



Section  
des  
Etudes Vietnamiennes

<http://www.inalco.fr>

11ème conférence NooJ

Budapest

8-10 juin 2008

## NooJ'08 Conference

---

A Vietnamese module for NooJ :  
Modelization, realization and perspectives

Philippe Lambert  
*ATER, INALCO*

Michel Fournié  
*P.U., INALCO*

National Institute for oriental Studies - INALCO  
2 rue de Lille  
75007 PARIS – FRANCE  
[philambert@gmail.com](mailto:philambert@gmail.com)



**Section  
des  
Etudes Vietnamiennes**

<http://www.inalco.fr>

**11ème conférence NooJ  
Budapest  
8-10 juin 2008**

# Outline

---

- The Vietnamese language
- A Vietnamese module for NooJ
- The perspectives



**Section  
des  
Etudes Vietnamiennes**

<http://www.inalco.fr>

**11ème conférence NooJ  
Budapest  
8-10 juin 2008**

- **The Vietnamese language**
- A Vietnamese module for NooJ
- The perspectives



Section  
des  
Etudes Vietnamiennes

<http://www.inalco.fr>

11ème conférence NooJ  
Budapest  
8-10 juin 2008

# Vietnamese language

## Overview

- 90 million of speakers all over the world





Section  
des  
Etudes Vietnamiennes

<http://www.inalco.fr>

11ème conférence NooJ  
Budapest  
8-10 juin 2008

# Vietnamese language

---

## Its characteristics

- Viet-Muong linguistic group, Môn-Khmer branch, Austro-Asian family
- a tonal tongue (6 tones : a á à ả ã ạ )
- type : analytic
- monosyllabic
- dextrogyre syntactic structure
- **historically** : (André- Georges Haudricourt) :
  - Origin : Môn-Khmer,
  - Chinese influence : ideographic scripture and lexicon
  - Colonial impact : Roman alphabet, lexicon, grammar



Section  
des  
Etudes Vietnamiennes

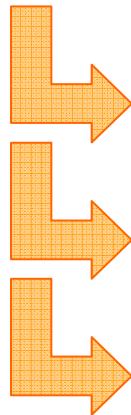
<http://www.inalco.fr>

11ème conférence NooJ  
Budapest  
8-10 juin 2008

# Vietnamese language

---

## an analytic tongue



Words invariability

Positional syntax

Combinatory constraints



Section  
des  
Etudes Vietnamiennes

<http://www.inalco.fr>

11ème conférence NooJ  
Budapest  
8-10 juin 2008

## Examples :

- Sao(*int.pron.*) không (*neg.*) bảo(*v.*) nó(*pers. pron.*) đến (*v.*)?  
*Why don't you inform me that he comes ?*
- Sao bảo nó không đến ? *Why are you telling that he will not come ?*
- Sao đến nó không bảo ? *Why did he come without telling us ?*
- Bảo nó đến không sao : *No matter if you tell him to come.*
- etc...



Section  
des  
Etudes Vietnamiennes

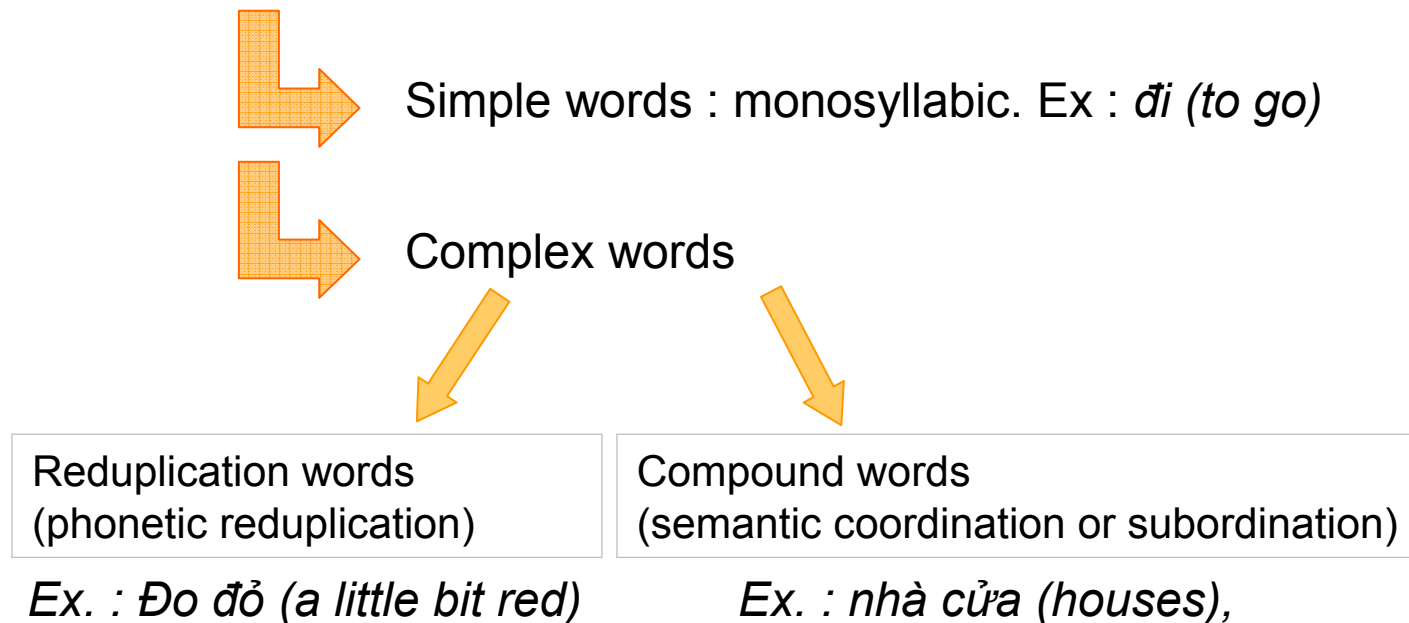
<http://www.inalco.fr>

11ème conférence NooJ  
Budapest  
8-10 juin 2008

# Vietnamese language

## a monosyllabic tongue

Two main types of lexical unit





Section  
des  
Etudes Vietnamiennes

<http://www.inalco.fr>

11ème conférence NooJ  
Budapest  
8-10 juin 2008

- The Vietnamese language
- **A Vietnamese module for NooJ**
- The Perspectives



Section  
des  
Etudes Vietnamiennes

<http://www.inalco.fr>

11ème conférence NooJ  
Budapest  
8-10 juin 2008

## The vietnamese module for NooJ

---

### Textual data of the test

- 1. The initiation module to learn vietnamese language**  
(vietnamese courses, INALCO)
- 2. Three texts from the electronic newspapers**



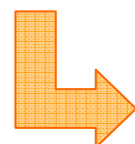
Section  
des  
Etudes Vietnamiennes

<http://www.inalco.fr>

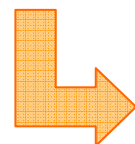
11ème conférence NooJ  
Budapest  
8-10 juin 2008

# The vietnamese module for NooJ

## Unicode encoding problem



A dozen of vietnamese writing systems  
(unicode, VPS, VNI, ABC, ...)



**(2) Enter a Text:**

ee

Value (decimal):

101 204 163 32 225 186 185



Normalize the system for NooJ by transcoding  
dictionaries and textual data to uniformize them



Section  
des  
Etudes Vietnamiennes

<http://www.inalco.fr>

11ème conférence NooJ

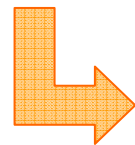
Budapest

8-10 juin 2008

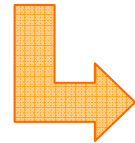
# The vietnamese module for NooJ

## Tagset for vietnamese

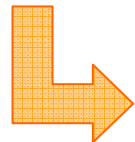
- Tagging vietnamese



Lack of determined grammatical categories



Ph.D. Thesis of  
Nguyễn Thị Minh Huyền (*Loria – France*)  
Lexicalized Tree Adjoining Grammar formalism



Tagset built : 1./ ACRONYM, 2./ QUALIFIER,  
3./ ADVERB, 4./ ADVERBIAL LOCUTION,  
5./ AFFIX, 6./ APPELLATIVE, 7./ CLASSIFIER, 8./ SPECIFICATIVE,  
9./ CONJUNCTION, 10./ TOOL WORD, 11./ POST-VERB,  
12./ NOUN, 13./ VERB.



Section  
des  
Etudes Vietnamiennes

<http://www.inalco.fr>

11ème conférence NooJ  
Budapest  
8-10 juin 2008

**Category**

**Sub category**

<b>ABREVIATION</b>	<b>ABR</b>	
<b>QUALIFICATIF</b>	<b>QLF</b>	
<b>ADVERBE</b>	<b>ADV</b>	TEMPS
<b>LOCUTION ADVERBIALE</b>	<b>LOCADV</b>	
<b>AFFIXE</b>	<b>AFF</b>	
<b>APPELLATIF</b>	<b>APP</b>	
<b>CLASSIFICATEURS</b>		CL
<b>SPECIFICATIFS</b>		
<b>CONJONCTION</b>	<b>CONJ</b>	LOCALISATION COORDINATION
<b>MOT-OUTILS</b>	<b>MO</b>	TOTALITE PLURALITE DISTANCE RESTRICTION APPARTENANCE PERFECTIF TEMPOREL NEGATION BUT ACHEVEMENT COPULE
<b>POST-VERBE</b>	<b>PV</b>	MOUVEMENT REPETITION RESULTATIF
<b>NOM</b>	<b>N</b>	
	<b>NPR</b>	NOM PROPRE
<b>VERBE</b>	<b>V</b>	
	<b>VA</b>	VERBE ACTION
	<b>VE</b>	VERBE ETAT



Section  
des  
Etudes Vietnamiennes

<http://www.inalco.fr>

11ème conférence NooJ  
Budapest  
8-10 juin 2008

## The vietnamese module for NooJ

# Dictionaries

- A **general dictionary** including the integrality of tagged terms with 1061 entries ;
- An **economic** dictionary with 818 locutional entries.  
(tagged as *syntagmatic locutions*) ;
- A dictionary of **Vietnam's provincial names** with 64 entries ;
- An exhaustive dictionary of **vietnamese family names** with 137 entrées;
- A thematic dictionary about **Information Technologies** containing 424 entries.
- A dictionary of **appellatives** (personal pronouns)with 49 entries ;
- A dictionary of **specificatives** (subclasses of classifiers) comprenant 143 entrées.



Section  
des  
Etudes Vietnamiennes

<http://www.inalco.fr>

11ème conférence NooJ  
Budapest  
8-10 juin 2008

Entry	S-Lemma	Category	SynSem
An	An	NPR	-
an ninh	an ninh	N	UNAMB
an toàn	an toàn	N	UNAMB
an toàn	an toàn	QLF	-
An Thành Thủy	An Thành Thủy	NPR	UNAMB
anh	anh	APP	-
anh	anh	N	-
Anh	Anh	NPR	-
ảnh	ảnh	N	-
Ánh	Ánh	N	-
ăn	ăn	V	VA
ấm	ấm	QLF	-
ấp	ấp	N	-
Ấu Cơ	Ấu Cơ	NPR	UNAMB
ây	ây	DEM	-
ba	ba	NB	-

bản đồ đẳng dụng,SYNTLOC+ECO+UNAMB
bản phá giá,SYNTLOC+ECO+UNAMB
bản quản lý,SYNTLOC+ECO+UNAMB
hội đồng quản trị,SYNTLOC+ECO+UNAMB
bản quyết toán,SYNTLOC+ECO+UNAMB
bản tay vô hình,SYNTLOC+ECO+UNAMB
bản tổng kết tài sản,SYNTLOC+ECO+UNAMB
bản thu nhập,SYNTLOC+ECO+UNAMB
bản vị đô la,SYNTLOC+ECO+UNAMB
bản vị tiền,SYNTLOC+ECO+UNAMB
bản vị vàng,SYNTLOC+ECO+UNAMB
bảng cân đối,SYNTLOC+ECO+UNAMB
bảng quan với rủi ro,SYNTLOC+ECO+UNAMB
bảng tính,SYNTLOC+ECO+UNAMB
báo cáo luồng tiền,SYNTLOC+ECO+UNAMB
báo cáo ngân lưu,SYNTLOC+ECO+UNAMB
báo cáo tài chính,SYNTLOC+ECO+UNAMB
bảo chứng vàng ,SYNTLOC+ECO+UNAMB

bà,APP+ FAM+S
bác,APP+ FAM+S
bầy đàn,APP+ GEN+PLUR
bọn,APP+ GEN+PLUR
bố,APP+ FAM+S
câu,APP+ FAM+S
con,APP+ FAM+S
cô,APP+ FAM+S



Section  
des  
Etudes Vietnamiennes

<http://www.inalco.fr>

11ème conférence NooJ  
Budapest  
8-10 juin 2008

## The vietnamese module for NooJ

---

# Grammars

- A **number** syntactic grammar with 3 graphes for numeric values from 0 to 999 999 ;
- a graph for the **nominal syntagm** ;
- A **dates** graph;
- A syntactic graph for vietnamese **question structures** ;  
(open and closed questions, emphatic) ;
- A graph for **compound verbs** .

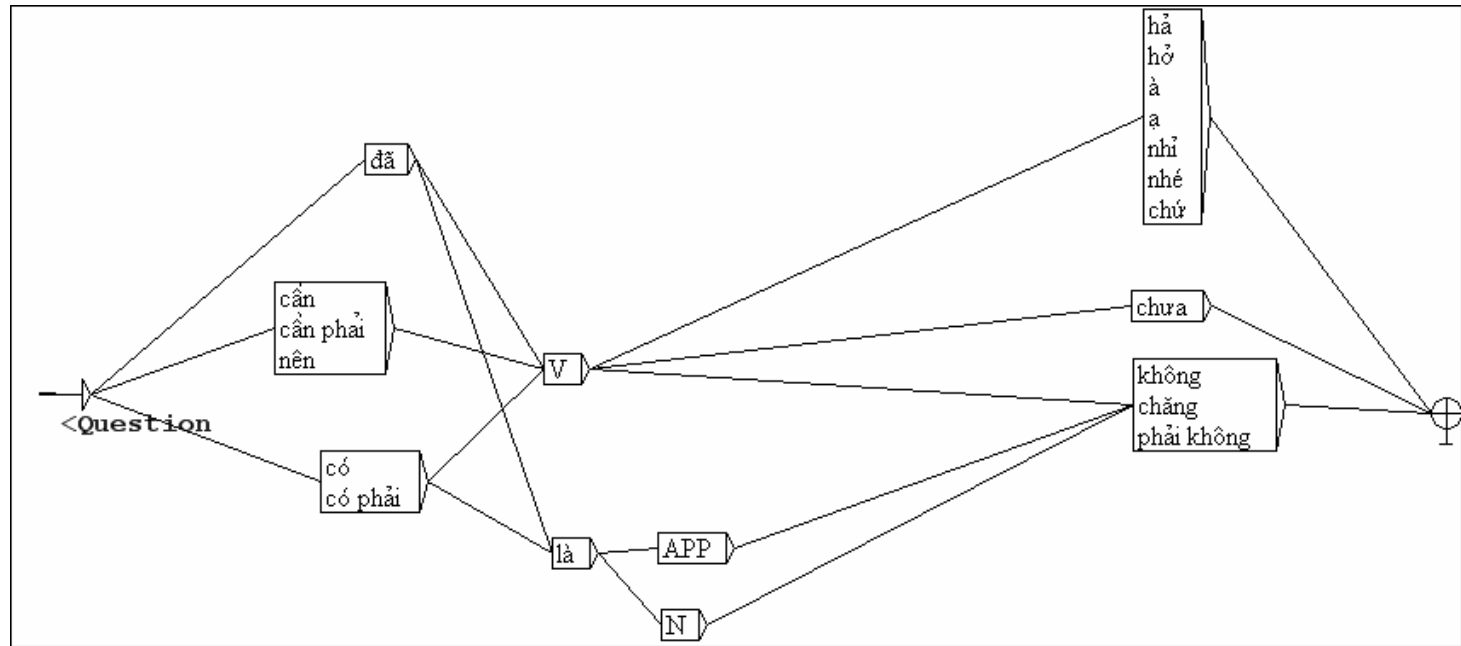


Section  
des  
Etudes Vietnamiennes

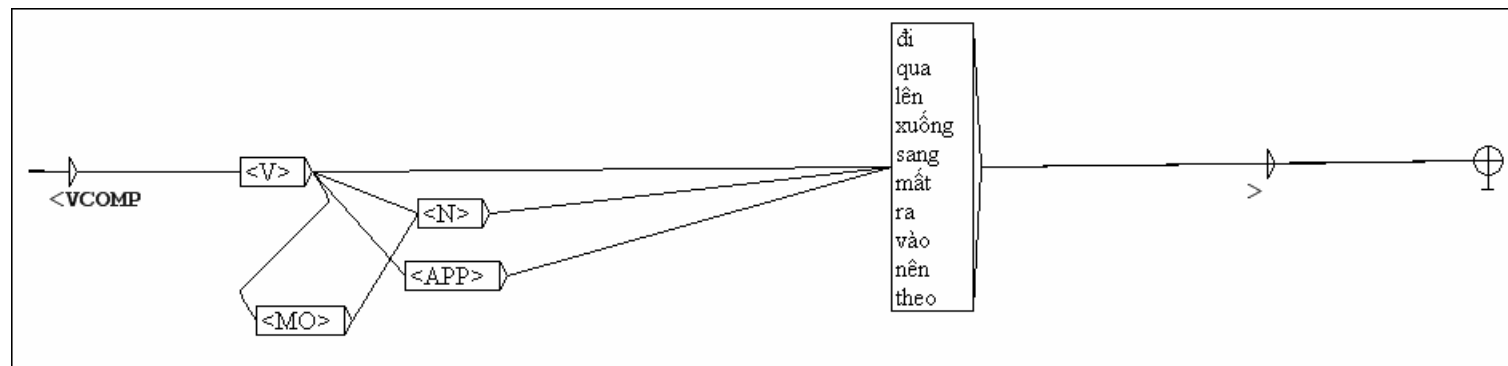
<http://www.inalco.fr>

11ème conférence NooJ  
Budapest  
8-10 juin 2008

Question structure graph



Compound verb graph





**Section  
des  
Etudes Vietnamiennes**

<http://www.inalco.fr>

**11ème conférence NooJ  
Budapest  
8-10 juin 2008**

- The Vietnamese language
- Vietnamese module for NooJ
- **The Perspectives**



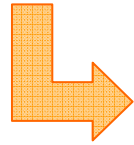
**Section  
des  
Etudes Vietnamiennes**

<http://www.inalco.fr>

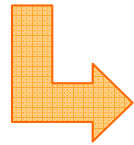
**11ème conférence NooJ  
Budapest  
8-10 juin 2008**

# Perspectives

---



Didactic : teaching vietnamese for beginners  
(morphological and lexical parsing)



Content analysis for literacy studies and traductology



**Section  
des  
Etudes Vietnamiennes**

<http://www.inalco.fr>

**11ème conférence NooJ  
Budapest  
8-10 juin 2008**

...Thank you for your attention