Challenges and Solutions for Consistent Annotation of Vietnamese Treebank

Quy T. Nguyen[1&2], Yusuke Miyao[1&2], Ha T.T. Le[3], Ngan L.T. Nguyen[4]
[1]The Graduate University for Advanced Studies (SOKENDAI), Japan
[2]National Institute of Informatics, Japan
[3]University of Social Sciences and Humanities, Vietnam
[4] University of Information Technology, Vietnam
quynt@nii.ac.jp, yusuke@nii.ac.jp, trucha.ussh@gmail.com, ngannlt@uit.edu.vn

Abstract

Treebanks are important resources for research in natural language processing, speech recognition, theoretical linguistics, etc. To strengthen the automatic processing of the Vietnamese language, a Vietnamese treebank has been built. However, the quality of this treebank is not satisfactory and is a possible source for the low performance of Vietnamese language processing. We have been building a new treebank for Vietnamese with about 40,000 sentences annotated with three layers: word segmentation, part-of-speech tagging, and bracketing. In this paper, we describe several challenges of Vietnamese language and how we solve them in developing annotation guidelines. We also present our methods to improve the quality of the annotation guidelines and ensure annotation accuracy and consistency. Experiment results show that inter-annotator agreement ratios and accuracy are higher than 90% which is satisfactory.

Keywords: Vietnamese Treebank, Consistent Annotation, Challenges and Solutions

1. Introduction

Treebanks–corpora annotated with syntactic structures, are important resources for researchers in natural language processing (NLP). Treebanks provide important syntactic information in order to improve the quality of NLP tools. To strengthen the automatic processing of the Vietnamese language, Nguyen et al. (2009) have built a Vietnamese treebank, named VLSP treebank, containing 10,000 sentences. However, the quality of the VLSP treebank, including the quality of the annotation scheme, the annotation guidelines, and the annotation process, is not satisfactory and is a possible source for the low performance of Vietnamese language processing (Nguyen et al., 2012; Nguyen et al., 2013).

We have been building a new Vietnamese treebank with 3,000 texts (about 40,000 sentences) covering 14 topics collected from a Vietnamese online newspaper, Thanhnien news[1]. Our treebank is annotated with three layers: word segmentation (WS), part-of-speech (POS) tagging, and bracketing as showed in Figure 1[2]. We have found that ensuring the annotation consistency and accuracy is one of the most important considerations in the annotation of a treebank. This requires clear and complete annotation guidelines. The guidelines contain the annotation scheme, consistent principles to annotate linguistic phenomena, and sufficient examples. These documents are not only used to train annotators but also valuable sources serving the uses of the treebank.

We prepared three set of guidelines for the Vietnamese treebank: WS guidelines, POS tagging guidelines, and bracketing guidelines. In this paper, Section 2 describes the general characteristics of the Vietnamese language in comparison with other languages (e.g., English and Chinese) to indicate that building a high-quality Vietnamese treebank is a challenging problem. We also present our methodology to tackle the challenges in this section. We then discuss difficulties in WS, POS tagging, and bracketing, and how we solve them in developing the annotation guideline in Section 3, 4, and 5 respectively. Finally, in Section 6, we describe our annotation process, how we revise the guidelines during the annotation process, and methods to ensure the annotation consistency and accuracy.

This study is not only beneficial for the development of computational processing technologies for Vietnamese, a language spoken by over 90 million people, but also for similar languages such as Thai, Laos, and so on. This study also promotes the computational linguistic studies on how to transfer methods developed for a popular language, like English, to a language that has not yet intensively studied.

Original sentence:

Nam kể về tai nạn hôm qua.
{Nam tells about the yesterday's accident.}

1. **Word segmentation:**

Nam kể$_{\text{to tell}}$ *về*$_{\text{about}}$ *tai_nạn*$_{\text{accident}}$ *hôm_qua*$_{\text{yesterday}}$.

2. **POS tagging:**

Nam/Nr kể/Vv về/Cs tai_nạn/Nn hôm_qua/Nt ./PU

3. **Bracketing:**

```
(S
    (NP-SBJ (Nr-H Nam))
    (VP (Vv-H kể)
        (PP-DOB (Cs-H về)
            (NP (Nn-H tai_nạn)
                (NP-TMP (Nt-H hôm_qua)))))
    (PU .))
```

Figure 1: An example to illustrate process of treeing a Vietnamese sentence.

[1]http://thanhnien.vn

[2]Underscore "_" is used to link syllables of Vietnamese multi-syllable words. Translation for the Vietnamese word is given as a subscript. If the Vietnamese word does not have a translatable meaning, the subscript is blank. Translation for a Vietnamese sentence is given in curly brackets below the original text.

Meaning: The construction unit is too slow.

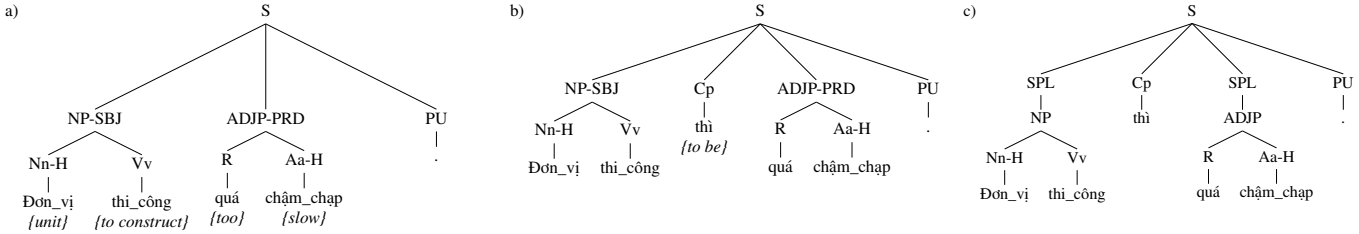


Figure 2: Examples showing ambiguity of annotating a sentence in Vietnamese.

2. Characteristics of Vietnamese language and methodology for guideline preparation

Unlike Western languages, in which blank spaces denote word delimiters, in Vietnamese, blank spaces play the roles of not only word delimiters but also syllable delimiters (Diep, 2005; SCSSV, 1983) that cause difficulties in defining words. In addition, unlike English and Japanese, Vietnamese is not an inflectional language for which morphological forms can provide useful clues for word segmentation and POS tagging. While similar problems also occur with Chinese (Xia et al., 2000), annotating Vietnamese words may be more difficult, because the modern Vietnamese writing system is based on Latin characters, which represent the pronunciation but not the meaning of words, resulting in many homonyms.

Difficulties in Vietnamese occur in not only determining words as mentioned above but also bracketing phrases. One of the reasons is that there are many expressions having the same POS sequence but different phrase types in Vietnamese. Other difficulties are caused by the fact that word order in Vietnamese is very flexible.

Moreover, there is little consensus in community about how to define words, phrases and grammatical structures. Though people agree that Vietnamese is the subject-verb-object (SVO) language, Figure 2a shows a sentence in Vietnamese that the head word of the predicate is not a verb. For sentences that do not have the main verb, we can use the conjunction *thì* to link the subject and the predicate as shown in Figure 2b. However, when the conjunction *thì* is used, linguists disagree about how to bracket this sentence. Diep (2005) considered this sentence as a single sentence (Figure 2b), where the conjunction *thì* is used to link the subject and the predicate. SCSSV (1983), in contrast, considered this sentence as a subordinate compound sentence (Figure 2c) because they said that the conjunction *thì* is used to link two clauses of a subordinate compound sentence.

We prepared the guidelines for the Vietnamese treebank including three sets: word segmentation guidelines, POS tagging guidelines, and bracketing guidelines. The problems were tackled on the basis of the following approaches:

- We refer to Vietnamese grammar books (SCSSV, 1983; Diep, 2005) and discuss with our collaborators, who are Vietnamese linguistics experts, to solve the ambiguities and difficulties.
- We study the guidelines of Chinese Penn Treebank (Xia, 2000b; Xia, 2000a; Xue et al., 2000), English Penn Treebank (Santorini, 1990; Bies et al., 1995), and VLSP treebank (Nguyen et al., 2010b; Nguyen et al., 2010a; Nguyen et al., 2010c) and adapt them to our guidelines if possible.
- During the annotation process, annotators[3] are requested to discuss with us about the constructions that they cannot annotate or feel ambiguous. These constructions are important clues to revise the guidelines.
- We conduct nine rounds of measurement of inter-annotator agreement and accuracy, for which two annotators annotate the same data. The inconsistencies and annotation errors found in each round are important clues to improve annotation guidelines and to train annotators again.

Details of applying these approaches during the process of building the Vietnamese treebank are explained in the following sections.

3. Word segmentation guidelines

3.1. Challenges of word segmentation

Words are the most basic units of a treebank (Sciullo and Williams, 1987), and defining words is the first step in the annotation process. (Xia, 2000b; Xia, 2000a; Sornlertlamvanich et al., 1999). For languages like English, defining words is almost trivial, because the blank spaces denote word delimiters. However, it is a difficult problem in Vietnamese even for a native speaker. Although most linguists agree that the Vietnamese language has two types of words, single-syllable words (single words) and multi-syllable words (compound words), distinguishing between single and multi-syllable words involves much ambiguity.

The ambiguities of Vietnamese WS occur for the following reasons. First, in Vietnamese, blank spaces play the roles of not only word delimiters but also syllable delimiters. Second, there are no morphological marks to act as important clues to identify words. Third, the Vietnamese writing system is based on Latin characters, which represent the pronunciation but not the meaning of words. Expressions that have the same surface form but different word segmentation appear frequently in Vietnamese. Rows 1 and 2 in Table 1, for instance, show two different segmentation

[3] Our treebank is annotated by two annotators who are graduate linguistics students.

No.	Expression (A B)	Meaning	WS
1	quần$_{trousers}$ áo$_{shirt}$	clothes	a word
2	quần$_{trousers}$ áo$_{shirt}$	trousers and shirt	2 words
3	ăn$_{eat}$ nói$_{speak}$	to speak	a word
4	tìm$_{find}$ kiếm$_{find}$	to find	a word
5	nồi$_{pot}$ đồng$_{copper}$	copper pot	2 words
6	nồi$_{pot}$ bằng$_{by}$ đồng$_{copper}$	copper pot	3 words
7	đen$_{black}$ đúa	black	a word
8	cá$_{fish}$ heo$_{pig}$	dolphin	a word
9	cá$_{fish}$ lia_thia$_{betta fish}$	betta fish	2 words
10	nghiên_cứu$_{research}$ viên$_{-er}$	researcher	2 words
11	nhà$_{-er}$ nghiên_cứu$_{research}$	researcher	2 words

Table 1: Examples to illustrate the principles of word segmentation.

types of the expression *quần áo*. Fourth, there is little consistency in segmenting the expressions. For example, some linguists consider the expression *cá$_{fish}$ rô$_{anabas}$ {anabas}* as a compound word but *bệnh$_{illness}$ sởi$_{measles}$ {measles}* as two words (Hoang, 1998; Diep, 2005). However, these expressions have a similar construction: the combination of a categorization noun[4] and a specific noun.

3.2. Policy for annotation of word segmentation

As mentioned above, our purpose for word segmentation is to build a treebank for Vietnamese. Therefore, we consider a word as the smallest syntactic unit having a complete meaning and preventing syntactic rules from analyzing word structure (Sciullo and Williams, 1987). On the basis of this word definition, we propose the following rules to solve the difficulties in Vietnamese word segmentation:

- If A and B[5] have different meanings and the meaning of the combination form (A_B) is different from the split form (A B), we select the form that has a meaning more appropriate for the context. Examples 1 and 2 in Table 1 show an expression having two different meanings because of different word segmentation.
- If A and B have different meanings and A_B has the same meaning as A or B, the combination form is selected. The example is given in row 3 of Table 1.
- If A and B have the same meaning, the combination form is selected (example 4 in Table 1).
- If another syllable can be inserted between A and B, we select the split form (examples 5 and 6 in Table 1).
- If A is a word and B is not (or vice versa), we select the combination form. Example 7 in Table 1 shows that if *đúa* is considered as a single word, its meaning is undefined. Therefore, it is considered as part of a multi-syllable word.
- For the expression of a categorization noun (A) and a specific noun (B), if B indicates something different from what the expression indicates, A_B is considered as a compound word. In contrast, if B has a similar meaning to A B, A and B are considered as two words (examples 8 and 9 in Table 1).
- An expression of one or more Sino-Vietnamese syllables and an original Vietnamese word, in which the Sino-Vietnamese syllables are the elements used to create the new words, is not considered as a word (example 10 in Table 1).
- Special classifier nouns are considered as single words (example 11 in Table 1).

It should be noted that these rules do not necessarily conform to the rules used by linguists. For example, Diep (2005) considers the Sino-Vietnamese syllable *viên$_{-er}$* in example 10 in Table 1 as a component of the compound word and considers the special classifier noun *nhà$_{-er}$* as a single word. We, on the other hand, consider both *viên$_{-er}$* and *nhà$_{-er}$* as single words because we found that they both have the same grammatical function that is forming new words. However, in our guidelines, the word types for which there is little consensus between linguists for segmenting them are annotated with additional information so that such words can be automatically converted according to the need.

4. Part-of-speech tagging guidelines

4.1. Challenges of POS tagging

Tagging POS for Vietnamese words is not a trivial problem because they are not marked with morphological features, such as tense, number, gender, etc. While the same problem also appears with Chinese, Vietnamese may be more difficult, because the Vietnamese writing system is based on Latin characters, which represent the pronunciation, but not the meaning of words.

Words that have the same surface form and pronunciation but different meanings and grammar functions occur frequently in the text. For example, we can understand the word *mới* in accordance with two meanings shown in rows 1 and 2 of Table 2. If we consider *mới* as an adjective modifying the preceding word, the noun *nghiên_cứu$_{research}$*, it means *new*; The word *mới* means *recently* or *just* if we consider it as an adjunct modifying the following word, the verb *thực_hiện$_{to\ conduct}$*.

Determining POS of the words having the same surface form may be more ambiguous because a verb or an adjective can appear in the position of a noun as in the case of *báo cáo* in rows 3 and 4 of Table 2. Solely referring to the sentence, we do not have any clue to determine if *báo cáo* belongs to the verb class or noun class. *Báo cáo* means *defend* if it is considered as a verb (row 3) and *thesis* if it is considered as a noun (row 4).

Ambiguity of the POS tagging is also caused by the omission of words which happens frequently in Vietnamese. For example, if a verb or an adjective plays the same roles as a noun, it is actually preceded by a special classifier noun

[4]Categorization nouns indicate general entities, such as *cá$_{fish}$* and *cây$_{tree}$*.

[5]Without loss of generalization, we assume the expression we want to segment is A B, where A and B can be syllables or words.

No.	Word in context	Word	POS
1	Một nghiên cứu mới thực hiện tại Nhật. *{A new reseach conducted in Japan.}*	mới$_{new}$	Adjective
2	Một nghiên cứu mới thực hiện tại Nhật. *{A research has just conducted in Japan.}*	mới$_{just}$	Adjunct
3	Báo cáo tốt nghiệp của cô ấy rất tốt. *{Her final defense is very good.}*	báo cáo *{defense}*	Verb
4	Báo cáo tốt nghiệp của cô ấy rất tốt. *{Her thesis is very good.}*	báo cáo *{thesis}*	Noun
5	Việc báo cáo tốt nghiệp của cô ấy rất tốt. *{Her final defense is very good.}*	việc báo cáo *{defense}*	Verb
6	Cuốn báo cáo tốt nghiệp của cô ấy rất tốt. *{Her thesis is very good.}*	cuốn báo cáo *{thesis}*	Noun
7	Bạn sẽ đẹp nhất đêm nay. *{You will be the most beautiful girl tonight.}*	sẽ$_{will}$	Adjunct
8	Tôi sẽ đi Nhật vào tối nay. *{I will go to Japan tonight.}*	sẽ$_{will}$	Adjunct

Table 2: Examples illustrating the challenges of POS tagging.

No.	POS tag	Meaning of tag	No.	POS tag	Meaning of tag
1	SV	Sino-Vietnamese syllable	17	NA	Noun-adjective
			18	Vcp	Comparative verb
2	Nc	Classifier noun	19	Vv	Other verb
3	Ncs	Special classifier noun	20	An	Ordinal number
4	Nu	Unit noun	21	Aa	Other adjective
5	Nun	Administrative unit noun	22	Pd	Demonstrative pronoun
6	Nw	Quantifier indicating the whole	23	Pp	Other pronoun
			24	R	Adjunct
7	Num	Number	25	Cs	Preposition or conjunction introducing a clause
8	Nq	Other quantifier			
9	Nr	Proper noun	26	Cp	Other conjunction
10	Nt	Noun of time	27	ON	Onomatopoeia
11	Nn	Other noun	28	ID	Idioms
12	Ve	Exitting verb	29	E	Exclamation word
13	Vc	Copula "là" verb	30	M	Modifier word
14	D	Directional verb	31	FW	Foreign word
15	VA	Verb-adjective	32	X	Unidentified word
16	VN	Verb-noun	33	PU	Punctuation

Table 3: POS tag set designed for our treebank.

(as the case of *báo cáo* in rows 5[6] of Table 2). Otherwise, a noun is preceded by a classifier noun[7] (the noun *báo cáo* in row 6 of Table 2 follows the classifier noun *cuốn*). However, such useful nouns are usually omitted in Vietnamese sentences which causes ambiguity of tagging words.

Some linguists (SCSSV, 1983; Diep, 2005) have claimed that POS can be recognized by referring to the adjuncts modifying the words. For example, adjuncts indicating degree and tenses modify adjectives and verbs, respectively. However, this method does not necessarily work sufficiently with real texts. In practice, many verbs and adjectives in Vietnamese can be modified by the same adjunct. For example, the adjunct indicating tense, *sẽ*$_{will}$ shown in Table 2 can modify both the adjective *đẹp*$_{beautiful}$ (row 7) and the verb *đi*$_{to\ go}$ (row 8).

Because of the above characteristics of Vietnamese, it is difficult not only to define the POS tag set but also to tag each word in context. In addition, there is still little consensus between linguists as to methodology for classifying words in Vietnamese. For instance, both Diep (2005) and SCSSV (1983) classified the words based on their meanings, their combination ability, and their syntactic functions. However, Diep (2005) considered the words expressing the whole, such as *cả*$_{all}$, *tất_cả*$_{all}$, *toàn_bộ*$_{all}$, etc. as pronouns, while SCSSV (1983), in contrast, considered them as nouns, and Hoang (1998) considered *cả* as a pronoun and *tất_cả* as a noun in all contexts.

4.2. Building part-of-speech tag set

In previous work, Nguyen et al. (2009) classified the words on the basis of their combination ability and syntactic function. They created a POS tag set for Vietnamese including a total of 17 tags (except the tags for unknown words and the punctuation). However, this tag set cannot cover all the combination abilities as well as the syntactic functions of the Vietnamese words. For example, they used the tag *P* to annotate all pronouns. However, the pronouns used to express space or time (demonstrative pronouns) such as *này*$_{this}$ and *đó*$_{that}$ can be modifiers of the head nouns in noun phrases. Personal pronouns, in contrast, always play the roles of the head words of noun phrases.

Therefore, in this work, we created a new POS tag set for Vietnamese. Our criteria to classify the words are also based on the combination abilities and the syntactic functions of the words, like those of the VLSP treebank. However, we referred to the linguistics literature, carefully analyzed the roles of words and discussed with our linguistics colleagues to create a new POS tag set for Vietnamese with 33 tags which are shown in Table 3. Using our POS tags, we can recognize the role of a word in a phrase or sentence. For example, the demonstrative pronouns modifying head words of noun phases are annotated with the *Pd* label, and personal pronouns that are head words of noun phrases are annotated with the *Pp* label.

4.3. Policy for annotation of part-of-speech

In our POS tagging guidelines, the words are tagged on the basis of the following criteria:

- Combination ability of the word. For example, *khó_khăn* can be understood as *difficulty* or *difficult*. However, if it is a noun, it cannot combine with the adjunct *rất*$_{very}$. If it is an adjective, it cannot combine with the quantifier *những*$_{-s/-es}$.

- Syntactic function of the word. For example, if the quantifier indicating the whole modifies a noun, it will be annotated with an Nw tag. The quantifier indicating the whole will be annotated with a Pp tag if it is head word of a noun phrase.

- Meaning of the word in the sentence. For example, the combination ability of the verb *đi*$_{to\ go}$ and the adjective *đẹp*$_{beautiful}$ mentioned above is the same, they are modified by the adjunct *sẽ*. They also have the same syntactic function which is head word of predicates. However, their meanings are different: the adjective expresses the quality, and the verb expresses the action.

[6] *Việc* is a special classifier noun that is understood as *-ion, -ment, -ing, -ity, -ness*, or so on when it comes before verbs or adjectives. An expression of the special classifier noun *việc* and a verb or adjective is understood as a noun in English. For example, *học_tập* means *to learn*, so to express *learning*, we can say *việc học_tập*.

[7] Classifier nouns indicate two types of things, animate things and inanimate things.

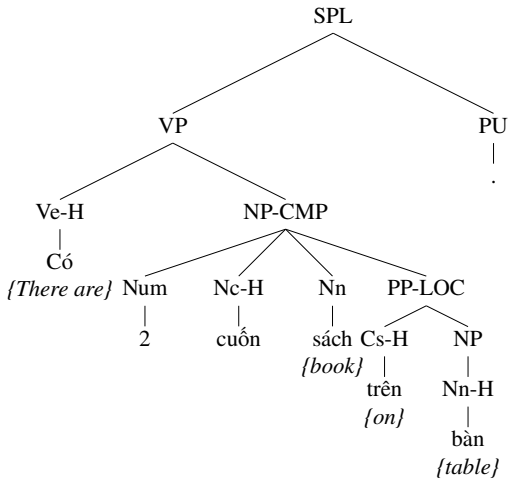


Figure 3: Example of bracketing a special sentence in Vietnamese.

No.	Pair of phrases	POS sequense	Example	Bracketing
1	NP and ADJP	Noun follows adjective	nhiều$_{a\ lot}$ kinh_nghiệm$_{experience}$ *{a lot of experiences}* nhiều$_{a\ lot}$ kinh_nghiệm$_{experience}$ *{experienced}*	NP ADJP
2	VP and ADJP	Verb follows adjetive	lầm_lũi$_{silently}$ sống$_{live}$ *{live silently}* ít$_{little}$ học$_{learn}$ *{unlearned}*	VP ADJP
3	S and NP	Adjective follows noun	lá$_{leaf}$ vàng$_{yellow}$ *{the leaf is yellow}* lá$_{leaf}$ vàng$_{yellow}$ *{the yellow leaf}*	S NP
4	S and NP	Verb follows noun	Chim$_{bird}$ hót$_{sing}$ *{The bird sings}* nguyên_tắc$_{principle}$ hoạt_động$_{to\ operate}$ *{operating principle}*	S NP

Table 4: Types of expressions that have the same POS sequence.

In addition, for each tag, the guidelines describe ambiguous cases and ways to distinguish among them. There are words that give us no clues to determine their POS if we only refer to single sentences as in the case of *báo_cáo* mentioned above. In these cases, the contexts of the words can be determined by referring to the surrounding text. Therefore, our annotation tool is designed to allow annotators to view the text to which the sentence belongs. For the words that give us no clues to determine their POS accurately, we decided to tag them on the basis of their combination ability, their syntactic function, or their meaning in the immediately-preceding phrase. For example, we tagged *mới* mentioned in Table 2 as an adjective based on its syntactic function in the phrase *một$_a$ nghiên_cứu$_{research}$ mới$_{new}$ {a new research}*.

In Vietnamese, several types of words are still little consensus on how to determine POS tags. For example, emotional verbs can be considered as adjectives, while some people said that they have two POSs. For these cases, we tagged them with double-POS tags so that they can be automatically changed to others.

5. Bracketing guidelines

5.1. Representation scheme

Our scheme is built on the basis of the VLSP treebank (Nguyen et al., 2009). We use the following four types of labels: constituency labels indicating syntactic categories of the phrases, functional labels indicating syntactic functions and meanings (if any) of the phrases, null elements to mark ellipses, and reference indices to mark syntactic movement. We also use the label H to tag the head words of the phrases. In addition, we refer to the scheme of English Penn Treebank, the scheme of Chinese Penn Treebank, and linguistics literature to complete the annotation scheme for Vietnamese. For example, Figure 3 shows a Vietnamese sentence that has only a verb phrase. This type of sentence was not distinguished from the sentences that have the standard structure[8] in the VLSP treebank. In our treebank, the sentences that do not have the standard structure will be bracketed with the label *SPL* so that we can distinguish them from the sentences that include a subject and a predicate, which are bracketed with the label *S*.

5.2. Policy for annotation of bracket

In this section, we will discuss two typical confusing cases of Vietnamese bracketing. The first case is to differentiate between the expressions that have the same POS sequence. We classify these expressions into four types shown in Table 4.

These ambiguities occur for the following two reasons.

1. In Vietnamese phrases, the lexical words modifying the head words commonly follow the head words. However, there are also the adjectives that can come before or follow the nouns and the verbs in the noun phrases and the verb phrases. This causes the ambiguities for recognizing whether a phrase in which an adjective comes before a verb is an adjective phrase or a verb phrase, and the phrase in which an adjective comes before a noun is an adjective phrase or a noun phrase, such as the phrases shown in rows 1 and 2 of Table 4.

2. The words are not marked with tense, number, case, etc. and they are expressed through the adjunct. However, the adjunct is dropped frequently in the text. This causes the ambiguities of distinguishing between the clauses and the phrases. Row 3 of Table 4 shows two ambiguities of distinguishing between sentences and phrases.

To solve the above ambiguities, we propose the following principles:

- For a noun phrase and an adjective phrase that have the same structure, if the phrase modifies a verb about quantity, it is bracketed with an NP (example 1 in Table 5). Conversely, if the phrase modifies a noun about quality or is the predicate of the sentence, the phrase is bracketed with an ADJP (example 2 in Table 5).

- For a verb phrase and an adjective phrase that have the same structure, if the words can be inverted without changing the meaning, the phrase is annotated with a VP label (examples 3 in Table 5). Otherwise, it will

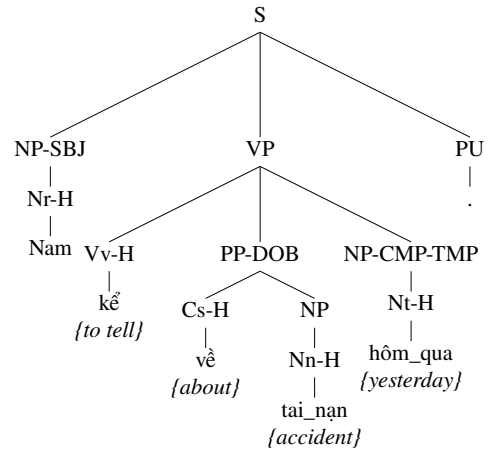
[8] A single sentence that has the standard structure has two main components: subject and predicate.

No.	Ambiguity	Expression in context	Expression	Bracketing	Reason of bracketing
1	NP or ADJP	Tôi có nhiều kinh nghiệm. *{I have a lot of experiences.}*	nhiều$_{a\ lot}$ kinh_nghiệm$_{experience}$ *{a lot of experiences}*	NP	Phrase *nhiều kinh_nghiệm* modifies the verb *có*$_{have}$ about quantity.
2	NP or ADJP	Tôi là người nhiều kinh nghiệm. *{I am an experienced person.}*	nhiều$_{a\ lot}$ kinh_nghiệm$_{experience}$ *{experienced}*	ADJP	Phrase *nhiều kinh_nghiệm* modifies the noun *người*$_{person}$ about quality.
3	VP or ADJP	Anh ấy lầm lũi sống. *{He lives silently.}* Anh ấy sống lầm lũi. *{He lives silently.}*	lầm_lũi$_{silently}$ sống$_{live}$ *{lives silently}* sống$_{live}$ lầm_lũi$_{silently}$ *{lives silently}*	VP VP	Inverting the adjective *lầm_lũi*$_{silently}$ and the verb *sống*$_{live}$ does not cause meaning change.
4	VP or ADJP	Tôi học ít. *{I learn little.}* Tôi ít học. *{I am unlearned}*	học$_{learn}$ ít$_{little}$ *{learn little}* ít$_{little}$ học$_{learn}$ *{unlearned}*	VP ADJP	Inverting the adjective *ít*$_{little}$ and the verb *học*$_{learn}$ causes meaning change.
5	S or NP	Cây này lá vàng. *{This tree's leaves are yellow.}* Cây này lá đã vàng. *{This tree's leaves have been yellow.}*	lá$_{leaf}$ vàng$_{yellow}$ *{the leaf is yellow}*	S	We can add the adjunct indicating past tense *đã* as a modifier of the adjective *vàng*$_{yellow}$.
6	S or NP	Cây này có lá vàng. *{This tree has the yellow leaves.}*	lá$_{leaf}$ vàng$_{yellow}$ *{the yellow leaf}*	NP	We cannot add the adjunct indicating tense as a modifier of the adjective *vàng*$_{yellow}$.
7	S or NP		Chim$_{bird}$ hót$_{sing}$ *{The bird sings}* nguyên_tắc$_{principle}$ hoạt_động$_{tooperate}$ *{operating principle}*	S NP	The bird can sing. However, the principle cannot operate.

Table 5: Examples of bracketing the expressions that have the same POS sequence.

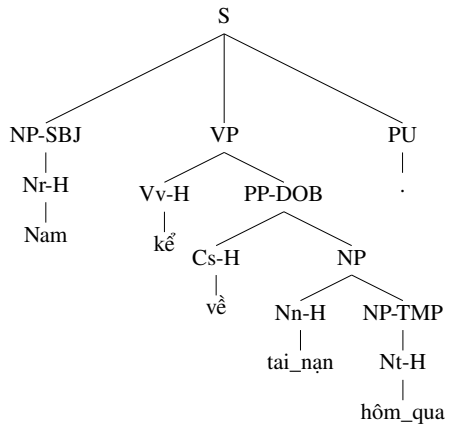
a) Original text:
Nam kể về tai nạn hôm qua.

b) *hôm_qua*$_{yesterday}$ is bracketed as a post-modifier of the verb *kể*$_{to\ tell}$



{Nam told about the accident yesterday.}

c) *hôm_qua*$_{yesterday}$ is bracketed as a post-modifier of the noun *tai_nạn*$_{accident}$



{Nam tells about the yesterday's accident.}

Figure 4: Example for the confusion caused by various judgements of the phrase.

be bracketed with a VP label if the verb precedes the adjective and bracketed with a ADJP label if the verb follows the adjective (example 4 in Table 5).

- For a clause and a noun phrase in which the noun comes before the adjective (as mentioned in example 3 in Table 4), if we can insert the adjunct indicating tense as a pre-modifier of the adjective, the expression should be bracketed with an S label (example 5 in Table 5). In contrast, the expression will be bracketed as a noun phrase (example 6 in Table 5).

- For a clause and a noun phrase in which the noun comes before the verb (as mentioned in example 4 in Table 4), if the noun is not the subject of the action stated by the verb, the expression is bracketed with an NP label (example 7 in Table 5).

The second confusing case is annotation of the ambiguous sentences that can be bracketed with various structures. These ambiguities occur for the following reasons:

1. One phrase can be interpreted by different valid structures. Figure 4 is an example for this. In this example, we can understand *hôm_qua*$_{yesterday}$ as an adverb phrase modifying the verb *kể*$_{to\ tell}$ (Figure 4b) or a phrase modifying the noun *tai_nạn*$_{accident}$ (Figure 4c).

2. Ellipses occur frequently. For example, Diep (2005) considered the sentence in Figure 5 as a single sentence, where the expression before the comma is a subordinate component of the sentence that expresses the manner (Figure 5b). However, this sentence can be understood as a subordinate compound sentence (SCSSV, 1983) in which the subject of the first clause is dropped because it is the same as the subject of the second clause (Figure 5c).

3. Many words in Vietnamese were annotated with a double-POS tag, which caused ambiguities in selecting the constituent label to bracket them.

To disambiguate these cases, we refer to the context to find their actual meaning and structure. The cases in which there is no clue for disambiguation are bracketed as follows: (1) If

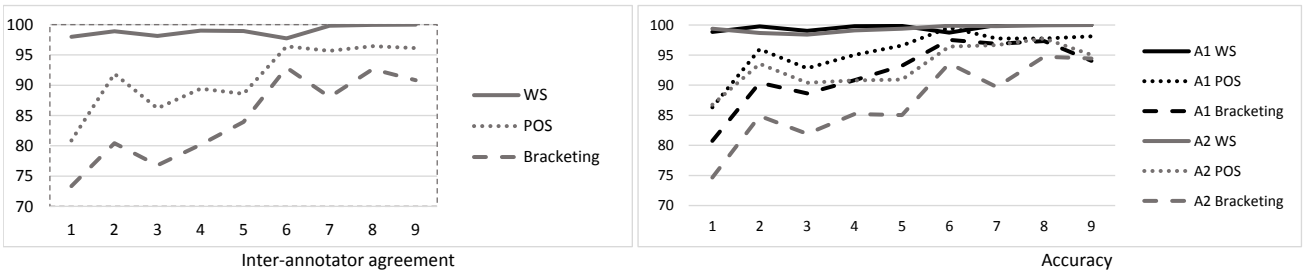
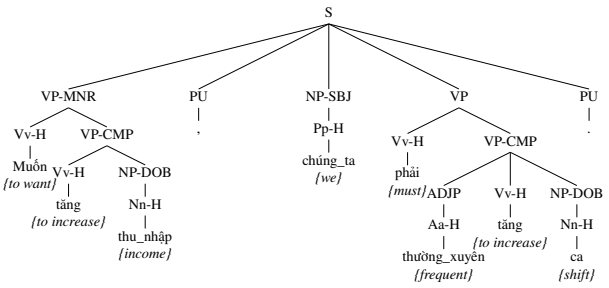


Figure 6: Results for nine rounds of measurement of inter-annotator agreement and accuracy.

a) Original text:
Muốn tăng thu nhập, chúng ta phải thường xuyên tăng ca.
{To increase the income, we must work overtime frequently.}

b) Bracketed as a single sentence



c) Bracketed as a compound sentence

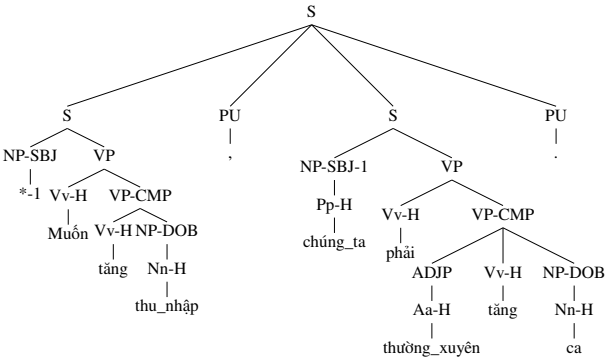


Figure 5: Example for confusion caused by ellipsis.

one phrase can be interpreted by different valid structures, the phrase will be bracketed with all valid structures; (2) For ambiguities caused by ellipses, we annotate each type of ellipsis in such a structure that maintains meaning of the sentence. For example, we bracket the sentence in Figure 5 as a single sentence (Figure 5b); (3) For ambiguities caused by the double-POS words, we also bracket each sentence with a unique structure. However, the sentences can be converted into other structures on the basis of the POS tags.

6. Annotation process and quality control

Although we tried to write the guidelines as completely as possible before the annotation process began, revising the guidelines during the annotation process is unavoidable because real text is far more complicated than the examples mentioned in the literature. Therefore, in this section, we will discuss our method to improve the quality of annotation guidelines and to ensure correct and consistent annotation.

After finishing the drafts of annotation guidelines, we trained two annotators and asked the annotators to annotate 600 texts (about 8,000 sentences) (preliminary annotation). In this annotation stage, the annotators were asked to discuss about the constructions which they found difficult to annotate because of ambiguities or other reasons. Based on these discussions. we revised the guidelines for the instructions that cannot be applied to new data and the constructions that are not covered by the guidelines. After revising the guidelines, we retrained annotators with the second version of the guidelines. Then, we carried out nine measurement rounds to calculate inter-annotator agreement scores and accuracies. Each round includes the following steps:

- We randomly select three texts (about 40 syntactic trees) from the results of the preliminary annotation;
- Each annotator re-annotates the texts independently;
- We compare the annotation results of each annotator to the benchmark data annotated by us and those of the other annotator;
- We discuss with annotators about the annotation errors and the inconsistencies, and revise the annotation guidelines (if necessary).

Figure 6 shows the inter-annotator agreement scores and the accuracies of three annotation layers. The left figure shows the agreement between two annotators; the right one shows the accuracy of each annotator (denoted by A1 and A2) compared to the benchmark data. This figure shows that from the sixth round, the agreement ratios and accuracies were higher than 90%, which indicates that the annotation is reliable.

After we finished the ninth measurement round, our annotators edited 600 texts. Then, the annotation results of each annotator was checked and edited by the other annotator. Finally, to clean up the corpus, we ran tools to detect annotation errors. These errors were manual edited by our annotators before our corpus is released.

Our observations on the inconsistent annotations and errors revealed that most of the inconsistencies were caused by the ambiguous expressions. There are three main reasons for the ambiguous expressions: (1) there is no infection in

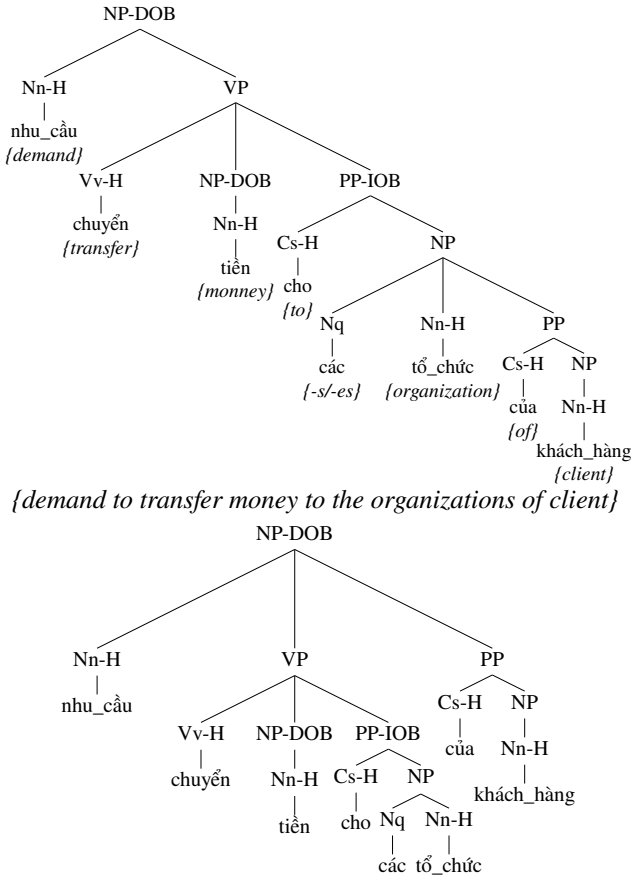


Figure 7: An inconsistent annotation between the two annotators.

Vietnamese; (2) word order is very flexible; (3) a sentence can have many meanings. Figure 7 shows an example that we can understand a sentence by two different meanings. Although our annotation guidelines contain many examples of ambiguous expressions as well as their correct annotations, real texts are complicated. Ambiguous expressions appear in various forms and difficult to recognize all structures that can be annotated. Therefore, to achieve a high agreement ratio, the annotators need to be trained carefully and to practice the annotation more on the basis of real texts so that they become familiar with annotation and analyzing the texts following the guidelines; the guidelines also need to be updated for new constructions throughout the annotation process.

7. Conclusion

We have solved the challenges in building a Vietnamese treebank, namely, developing WS guidelines, POS tagging guidelines, and bracketing guidelines, as well as ensuring the annotation consistency and accuracy. Our guidelines were developed based on not only the linguistics literature but also the analysis of the linguistic phenomena on real texts. Moreover, we discussed with linguistic experts to solve the difficulties. So far, we have annotated 600 texts. In future, we will annotate the rest of Vietnamese treebank, which includes 2,400 texts and revise the guidelines for new structures (if any). We plan to complete and publicize the annotated corpus and the annotation guidelines at the end of 2016.

8. Bibliographical References

Bies, A., Ferguson, M., Katz, K., MacIntyre, R., Tredinnick, V., Kim, G., Marcinkiewicz, M. A., and Schasberger, B. (1995). Bracketing guidelines for treebank ii style penn treebank project. *University of Pennsylvania*, 97:100.

Diep, B. Q. (2005). *Vietnamese grammar*. Vietnam Education Publisher.

Hoang, P. (1998). *Vietnamese Dictionary*. Scientific & Technical Publishing.

Nguyen, T. P., Vu, L. X., Nguyen, H. T. M., Nguyen, H. V., and Le, P. H. (2009). Building a large syntactically-annotated corpus of vietnamese. In *Proceedings of the Third Linguistic Annotation Workshop*, pages 182–185. Association for Computational Linguistics.

Nguyen, T. P., Vu, L. X., and Nguyen, H. T. M., (2010a). *Vietnamese part-of-speech tagging guidelines*. Ministry of Education and Training (Vietnam), Vietnam.

Nguyen, T. P., Vu, L. X., and Nguyen, H. T. M., (2010b). *Vietnamese word segmentation guidelines*. Ministry of Education and Training (Vietnam), Vietnam.

Nguyen, T. P., Vu, L. X., Nguyen, H. T. M., Dao, T. M., Dao, N. T. M., and Le, N. K., (2010c). *Vietnamese bracketing guidelines*. Ministry of Education and Training (Vietnam), Vietnam.

Nguyen, Q. T., Nguyen, N. L., and Miyao, Y. (2012). Comparing different criteria for vietnamese word segmentation. In *Proceedings of 3rd Workshop on South and Southeast Asian Natural Language Processing (SANLP)*, pages 53–68. Citeseer.

Nguyen, Q. T., Nguyen, N. L., and Miyao, Y. (2013). Utilizing state-of-the-art parsers to diagnose problems in treebank annotation for a less resourced language. In *Proceedings of the 7th Linguistic Annotation Workshop & Interoperability with Discourse*, pages 19–27. Association for Computational Linguistics.

Santorini, B. (1990). Part-of-speech tagging guidelines for the penn treebank project (3rd revision).

Sciullo, A.-M. D. and Williams, E. (1987). *On the definition of word*, volume 14. Springer.

SCSSV. (1983). *Vietnamese grammar*. Social Sciences Publishers.

Sornlertlamvanich, V., Takahashi, N., and Isahara, H. (1999). Building a thai part-of-speech tagged corpus (orchid). *Journal of the Acoustical Society of Japan (E)*, 20(3):189–198.

Xia, F., Palmer, M., Xue, N., Okurowski, M. E., Kovarik, J., Chiou, F.-D., Huang, S., Kroch, T., and Marcus, M. P. (2000). Developing guidelines and ensuring consistency for chinese text annotation. In *LREC*.

Xia, F. (2000a). The part-of-speech tagging guidelines for the penn chinese treebank (3.0).

Xia, F. (2000b). The segmentation guidelines for the penn chinese treebank (3.0).

Xue, N., Xia, F., Huang, S., and Kroch, A. (2000). The bracketing guidelines for the penn chinese treebank (3.0).