

Developing Tools and Building Linguistic Resources for Vietnamese Morpho-Syntactic Processing

Thanh Bon Nguyen ⁽¹⁾, Thi Minh Huyen Nguyen ⁽²⁾

Laurent Romary ⁽²⁾, Xuan Luong Vu ⁽³⁾

(1) IFI, Hanoi - ntbon@ifi.edu.vn

(2) LORIA, Nancy - nguyen@loria.fr, romary@loria.fr

(3) Vietnam Lexicography Centre, Hanoi - vuluong@vietlex.com

Abstract

Vietnamese is spoken by about 80 millions people around the world, yet very few concrete works on this language have been noticed in Natural Language Processing (NLP) until now. The fundamental problems in automatic analysis of Vietnamese, such as part-of-speech (POS) tagging, parsing, etc. are extremely difficult due to the lack of formal linguistic knowledge on one hand, and the specificities of isolating languages on the other hand. In this paper we present our efforts to develop a set of tools permitting the construction and management of language resources for Vietnamese in a normalized framework, whose aim is to be largely distributed and usable for research purposes in NLP. We first define a tagset by constructing Vietnamese morpho-syntactic descriptors that fit in a model compatible with MULTTEXT¹, so as to account for possible multilingual applications as well as the reusability of defined tagsets. We then implement a system undertaking the tasks of word segmentation and POS tagging. Our system ensures a representation format of linguistic resources that is currently considered in the framework of ISO TC37 SC4². Finally we attempt to construct a formal syntactic description of nominal groups using the Tree Adjoining Grammar (TAG) formalism.

Introduction

Vietnamese is spoken by about 80 millions people around the world, yet very few concrete works on this language have been noticed in Natural Language Processing (NLP) until now. Neither tools nor language resources are shared in public research. Except a few works carried out on English to Vietnamese, research in this domain has not until recently raised much attention amongst the scientific community in Vietnam.

Moreover, Vietnamese linguists are not involved in computational linguistics yet. The fundamental problems in automatic analysis of Vietnamese, such as part-of-speech (POS) tagging, parsing, etc. are extremely difficult due to the lack of formal linguistic knowledge. There does not even exist any recognizable standard for Vietnamese word categories (Cao X. Hào, 2000; Ủy ban KH XHVN, 1983). This actually comes from unclear limits between the grammatical roles of many words in Vietnamese. In general, only researches on main POS categories are found in Vietnamese linguistic literature, many tool words are not well described and classified.

Another major difficulty for the automatic processing of Vietnamese comes from the fact that it is an isolating language in which almost every word is monosyllabic and there is no morphological variation. All grammatical relations are determined by word order and tool words. Compound words, derivatives, expressions, etc. are composed from monosyllabic words and frequently appear in the texts. These specificities make the tasks of word segmentation and morpho-syntactic annotation of Vietnamese extremely difficult.

In this paper, we present our efforts to develop a set of tools permitting the construction and management of language resources for Vietnamese in a normalized framework (cf. Nancy Ide & Laurent Romary, 2001), whose purpose is to be largely distributed and usable for research purposes in NLP.

After a brief reminder of the linguistic characteristics of Vietnamese (section 2), we present our system for morpho-syntactic annotation (section 3). We define a tagset by constructing Vietnamese morpho-syntactic descriptors that fit in a model compatible with MULTTEXT, so as to account for possible multilingual applications as well as the reusability of defined tagsets. We also implement a system undertaking the tasks of word segmentation (semi-automatically) and POS tagging (with an editor to validate the results). Our system ensures a representation format of linguistic resources that is currently considered in the framework of ISO subcommittee TC37 SC4. This system also contains a concordancer helping the linguists study the grammatical usage of words found in Vietnamese corpora. The annotated corpus and the lexicon that we have produced are accessible from our team website³.

For the syntactic processing (section 4), we attempt to construct a formal syntactic description of noun phrases using the Tree Adjoining Grammar (TAG) formalism. This work is undertaken in the perspective of building a Vietnamese syntactic lexicon for a TAG parser.

Some Characteristics of Vietnamese

To begin with, we remind some important specificities of Vietnamese (Thanh Bon Nguyen et al., 2004).

Vocabulary

Vietnamese has a special unit called "tiếng" that corresponds at the same time to a syllable with respect to phonology, a morpheme with respect to morpho-syntax, and a word with respect to sentence constituent creation. For convenience, we call these "tiếng" syllables. The Vietnamese vocabulary contains:

- Simple words, which are monosyllabic.
- Reduplicated words composed by phonetic reduplication (e.g. trắng/white - trắng trắng / whitish).
- Compound words composed by semantic coordination (e.g. quần/trousers, áo/shirt - quần áo/clothes).
- Compound words composed by semantic subordination (e.g. xe/vehicle, đạp/pedal - xe đạp/bicycle).

- Some compound words whose syllable combination is no more recognizable (bồ nông/pelican).
- Complex words phonetically transcribed from foreign languages (cà phê/coffee).

Grammar

As with other isolating languages, the most important syntactic information source in Vietnamese is word order. The basic word order is Subject - Verb - Object. The other syntactic means are tool words, the reduplication, and the intonation.

Vietnamese belongs to the class of topic-prominent languages (Charles N. Li & Sandra A. Thompson, 1976). In these languages, topics are coded in the surface structure and they tend to control co-referentiality (cf. Cây đó lá to nên tôi không thích / Tree that leaves big so I not like, which means This tree, its leaves are big, so I don't like it); the topic-oriented "double subject" construction is a basic sentence type (cf. Tôi tên là Nam, sinh ở Hà Nội / I name be Nam, born in Hanoi, which means My name is Nam, I was born in Hanoi), while such subject-oriented constructions as the passive and "dummy" subject sentences are rare or non-existent (cf. There is a cat in the garden should be translated in Có một con mèo trong vườn / exist one <animal-classifier> cat in garden).

Corpus Morpho-Syntactic Annotation

Many applications in the domain of NLP make use of annotated corpora. Yet the construction of these corpora is expensive, so it would be very counter-productive for each research team to gather and prepare its resource from scratch. In Vietnam, works in language engineering have only very recently been motivated. As our project is one of the first projects in this domain, our objective is to make available in the research community a set of tools permitting the construction and management of language resources for Vietnamese in a normalized framework. In this section we focus on the system of morpho-syntactic annotation. In the next section we introduce a modelisation of Vietnamese syntax.

Lexical Resource

One of the necessary resources for language processing is lexical resource. For the fundamental tasks of Vietnamese text analysis, i.e. the text segmentation and the POS tagging, we need a list of all syllables in Vietnamese and a lexicon containing morpho-syntactic information. Our initial sources of information were a syllable base (about 6700 entries, provided by the Vietnam Lexicography Centre) and a print dictionary (Hoàng Phê, 2002), in which each headword is defined by descriptions of its possible meanings and the respective grammatical categories. The set of 8 categories used in this dictionary is a compromise accepted by the Vietnam Committee of Social Sciences (Ủy ban KHXHVN, 1983).

As described in (Thanh Bon Nguyen et al., 2004), we have built a morpho-syntactic lexicon containing all headwords in the above print dictionary. Each headword can correspond to several entries in our lexicon, upon to its number of morpho-syntactic descriptions. We make use of 11 main grammatical categories instead of 8 categories in the print dictionary, in maintaining the coherence with the descriptions in (Ủy ban KHXHVN, 1983). The subcategorisation of each category is described by a feature structure, of which each feature is chosen from different discussions on Vietnamese grammar in the literature. Convinced by multilingual application benefit, this description model is compatible with the MULTTEXT model for Western and Eastern European languages. Another important aspect that we pay much attention to is the lexicon representation, in such a way that it can be easily exploited and updated. A proposition for lexical markup normalization is being discussed in the framework of the ISO TC 37 SC 4 "Language Resource Management". For our morpho-syntactic lexicon, we choose for the time being a simple representation with explicit XML tags (cf. Thanh Bon Nguyen et al., 2004).

The morpho-syntactic descriptions elaborated in the lexicon give us the capacity to define tagsets with 1-1 mappings from the description space to the tag space. That makes it possible to compare or to reuse annotated corpora with different tagsets. We now present the developed tools for the task of morpho-syntactic annotation.

Tools for Annotated Corpora Building

Our system ensures the annotation of the corpus in two principal steps: the text tokenization and the POS tagging.

The tokens in question are lexical units that are supposed present in the system lexicon. As compound words are very frequent in Vietnamese, the tokenization cannot simply be obtained by segmenting the text using white spaces and punctuation as separation marks. The tokenizer is developed in two phases.

In the first phase, we make use of the syllable base and the lexicon to recognize all possible segmentations for each text segment (separated by the punctuation). The strategy to select the good solutions is to choose the segmentation having the smallest number of words. This idea, coming from empiric observation, works well enough. Ambiguity cases are solved manually (by the interface of the system).

In the second phase, equipped with tagged corpus, we replace the user intervention of the first tokenizer with the automatic choice using tag sequence probabilities. The considered tagset is the one comprising 11 main grammatical categories of the system lexicon. A manual correction is of course always necessary to achieve a perfect result.

The POS tagging task is ensured by a stochastic tool, as described in (Thi Minh Huyen Nguyen et al., 2003).

In the same way as the lexical resource building, all the produced resources of the system are marked up in a structure equivalent with the propositions considered by the ISO TC 37 SC 4. Here are a simple example of tokenized text and tagged text encoding.

Input text:

Ông già đi rất nhanh.

Intermediate tokenized text based on the syllable list (each token corresponds to a syllable or a punctuation):

```
<token id = "t1">Ông</token>
<token id = "t2">già</token>
<token id = "t3">đi</token>
<token id = "t4">rất</token>
<token id = "t5">nhanh</token>
<token id="t6">.</token>
```

Tokenized and tagged texts based on the lexicon (each token corresponds to an entry in the lexicon):

- Solution 1:

```
<wordForm entry = "Ông già" tokens = "t1 t2" tag = "pos@N" /> <!-- old man -->
<wordForm entry = "đi" tokens = "t3" tag = "pos@V" /> <!-- walk / die -->
<wordForm entry = "rất" tokens = "t4" tag = "pos@R" /> <!-- very -->
<wordForm entry = "nhanh" tokens = "t5" tag = "pos@A" /> <!-- quick -->
<wordForm entry = "." tokens = "t6" tag = "pos@dot" /> <!-- The old man walks/dies very quickly -->
```

- Solution 2:

```
<wordForm entry = "Ông" tokens = "t1" tag = "pos@P" /> <!-- you (man) -->
<wordForm entry = "già" tokens = "t2" tag = "pos@A" /> <!-- old -->
<wordForm entry = "đi" tokens = "t3" tag = "pos@R" /> <!-- grow -->
<wordForm entry = "rất" tokens = "t4" tag = "pos@R" /> <!-- very -->
<wordForm entry = "nhanh" tokens = "t5" tag = "pos@A" /> <!-- quick -->
<wordForm entry = "." tokens = "t6" tag = "pos@dot" /> <!-- You grow old very quickly -->
```

The system also provides interfaces to manipulate the tokenizing and the tagging results. In addition, a concordancer is available for different statistical analyses of tagged corpora.

TAG Description of Nominal Groups

As mentioned in the introduction, Vietnamese linguists have not been involved in computational linguistics yet. Therefore there does not exist any valid formal grammar for Vietnamese until now. The formalism that we choose in our project for Vietnamese parsing is Tree Adjoining Grammar (TAG), which is well studied for French and English grammars (Xtag, 2001; Abeillé,

2002). This choice is justified by two factors. Theoretically, the syntax/semantic interface is simpler in TAGs than in context free grammars, thanks to the extended locality domain provided by TAGs; however the worst case complexity for TAG parsing remains polynomial ($O(n^6)$). Empirically, the generic tools for TAG-based parsing system are ample (e.g. XTAG⁴, Dyalog⁵) and also well developed at LORIA (Crabbé et al., 2003). Furthermore, a normalized format for resources is available: TAGML⁶ (Bonhomme and Lopez, 2000).

In this section we present a tentative TAG modelisation of Vietnamese noun phrases (NP). This work is in the perspective to build Vietnamese syntactic resources for a parser using TAG formalism.

Description of NPs in Vietnamese

Nguyễn Tài Cẩn (1998) gives a detailed description of NPs in Vietnamese. We summarize it here in a few lines. Generally speaking, the full structure of a NP can contain the following parts:

$C_1 C_2 C_3 N_1 N_2 C_4 C_5$, in which

N_1 is a unit noun (cf. Thanh Bon Nguyen et al., 2004);

N_2 is a mass noun (N_1 is a measure unit or a classifier of N_2);

C_1 is a total quantifier (e.g. tất cả / *all*);

C_2 is a numeral or a determiner;

C_3 is the special strengthening particle cái;

C_4 is a complement sequence, in which each complement can be a noun, an adjective, a verb or their respective syntagm, a preposition, a number;

C_5 is a demonstrative pronoun.

Due to the limited space, instead of detailing each part, we just give some examples and then a list of attributes that constraint the collocation of these parts.

Examples

1) A NP with a full structure:

tất cả [C_1] năm [C_2] cái [C_3] quyển [N_1] sách [N_2] cũ [C_4] này [C_5]

= all | five | <strengthening particle> | <classifier noun> | book | old | this

= *all of these five old books*

2) A NP of one part with the generic degree of determination:

sách [N_2] = *book*.

3) A NP without N_2 :

năm [C_2] quyển [N_1] cũ [C_4] này [C_5]

= *these five old books*

The importance differences of Vietnamese in comparison with a language such as English are the following:

- All words are morphologically invariable.

- There exists the noun couple N_1 and N_2 as “center” of the noun phrase. Three compositions are possible: $N_1 + N_2$; $N_1 + \emptyset$; $\emptyset + N_2$; in the first two cases, N_1 is the syntagm head, and in the last case the head is N_2 .
- The particle *cái* has no equivalence in English.

Attribute list

- Attributes for NP:

generic = [+ , -] (a negative value constraints the appearance of N_1)

quantity = [+ , -] (a positive value constraints the appearance of N_1 and favors the appearance of C_2 and C_3)

demonstrative = [+ , -] (a positive value constraints the appearance of C_5 and favors the appearance of C_3)

- Attributes for N_1 and N_2 :

countable = + value for N_1 , - value for N_2 (in general)

classified = + , - (a positive value constraints the collocation of different classes of nouns for N_1 and N_2)

sense = empty, full (an “empty” value constraints the obligation of the presence of C_4 or C_5)

unit = human [classifier], thing [classifier], exact [measure], inexact [measure] (for N_1), - (for N_2)

- Attributes for C_2 :

number = singular, plural (in case C_2 is a determiner)

definite = + , - (in case C_2 is a determiner)

The constraints on the presence of C_1 change upon each value of C_1 , so we will not discuss them.

The constraints about the order of complements in the sequence C_4 are also ignored for lack of space.

NPs in TAG

Without any ambition to construct here a complete TAG model of NPs for Vietnamese, we consider two examples very simple for illustration. As in (Abeillé, 1993), we make use of N (Noun) for the root node of the NP tree.

Example 1

[Tôi đang đọc] sách = [I am reading] books.

NP = sách = N_2 (cf. the above description)

In this case N_2 is the head of the syntagm.

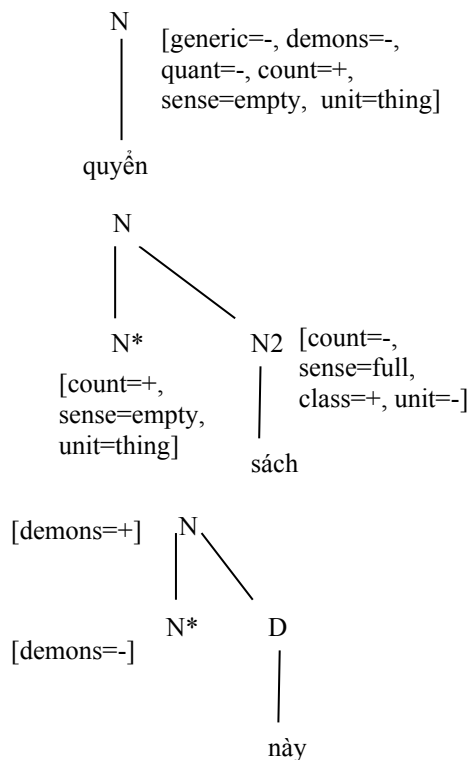
N	[generic=+,
	demons=-, quant=-,
	count=-, sense=full,
	class=+, unit=-]
sách	

Example 2

[Tôi đang đọc] **quyển sách này** = [I am reading] this book.

NP = quyển sách này = N₁ N₂ C₅

In this case N₁ is the head of the syntagm.



(For more precise descriptions, please refer to the documentation on our website³)

Conclusions

We have presented our work in the objective to construct a set of basis tools such as tokenizer, POS tagger with validation editors and concordancer for Vietnamese text analysis, as well as to build a database of linguistic resources such as morpho-syntactic lexicon, annotated corpus and syntactic lexicon for TAG formalism. Undertaken in a normalized framework that is considered by the subcommittee ISO TC 37 SC 4, these tools and resources are extensible and freely accessible for research purposes. We are now in the process of extending the annotated corpus with tagset evaluation and improvement. A lot of work also remains to be done to obtain a complete Vietnamese grammar in TAG formalism, and its syntactic lexicon.

Acknowledgements

This work would not have been possible without the enthusiastic collaboration of all the linguists at the Vietnam Lexicography Centre, especially Hoang T. T. Linh, Dang T. Hoa, Dao M. Thu and Pham T. Thuy. The research reported here was partially sponsored by Vietnamese national program of Sciences and Technology 2001-2003 (KC01).

References

- Anne Abeillé. 1993. *Les nouvelles syntaxes*. Armand Colin Editeur, Paris, FR.
- Anne Abeillé. 2002. *Une grammaire d'arbres adjoints pour le français*. Editions du CNRS, Paris, FR.
- Patrice Bonhomme et Patrice Lopez. 2000. *TAGML: codage XML et ressources pour les grammaires d'arbres adjoints lexicalisés*. LREC 2000, Athènes, GR.
- Cao Xuân Hạo. 2000. *Tiếng Việt - mấy vấn đề ngữ âm, ngữ pháp, ngữ nghĩa (Vietnamese - Some Questions on Phonetics, Syntax and Semantics)*. NXB Giáo dục, Hanoi, VN.
- Benoît Crabbé, Bertrand Gaiffe et Azim Roussanaly. 2003. *Une plate-forme de conception et d'exploitation d'une grammaire d'arbres adjoints lexicalisés*. The TALN Conference, Batz-sur-mer, FR.
- Hoàng Phê. 2002. *Từ điển tiếng Việt (Vietnamese Dictionary)*. Vietnam Lexicography Centre, NXB Đà Nẵng, VN.
- Nancy Ide, Laurent Romary. 2001. *Standards for Language Resources*. Proceedings of the IRCS Workshop on Linguistic Databases, Philadelphia, US.
- Charles N. Li, Sandra A. Thompson. 1976. *Subject and Topic: A new Typology of Language*. In Charles N. Li (ed.). *Subject and Topic*, London/New York: Academic Press, pp. 457-489.
- Nguyễn Tài Cẩn. 1998. *Ngữ pháp tiếng Việt (Vietnamese Grammar)*, NXB Đại học Quốc gia, Hanoi, VN.
- Thi Minh Huyen Nguyen, Laurent Romary, Xuan Luong Vu. 2003. *Une étude de cas pour l'étiquetage morpho-syntaxique de textes vietnamiens*. The TALN Conference, Batz-sur-mer, FR.
- Thanh Bon Nguyen, Thi Minh Huyen Nguyen, Laurent Romary, Xuan Luong Vu. 2004. *Lexical descriptions for Vietnamese language processing*. Proceedings of the Asian Language Resources Workshop (to appear), IJC-NLP 2004, Hainan, CN.
- Nguyen Thi Minh Huyen, Le Hong Phuong, Vu Xuan Luong. 2003. *A case study of the probabilistic tagger QTAG for Tagging Vietnamese Texts*. Proceedings of ICT.rda'03 (The First National Symposium on Research, Development and Application of Information and Communication Technology), Hanoi, VN.
- Ủy ban Khoa học Xã hội Việt Nam. 1983. *Ngữ pháp tiếng Việt (Vietnamese Grammar)*. NXB Khoa học Xã hội, Hanoi, VN.
- XTAG Research Group. 2001. *A Lexicalized Tree Adjoining Grammar for English*. IRCS, University of Pennsylvania, num. IRCS-01-03.

¹ <http://www.lpl.univ-aix.fr/projects/multext>

² <http://www.tc37sc4.org>

³ <http://www.loria.fr/equipes/led/outils.php> (vnACCMS)

⁴ <http://www.cis.upenn.edu/~xtag/>

⁵ <http://atoll.inria.fr/~clerger/DyALog/DyALog.fr.html>

⁶ <http://www.loria.fr/~azim/LLP2/help/fr/tagml2/>