A lexicon for Vietnamese language processing

Thi Minh Huyen Nguyen, Laurent Romary, Mathias Rossignol, Xuan Luong Vu

A Lexicon for Vietnamese Language Processing

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

Laurent Romary (romary@loria.fr)
LORIA, Nancy, France

Mathias Rossignol (mathias.rossignol@mica.edu.vn)
International Research Center MICA, Hanoi, Vietnam

Xuân Lương Vũ (vuluong@vietlex.com)
Vietnam Lexicography Center, Hanoi, Vietnam

Abstract. Only very recently have Vietnamese researchers begun to be involved in the domain of Natural Language Processing (NLP). As there does not exist any published work in formal linguistics nor any recognizable standard for Vietnamese word definition and word categories, the fundamental tasks for automatic Vietnamese language processing, such as part-of-speech tagging, parsing, *etc.*, are very difficult tasks for computer scientists. The fact that all necessary linguistic resources have to be built from scratch by each research team is a real obstacle to the development of Vietnamese language processing. The aim of our projects is thus to build a common linguistic database that is freely and easily exploitable for the automatic processing of Vietnamese. In this paper, we present our work on creating a Vietnamese lexicon for NLP applications. We emphasize the standardization aspect of the lexicon representation. We especially propose an extensible set of Vietnamese syntactic descriptions that can be used for tagset definition and morphosyntactic analysis. These descriptors are established in such a way as to be a reference set proposal for Vietnamese in the context of ISO subcommittee TC 37/SC 4 (Language Resource Management).

Keywords: lexicon, linguistic resources, part-of-speech, standardization, syntactic description, Vietnamese

1. Introduction

Over the last 20 years, the field of Natural Language Processing (NLP) has seen numerous achievements in domains as diverse as part-of-speech (POS) tagging, topic detection, or information retrieval. However, most of those works were carried out for occidental languages (roughly corresponding to the Indo-European family) and lose much of their validity when applied to other language families. Thus, there clearly exists today a need to develop tools and resources for those other languages. Furthermore, an issue of great interest is the reusability of these linguistic resources in an increasing number of applications, and their

comparability in a multilingual framework. This paper focuses on the case of Vietnamese.

Only very recently have Vietnamese researchers begun to be involved in the domain of NLP. As there does not exist any published work in formal linguistics nor any recognizable standard for Vietnamese word definition and word categories, the fundamental tasks for automatic Vietnamese language processing, such as part-of-speech tagging, parsing, *etc.*, are very difficult for computational linguists. The fact that all necessary linguistic resources have to be built from scratch by each research team is a real obstacle to the development of Vietnamese language processing.

The aim of our project is therefore to build a common linguistic database that is freely and easily exploitable for the automatic processing of Vietnamese. In this paper, we present our work on creating a Vietnamese lexicon for NLP applications. We emphasize the standardization aspect of the lexicon representation. We especially propose an extensible set of Vietnamese syntactic descriptions that can be used for tagset definition and morphosyntactic analysis. These descriptors are established in such a way as to be a reference set proposal for Vietnamese in the context of ISO subcommittee TC 37/SC 4 (Language Resource Management).

We begin with an overview of the specificities of the Vietnamese language and of the context of our research (Section 2). We then present the lexicon model (Section 3) and detail the lexical descriptions used in our lexicon (Section 4). We finally introduce in section 5 our ongoing work to build an extended lexicon in which each lexical entry is enriched with more elaborate syntactic information.

2. Overview of Vietnamese Language Resources for NLP

In this section, we first present some general characteristics of the Vietnamese language. We then introduce the current status of language resources construction for Vietnamese language processing.

2.1. Characteristics of Vietnamese

The following basic characteristics of Vietnamese are adopted from Cao (2000) and Hữu et al.(1998).

2.1.1. *Language family*

Vietnamese is classified in the VietMuong group of the Mon-Khmer branch, that belongs to the Austro-Asiatic language family. Vietnamese is also known to have a similarity with languages in the Tai family. The

Vietnamese vocabulary features a large amount of Sino-Vietnamese words. Moreover, by being in contact with the French language, Vietnamese was enriched not only in vocabulary but also in syntax by the calque (or loan translation) of French grammar.

2.1.2. *Language Type*
Vietnamese is an isolating language, which is characterized by the following specificities:

- it is a monosyllabic language;
- its word forms never change, contrary to occidental languages that make use of morphological variations (plural form, conjugation...);
- hence, all grammatical relations are manifested by word order and function words.

2.1.3. *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* – `trăng 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` / *to pedal* – `xe đạp` / *bicycle*);
- some compound words whose syllable combination is no longer recognizable (*e.g.* `bồ nông` / *pelican*);
- complex words phonetically transcribed from foreign languages (*e.g.* `cà phê` / *coffee*, from the French *café*).

2.1.4. *Grammar*
The issue of syntactic category classification for Vietnamese is still in debate amongst the linguistic community (Cao, 2000; Hữu et al.,

1998; Diệp and Hoàng, 1999; Uỷ ban KHXHVN, 1983). That lack of consensus is due to the unclear limit between the grammatical roles of many words as well as the very frequent phenomenon of syntactic category mutation, by which a verb may for example be used as a noun, or even as a preposition. Vietnamese dictionaries (Hoàng, 2002) use a set of 8 parts of speech proposed by the Vietnam Committee of Social Science (Uỷ ban KHXHVN, 1983). We discuss precisely of these parts of speech in Section 4.

As for other isolating languages, the most important syntactic information source in Vietnamese is word order. The basic word order is Subject - Verb - Object. There are only prepositions but no postpositions. 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, intonation.

From the point of view of functional grammar, the syntactic structure of Vietnamese follows a topic-comment structure. It belongs to the class of topic-prominent languages as described by Li and Thompson 1976. In those languages, topics are coded in the surface structure and they tend to control co-referentiality (*e.g.* `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 (*e.g.* `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 (*e.g. There is a cat in the garden* should be translated as `Có một con mèo trong vườn` / *exist one <animal-classifier> cat in garden*).

2.2. Building Language Resources for Vietnamese Processing

While research in machine translation in Vietnam started in the late 1980's (Dien and Kiem, 2005), other works in the domain of NLP for Vietnamese are still very sparse. Moreover, linguists in Vietnam are not yet involved in computational linguistics.

Dien *et al.* (Dien et al., 2001; Dien and Kiem, 2003; Dien et al., 2003) mainly work on English - Vietnamese translation. Concerning the processing of Vietnamese, the authors published some papers on word segmentation, POS tagging for English-Vietnamese corpus, and the building of a machine-readable dictionary. Due to the lack of linguistic resources for Vietnamese and standard word classifications, the authors

make use of available word categories in print dictionaries, and also project English tags onto Vietnamese words. However, the developed tools and resources are not shared in the public research, which makes it difficult to evaluate their actual relevance.

Some other groups working on Vietnamese text processing focus their research on technical aspects and frequently meet the problem of lacking language resources such as lexicon and annotated corpora.

In 2001, we participated in the first national research project for Vietnamese language processing ("Research and development of technology for speech recognition, synthesis and language processing of Vietnamese", Vietnam Sciences and Technologies Program KC 01-03). In (T. M. H. Nguyen et al., 2003), we present our work on the POS tagging of Vietnamese corpora. Starting from a standardization point of view, we make use for the tagger of a tagset defined by considering a lexical description model compatible with the MULTEXT model (*cf.* Section 3.3). The tools (tokenizer, tagger), the tagged lexicon and corpus are distributed on the website of LORIA[1].

We now present the lexicon that we built in collaboration with the Vietnam Lexicography Centre (Vietlex), thanks to the grant of the KC 01-03 project.

3. Lexicon Model

Our NLP lexicon is based on a print dictionary (Hoàng Phê, 2002). As our objective is to build a lexicon that can be shared for public research, we pay much attention to resource standardization.

There have recently been many efforts to establish common formats and frameworks in the domain of NLP, in order to maximize the reusability of data, tools, and linguistic resources. In particular, the ISO subcommittee TC 37/SC 4, launched in 2002, aims at preparing various standards by specifying principles and methods for creating, coding, processing and managing language resources, such as written corpora, lexical corpora, speech corpora, dictionary compilations and classification schemes. Among several subjects, the LMF (*Lexical Markup Framework*) project is dedicated to lexicon representation.

In this section, we first present the structure of the print dictionary upon which our lexicon is based, and then introduce the LMF-based model of our NLP lexicon.

[1] Laboratoire Lorrain de Recherche en Informatique et ses Applications http://led.loria.fr/outils.php

yêu$_1$ d. *(id.)*. Vật tưởng tượng trong cổ tích, thần thoại, hình thù kì dị, chuyên làm hại người.

yêu$_2$ **1** đg. 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. *Mẹ yêu con. Yêu nghề. Yêu đời. Trông thật đáng yêu. Yêu nên tốt, ghét nên xấu (tng.).* **2** đg. Có tình cảm thắm thiết dành riêng cho một người khác giới nào đó, muốn chung sống và cùng nhau gắn bó cuộc đời. *Yêu nhau. Người yêu.* **3** đg. Từ dùng sau một động từ trong những tổ hợp tả một hành vi về hình thức là chê trách, đánh mắng một cách nhẹ nhàng, nhưng thật ra là biểu thị tình cảm thương yêu. *Mẹ mắng yêu con. Nguýt yêu. Tát yêu.*

Figure 1. Two entries of the morpheme " yêu " in the print dictionary

3.1. Vietnamese print dictionary

Vietlex owns the electronic version of the dictionary, in MS Word format. It contains 39 924 entry words, each of which may have several related meanings. Each of those numbered meanings is associated with a part-of-speech, an optional usage or domain note, a definition, and examples of use. For example, the morpheme " yêu " corresponds to two entries in the dictionay, as shown in Figure 1.

To facilitate the management of this resource, we convert the dictionary into XML format, by using the guidelines for print dictionary encoding proposed by the TEI (*Text Encoding Initiative*) project (Ide and Véronis, 1995). Reusing elements proposed by the TEI for dictionary encoding, we have defined a specialized DTD for the representation of the information contained in the Vietlex Centre Vietnamese dictionary. The data for each entry are automatically extracted based on the typographic indications in the original document. Since our focus is currently mainly on orthography and syntactic categories, the markup scheme remains very simple. The encoding of elements such as examples of use shall be further sophisticated in the future.

Figure 2 shows the XML representation of the information presented in the previous example for the morpheme " yêu ".

We now introduce the LMF project and our LMF-based lexicon representation model.

3.2. LMF-based lexicon representation model

3.2.1. *LNF (Lexical Mark-up Framework)*

LMF (ISO 24613, 2006) is an abstract meta-model providing a framework for the development of NLP-oriented lexicons. Its aim is to define a generic standard for the representation of lexical data, to facilitate their exchange and management. Its definition is inspired by several pre-normative international projects such as Eagles, Isle or Parole.

The approach chosen in LMF for the description of lexical entries is to systematically link syntactic behaviour and semantic description of the meaning of the word (Romary et al., 2004). That choice is

```
<superEntry>
    <entry n="1">
        <form><orth>yêu</orth></form>
        <sense> <!-- demon -->
            <gramGrp><pos>d.</pos></gramGrp>
            <usg type="style">(id.).</usg> <!-- not frequent -->
            <def>Vật tưởng tượng trong cổ tích, thần thoại, hình thù kì dị, chuyên làm
            hại người.</def>
        </sense>
    </entry>
    <entry n="2">
        <form><orth>yêu</orth></form>
        <sense n="1"> <!-- love (general love)-->
            <gramGrp><pos>đg.</pos></gramGrp>
            <def>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.</def>
            <eg>Mẹ yêu con. Yêu nghề. Yêu đời. Trông thật đáng yêu. Yêu nên tốt,
            ghét nên xấu (tng.).</eg>
        </sense>
        <sense n="2"> <!-- love (romantic love) -->
            <gramGrp><pos>đg.</pos></gramGrp>
            <def>Có tình cảm thắm thiết dành riêng cho một người khác giới nào đó,
            muốn chung sống và cùng nhau gắn bó cuộc đời.</def>
            <eg>Yêu nhau. Người yêu.</eg>
        </sense>
        <sense n="3"> <!-- love – modifier of another verb to express a tender
        action, "not serious" -->
            <gramGrp><pos>đg.</pos></gramGrp>
            <def>Từ dùng sau một động từ trong những tổ hợp tả một hành vi về hình
            thức là chê trách, đánh mắng một cách nhẹ nhàng, nhưng thật ra là biểu thị
            tình cảm thương yêu.</def>
            <eg>Mẹ mắng yêu con. Nguýt yêu. Tát yêu.</eg>
        </sense>
    </entry>
</superEntry>
```

Figure 2. Two dictionary entries for the morpheme " yêu ", in XML format

linguistically motivated, in particular by Saussure's work, according to which a word is defined by a signifier/signified pair, corresponding to a morphological/semantic description.

The LMF model proposes to develop a lexical database potentially gathering several lexicons, each of which is composed of a kernel around which are built lexical extensions corresponding to morphological, syntactic, semantic and inter-linguistic information, as presented on Figure 3. For instance, the extension for NLP syntax is represented in the diagram shown on Figure 4.

In accordance with the general principles of ISO/TC 37/SC 4 (Ide and Romary 2001, 2003), that information is described using elementary data categories defined in the central DCR (*Data Category Registry*) of TC 37. The development process of a LMF-conformant lexicon is presented on Figure 5.

3.2.2. *A LMF-based lexicon model for Vietnamese*

Our lexicon is organized as follows:

– each word form corresponds to a single lexical entry;

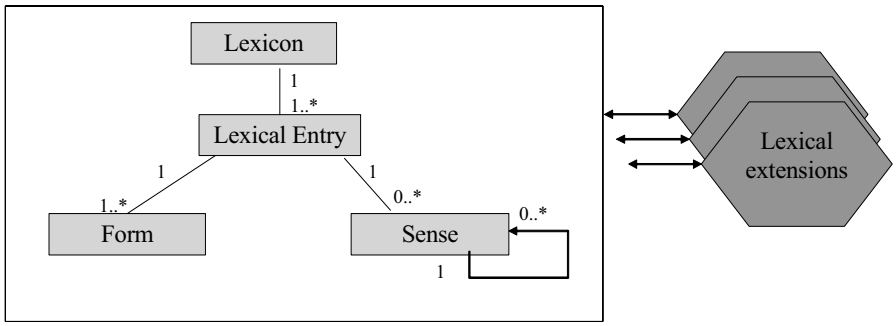

Figure 3. Principles of the LMF model

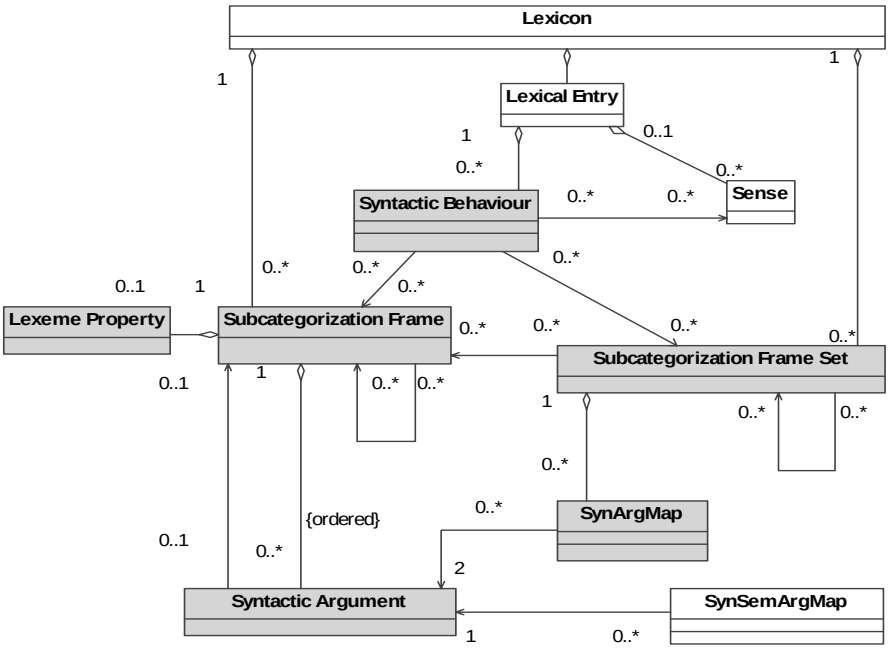

Figure 4. LMF extension for NLP syntax (ISO 24613, 2006)

– the senses of each lexical entry are organized following the sense hierarchy in the print dictionary (Hoàng Phê, 2002);

– with each sense is associated the corresponding definitions, examples, grammatical descriptions, *etc.*

This structure permits us to easily extract all information contained in the print dictionary we have presented. The information that we do not have concerns more precise grammatical descriptions of each word-meaning pair. As the first application of our lexicon is for the task of POS tagging, we need to provide the syntactic informations in such a

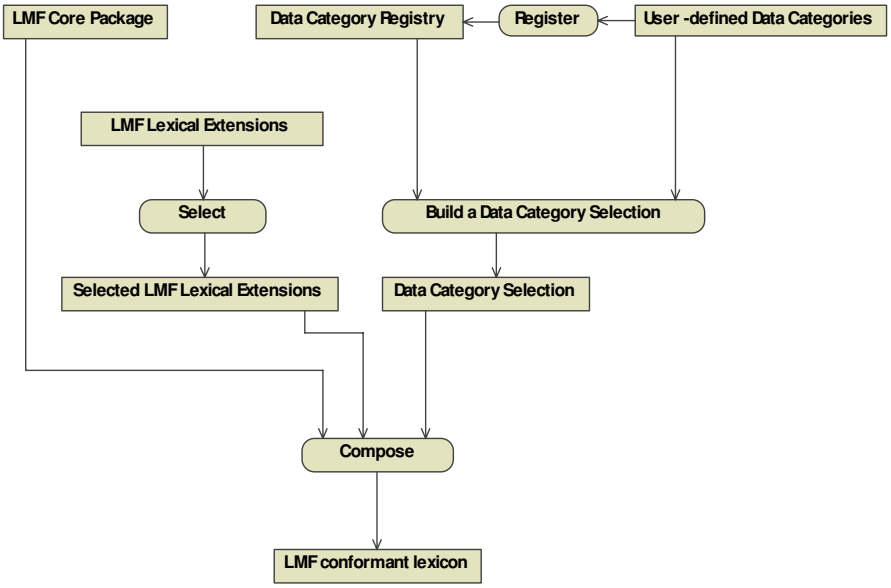

Figure 5. LMF usage

way that lexicon users can learn the possible tags of each word. We propose to use the model discussed hereafter.

3.3. The two-layer model of lexical descriptions

One of the sources of inspiration of TC 37/SC 4 is the Multext (*Multilingual Text Tools and Corpora*) project (Ide and Véronis, 1994). It has developed a morphosyntactic model for the harmonization of multilingual corpus tagging as well as the comparability of tagged corpora. It puts emphasis on the fact that in a multilingual context, identical phenomena should be encoded in a similar way to facilitate multiple applications (*e.g.* automatic alignment, multilingual terminological extraction, *etc.*). One principle of the model is to separate lexical descriptions, which are generally stable, from corpus tags. For lexical descriptions, the model uses two layers, the kernel layer and the private layer, as described below.

The kernel layer contains the morpho-syntactic categories common to most languages. The Multext model for Western European languages consists of the following categories: Noun, Verb, Adjective, Pronoun, Article/Determiner, Adverb, Adposition, Conjunction, Numeral, Interjection, Unique Membership Class, Residual, Punctuation (Ide and Véronis, 1994; Erjavec et al., 1998).

The private layer contains additional information that is specific to a given language or application. The specifications in this layer are

represented by attribute-value couples for each category described in the kernel layer. For instance, the English noun category is specified by three attributes: Type, Number and Gender, to which the following values can be assigned: common or proper (for Type), singular or plural (for Number), masculine or feminine or neuter (for Gender). Note that an extension of specifications in this layer is possible so as to be relevant for various text-processing tasks.

Possessing these fine descriptions, one can create a tagset, up to specific applications, by defining a mathematical map from the lexical description space to the corpus tag space, while maintaining the comparability of the tagsets.

In the next section, we present our lexical specifications proposal, which fits the MULTEXT scheme, for Vietnamese language, by building upon work published in (Nguyen et al., 2003). The lexical resources built in the framework of the KC 01-03 project are freely accessible[2] for research purposes, and all contributions are welcome.

4. Syntactic Category Descriptions

As we all know, linguistic theories first developed descriptions of Indo-European languages, which are inflecting languages where morphological variations strongly reflect the syntactic roles of each word. The distinction between categories like noun, verb, adjective, *etc.* in the kernel layer of MULTEXT is relatively clear. Meanwhile, with respect to analytic languages like Vietnamese, the syntactic category classification is far from perfect due to the absence of any morphological information. Many discussions are still going on about that matter amongst the linguistic community. In order to build a descriptor set comparable with the MULTEXT model, we start in (Nguyen et al., 2003) with the classification presented by the Vietnam Committee of Social Science (Uỷ ban KHXHVN, 1983), which is taken into account in the Vietnamese dictionary (Hoàng Phê, 2002). By analyzing eight categories found in the literature (noun, verb, adjective, pronoun, adjunct, conjunction, modal particle, interjection), we have tried to align them with those employed in the kernel layer of MULTEXT. Then, following the MULTEXT principle, each category is characterized by attribute-value couples in the private layer.

Our task is to develop the above work by improving and detailing the description of each layer and constructing a lexicon in which every entry is encoded with these specifications. In addition to the mentioned

[2] However, due to copyright restrictions, we cannot publish other information from the print dictionary, such as the definitions, examples, *etc.*

theoretical considerations, this work has been led in parallel with research concerning the development of tools for the morphosyntactic and syntactic analysis of Vietnamese (Nguyen et al., 2003; Nguyễn, 2006), thus ensuring that the chosen categories do have practical applicability to actual Vietnamese text data.

4.1. Kernel Layer

The Vietnamese alphabet is an extension of the Latin one. The notions of punctuation and abbreviation for Vietnamese are the same as for English, and we keep for them the descriptions proposed by the Multext project. Therefore in this section we only discuss the syntactic categories of words in the vocabulary: Noun, Verb, Adjective, Pronoun, Article/Determiner, Adverb, Adposition, Conjunction, Numeral, Interjection, Modal Particle, Unique Membership Class, Residual. Only the modal particle class is added in comparison with Multext. Although classifier words play an important role in Vietnamese, like in most Asian languages, their use and morphology are very similar to nouns. That is why we do not define a specific "Classifier" part-of-speech, but address them in the private layer.

For each category we give a definition and some characteristics (grammatical roles) with illustrating examples if necessary. The characterization of words in the private layer is based on their combination ability with respect to grammatical roles.

4.1.1. *Nouns*

The Noun category contains words or groups of words used to designate a person, place, thing or concept (*e.g.* `người` / *person*; `xe đạp` / *bicycle*). The grammatical roles that a Vietnamese noun (or noun phrase) can play are: grammatical subject in a sentence; predicate in a sentence when preceded by the copula verb là (to be); complement of a verb or an adjective; adjunct; adverbial modifier.

4.1.2. *Verbs*

A verb is a word used to express an action or state of being (*e.g.* `đi` / *to go*; `cười` / *to laugh*). In Vietnamese, a verb (or verb phrase) can play the following grammatical roles: predicate in a sentence; sometimes grammatical subject; restrictive adjunct (*e.g.* `thuốc uống` / *medicine drink*, meaning *orally administered drug*; `bàn ăn` / *table eat*, meaning *dining-table*); complement or adjectival modifier in a verb phrase (*e.g.* `tập viết` / *practice write*, meaning *writing practice*, `bước vào` / *step enter*, meaning *step into*).

4.1.3. *Adjectives*
This category consists of words used to describe or qualify a noun (*e.g.* cao / *tall*; xinh đẹp / *beautiful*). The grammatical roles of adjectives (or adjectival phrases) in Vietnamese can be: predicate in a sentence (without a preceding copula verb); sometimes grammatical subject; restrictive modifier of a noun or a verb (*e.g.* áo trắng / *dress white*, meaning *white dress*, nghe rõ / *hear clear*, meaning *hear clearly*).

4.1.4. *Pronouns*
The pronoun class contains words used in place of a noun that is determined in the antecedent context (*e.g.* tôi / *I*; chúng ta / *we*). Consequently, a pronoun plays the grammatical role of the word it replaces.

4.1.5. *Determiners/Articles*
These are the grammatical words used to identify a noun's definite or indefinite reference and/or quantity reference. For example: 1) những (indefinite plurualizer) 2) một (one, *i.e.* "a" article) 3) các (definite pluralizer).

These determiners are often categorized as numeral or even as noun in print dictionaries. They can also be described in the literature as a subcategory of numerals (Nguyễn Tài Cẩn, 1998), while analyzing the structure of the noun phrase.

4.1.6. *Adverbs*
An adverb is a word used to describe a verb, adjective, or another adverb (*e.g.* đã / past tense indicator; mãi mãi / *forever*).

4.1.7. *Adpositions*
In Vietnamese, only prepositions exist (*e.g.* trên / *on*; đến / *to*); they 1) occur before a complement composed of a noun phrase, noun, pronoun, or clause functioning as a noun phrase, and 2) form a single structure with the complement to express its grammatical and semantic relation to another unit within a clause.

4.1.8. *Conjunctions*
A conjunction is a word that syntactically links words or larger constituents, and expresses a semantic relationship between them (*e.g.* và / *and*; để / *in order to*). In many works and print dictionaries, the prepositions (adpositions) and conjunctions constitute the conjunction (or linking word) category probably because some words can play both roles. Still, their distinction can be identified in various sub-categories of the linking word category.

4.1.9. *Numerals*
A numeral is a word that expresses a number or a rank (*e.g.* hai / *two*; nhất / *first*). Numerals are assigned to the Noun class by some authors, but the morpho-syntactic distinction between these words and other nouns is clear enough to separate them into a new class.

4.1.10. *Interjections*
An interjection is a word or a sound that expresses an emotion (*e.g.* ồ / *oh*). These words function alone and have no syntactic relation with other words in the sentence.

4.1.11. *Modal Particles*
This category contains words added to a sentence in order to express the speaker's feelings (intensification, surprise, doubt, joy, *etc.*). Modal particles can create different sentence types (interrogative, imperative, *etc.*). For instance: nhỉ is often added to the end of a sentence with the meaning of "isn't it" or "doesn't it"; nhé added to the end of a sentence makes that sentence imperative.

4.1.12. *Non-autonomous elements*
This category corresponds to the Unique Membership Class of the Multext model. The unique value is applied to categories with a unique or very small membership, and which are "unassigned" to any of the standard part-of-speech categories. In Vietnamese these are some lexical elements, often come from Chinese, and never stand-alone, which express negation (*e.g.* bất in bất quy tắc / *irregular*) or transformation (*e.g.* hoá in công nghiệp hoá / *industrialize*), *etc.*. Those words may not appear as independent entries in print dictionaries.

4.1.13. *Residuals*
The residual value is assigned to classes of text-words that lie outside the traditionally accepted range of grammatical classes, although they occur quite commonly in many texts and very commonly in some. That is for example the case of foreign words, or mathematical formulae.

In the next subsection, we concentrate on the descriptions, specific for Vietnamese and represented by attribute-value couples, of the most important categories: Noun, Verb, Adjective, Pronoun, Determiner/Article, Adverb, Adposition, Conjunction, Numeral, Interjection, and Modal Particle.

4.2. Private Layer

The choice of attributes for each category of the kernel layer is made by taking into account the ability of a word to combine with others

in various sentence constituents. This consideration, together with the absence of morphological information in Vietnamese, leads us to define attributes that are closer to semantic information than is usually the case in the private layers for occidental languages, whether explicitely, using a "Meaning" attribute, or indirectly, when specifying the subcategorization frame of verbs. We list below the defined attributes with their values between square brackets. For each attribute value, we provide, when possible, an English word representative of the concept. When no English word is relevant, an explanation is given after the list of values.

4.2.1. *Nouns (N)*

- Countability [countable (seed), partially countable, non-countable (rice)] – countable nouns are those that can be employed directly with a numeral. Nouns that are generally non-countable but can directly combine with numerals in certain specific contexts are called "partially countable".

- Unit [classifier, natural (handful), conventional (meter), collective (herd), administrative (county)] – provides attributes relevant for unit nouns, including classifier nouns. The latter appear here because in Vietnamese they usually behave like unit nouns.

- Meaning [object (table), plant (tree), animal (cow), part (head), material (fabric), perception (color), location (place), time (month), turn, substantivizer, abstract (feeling), other] – *turn* is defined for words such as `lần` (*time* in *Repeat 5 times*) or `lượt` (*turn* in *It is my turn*); *substantivizer* describes words used to turn a verb into a nominal group (*e.g.* "the action of . . . ing"). This attribute reflects the combination abilities within various nouns. The specification could be finer-grained, but we have no ambition to go any further for the time being.

4.2.2. *Verbs (V)*

- Transitivity [intransitive, transitive, any].

- Grade [gradable, non-gradable] – a gradable verb can be used with an adverb of degree (*e.g.* very).

- Frame [copula (be), modal (can), passive (undergo), existence (remain), transformation (become), process stage (begin), comparison (equal), opinion (think), imperative (order), giving (offer), directive movement (enter), non directive movement (go), moving (push), other transitive, other intransitive] – this Frame attribute

encodes the distinction of verb valence (number of complements) and categories (noun, verb, clause, *etc.*) of the complements in the verb phrases.

4.2.3. *Adjectives (A)*

- Type [qualitative (nice), quantitative (high)] – a quantitative adjective can have a complement specifying a quantity (*e.g.* "high two meters"), and in that case it cannot be used with adverbs of degree (*e.g.* very).
- Grade [gradable (good), non-gradable (absolute)] – *cf.* the Grade attribute of Verb.

4.2.4. *Pronouns (P)*

- Type [personal (he), pronominal (myself), indefinite (one), time (that moment), amount (all), demonstrative (that), interrogative (who), predicative (that), reflexive (one another)].
- Person [first, second, third].
- Number [singular, plural].

4.2.5. *Determiners/Articles (D)*

- Type [definite, indefinite].
- Number [singular, plural].

4.2.6. *Numerals (M)*

- Type [cardinal (four), approximate (dozen), fractional (quarter), ordinal (fourth)].

4.2.7. *Adverbs (R)*

- Type [time (already), degree (very), continuity (still), negation (not), imperative, effect, other (suddenly)].
- Position [pre, post, undefined].

4.2.8. *Adpositions (S)*

- Type [locative (in), directive (across), time (since), aim (for), destination (to), relative (of), means (by)].

4.2.9. *Conjunctions (C)*
- Type [coordinating (however), consequence (if ... then), enumeration (..., ..., and ...)].

- Position [initial, non-initial] – necessary in case of discontinuous conjunctions.

4.2.10. *Interjections (I)*
- Type [exclamation, onomatopoeia].

4.2.11. *Modal Particles (T)*
- Type [global, local] – reflects the scope of a particle: whole sentence or one word only.

- Meaning [opinion, strengthening, exclamation, interrogation, call, imperative] – reflects different sentence types (exclamation, interrogation, *etc.*), determined by these particles.

4.3. Data Examples

Making use of the descriptors presented above, we have built a lexicon in which with each entry is associated its lexical descriptions. This construction is, for the private layer, performed manually by the linguists of the Vietnam Lexicography Centre, based on the descriptions of each entry in the print Vietnamese dictionary (Hoàng, 2002). As presented in Section 3.1, each entry in the dictionary contains distinct information about its grammatical category and its description for various meanings, with examples. With respect to the kernel layer, we first automatically get the 8 categories recorded there, and then manually process with the categories that should be revised, as described in 3.1. The data have two formats: simple text, as in the Multext model, and XML format. We choose for the time being a simple XML scheme that represents explicitly the feature structure corresponding to the private layer.

Here are some entries illustrating the data encoded in XML format. Due to the already mentioned copyright restrictions, the given example, as well as the publicly available lexical database, do not feature word definitions and examples, although that information has been used to find the values of the various attributes. That is why the presented data is, as of now, incomplete with respect to the LMF specification, since it cannot include the "Sense" structure.

Example 1. The word chạy in three uses: 1) *run* in *the horse runs*, 2) *run* in *run ultra-violet rays*, 3) *good* in *the sale is very good*.

```
<struct type='lexicalEntry'>
   <feat type='form'>chạy</feat>
   <struct type='grammaticalDescriptionGroup'>
       <struct type='grammaticalDescription'>
          <feat type='grammaticalCategory'>verb</feat>
          <struct type='subcategoryDescription'>
             <feat type='transitivity'>intransitive</feat>
             <feat type='grade'>non-gradable</feat>
             <feat type='frame'>non-directive movement</feat>
          </struct>
      </struct>
      <struct type='grammaticalDescription'>
         <feat type='grammaticalCategory'>verb</feat>
         <struct type='subcategoryDescription'>
            <feat type='transitivity'>transitive</feat>
            <feat type='grade'>non-gradable</feat>
            <feat type='frame'>moving</feat>
         </struct>
      </struct>
      <struct type='grammaticalDescription'>
         <feat type='grammaticalCategory'>adjective</feat>
         <struct type='subcategoryDescription'>
            <feat type='type'>qualitative</feat>
            <feat type='grade'>gradable</feat>
         </struct>
      </struct>
   </struct>
</struct>
```

Example 2. The syllable hoá has the same role as the suffix *ize* (*e.g.* in *industrialize*) in English.

```
<struct type='lexicalEntry'>
   <feat type='form'>hoá</feat>
   <struct type='grammaticalDescription'>
        <feat type='grammaticalCategory'>U_hoa</feat>
        <feat type='position'>post</feat>
   </struct>
</struct>
```

5. Ongoing Work: Building a Syntactic Lexicon

As the NLP community in Vietnam grows rapidly, the needs for linguistic resources are more and more apparent. In this context, we have obtained a large agreement amongst different research groups in Vietnam to submit a new national project called VLSP (*Vietnamese Language and Speech Processing*).

The VLSP project has just started in August 2006 [3]. The objective of this project is to create various essential language resources and tools for Vietnamese text and speech processing. The construction of a morpho-syntactic and syntactic lexicon is obviously one of the important tasks of the project.

As shown in Section 3.2, a lexicon model having the lexical extension for the syntax associates with each sense of an entry its syntactic behaviour information. That information gathers the descriptions of possible subcategorization frame sets.

For that task, two complementary approaches will be followed: the first one is to record the basic construction sets described in Vietnamese grammar documents. Based on the existing lexicon presented in the previous sections, we can automatize the process of linking the basic subcategorization frame sets to each lexical entry. For example, with the " Frame " attribute of a verb, we are able to link that verb to the corresponding subcategorization frame set that is common to other verbs having the same Frame value. The second approach is to learn other construction sets from corpora. For this task, we are also developping tools for corpus annotation. Moreover, we aim at creating online tools for the access and contribution to the construction of all the resources by the NLP community, for research purpose. We finally intend to complement the lexicon with new meaning descriptions independent of the copyrighted material we have relied on so far, in order to develop a fully LMF-conformant publicly available lexicon.

Another direction for future works concerns the integration of our proposal for lexicon attributes into ISO standards. Indeed, the isolating, non-flexional nature of Vietnamese has led us to define specific attributes to specify word roles, more semantic than what is commonly used for western languages. Hence most of the attributes that we propose to use are absent from the current ISO 12620 Data Category Registry (DCR). In the next step, we intend to work in cooperation with specialists of other isolating languages to propose a consensual set of values for integration in the DCR.

[3] *cf.* the project forum at `http://www.viettreebank.com`

6. Conclusion

We have presented our proposal for a reference set of Vietnamese lexical descriptors by following the standardization activities of the ISO subcommittee TC 37/SC 4. These descriptors are expressed, for the time being, in a two-layer model comparable with the MULTEXT model, which is developed for various European languages. In the kernel layer, we have added the modal particle category that contains modal words appearing frequently in Vietnamese. The other categories remain the same. In the private layer, where specific features of Vietnamese are recorded, we proposed various attributes that are syntactically important for this analytic language in which morphology is not present to help us analyze syntactic structures. With the help of the Vietnam Lexicography Center, we applied all these descriptions to a lexicon that contains all the entries (about 40,000) of the Vietnamese dictionary (Hoàng, 2002). These resources are represented in a common format that ensures their extensibility and is widely adopted by the international research community, with the purpose of sharing them with all the researchers in the domain of NLP. This base can help us define tagsets for various applications using morpho-syntactically annotated corpora. We expect that the ongoing project in order to build a syntactic lexicon will be fruitful with the contribution of the NLP community.

Acknowledgements

This work would not have been possible without the enthusiastic collaboration of all the linguists at the Vietnam Lexicography Centre, especially Hoàng Thị Tuyền Linh, Đặng Thanh Hoà, Đào Minh Thu and Phạm Thị Thuỷ. Great thanks to them! Many thanks also to Nguyễn Thành Bôn for his contribution to the development of the various tools.

References

Cao Xuân Hạo: 2000, *Tiếng Việt - mấy vấn đề ngữ âm, ngữ nghĩa (Vietnamese - Some Questions on Phonetics, Syntax and Semantics).* Hà Nội, Viết Nam: NXB Giáo dục.

Dien, D., P. P. Hoi, and N. Q. Hung: 2003, 'Some Lexical Issues in Electronic Vietnamese Dictionary'. In: *PAPILLON-2003 Workshop on Multilingual Lexical Databases.* Hokaido University, Japan.

Dien, D. and H. Kiem: 2003, 'POS-Tagger for English - Vietnamese Bilingual Corpus'. In: *Workshop: Building and Using Parallel Texts: Data Driven Machine Translation and Beyond*. Edmonton, Canada.

Dien, D. and H. Kiem: 2005, 'State of the Art of Machine Translation in Vietnam'. *AAMT Journal, special issue on MT Summit X*.

Dien, D., H. Kiem, and N. V. Toan: 2001, 'Vietnamese Word Segmentation'. In: *Proceedings of the Sixth Natural Language Processing Pacific Rim Symposium (NLPRS2001)*. Tokyo, Japan.

Diệp Quang Ban and Hoàng Văn Thung: 1999, *Ngữ pháp tiếng Việt (Vietnamese Grammar)*, Vol. 1. Hà Nội, Viêt Nam: NXB Giáo dục.

Erjavec, T., N. Ide, and D. Tufis: 1998, 'Development and Assessment of Common Lexical Specifications for Six Central and Eastern European Languages'. In: *Proceedings of the First International Conference on Language Resources and Evaluation*. Granada, Spain.

Hoàng Phê (ed.): 2002, *Từ điển tiếng Việt (Vietnamese Dictionary)*. Viêt Nam: NXB Đà Nẵng.

Hữu Đạt, Trần Trí Dõi, and Đào Thanh Lan: 1998, *Cơ sở tiếng Việt (Basis of Vietnamese)*. Hà Nội, Viêt Nam: NXB Giáo dục.

Ide, N. and L. Romary: 2001, 'Standards for Language Resources'. In: *Proceedings of the IRCS Workshop on Linguistic Databases*. Philadelphia, US.

Ide, N. and L. Romary: 2003, 'Encoding Syntactic Annotation'. In: A. Abeillé (ed.): *Building and Using Parsed Corpora*. Dordrecht, Netherlands: Kluwer Academic Publishers.

Ide, N. and J. Véronis: 1994, 'MULTEXT: Multilingual Text Tools and Corpora'. In: *Proceedings of the 15th International Conference on Computational Linguistics (COLING 94)*. Kyoto, Japan.

Ide, N. and J. Véronis: 1995, 'Encoding dictionaries'. In: N. Ide and J. Véronis (eds.): *Text Encoding Inititative: Background and context*. Dordrecht, Netherlands: Kluwer Academic Publishers.

ISO 24613, Rev.13: 2006, 'Language resource management - Lexical markup framework (LMF)'. ISO, Geneva, Switzerland.

Li, C. N. and S. A. Thompson: 1976, 'Subject and Topic: A new Typology of Language'. In: C. N. Li (ed.): *Subject and Topic*. London/New York: Academic Press, pp. 457–489.

Nguyen, T. M. H., L. Romary, and X. L. Vu: 2003, 'Une étude de cas pour l'étiquetage morpho-syntaxique de textes vietnamiens'. In: *Actes de la Conférence francophone internationale sur le Traitement Automatique des Langues Naturelles (TALN 03)*. Batz-sur-mer, France.

Nguyễn, T. M. H.: 2006, 'Outils et ressources linguistiques pour l'alignement de textes multilingues français-vietnamiens'. Thèse de doctorat en informatique, Université Henri Poincaré, Nancy I, Nancy, France.

Nguyễn Tài Cẩn: 1998, *Ngữ pháp tiếng Việt (Vietnamese Grammar)*. Hà Nội, Viêt Nam: NXB Đại học Quốc gia.

Romary, L., S. Salmon-Alt, and G. Francopoulo: 2004, 'Standards going concrete: from LMF to Morphalou'. In: *Workshop Enhancing and using electronic dictionaries, The 20th International Conference on Computational Linguistics (COLING)*. Geneva, Switzerland.

Uỷ ban Khoa học Xã hội Việt Nam: 1983, *Ngữ pháp tiếng Việt (Vietnamese Grammar)*. Hà Nội, Viêt Nam: NXB Khoa học Xã hội.