HAL open science

Lexical descriptions for Vietnamese language processing

Thanh Bon Nguyen, Thi Minh Huyen Nguyen, Laurent Romary, Xuan Luong Vu

To cite this version:

Thanh Bon Nguyen, Thi Minh Huyen Nguyen, Laurent Romary, Xuan Luong Vu. Lexical descriptions for Vietnamese language processing. The 1st International Joint Conference on Natural Language Processing - IJCNLP'04 / Workshop on Asian Language Resources, 2004, Sanya, Hainan Island, China, 8 p. inria-00107760

HAL Id: inria-00107760

https://inria.hal.science/inria-00107760v1

Submitted on 1 Dec 2008

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Lexical descriptions for Vietnamese language processing

Thanh Bon Nguyen
IFI, Hanoi
ntbon@ifi.edu.vn

Thi Minh Huyen Nguyen
Laboratory LORIA, Nancy
nguyen@loria.fr

Laurent Romary
Laboratory LORIA, Nancy
romary@loria.fr

Xuan Luong Vu
Vietnam Lexicography Centre, Hanoi
vuluong@vietlex.com

Abstract

Only very recently have Vietnamese researchers begun to be involved in the domain of Natural Language Processing. As there does not exist any published work in formal linguistics or any recognizable standard for Vietnamese word categories, the fundamental works in Vietnamese text analysis such as part-of-speech tagging, parsing, etc. are very difficult tasks for computer scientists. All necessary linguistic resources have to be built from scratch, and until now almost no resources are shared in public research. The aim of our project is to build a common linguistic database that is freely and easily exploitable for the automatic processing of Vietnamese. In this paper, we propose an extensible set of Vietnamese syntactic descriptions that can be used for tagset definition and corpus annotation. These descriptors are established in such a way to be a reference set proposal for Vietnamese in the context of ISO subcommittee TC37/SC4 (*Language Resource Management*)[1].

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 their validity when applied to other language families. Today, there clearly exists 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 or any recognizable standard for Vietnamese word categories, the fundamental works in Vietnamese text analysis such as POS tagging, parsing, etc. are very difficult tasks for computer scientists. All necessary linguistic resources have to be built from scratch, and until now almost no resources are shared in public research.

The aim of our project is to build a common linguistic database that is freely and easily exploitable for the automatic processing of Vietnamese. In this paper, we propose an extensible set of Vietnamese morpho-syntactic descriptions that can be used for tagset definition and corpus annotation. These descriptors are established in such a way to be a reference set proposal for Vietnamese in the framework of ISO subcommittee TC37/SC4 (*Language Resource Management*).

Before detailing the set of descriptions we are proposing (Section 3), we present an overview of the specificities of the Vietnamese language and of the context of our research (Section 2).

[1] http://www.tc37sc4.org

2 Language Resources for Vietnamese

2.1 Characteristics of Vietnamese

To begin with, we present some basic characteristics of Vietnamese (Cao X. Hạo, 2000; Hữu Đạt et al., 1998).

Language family

Vietnamese is classified in the Viet-Muong 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.

Language Type

Vietnamese is an isolating language, which is characterized by the following specificities:

- It is a monosyllabic language.
- Its word forms never change, which is contrary to occidental languages that make use of morphological variations (plural form, conjugation...).
- Hence, all grammatical relations are manifested by word order and tool words.

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`/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

The issue of syntactic category classification for Vietnamese is still in debate amongst the linguistic community (Cao X. Hạo, 2000; Hữu Đạt et al., 1998; Diệp Q. Ban & Hoàng V. Thung, 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 frequent phenomenon of syntactic category mutation. Vietnamese dictionaries (Hoàng Phê, 2002) use a set of 8 parts of speech proposed by the Vietnam Committee of Social Science (Uỷ ban KHXHVN, 1983).

As 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 tool words, the reduplication, and the intonation.

From the viewpoint of functional grammar, the syntactic structure of Vietnamese follows a topic-comment structure. It belongs to the class of topic-prominent languages as referred to (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, the 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).

2.2 Building resources for Vietnamese

Automatic Processing of Vietnamese

Linguists in Vietnam are not yet involved in computational linguistics. Likewise, very few concrete initiatives for the automatic analysis of Vietnamese have been noticed until now. We hereafter introduce the two most significant published works in this domain we are aware of.

Dinh Dien et al. (2001, 2003a, 2003b) 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.

T. M. H. Nguyen et al. (2003) present a work on the POS tagging of Vietnamese corpora. Starting from a standardization point of view, the authors make use for the tagger of a tagset defined by considering a lexical description model compatible with the MULTEXT model reminded hereafter. The tools (tokenizer, tagger), the tagged lexicon and corpus are distributed on the team's website[1].

Standardization in Resource Construction

There are currently many efforts in establishing 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 TC37 / SC4, 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 compiling and classification schemes.

One of the projects that are the sources of inspiration in the framework of TC 37 SC 4 is MULTEXT (*Multilingual Text Tools and Corpora*). This project has developed a morphosyntactic model for the harmonization of multilingual corpus tagging as well as the comparability of tagged corpora. It is emphasized 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 two-layers model of lexical descriptions

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 (cf. Nancy Ide and Jean Véronis, 1994; Tomaž Erjavec et al, 1998).

The private layer contains additional information that is private to each language or proper to particular applications. 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, which can be assigned the following values: 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 will present our proposal of lexical specifications, which fits the MULTEXT scheme, for Vietnamese language, by developing the work published by T. M. H. Nguyen et al (2003). We have built a lexicon based on these descriptions. These resources can be freely accessible for research purposes and all contributions would be welcome.

[1] http://www.loria.fr/equipes/led/outils.php

3 Syntactic Category Descriptions

As we all know, linguistic theory was first developed to describe Indo-European languages, which are inflecting languages where the morphological variation strongly reflects 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, in (T. M. H. Nguyen et al., 2003), the authors start with the classification presented by the Vietnam Committee of Social Science (Uỷ ban KHXHVN, 1983), which is taken in 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), the authors try to align them with those employed in the kernel layer of MULTEXT. Then, in the same principle as MULTEXT, each category is characterized with 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.

3.1 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. 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.

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.

Nouns

Noun category contains word or group of words used to express the name of a person, place, thing or concept. 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.

Verbs

A verb is a word used to express an action or state of being. 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`, that means `orally administered drug`, `bàn ăn / table eat`, that means `dining-table`); complement or adjectival modifier in a verb phrase (e.g. `tập viết / practice write`, that means `writing practice`, `bước vào / step enter`, that means `step into`).

Adjectives

This category consists of words used to describe or qualify a noun. 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`, that means `white dress`, `nghe rõ / hear clear`, that means `hear clearly`).

Pronouns

The pronoun class contains words used in place of a noun that is determined in the antecedent context. Consequently, a pronoun plays the grammatical role of the word it replaces.

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 categorized as numeral or even as noun in print dictionaries. They can also

be described in the literature as a subcategory of numerals (cf. Nguyễn Tài Cẩn, 1998), while analyzing the structure of noun phrase.

Adverbs

An adverb is a word used to describe a verb, adjective, or another adverb.

Adpositions

In Vietnamese there only exist prepositions, which 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.

Conjunctions

A conjunction is a word that syntactically links words or larger constituents, and expresses a semantic relationship between them.

The prepositions (adpositions) and conjunctions constitute the conjunction (or linking word) category in many works and print dictionaries, probably because some words can play both roles. Still, their distinction can be identified in various subcategories of the linking word category.

Numerals

A numeral is a word that expresses a number or a rank. Numerals are assigned to the Noun class by some authors. But the morpho-syntactic distinction of these words from other nouns is clear enough to separate them into a new class.

Interjections

An interjection is a word or a sound that expresses an emotion. These words function alone and have no syntactic relation with other words in the sentence.

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 be imperative.

Unique Membership Class

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. (cf. Example 2 in 3.3). These words do not appear as independent entries in print dictionaries.

Residuals

The residual value is assigned to classes of text-word, which lie outside the traditionally accepted range of grammatical classes, although they occur quite commonly in many texts and very commonly in some. For example: 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.

3.2 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, lead us to defining a "Meaning" attribute for most important categories. Below we list the 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.

Nouns (N)

- Countability [countable (*seed*), partially countable, non-countable (*rice*)] - Note that amongst mass nouns are only material and aggregate nouns (e.g. people) that are absolutely non-countable; nouns that generally have a non-countable meaning but can di-

rectly combine with numerals in certain specific contexts are called "partially countable".

- Unit [natural (*cup*), conventional (*meter*), collective (*herd*), administrative (*county*)] - provides attributes relevant for unit nouns, including classifier 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 fact of...*). 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.

Verbs (V)

- Transitivity [intransitive, transitive, any]
- Grade [gradable, non-gradable] - a gradable verb can be used with an adverb of degree (e.g. *very*).
- Meaning [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 Meaning attribute encodes the distinction of verb valence (number of complements) and categories (noun, verb, clause, etc.) of the complements in the verb phrases.

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.

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]

Determiners/Articles (D)

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

Numerals (M)

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

Adverbs (R)

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

Adpositions (S)

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

Conjunctions (C)

- Type [coordinating (*however*), consequence (*if ... then*), enumeration (..., ..., *and* ...)]
- Position [initial, non-initial] - necessary in case of discontinuous conjunctions.

Interjections (I)

- Type [exclamation, onomatopoeia]

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.

3.3 Data examples

Making use of the descriptors presented above, we have constructed a lexicon in which with each entry is associated its lexical description. This construction, 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 Phê, 2002). Note that this print dictionary is previously converted to XML format from its MS Word format. 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.

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='lexical entry'>
   <feat type='form'>chạy</feat>
   <struct type='grammatical description group'>
         <struct type='grammatical description'>
               <feat type='grammatical category'>verb</feat>
               <struct type='subcategory description'>
                     <feat type='transitivity'>intransitive</feat>
                     <feat type='grade'>non-gradable</feat>
                     <feat type='meaning'>non-directive
                     movement</feat>
               </struct>
         </struct>
         <struct type='grammatical description'>
               <feat type='grammatical category'>verb</feat>
               <struct type='subcategory description'>
                     <feat type='transitivity'>transitive</feat>
                     <feat type='grade'>non-gradable</feat>
                     <feat type='meaning'>moving</feat>
               </struct>
         </struct>
         <struct type='grammatical description'>
               <feat type='grammatical category'>adjective</feat>
               <struct type='subcategory description'>
                     <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='lexical entry'>
   <feat type='form'>hoá</feat>
   <struct type='grammatical description'>
         <feat type='grammatical category'>U_hoa</feat>
         <feat type='position'>post</feat>
```

```
    </struct>
  </struct>
```

4 Conclusion

We have presented our proposal of a reference set for 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 the morphology does not help to 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 Phê, 2002). These resources are represented in a common format that ensures their extensibility and is widely adopted by the international research community, in the objective to share 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, as well as construct a syntactic lexicon for a given grammatical formalism.

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. Great thanks to them!

References

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.

Diệp Quang Ban, Hoàng Văn Thung. 1999. *Ngữ pháp tiếng Việt (Vietnamese Grammar),* volume 1. NXB Giáo dục, Hanoi, VN.

Dinh Dien, Hoang Kiem. 2003a. *POS-Tagger for English - Vietnamese Bilingual Corpus.* Workshop: Building and Using Parallel Texts: Data Driven Machine Translation and Beyond, Edmonton, CA.

Dinh Dien, Hoang Kiem, Nguyen Van Toan. 2001. *Vietnamese Word Segmentation.* Proceedings of the Sixth Natural Language Processing Pacific Rim Symposium (NLPRS2001), Tokyo, JP.

Dinh Dien, Pham Phu Hoi, Ngo Quoc Hung. 2003b. *Some Lexical Issues in Electronic Vietnamese Dictionary*, PAPILLON-2003 Workshop on Multilingual Lexical Databases, Hokkaido University, JP.

Tomaž Erjavec, Nancy Ide, Dan Tufis. 1998. *Development and Assessment of Common Lexical Specifications for Six Central and Eastern European Languages*. Proceedings of the First International Conference on Language Resources and Evaluation, Granada, SP.

Hữu Đạt, Trần Trí Dõi, Đào Thanh Lan. 1998. *Cơ sở tiếng Việt (Basis of Vietnamese)*. NXB Giáo dục, Hanoi, VN.

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, Philapdelphia, US.

Nancy Ide, Jean Véronis. 1994. *MULTEXT: Multilingual Text Tools and Corpora*. Proceedings of the 15th International Conference on Computational Linguistics, COLING'94, Kyoto, JP.

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.

Uỷ ban Khoa học Xã hội Việt Nam. 1983. *Ngữ pháp tiếng Việt (Vietnamese Grammar)*. NXB Khoa học Xã hội, Hanoi, VN.