”Equip-13.4.2” has only one syntactic frame: ”NP V NP with NP” (as shown in Table 1). In the Stage 1, this frame was directly translated into Portuguese: ”NP V NP *com* NP”. Since we have only one syntactic frame, we selected it to be the parameter to discard or to select a candidate member.

In the Stage 3, 38 candidate members were defined for the class ”Equip-13.4.2”. Searching in the results of Stage 2, only 12 verbs presented the syntactic frame defined in the Stage 1. However, only the verb *dotar* (to gift) presented a threshold higher than 0,05. Therefore, the Portuguese version of the class ”Equip-13.4.2” has one syntactic frame (as defined above) and only one member: the verb *dotar* (to gift). In order to verify if the verb *dotar* (to gift) was correctly selected, we evaluated the sentences in the corpus from which the syntactic frame was derived. Two sentences were found:

1. *A natureza dotara Aurélia com a inteligência viva e brilhante[...]* (Nature gifted Aurélia with a bright, vibrant intelligence.)
2. *Era tão universal e inventivo, que dotou a poesia malaia com um novo metro[...]* (He was so universal and creative that he has gifted malayan poetry with a new meter.).

The two sentences present the semantic of the class: X gives something to Y that Y needs. However, if we go back to the Table 1, some of the requirements are missed. For example, the first argument needs to be an animate Agent or an organization and in the first sentence the first argument (*A natureza* - Nature) is not animate neither an organization. This may be explained because Nature was used in a figurative way and took the place of an animate entity. This class is shown in Table 4.

Equip-13.4.2 - BR		
Thematic roles and Selectional restrictions: Agent [+animate — +organization], Theme and Recipient [+animate — +organization]		
Members: *dotar* (to gift)		
Frames:		
NP V NP PP	*Brown dotou Jones com uma câmera.*	Agent V Recipient com Theme
Semantic Predicates	(1) has_possession(start(E), Agent, Theme); (2) has_possession(end(E), Recipient, Theme); (3) transfer(during(E), Theme); (4) cause(Agent, E)	

Table 4: The structure of ”Equip-13.4.2” class of VerbNet.Br

Remove-10.1

Finally, for the class Remove-10.1, the two diathesis alternations (shown in Table 2) were translated from English into Portuguese: ”NP V NP” and ”NP V NP *de* NP”. To be a member, a verb needed to present two of these syntactic frames (the roof of 1.2 (0.6*2)), respecting the MLE measure.

In Stage 3, 151 verbs were identified. Looking at the results from Stage 2 , we found 85 verbs that present at least one of the syntactic frames. Selecting only verbs that present the two diathesis alternations defined for this class by using the threshold of 0.05, we found the verbs *arredar* (to move away), *destituir* (to oust), *diminuir* (to decrease), *dispensar* (to dismiss), *excluir* (to exclude), *isolar* (to isolate), *separar* (to separate) and *tirar* (to remove). Searching for sentences of verb *separar* (to separate) we found two examples:

1. *O vaqueiro separa escrupulosamente a grande maioria de novas cabeas pertencentes ao patrão[...]* (The cowboy carefully picks out most of the new cattle belonging to his master.)
2. *Cetonas em estado de triplete podem separar hidrogênios de grupos benzilas[...]* (Ketones in triplet states can separate hydrogen from benzyl groups.)

The semantic of this class is ”the removal of an entity from a location” (Levin, 1993). The sentences presented before follow this semantic and respect the restrictions defined for the thematic roles (shown in Table 2). This class is presented in Table 5.

Remove-10.1 - BR		
Thematic roles and Selectional restrictions: Agent [+int_control — +organization], Theme and Source [+location]		
Members: *arredar* (to move away), *destituir* (to oust), *diminuir* (to decrease), *dispensar* (to dismiss), *excluir* (to exclude), *isolar* (to isolate), *separar* (to separate) and *tirar* (to remove)		
Frames:		
NP V NP	Doug removeu as manchas.	Agent V Theme
Semantic Predicates	(1) cause(Agent, E) (2) location(start(E), Theme, ?Source) (3) not(location(end(E), Theme, ?Source))	
NP V NP PP.source	Doug removeu as manchas da toalha.	Agent V Theme +src Source
Semantic Predicates	(1) cause(Agent, E); (2) location(start(E), Theme, Source); (3) not(location(end(E), Theme, Source))	

Table 5: The structure of ”Remove-10.1” class of VerbNet.Br

Banish-10.2

The class ”Banish-10.2” has three syntactic frames (as shown in Table 3). In the Stage 1, we translated directly all of these: ”NP V NP”, ”NP V NP *de* NP” and ”NP V NP *para* NP”. To be a member, a verb needed to present two of these syntactic frames (the roof of 1.8 (0.6*3)), respecting the MLE measure.

In the Stage 3, 35 verbs were defined for this class. Searching in the results of the Stage 2, we found 18 verbs that present at least one of the syntactic frames. However only the verbs *excluir* (to exclude) and *tirar* (to remove) present at least 2 syntactic frames that have a threshold higher than 0.05. Both presented the same syntactic frames: NP V NP and NP V NP *de* NP.

Therefore, the Portuguese version of the class ”Banish-10.2” has two verbs, *excluir* (to exclude) and *tirar* (to remove), and presents two syntactic frames: NP V NP and NP V NP *de* NP. Searching for sentences of the verb *excluir* (to exclude), we found two examples:

1. *[...] outras espécies [...] excluem as espécies responsáveis pela mudança.* (Other species exclude the species responsible for the change.)

The semantic of this class is ”removal of an entity, typically a person, from a location” (Levin, 1993). The sentence presented fits in this semantics, but we could not find an example of the alternation ”NP V NP *de* NP” with the second NP (Theme) being animate. We only find sentences that fit in the semantic of Remove-10.1 class. This class is shown in Table 6 (the ? means that the sentence seems to be incorrect, according to the corpus we have used).

Banish-10.2 - BR		
Thematic roles and Selectional restrictions: Agent [+animate — +organization], Theme, Source [+location] and Destination [+location — -region]		
Members: *excluir* (to exclude) and *tirar* (to remove)		
Frames:		
NP V NP	*O rei excluiu o general.*	Agent V Theme
Semantic Predicates	(1) cause(Agent, E); (2) location(start(E), Theme, ?Source); (3) location(end(E), Theme, ?Destination)	
NP V NP PP.source	*O rei excluiu o general do exército.*	Agent V Theme +src Source
Semantic Predicates	(1) cause(Agent, E); (2) location(start(E), Theme, Source); (3) not(location(end(E), Theme, Source))	
NP V NP PP.destination	?*O rei excluiu o general para a ilha.*	Agent V Theme para Destination
Semantic Predicates	(1) cause(Agent, E); (2) location(start(E), Theme, ?Source); (3) location(end(E), Theme, Destination)	

Table 6: The structure of ”Banish-10.2” class of VerbNet.Br

5. Conclusions and Future Work

We have presented a semi-automatic method for building the VerbNet.Br and some preliminary results with three classes. The classes presented were ”Equip-13.4.2”, ”Remove-10.1” and ”Banish-10.2”. The second and the third ones are related, since they have almost the same meaning and differ only in some diathesis alternations. The thematic roles, selectional restrictions and semantic predicates will be directly inherited from English. As the proposed method uses existing resources in one language to build a new resource in another language, it is cross-linguistic, that is, the method explores the compatibilities between English and Portuguese languages. However, we can observe that a linguistic revision of the results of this semi-automatic method is highly desirable. Therefore, we are looking for collaborators interested in validating this resource.

As future work, we intend to finish stages one and four and apply the method for all the remaining classes. We will also change the thresholds used to evaluate the precision and recall. Besides that, we will evaluate how many verbs are defined as candidate members (result of Stage 3) and how many verbs are selected (result of Stage 4). This will be achieved by calculating the ratio of selected verbs to candidate verbs.

We will also use a completely automatic method to group verbs, by using machine learning. This method will use clustering to group verbs according to subcatecategorization frames. We intend to compare the resulting classes of this automatic method with classes of our semi-automatic method proposed. We have the hypothesis that the semi-automatic method will present classes with more precision. However, the automatic method is expected to have a best recall.

Since we expect that the automatic method will present more verbs, we will try to include these verbs in VerbNet.Br classes and improve the resource, similarly to the task carried out by Kipper (2005).

6. Acknowledgements

We thank the Brazilian Science Foundation FAPESP for financial support.

7. References

S. M. Aluísio, G. M. Pinheiro, A. M. P. Manfrim, L. H. M. Genovês Jr., and S. E. O. Tagnin. 2004. The Lácio-web: Corpora and Tools to Advance Brazilian Portuguese Language Investigations and Computational Linguistic Tools. In *Proceedings of the 4th International Conference on Language Resources and Evaluation*, pages 1779–1782, Lisbon, Portugal.

C. F. Baker, C. J. Fillmore, and J. F. Lowe. 2005. The Berkeley Framenet Project. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics*, pages 86–90, University of Montréal, Canadá.

L. Bentivogli, E. Pianta, and C. Girardi. 2002. Multiwordnet: developing an aligned multilingual database. In *Proceedings of the First International Conference on Global WordNet Conference*, pages 293–302, Mysore, India.

A. Bertoldi and R. L. O. Chishman. 2009. Desafios para a Criação de um Léxico baseado em Frames para o Português: um estudo dos frames Judgment e Assessing. In *Proceedings of the The 7th Brazilian Symposium in Information and Human Language Technology*, São Carlos, SP, Brazil.

E. Bick. 2005. *The Parsing System*. Ph.D. thesis.

D. Croch and T. H. King. 2005. Unifying Lexical Resources. In *Proceedings of Interdisciplinary Workshop on the Identication and Representation of Verb Features and Verb*, pages 32–37, Saarbruecken, Germany.

B. C. Dias da Silva, A. Di Felippo, and M. G. V. Nunes. 2008. The Automatic Mapping of Princeton Wordnet lexical-conceptual relations onto the Brazilian Portuguese Wordnet database. In *Proceedings of the 6th International Conference on Language Resources and Evaluation*, pages 1535–1541, Marrakech, Morocco.

M. S. Duran and S. M. Aluisio. 2011. Propbank-br: a Brazilian Portuguese corpus annotated with semantic role labels. In *Proceedings of The 8th Brazilian Symposium in Information and Human Language Technology*, Cuiabá, MT, Brazil.

A. Di Felippo and B. C. Dias da Silva. 2007. Towards na automatic strategy for acquiring the Wordnet.br hierarchical relations. In *Proceedings of the 5th Workshop in Information and Human Language Technology, in conjunction with XXVII Congresso da Sociedade Brasileira de Computao*, pages 1717–1720, Rio de Janeiro, RJ, Brazil.

C. Fellbaum. 1998. *WordNet: An electronic lexical database*. MIT Press, Cambridge, MA.

E. E. Ferrer. 2004. Towards a semantic classification of spanish verbs based on subcategorisation information. In *Proceedings of the Workshop on Student research, in conjunction with ACL 2004*, pages 163–170, Barcelona, Spain.

R. Girju, D. Roth, and M. Sammons. 2005. Token-level Disambiguation of Verbnet Classes. In *Proceedings of Interdisciplinary Workshop on the Identification and Representation of Verb Features and Verb Classes*, Saarbruecken, Germany.

R. Jackendoff. 1990. *Semantic Structures*. MIT Press, Cambridge, MA.

E. Joanis and S. Stevenson. 2003. A general feature space for automatic verb classification. In *Proceedings of the 10th conference on European chapter of the Association for Computational Linguistics*, pages 163–170, Budapest, Hungria.

K. Kipper. 2005. *Verbnet: A broad coverage, comprehensive verb lexicon*. Doctor of philosophy, University of Pennsylvania.

B. Levin. 1993. *English Verb Classes and Alternation, A Preliminary Investigation*. The University of Chicago Press, Chicago, IL.

P. Merlo and S. Stevenson. 2001. Automatic verb classification based on statistical distributions of argument structure. *Computational Linguistics*, 27(3):373–408.

P. Merlo, S. Stevenson, V. Tsang, and G. Allaria. 2002. A multilingual paradigm for automatic verb classification. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 207–214, Philadelphia, PA.

C. Messiant. 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*, pages 55–60, Columbus,OH.

M. Palmer, D. Gildea, and P. Kingsbury. 2005. The Proposition Bank: A Corpus Annotated with Semantic Roles. *Computational Linguistics*, 31(1):71–106.

M. M. Salomao. 2009. Framenet Brasil: Um trabalho em progresso. *Revista Calidoscópio*, 7(3):171–182.

C. Scarton. 2011. Verbnet.br: construção semiautomática de um léxico computacional de verbos para o Português do Brasil. In *Proceedings of the The 8th Brazilian Sym-*

posium in Information and Human Language Technology, Cuiabá, MT, Brazil.

S. Schulte im Walde. 2006. Experiments on the Automatic Induction of German Semantic Verb Classes. *Computational Linguistics*, 32(2):159–194.

L. Shi and R. Mihalcea. 2005. Putting Pieces Together: Combining Framenet, Verbnet and Wordnet for Robust Semantic Parsing. In *Proceedings of 6th International Conference on Computational Linguistics and Intelligent Text Processing*, pages 99–110, Mexico City, Mexico.

L. Sun and A. Korhonen. 2011. Hierarchical Verb Clustering Using Graph Factorization. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 1023–1033, Edinburgh, UK.

L. Sun, A. Korhonen, and Y. Krymolowski. 2008. Verb class discovery from rich syntactic data. In *Proceedings of the 9th international conference on Computational linguistics and intelligent text processing*, pages 16–27, Haifa, Israel.

L. Sun, A. Korhonen, and Y. Krymolowski. 2009. Improving verb clustering with automatically acquired selectional preferences. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 638–647, Singapore.

L. Sun, A. Korhonen, T. Poibeau, and C. Messiant. 2010. Investigating the cross-linguistic potential of Verbnet-style classification. In *Proceedings of the 23rd International Conference on Computational Linguistics*, pages 1056–1064, Beijing, China.

R. Swier and S. Stevenson. 2004. Unsupervised Semantic Role Labelling. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 95–102, Barcelona, Spain.

P. Vossen. 2004. Eurowordnet: a multilingual database of autonomous and language specific wordnets connected via an interlingual-index. *International Journal of Linguistics*, 17.

A. Zanette, C. Scarton, and L. Zilio. 2012. Automatic extraction of subcategorization frames from corpora: an approach to Portuguese. In *Proceedings of the 2012 International Conference on Computational Processing of Portuguese - Demo Session*, Coimbra, Portugal.

A. Zanette. 2010. *Aquisiçao de Subcategorization Frames para Verbos da Língua Portuguesa.* Projeto de diplomação, Federal University of Rio Grande do Sul.