Construction of a VerbNet style lexicon for Vietnamese

HA My Linh
University of Science,
Vietnam National University
Hanoi, Vietnam
hamylinh@hus.edu.vn

LE Van Cuong
University of Social
Sciences and Humanities
Hanoi, Vietnam
cuongle.ussh@gmail.com

NGUYEN Thi Minh Huyen
University of Science,
Vietnam National University
Hanoi, Vietnam
huyenntm@hus.edu.vn

Abstract

Lexical resources like VerbNet (Kipper et al., 2006) or similar lexicons play an important role in the applications involving semantic understanding. For Vietnamese, the currently available computational lexicon (VCL) includes morpho-syntactic information for each lexical entry, subcategorization frames and also some semantic constraints for each verb. However, the information related to verb meaning and behaviors is still very far from complete. In this paper, we present our work on the construction of a VerbNet style lexicon for Vietnamese, called *viVerbNet*, in order to make available a fundamental lexical resource for semantic analysis of Vietnamese language. Each verb entry in viVerbNet is extracted from VCL, and enriched with various information acquired automatically from Vietnamese corpora as well as manually from a comparative investigation of the English VerbNet. At the current stage, we have built semantic components for 50 verb groups resulted from the application of a clustering algorithm on Vietnamese verbs.

1 Introduction

Lexicon are amongst the most important linguistic resources for natural language processing (NLP). Depending on what kind of applications for which the lexicon is developed, the annotation of lexical entries can be one of the most time-consuming and laborious tasks. For an application at a deep level like semantic understanding, a computational lexicon should ideally includes information related to the meanings and the behaviors of each word.

For English, there are several projects carried on lexical resources, e.g. WordNet (Miller et al., 1990), FrameNet (Baker et al., 1998), and more recently VerbNet (Kipper et al., 2006). These resources, together with semantically annotated corpora constitute important linguistic resources for developing language understanding applications. To obtain the semantic representation of a sentence, we usually need to identify the predicate and its arguments in that sentence. A predicate can be a verb, an adjective, or even a noun. The most important and complicated class of predicates is verb. For this reason, we are particularly interested in VerbNet, as it provides us with rich syntactic and semantic patterns of each verb, which proved very useful in semantic role labeling (Shi and Mihalcea, 2005), (Giuglea and Moschitti, 2006), (Loper et al., 2007), or in verb sense disambiguation (Brown et al., 2011), (Kawahara and Palmer, 2014).

In this paper, we present our work on the construction of a VerbNet style lexicon for Vietnamese, called *viVerbNet*, in order to make available a fundamental lexical resource for semantic analysis of Vietnamese language. Each verb entry in viVerbNet is extracted from VCL (Nguyen et al., 2006), and enriched with various information acquired automatically from Vietnamese corpora as well as manually from a comparative investigation of the English VerbNet. A clustering technique is applied to obtain classes of verbs sharing semantic and syntactic behaviors. For each verb class, we investigate their characteristics basing on annotated corpora as well as a comparative study of the corresponding English verb class.

The paper is structured in 4 main sections as follows. Section 2 presents the related works inspiring our project. Section 3 introduces the workflow for the construction of viVerbNet. Section 4 shows the application of a clustering technique for acquiring verb classes. Finally, Section 5 discusses the specifications of each verb.

2 Related Works

As mentioned above, a number of large and meaningful lexical resources have been built for semantic processing of English language.

- Wordnet (Miller et al., 1990) is a large English lexical resource in which words are organized into groups of synonym senses called synsets. Synsets are linked with each other by means of conceptual-semantic and lexical relations. WordNet is a dominant lexicon useful for sense resolution and semantic tagging.

- FrameNet (Baker et al., 1998) is a database containing more than 13,000 lexical units accompanied by their semantic frames. Over 200,000 manually annotated sentences with more than 1,200 semantic frames in FrameNet provide a training dataset for many applications such as machine translation, sentiment analysis, information extraction, etc.

In the following, we will present in more detail VerbNet, another important lexical resource in which verbs are fully syntactically and semantically annotated. VCL (Nguyen et al., 2006), the only Vietnamese large-scale computational lexicon, will be equally introduced as a foundation for building viVerbNet.

2.1 VerbNet

VerbNet (Kipper et al., 2006) is the largest English verb network, linking syntactic and semantic types of more than 5,200 verbs and 237 verb classes. This hierarchical verb vocabulary is mapped directly to other resources such as WordNet, FrameNet, and PropBank (Kingsbury and Palmer, 2002). Verb classes in VerbNet are designed based on Levin's verb classification (Levin, 1993). An example of a verb class is shown in Table 1.

Table 1: A class in VerbNet

Class Put-9.1	
Roles & Restrictions	Agent [+*animate*] Theme [+*concrete*] Destination [+*location & -region*]
Members	arrange, emplace, immerse, implant, lodge, ...
Frames:	
Description	NP V NP PP.destination
Example	I put the book on/under/near the table.
Syntax	Agent V Theme *{{+loc}}* Destination
Semantics	motion(*during(E), Theme*) not(*Prep(start(E), Theme, Destination)*) Prep(*end(E), Theme, Destination*) cause(*Agent, E*)

A verb class in VerbNet is defined by a set of members, the thematic roles and selectional restrictions of the arguments subcategorized by these members, as well as the syntactic and semantic descriptions related to their frames.

2.2 Vietnamese Computational Lexicon (VCL)

The Vietnamese Computational Lexicon (VCL) is the only large-scale lexical resource for fundamental tasks of NLP. VCL (Nguyen et al., 2006) contains about 42000 lexical entries, structured following the Lexical Mark-up Framework (LMF) - an abstract meta model from ISO TC 37/SC 4 (Francopoulo et al., 2006) that provides a framework for the development of NLP oriented lexicons. This lexicon includes all the information (word senses, part-of-speech, definition, examples) from one of the best Vietnamese print dictionaries (Hoàng, 2003). In addition, each entry is described in three aspects: morphology, syntax, and semantics. As Vietnamese words are morphologically invariable, the morphological information in VCL is only related to the word formation: a word can be either single, or compound, or redoubled, otherwise it can be a loan word, or an abbreviation, or a symbol. Table 2 shows an example of an entry in VCL, illustrating the informa-

tion not only at morphology level, but also at syntactic and semantic levels.

Table 2: A meaning of word "yêu" (*love*) in VCL

yêu (love)		
Morp	*simple word*	
Syntactic	Category	*V*
	Subcategory	*Vt*
	FrameSet	*Sub+V+Dob*
	Before	*R: rất (very)*
Semantic	Logical constraint	Categorial Meaning: *Emotion*
		Antonym: *Ghét (hate)*
	Semantic constraint	Sub: *Agt{Person}*
Definition	có tình cảm dễ chịu khi tiếp xúc với một đối tượng nào đó, muốn gần gũi và thường sẵn sàng vì đối tượng đó mà hết lòng	
Example	tôi yêu mẹ (*I love mom*)	

VCL is a very useful lexical resource for the fundamental NLP tasks. Its design allows easy update and extension, as well as a good exchangeability with other languages. Regarding the verbs, VCL still has some limitations in comparison with VerbNet as presented below.

- VCL contains 6652 verbs (8689 senses) and a total of 20 subcategorization frames associated to these verbs. But this information is far from complete: Most verbs are only attached to one frame, and the information about each frame is usually incomplete syntactically and semantically.
- VCL makes use of a set of 16 semantic roles such as: Agent, Experiencer, Possessor, *etc*. This set is quite limited compared to about 30 semantic roles in VerbNet.
- The semantic and logical constraints were manually built, however it remains several cases which have not been covered in the lexicon.

From these observations, we choose to build a VerbNet style lexicon for Vietnamese based on the verb entries available in VCL, in enriching them with other sources of information. This new lexicon is called viVerbNet.

3 Building viVerbNet

In order to acquire an equivalent resource to VerbNet for Vietnamese, we noted the need to revise and gather additional information such as thematic roles, selectional and syntax restrictions, syntactic frame, and semantic predicate for the verbs present in VCL.

As examples, we inspected in detail the thematic roles and components of a transitive verb (viết - *write*), an intransitive verb (đi - *go*), and an emotional verb (yêu - *love*) from the VCL and compared them to their translations in VerbNet. Some observations are made as follows.

- The semantic roles used in VCL are not equivalent to these in VerbNet.
  - VCL uses the semantic role *Content*, which is specialized into more concrete roles in Verbnet: topic, cause, goal, *etc*.
  - In many cases, the semantic roles are not defined in a similar way within the same context. For example, in VCL the subject argument of the verb "yêu (*love*)" is specified as *Agent* {Person}, while in VerbNet it is labeled as *Experiencer*.
- The selectional restrictions for semantic roles in VCL are quite incomplete.

Consequently, from VCL we cannot reconstitute the verb classes and the accompanied descriptions comparable to these in VerbNet. To build viVerbNet, we need to find a way for classifying verbs into groups of verbs having similar behaviors and describe these behaviors of each class. A clustering method applied on a large corpus can be useful for identifying the verb classes. In addition, we should revise the definition of the semantic roles and the selectional restrictions in VCL, insuring their compatibility with the same concepts in VerbNet.

Annotated corpora are equally important resources for extracting the specifications of each verb class:

- Viettreebank (Nguyen et al., 2009) is a constituency treebank with over 10,000 sentences. Subcategorization frames of several verbs can be extracted from this corpus. More detailed information can be equally extracted from a subset of this corpus: the Vietnamese dependency treebank (Nguyen et al., 2013).

- The Vietnamese Propbank (Ha et al., 2015) contains over 5000 sentences from VietTreeBank with labeled semantic roles compatible with the English Propbank. The semantic role labels in this corpus can be used to specify themantic roles for verbs in viVerbNet.

Beside the information acquired from corpora, an investigation of similar verb classes in the English VerbNet helps to determine comparable specifications of Vietnamese verb classes.

Figure 1 shows the summary of our workflow for constructing viVerbNet. In the next sections, we will describe in detail about the clustering of Vietnamese verbs and the specifications of the syntactic and semantic components for each verb group.

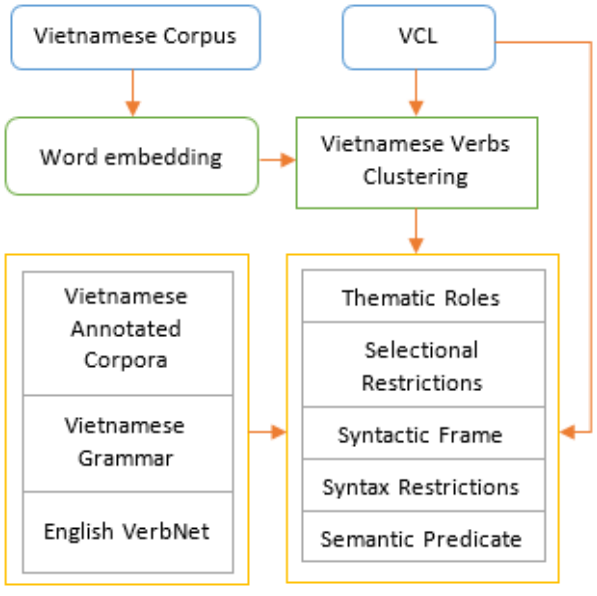

Figure 1: viVerbNet construction

4 Clustering Vietnamese Verbs

4.1 Clustering Method

In order to identify verb classes for Vietnamese, we first extract all the verbs from VCL, then apply an automatic clustering method on these verbs based on a large Vietnamese corpus.

In this paper, we use a hierarchical clustering algorithm (HCA) (Day and Edelsbrunner, 1984) to cluster 6652 verb entries extracted from the VCL dictionary. For any clustering algorithm, the most important questions are the data representation and the similarity measure. In our case, each verb is represented by a vector generated by word embedding models and we use the cosine distance as similarity measure.

Word embedding models. We have experimented with two different word embedding models generated by the word2vec algorithm (Mikolov et al., 2013): one (called *word2vec1*) is trained[1] on our own corpus, the other (called *word2vec2*) is pretrained in (Vu et al., 2019). In addition, we also experimented another pre-trained BERT word embeddings called PhoBERT (Nguyen and Nguyen, 2020), which is currently the best language model for Vietnamese in several problems like POS tagging, dependency parsing, and named-entity recognition.

4.2 Experimentation Results

From VCL, we have extracted 6652 verbs with a total of 8689 senses. The *word2vec1* model, trained on a 379 MB Vietnamese word-segmented corpus containing 99,531 articles collected from two news websites, covers 6039 verbs. The *word2vec2* model, trained on a 7.1 GB Vietnamese news corpus, covers 6388 verbs. PhoBERT, trained on a 20 GB corpus including 1 GB from Wikipedia, and 19 GB from various news websites, has been applied on our corpus to generate the word vectors of 6039 verbs.

Using the HCA algorithm on these three sets of verb embeddings, we get the best number of clusters given by the Silhouettes measure (Rousseeuw, 1987) for our data which is 283. However, for a finer clustering of verbs, we choose 1000 as the number of clusters. As the word embeddings take in account the context of each word, verbs in the same cluster

[1] using Gensim library `https://radimrehurek.com/gensim/models/word2vec.html`

have close meaning and similar behaviors and usage. Consequently these clusters can be served for defining verb classes for viVerbNet.

The experimentation show that two models word2vec1 and word2vec2 give similar results, which is explainable as the two models are trained with the same algorithm and the same style of news corpus, even if our corpus is much more smaller. In the meanwhile, the PhoBERT model gives more different results as it allows to distinguish different senses of a same verb. Some examples resulted from three word embedding models are given in Table 3. These example clusters contain verbs with very close meaning (for PhoBERT we show only the verbs present in the two other models).

Table 3: Some verb cluters

Verbs meaning "give-birth"		
word2vec1	**word2vec2**	**PhoBERT**
đẻ	đẻ	đẻ
sinh_nở	sinh_nở	sinh_sản
sinh_đẻ	sinh_đẻ	sinh
chuyển_dạ	chuyển_dạ	
thai_ngén	thai_ngén	
Verbs meaning "die"		
word2vec1	**word2vec2**	**PhoBERT**
bị_thương	bị_thương	đi
thiệt_mạng	thiệt_mạng	thiệt_mạng
mất_tích	mất_tích	
tử_nạn	tử_nạn	
chết	chết	chết
chết_đuối	chết_đuối	
bỏ_mạng	lâm_nạn	
mắc_kẹt	thương_vong	

A detailed evaluation of the results is being undertaken in investigating the behaviors of each verb in the available annotated corpora mentioned in Section 3. We equally proceed to study syntactic and semantic descriptions of verb classes, in choosing the clusters of verbs having the same meaning.

5 Verb Specification in viVerbNet

In this section, we present the components of VerbNet and focus on the discussion of specifications of Vietnamese language in comparison with English. This comparative study help us to build the components for viVerbNet in a compatible schema with the English VerbNet.

5.1 Thematic Roles

The semantic roles describe the basic semantic relationships between predicates and their arguments. We rely on 24 semantic roles from the Vietnamese Propbank to build thematic roles for viVerbNet. As this PropBank is designed in assuring the compatibility with the English ProbBank, the mapping from these 24 semantic roles to 39 themantic roles of VerbNet is facilitated.

The examples below show different syntactic distributions of the verb "`hoàn_thành`", belonging to a verb class comparable to the VerbNet verb class "*complete*". The contents inside the brackets following a word include the English translation and/or the role of that word. For all examples from now on, V stands for VERB.

- Active voice: `Tôi [`*I*`/Agent] đã [`*temporal marker*`] hoàn_thành [`*finish*`/V] bài_tập [`*exercise*`/Patient]` (*I have finished the exercises*).
- Passive voice: `Bài_tập [Patient] đã được [`*passive marker*`] tôi [Agent] hoàn_thành [V]` (*The exercises have been finished by me*).
- Passive voice without agent: `Bài_tập [Patient] đã hoàn_thành [V]` (*The exercises have been finished*).

We can see that Vietnamese has the same basic syntactic order Subject-Verb-Object as English in active voice. Attention should be paid to the passive voice, where the grammatical calque can produce the passive sentence "`Bài_tập [Patient] được [`*passive marker*`] hoàn_thành[V] bởi [`*by*`] tôi[Agent]`", but it sounds unnatural and is rarely used in good practice. In the case of passive voice

without agent, the passive marker can be absent as shown in the example. For this reason, the category and position of a word in a sentence are not enough for identifying its semantic role. That proves the importance of the selectional restrictions associated to each role.

5.2 Selectional Restrictions

Selectional restrictions determine semantic constraints on semantic roles. These restrictions indicate the existence (+) or absence (-) of semantic attributes such as [concrete], [animate], [organization], *etc*. Logical operators (| (OR) and & (AND)) are used to combine multiple restrictions.

For example, the selectional restrictions of the verb cluster "cấm, đình_chỉ, hoãn, nghiêm_cấm" corresponding to the VerbNet verb class "*forbid-64.4*" are as follows:

```
Agent[+animate|+organization]
Theme[]
Recipient[+animate|+organization].
```

For the sake of interoperability, we mapped 75 semantic classes used for semantic constraints to the set of 37 selectional restrictions in VerbNet.

5.3 Syntactic Frames and Restrictions

Each verb is associated to one or more syntactic frames. A syntactic frame briefly describes the surface structure of sentence constituents. It also specify semantic roles around verbs and syntax restrictions expressing the constraints on sentence constituents associated to these roles, such as plural, sentential, *etc* as illustrated in the following patterns:

1. `Agent V Patient<+plural>`
2. `Pivot V Theme <+np_to_inf>`
3. `Agent V Theme <+sc_ing>`
4. `Pivot V Theme <+ac_ing>`

The first pattern shows the restriction on the number of the patient role (plural), as in the sentence "*The merger associated the two companies*". In Vietnamese, the plural number is expressed by function words for plural markers like `những, các` or by numeral nouns:

company - `công ty`;
companies - `các công ty`;
two companies - `hai công ty`.

The second pattern covers this kind of sentence "*I needed him to go.*", while the third pattern corresponds to the syntactic structure in "*He rehearsed singing the song.*". The sentence "*I need him cooking.*" is an example of the fourth pattern.

Regarding two last patterns related to the gerund construction V_ing in English, it is worth to emphasize the phenomenon of nominalization in Vietnamese. In Vietnamese, we can observe two types of verb nominalization. The first type consists of the verb-noun categorical mutation, where a verb and its verbal noun have exactly the same word form. For example:

- `Tôi đã thỏa_thuận với anh ấy` (*I made deals with him*), where `thỏa_thuận` is a verb meaning *make deals*.
- `Anh ấy và tôi có hai thỏa_thuận` (N) (*He and I have two deals*), where `thỏa_thuận` is a noun.

The second type of verb nominalization consists of adding a function word like "`sự`", "`việc`", meaning "*the fact of*" or a classifier noun like "`cái`", "`kẻ`" in front of that verb. For examples:

- `Kinh_tế nước_nhà phát_triển mạnh` (*The country's economy has developed strongly*), where `phát_triển` is a verb;
- `Sự phát_triển của kinh_tế đã mang lại một bộ_ mặt mới cho đất_nước` (*The development of economy has brought a new face to the country*), where `sự phát_triển` is equivalent to a noun.

In Viettreebank, verb nominalization with classifiers is frequently observed. More than 200 occurences of the pattern "`<classifier> + Verb`" can be found, for example "`cái ăn`" (literal translation <*classifier*> + *to eat*, i.e. *the food*), "`người đọc`" (literal translation <*classifier*> + *to read*, i.e. *reader*), *etc*.

All these specialities of Vietnamese language have been taken in account when we describe the syntactic frames and restrictions in viVerbNet.

While building viVerbNet, we use the same representation format of syntax as VerbNet. Allowed

prepositions in the syntax description are put between curly brackets. The following shows usage example and descriptions of a syntactic frame of the verb "đi" in the sense "*Move from one place to another*" (*) (Hoàng, 2003). This verb entry belongs to the verb cluster "đi, chạy, xuôi" that can be mapped to a subclass of the verb class *"attend"*.

EXAMPLE
Tôi đi chợ (*I go to market*)
DESCRIPTION
NP V Destination
SYNTAX
Agent V Destination
SEMANTICS
Motion(During(E), Theme) Location(End(E), Destination)

The SEMANTICS component is introduced in the next section.

5.4 Semantic Predicate

Each syntactic frame is associated to a conjunction of semantic predicates such as cause, manner, contact, *etc*. Each semantic predicate represents the relationship between participants and events to indicate the core meaning of the sentence.

Several predicates are used for describing different stages in the process of an event: the preparatory (Start (E)), the culmination (During (E)), and the consequent (End (E)) stages of an event. This clear representation helps fully describe the core semantic components as well as changes in complex event structures.

Operators can also be added in semantic predicate construction such as negation (NOT) and the absence (?) of certain roles in the described structure .

For example, here is the semantic predicate for "confine" verb class:

- Not (Location (Start(E), Theme, ?Destination)) Location (End(E), Theme, ?Destination) Confine (Result(E), Theme) Cause(Agent, E)

For the semantic component in viVerbNet, we use the same set of semantic predicates as VerbNet.

6 Conclusions

In this paper, we have presented the ongoing project on the construction of viVerbNet, an English VerbNet style lexicon for Vietnamese. We proposed to implement a clustering algorithm for grouping Vietnamese verbs extracted from the available Vietnamese computational lexicon in similar classes. We focused first on describing a small number of major verb classes, before continuing to for similar verb classes.

At the current stage, we have studied 50 verb classes amongst 1000 obtained clusters, in doing an comparative investigation of these classes with English verb classes with similar meanings. Annotated corpora are equally explored for extracting syntactic and semantic information of each verb entry.

The built viVerbNet is designed in a way to be compatible with the English VerbNet. The resources will be freely available for research purposes. We are developing a platform for a collaborative revision of this verb lexicon. In addition, we plan to develop syntactic frames for adjective and noun predicates as well.

viVerbNet will be an important linguistic resources that can be applied in several problems such as semantic role labeling, deep semantic parsing, or question answering for Vietnamese.

Acknowledgments

HA My Linh was funded by Vingroup Joint Stock Company and supported by the Domestic Master/ PhD Scholarship Programme of Vingroup Innovation Foundation (VINIF), Vingroup Big Data Institute (VINBIGDATA), code VINIF.2020.TS.20.

References

Collin F. Baker, Charles J. Fillmore, and John B. Lowe. 1998. The Berkeley FrameNet project. In *Proceedings of 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, Volume 1*, pages 86–90, Montreal, Quebec, Canada, August. Association for Computational Linguistics.

Susan Brown, Dmitriy Dligach, and Martha Palmer. 2011. Verbnet class assignment as a wsd task. volume 47, pages 85–94, 01.

William H. E. Day and Herbert Edelsbrunner. 1984. Efficient algorithms for agglomerative hierarchical clustering methods. *Journal of Classification*, 1:7–24.

Gil Francopoulo, Monte George, Nicoletta Calzolari, Monica Monachini, Nuria Bel, Mandy Pet, and Clau-

dia Soria. 2006. Lexical markup framework (LMF). In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, Genoa, Italy, May. European Language Resources Association (ELRA).

Ana-Maria Giuglea and Alessandro Moschitti. 2006. Semantic role labeling via FrameNet, VerbNet and PropBank. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 929–936, Sydney, Australia, July. Association for Computational Linguistics.

My Linh Ha, Thi Luong Nguyen, Viet Hung Nguyen, Thi Minh Huyen Nguyen, Hong Phuong Le, and Thi Hue Phan. 2015. Building a semantic role annotated corpus for vietnamese. In *Proceedings of the National Symposium on Research, Development and Application of Information and Communication Technology*, pages 409–414.

Phê Hoàng. 2003. *Từ điển tiếng Việt*. Nhà xuất bản Đà Nẵng, Việt Nam.

Daisuke Kawahara and Martha Palmer. 2014. Single classifier approach for verb sense disambiguation based on generalized features. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 4210–4213, Reykjavik, Iceland, May. European Language Resources Association (ELRA).

Paul Kingsbury and Martha Palmer. 2002. From TreeBank to PropBank. In *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC'02)*, Las Palmas, Canary Islands - Spain, May. European Language Resources Association (ELRA).

Karin Kipper, Anna Korhonen, Neville Ryant, and Martha Palmer. 2006. Extending VerbNet with novel verb classes. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, Genoa, Italy, May. European Language Resources Association (ELRA).

Beth Levin. 1993. *English verb classes and alternations : a preliminary investigation*.

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

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. In Yoshua Bengio and Yann LeCun, editors, *Proceedings of 1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2-4, 2013*.

George A. Miller, Richard Beckwith, Christiane Fellbaum, Derek Gross, and Katherine J. Miller. 1990. Introduction to WordNet: An On-line Lexical Database*. *International Journal of Lexicography*, 3(4):235–244, 12.

Dat Quoc Nguyen and Anh Tuan Nguyen. 2020. PhoBERT: Pre-trained language models for Vietnamese. *arXiv preprint*, arXiv:2003.00744.

Thi Minh Huyen Nguyen, Laurent Romary, Mathias Rossignol, and Xuan Luong Vu. 2006. A lexicon for vietnamese language processing. *Language Resources and Evaluation*, 40:291–309, 12.

Phuong Thai Nguyen, Luong Vu Xuan, Thi Minh Huyen Nguyen, Van Hiep Nguyen, and Phuong Le-Hong. 2009. Building a large syntactically-annotated corpus of Vietnamese. In *Proceedings of the 3rd Linguistic Annotation Workshop, ACL-IJCNLP*, Singapore.

Thi Luong Nguyen, My Linh Ha, Viet Hung Nguyen, Thi Minh Huyen Nguyen, and Hong Phuong Le. 2013. Building a treebank for vietnamese dependency parsing. *Proceedings of the 2013 RIVF International Conference on Computing & Communication Technologies - Research, Innovation, and Vision for Future (RIVF)*, pages 147–151.

Peter Rousseeuw. 1987. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math.*, 20(1):53–65, November.

Lei Shi and Rada Mihalcea. 2005. Putting pieces together: Combining framenet, verbnet and wordnet for robust semantic parsing. volume 3406, pages 100–111, 02.

Xuan-Son Vu, Thanh Vu, Son N. Tran, and Lili Jiang. 2019. Etnlp: A visual-aided systematic approach to select pre-trained embeddings for a downstream task. *Proceedings of the International Conference Recent Advances in Natural Language Processing (RANLP)*.