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 - ntb@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 MULTEXT, 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 SC4.

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. Hao, 2000; Uy ban KHHVN, 1983).

Some Characteristics of Vietnamese

- Reduplicated words composed by phonetic reduplication (e.g. trăng/white - trăng/trăng / 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 coreferentiality (cf. Tôi và 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. Có 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 (Uy 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 (Uy 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">di</token>
<token id = "t4">rất</token>
<token id = "t5">nhanh</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" /></!
  
  `wordForm entry = "di" tokens = "t3" tag = "pos@V" /></!
  
  `wordForm entry = "rất" tokens = "t4" tag = "pos@R" /></!
  
  `wordForm entry = "nhánh" tokens = "t5" tag = "pos@A" /></!
  
  `wordForm entry = "." tokens = "t6" tag = "pos@dot" /></!

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. XTAG, Dialog) and also well developed at LORIA (Crabbé et al., 2003). Furthermore, a normalized format for resources is available: TAGML (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<sub>1</sub>, C<sub>2</sub>, C<sub>3</sub>, N<sub>1</sub>, N<sub>2</sub>, C<sub>4</sub>, C<sub>5</sub>, in which N<sub>1</sub> is a unit noun (cf. Thanh Bon Nguyen et al., 2004);
- **N<sub>2</sub>** is a mass noun (N<sub>1</sub> is a measure unit or a classifier of N<sub>2</sub>);
- C<sub>1</sub> is a total quantifier (e.g. tất cả / all);
- C<sub>2</sub> is a numeral or a determiner;
- C<sub>3</sub> is the special strengthening particle cái;
- C<sub>4</sub> 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<sub>5</sub> 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:

```
```

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

```
```

3) A NP without N:

```
```

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<sub>1</sub> and N<sub>2</sub> as “center” of the noun phrase. Three compositions are possible: N<sub>1</sub> + N<sub>2</sub>; N<sub>1</sub> + O<sub>1</sub>; Ø + N<sub>2</sub>; in the first two cases, N<sub>1</sub> is the syntagm head, and in the last case the head is N<sub>2</sub>.
- The particle cái has no equivalence in English.

**Attribute list**

- Attributes for NP:
  - generic = [+,-] (a negative value constraints the appearance of N<sub>1</sub>)
  - quantity = [+,-] (a positive value constraints the appearance of N<sub>1</sub> and favors the appearance of C<sub>2</sub> and C<sub>3</sub>)
  - demonstrative = [+,-] (a positive value constraints the appearance of C<sub>5</sub> and favors the appearance of C<sub>5</sub>)
- Attributes for N<sub>1</sub> and N<sub>2</sub>:
  - countable = + value for N<sub>1</sub>, - value for N<sub>2</sub> (in general)
  - classified = +, - (a positive value constraints the collocation of different classes of nouns for N<sub>1</sub> and N<sub>2</sub>)
  - sense = empty, full (an “empty” value constraints the obligation of the presence of C<sub>4</sub> or C<sub>5</sub>)
  - unit = human [classifier], thing [classifier], exact [measure], inexact [measure] (for N<sub>1</sub>), - (for N<sub>2</sub>)
- Attributes for C<sub>2</sub>:
  - number = singular, plural (in case C<sub>2</sub> is a determiner)
  - definite = +, - (in case C<sub>2</sub> is a determiner)

The constraints on the presence of C<sub>1</sub> change upon each value of C<sub>1</sub>, so we will not discuss them. The constraints about the order of complements in the sequence C<sub>4</sub> 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 (cf. the above description)
In this case N is the head of the syntagm.

N [generic=+, demons=+, quant=+, count=+, sense=full, unit=thing]

Example 2
[Tôi đang đọc quyén sách này = [I am reading] this book.
NP = quyén sách này = N1 N2 C
In this case N is the head of the syntagm.

N [generic=-, demons=-, quant=-, count=+, sense=empty, unit=thing]

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

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 sub-committee 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
Benoît Crabbé, Bertrand Gaffè 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.
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 morphosyntaxique de textes vietnamiens. The TALN Conference, Batz-sur-mer, FR.

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/