A Metagrammar for Vietnamese LTAG

Lê Hồng Phương
LORIA/INRIA Lorraine
Nancy, France
lehong@loria.fr

Nguyễn Thị Minh Huyền
Hanoi University of Science
Hanoi, Vietnam
huyenntm@vnu.edu.vn

Azim Roussanaly
LORIA/INRIA Lorraine
Nancy, France
azim@loria.fr

Abstract

We present in this paper an initial investigation into the use of a metagrammar for explicitly sharing abstract grammatical specifications for the Vietnamese language. We first introduce the essential syntactic mechanisms of the Vietnamese language. We then show that the basic subcategorization frames of Vietnamese can be compactly represented by classes using the XMG formalism (eXtensible MetaGrammar). Finally, we report on the implementation the first metagrammar producing verbal elementary trees recognizing basic Vietnamese sentences.

1 Introduction

Metagrammars (MG) have recently emerged as a means to develop wide-coverage LTAG for well-studied languages like English, French and Italian (Candito, 1999; Kinyon, 2003). MGs help avoid redundancy and reduce the effort of grammar development by making use of common properties of LTAG elementary trees.

We present in this paper an initial investigation into the use of a metagrammar for explicitly sharing abstract grammatical specifications for the Vietnamese language. We use the eXtensible MetaGrammar (XMG) tool which was developed by Crabbé (Crabbé, 2005; Parmentier and L. Roux, 2005) to compile a TAG for Vietnamese. The built grammar is called **vnMG** and is made available online for free access[1].

Only in recent years have Vietnamese researchers begun to be involved in the domain of natural language processing in general and in the task of parsing Vietnamese in particular. No work on formalizing Vietnamese grammar is reported before (Nguyen et al., 2004). In (Lê et al., 2006), basic declarative structures and complement clauses of Vietnamese sentences have been modeled using about thirty elementary trees, representing as many subcategorization frames. We show in this paper that these basic subcategorization frames can be compactly represented by classes in XMG formalism.

We first introduce the essential syntactic mechanisms of the Vietnamese language. We then show that the basic subcategorization frames of Vietnamese can be compactly represented by classes using the XMG formalism. We then report on the implementation the first metagrammar producing verbal elementary trees recognizing basic Vietnamese sentences, before concluding.

2 Vietnamese Subcategorizations

As for other isolating languages, the most important syntactic information source in Vietnamese is word order. The basic word order is Subject – Verb – Object. A verb is always placed after the subject in both predicative and question forms. In a noun phrase, the main noun precedes the adjectives and the genitive follows the governing noun. The other syntactic means are function words, reduplication, and, in the case of spoken language, prosody (Nguyễn et al., 2006).

From the point of view of functional grammar, the syntactic structure of Vietnamese follows a topic-oriented structure. It belongs to the topic-prominent languages as described by (Li and Thompson, 1976). In those languages, topics are

[1] *http://www.loria.fr/∼lehong/tools/vnMG.php*

coded in the surface structure and they tend to control co-referentiality. The topic-oriented "double subject" construction is a basic sentence type. For example, "*Cậu ấy khoẻ mạnh, là sinh viên y khoa / He strong, be student medicine*", which means that "*He is strong, he is medicine student*". In Vietnamese, passive voice and cleft subject sentences are rare or non-existent.

In general, Vietnamese predicates may be classified into three types depending on the need of a copula connecting them with their subjects in the declarative and negative forms (Nguyễn, 2004). Complex predicates can be constructed to form coordinated predicative structures starting from these basic types of predicates. We present briefly these three types of Vietnamese predicates in the following subsections.

2.1 First Type Predicates

The first type predicates are predicates which connect directly to their subjects without the need of a copula in both of the declarative and negative forms. For example

- Declarative form: *Tôi đọc sách. / I am reading books.*
- Negative form: *Tôi không đọc sách. / I am not reading books*.

These predicates are assumed by verbal phrases or adjectival phrases. The fact that an adjective can be a predicate is a specificity of Vietnamese in comparison with predicates of occidental languages. In English or French for instance, only verbal phrases can be predicates, adjectives in these languages always signify properties of subjects and they are always followed the verb "*to be*" in English or "*être*" in French.

2.2 Second Type Predicates

The second type predicates are predicates which are connected to their subjects by the copula "*là*" in the declarative form and by copulas "*không là*" or "*không phải*", or "*không phải là*" in the negative form. Predicates of this type are rather rich. They can be:

- Nouns or noun phrases: *Tôi là sinh viên. / I am student.*
- Verbs, adjectives, verbal phrases or adjectival phrases: *Van xin là yếu đuối. / Begging is feeble.*, *Học cũng là làm việc / To study is to work.*

2.3 Third Type Predicates

The third type predicates are predicates which connect directly to their subjects in the declarative form; however in the negative form, they are connected to their subjects by a copula. Predicates of this type are usually

- A clause: *Nó vẫn tên là Quýt. / His name is still Quýt.*
- A composition of a numeral and a noun: *Lê này mười ngàn đồng. / This pear costs ten thousand dongs.*
- A composition of a preposition and a noun: *Lúa này của chị Hoa. / This is the rice of Ms. Hoa.*
- An expression: *Thằng ấy đầu bò đầu bướu lắm. / That guy is very stubborn.*

2.4 Subcategorizations

In the first grammar LTAG for Vietnamese presented in (Lê et al., 2006), each subcategorization is represented by the same structure of elementary trees associcated with a considered predicate. We view that the suject is subcategorized in the same way like arguments. The verbs anchor thus elementary trees composed of a node for the subject and one or more nodes for each of its essential complements.

We follow the de facto standard that in TAG, in which each subcategorization is represented by a family of elementary trees. We define families of verbal elementary trees in the Table 1.

We present in the next section a metagrammar that generates this set of elementary trees.

3 A Metagrammar for Verbal Trees

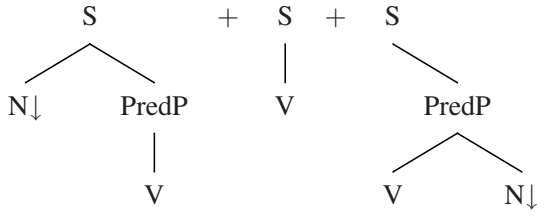
The subcategorizations of elementary trees describe only "canonical" constructions of predicative elements without taking into account for relative or question structures. For the purpose of investigation, we constraint ourselves in developing at the first stage only the verb spines and argument realizations shown in the subcategorizations presented in the previous section.

We have developed a XMG metagrammar that consists of 11 classes (or tree fragments). The

Subcategorizations	Families	Examples
Intransitive	N_0V	*ngủ/sleep*
With a nominal complement	N_0VN_1	*đọc/to read*
With a clausal complement	N_0VS_1	*tin/to believe*
With modal complement	$N_0V_0V_1$	*mong/to wish*
Ditransitive	$N_0VN_1N_2$	*cho/to give*
Ditransitive with a preposition	$N_0VN_1ON_2$	*vay/to borrow*
Ditransitive with a verbal complement	$N_0V_0N_1V_1$	*lãnh đạo/to lead*
Ditransitive with an adjectival complement	N_0VN_1A	*làm/to make*
Movement verbs with a nominal complement	$N_0V_0V_1N_1$	*ra/to go out*
Movement verbs with an adjectival complement	$N_0V_0AV_1$	*trở nên/to become*
Movement ditransitive	$N_0V_0N_1V_1N_2$	*chuyển/to transfer*

Table 1: Subcategorizations of Vietnamese verbs

metagrammar is currently able to produce the same set of elementary trees described in Table 1 including intransitive, transitive, ditransitive families with and/or without optional complements. As an illustration, the declarative transitive structure in Figure 1 can be defined by combining a canonical subject fragment with an active verb and a canonical object fragment.



This combination is conveniently expressed by a statement in terms of XMG language as usual:

$$\text{TransitiveVerb} = \text{Subject} \wedge \text{ActiveVerb} \wedge \text{Object}.$$

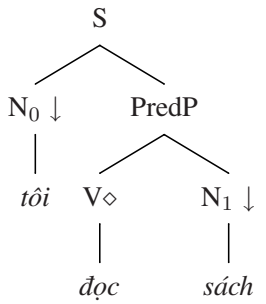


Figure 1: Declarative transitive structure $\alpha n_0 V n_1$

4 Conclusion and Future Work

This paper presents an initial investigation into the use of XMG formalism for developing a first metagrammar producing a LTAG for Vietnamese which recognizes basic verbal constructions. We have shown that the essential subcategorization frames of Vietnamese predicates can be effectively encoded by means of XMG classes while retaining basic properties of the realized verbal trees. This confirms that various syntactic phenomena of Vietnamese can be covered in a Vietnamese MG.

The first evaluation of the MG for Vietnamese is promising but the lexical coverage has to be improved further. Moreover, the grammar coverage needs to be revised by refining the constraints of agrammatical syntactic constructions. Although there are not many tree fragments in the current metagrammar, we find that the current MG overgenerates some undesired structures. The MG will also be extended to deal with constructions not yet covered like adjectival and noun phrase constructions. We also intend to generate a test suite to document the grammars and perform realistic evaluations.

There is an existing work on the development of metagrammars for not frequently studied languages like Korean and Yiddish and their relations to a German grammar (Kinyon, 2006). They showed that cross-linguistic generalizations, for example the verb-second phenomenon, can be incorporated into a multilingual MG. We think that a comparison of the Vietnamese MG with this work would be useful. In particular, a study of the relative position of verbs and arguments of Vietnamese and relate it to this work would be beneficial.

References

Marie-Hélène Candito. 1999. *Représentation modulaire et paramétrable de grammaires électroniques lexicalisées : application au français et à l'italien*. Doctoral Dissertation, Université Paris 7.

Benoit Crabbé. 2005. *Représentation informatique de grammaires fortement lexicalisées*. Doctoral Dissertation, Université Nancy 2.

Nguyễn Thị Minh Huyền, Laurent Romary, Mathias Rossignol and Vũ Xuân Lương. 2006. *A Lexicon for Vietnamese Language Processing*. Language Resources and Evaluation, Vol. 40, No. 3–4.

Kinyon A. and Rambow O. 2003. *Using the MetaGrammar to generate cross-language and cross-framework annotated test-suites*. In Proc. LINC-EACL, Budapest.

Alexandra Kinyon and Carlos A. Prolo. 2002. *A Classification of Grammar Development Strategies*. Proceedings of the Workshop on Grammar Engineering, Taipei, Taiwan.

Kinyon, Alexandra and Rambow, Owen and Scheffler, Tatjana and Yoon, SinWon and Joshi, Aravind K. 2006. *The Metagrammar Goes Multilingual: A Cross-Linguistic Look at the V2-Phenomenon*. Proceedings of the Eighth International Workshop on Tree Adjoining Grammar and Related Formalisms, Sydney, Australia

Lê Hồng Phương, Nguyễn Thị Minh Huyền, Laurent Romary, Azim Roussanaly. 2006. *A Lexicalized Tree-Adjoining Grammar for Vietnamese*. Proceedings of LREC 2006, Genoa, Italia.

Thanh Bon Nguyen, Thi Minh Huyen Nguyen, Laurent Romary, Xuan Luong Vu. 2004. *Developing Tools and Building Linguistic Resources for Vietnamese Morpho-Syntactic Processing*. Proceedings of LREC 2004, Lisbon, Portugal.

Charles N. Li and 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..

Yannick Parmentier and Joseph L. Roux. 2005. *XMG: a Multi-formalism Metagrammar Framework*. Proceedings of the Tenth ESSLLI Student Session.

Nguyễn Minh Thuyết and Nguyễn Văn Hiệp. 2004. *Thành phần câu tiếng Việt*. NXB Giáo dục, Hà Nội, Vietnam.