Current status of *Vietnamese Treebank* usefulness of collaboration with **A**sian **L**anguage **T**reebank

VU TAT THANG

Dept. of Multimedia Human-Machine Language Technology,

Institute of Information Technology,
Vietnam Acadamy of Science and Technology.

Content

- IOIT and International Collaborations
- Vietnamese Language
- VLSP Standard
- Current status of Vietnamese Processing
- Propose idea

Content

- IOIT and International Collaborations
- Vietnamese Language
- VLSP Standard
- Current status of Vietnamese Processing
- Propose idea

IOIT – a member of ASEAN MT

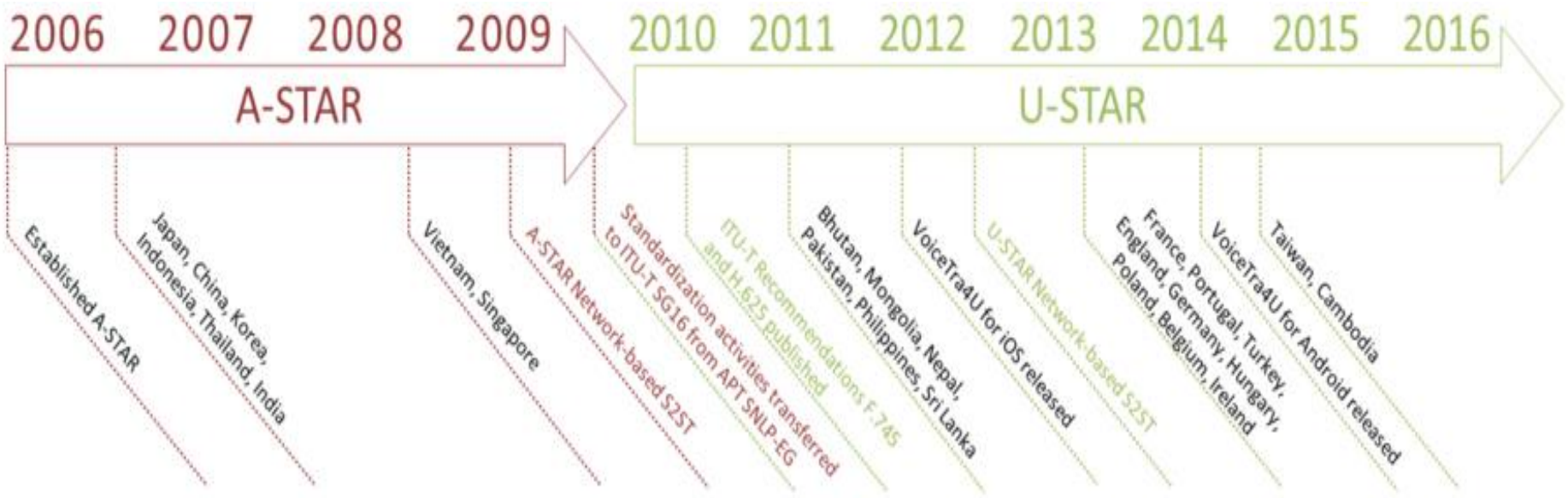
- Member of "**Network-based ASEAN Languages Translation Public Service Project**", 2012- 2015. Lead by NECTEC – Thailand
  - The communication among people in the ASEAN region has increased gradually and will become extreme especially after 2015 when the ASEAN Community begins. The automatic machine translation (MT) system has become more and more important to facilitate the cross-language communication, but has been limited for ASEAN countries.
  - Sharing language data
  - Develop platform
  - Integration of translation system

IOIT – a member of A-STAR (U-STAR)

- A-STAR (Asian Speech Translation Advanced Research), *2008-2010*

  U-STAR (Universal Speech Translation Advanced Research),

  2010 – till now



Mechanism of S2s system

Vietnamese

Multilingual Speech Recognition

Spoken Language Translation

Multilingual Speech Synthesis

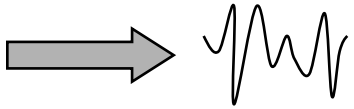
English

I go to school

Tôi đi đến trường

Tôi đi đến trường

I go to school



Statistical Speech Recognition

Statistical Machine Translation + Multi-engine approach

Corpus-based Speech Synthesis

Corpora

Large Scale Vietnamese Speech Corpora

Large Scale VietnameseText Corpora

Large Scale Parallel Corpora of Vietnamese and English

Large Scale English Text Corpora

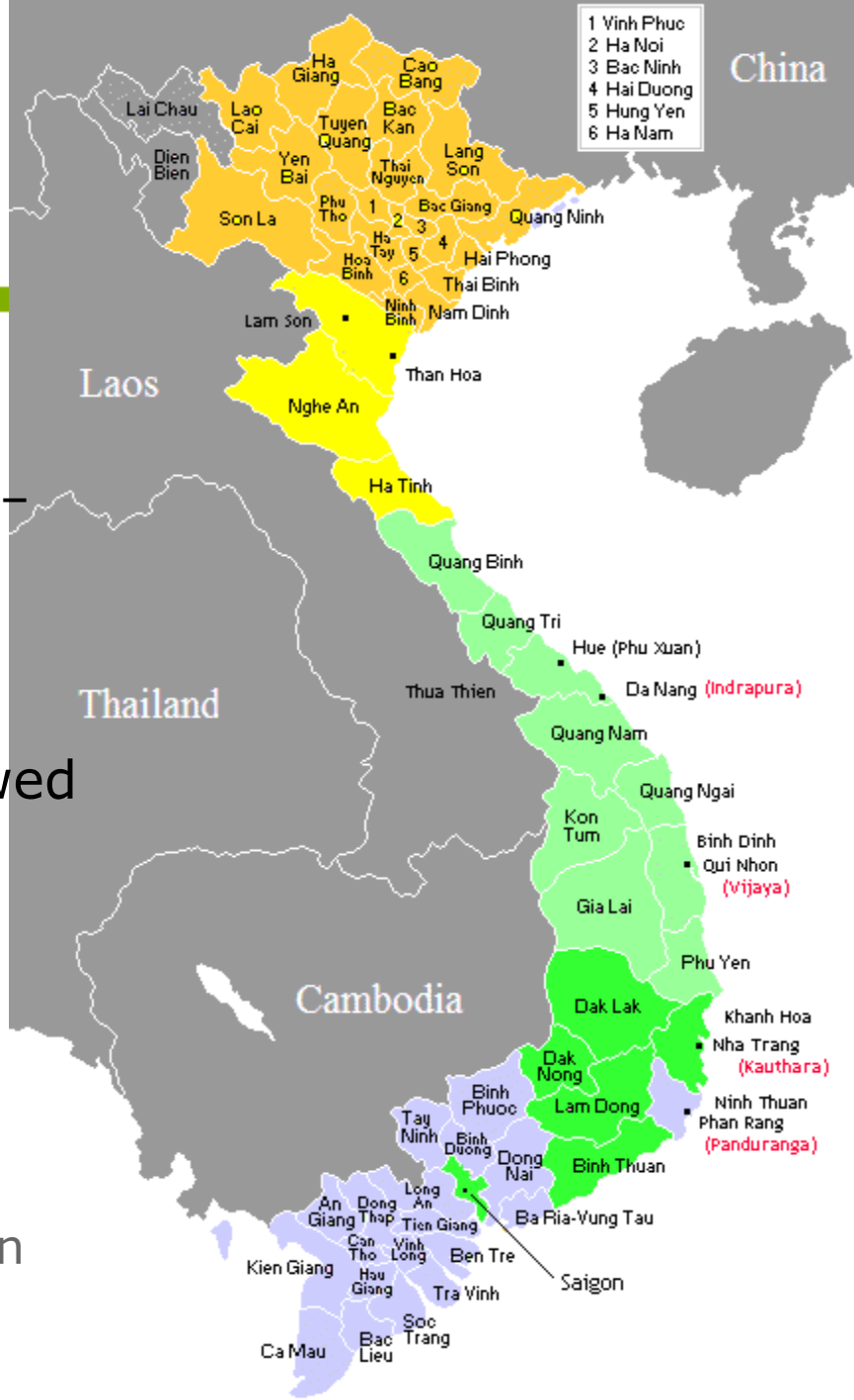
Large Scale English Speech Corpora

Content

- IOIT and International Collaborations
- Vietnamese Language
- VLSP Standard
- Current status of Vietnamese Processing
- Propose idea

Vietnamese Language

- Spoken as mother tongue by
  - 86% of Vietnam's population
  - ~ 3 million overseas Vietnamese – most live in US
- It is part of the Austro-asiatic language family (168 languages)
- Many vocabulary has been borrowed from Chinese
- Writing system:
  - Formerly, Chinese writing system
  - Today: Latin alphabet, with additional diacritics for tones and certain letters
- Dialects: Northern, Central, Southern



Vietnamese language

- Vietnamese language was established a long time ago
- Chinese characters was used for a long time
- Unique writing system of Vietnam called Chu Nom (字喃) in the 10th century
- Romanced script to represent the Quốc Ngữ since the beginning of the 20th century

Nam quốc sơn hà Nam đế cư
南 国 山 河 南 帝 居
Over Mountains and Rivers of the South, Reigns the Emperor of the South

Content

- IOIT and International Collaborations
- Vietnamese Language
- VLSP Standard
- Current status of Vietnamese Processing
- Propose idea

Setting up the VLSP “standards” for the public

- Importance of “standards” in VLSP: choose an unified view from various schools on Vietnamese language
- Guide for words recognition and description: morphological, syntactic, semantic criteria
- Guide for constituent labeling: noun phrase, verb phrase, clause, etc.
- Guide for sentence split
- Others

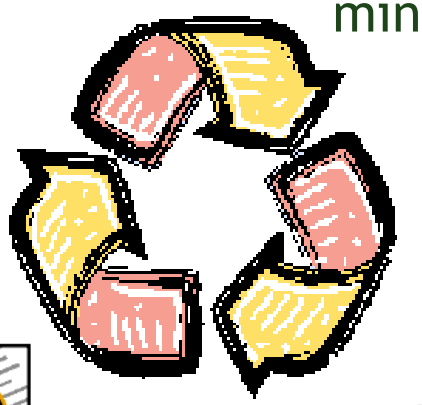
VLSP national project

National project with eleven active research groups on VLSP (Vietnamese Language and Speech Processing)

Building VLSP infrastructure, especially indispensable resources and tools for the VLSP development.

Building and developing several typical VLSP products for public end-users.

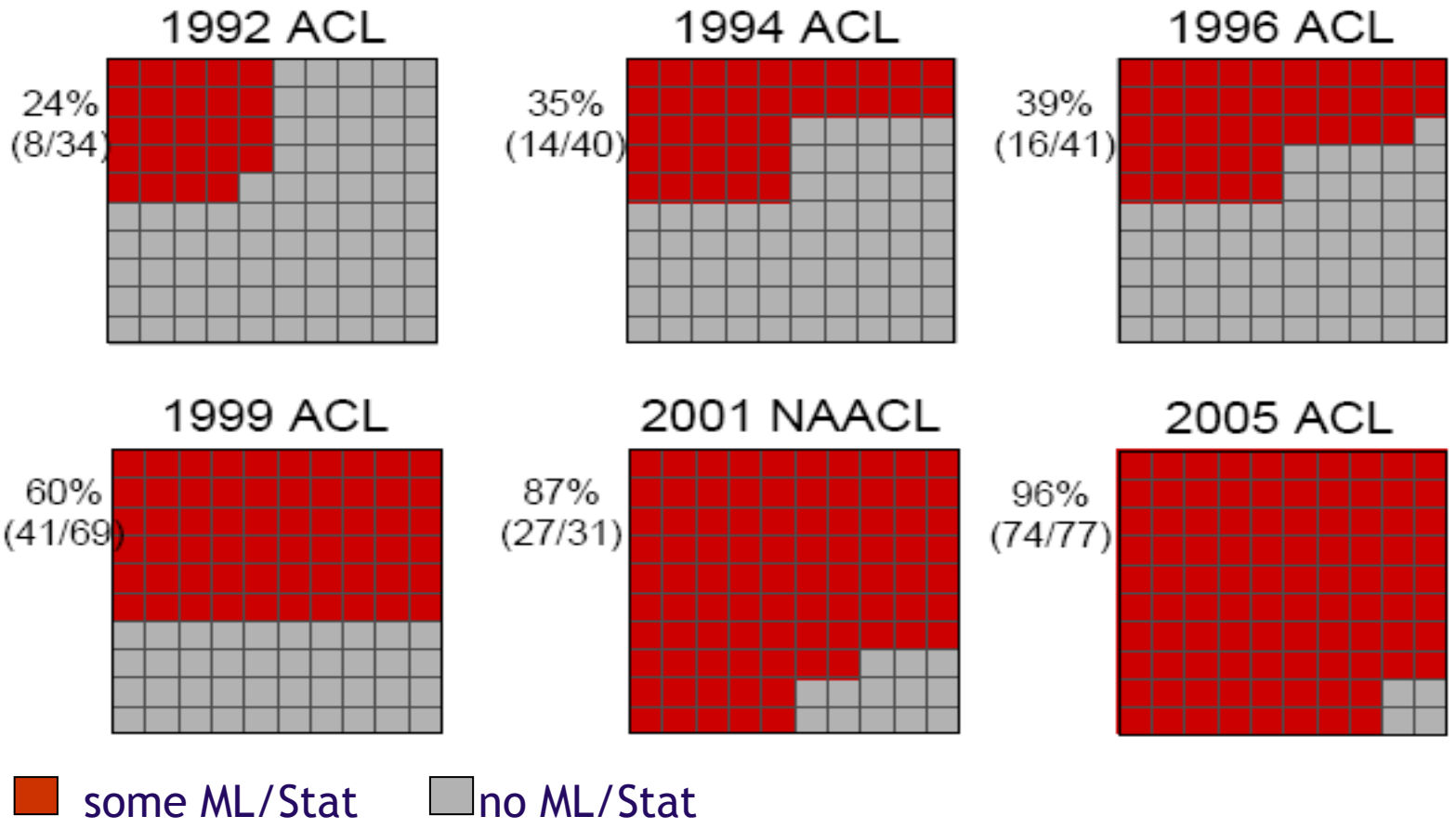
Pragmatics: Speech, text and Web data mining



Natural language processing methods

Tools, corpora, resources

ML and statistical methods in NLP

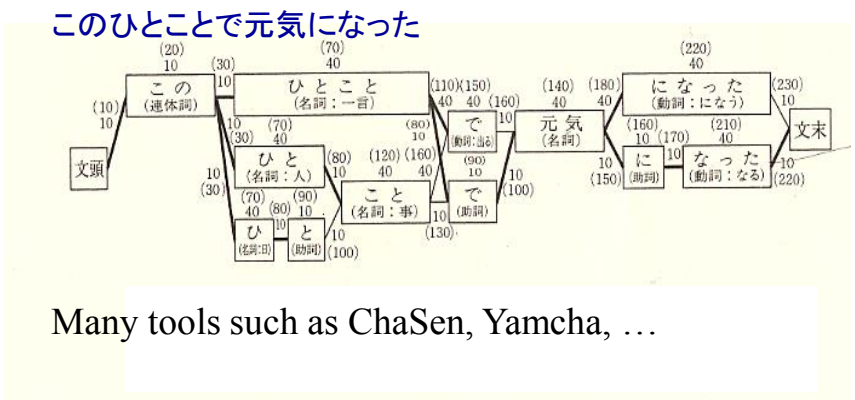


Pages 11-12 from Marie Claire, ECML/PKDD 2005

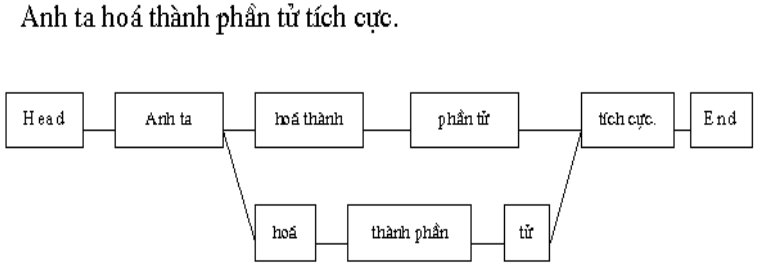
Word Segmentation

- Considering words "nhà cửa", "sắc đẹp", "hiệu sách". They are words in the following sentences:
  a. Nhà cửa bề bộn quá
  b. Cô ấy giữ gìn sắc đẹp.
  c. Ngoài hiệu sách có bán cuốn này

- And they are not words in:
  a. Ở nhà cửa ngõ chẳng đóng gì cả.
  b. Bức này màu sắc đẹp hơn.
  c. Ngoài cửa hiệu sách báo bày la liệt.



Many tools such as ChaSen, Yamcha, …



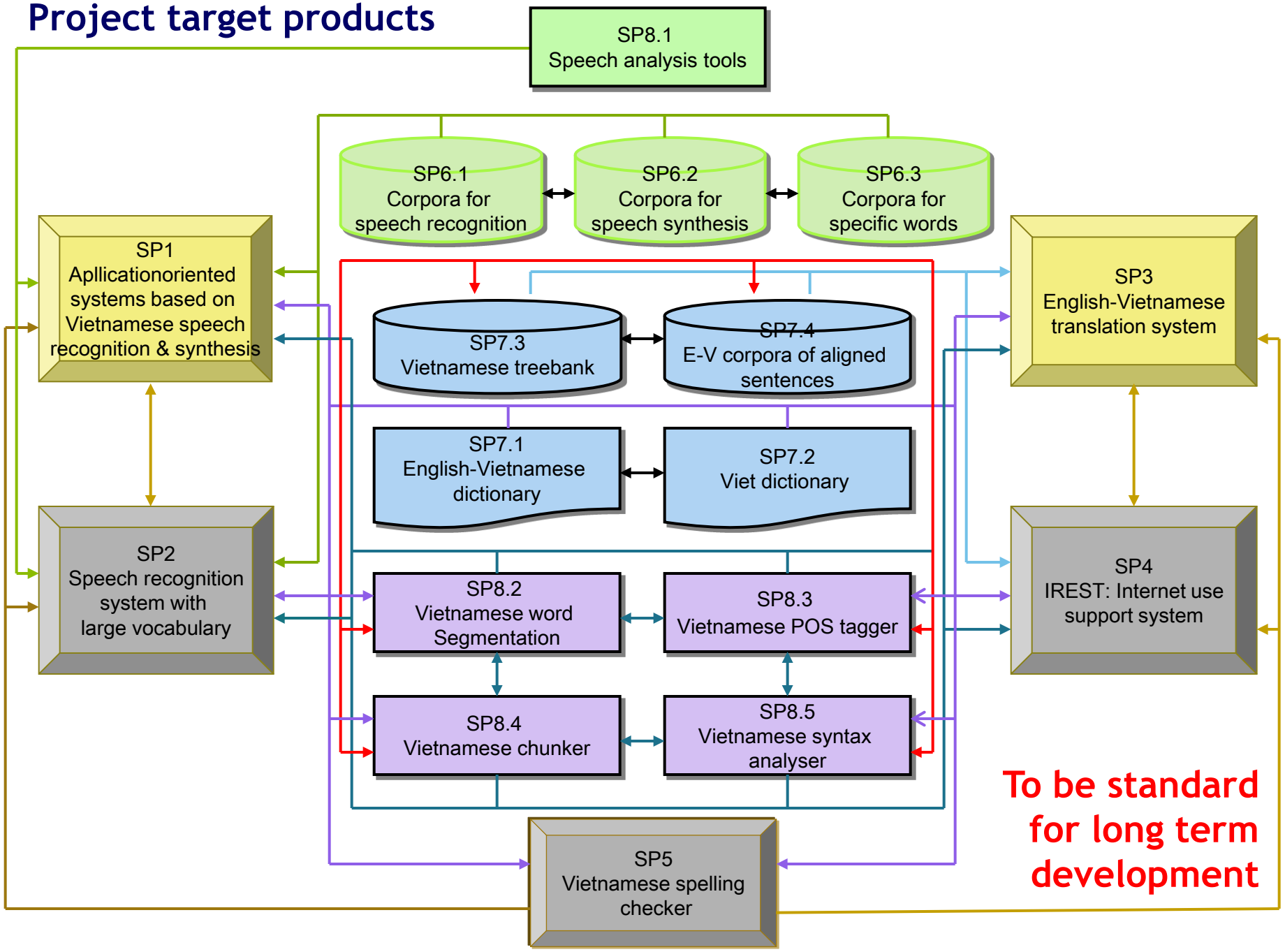
to do such a simple task

Example: Guideline for POS tagging

- 36 word labels in English, from Penn Treebank (1989)
- 30 word labels in Chinese, from Chinese TreeBank (1998)
- 47 word labels in Thai, from Orchid corpus (1997)
- How many for Vietnamese?

	Chinese Treebank	Penn Treebank
Tổng số thẻ	33	36
Danh từ (nouns)	3	4
Danh từ chỉ thời gian (temporal nouns)	NT	NN, NNS
Danh động từ (verbal nouns)	NN	NN, NNS
Danh từ riêng (proper nouns)	NR	NNP, NNPS
Các danh từ khác	NN	NN, NNS
Localizer	1(LC)	0
Đại từ (pronouns)	1(PN)	4(PRP, PRP$, WP, WP$)
Động từ	4	7
Động từ khuyết thiếu (modal)	VV	MD
Các động từ khác	VV, VA, VC, VE	VB, VBD, VBG, VBN, VBP,VBZ
Trạng từ	1(AD)	4(RB, RBR, RBS, WRB)
Giới từ	1(P)	1(IN)
DP-related	4	4
Định từ (Determiner)	DT	DT, WDT, PDT
Số	CD, OD	CD
Measure word	M	-
Liên từ	2 (CC, CS)	2 (CC, IN)
Tiểu từ (particles)	8	0
Khác	8	11
Thán từ	IJ	UH
Sound word (từ tượng thanh??)	ON	-
Punctuation	PU	-

Project target products



Content

- IOIT and International Collaborations
- Vietnamese Language
- VLSP Standard
- Current status of Vietnamese Processing
- Propose idea

NLP tools + resources

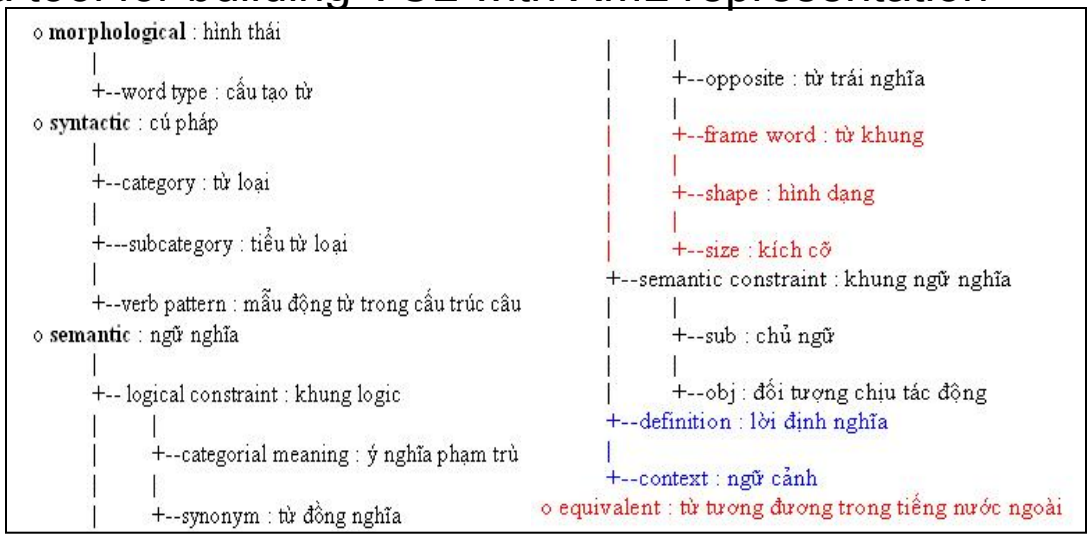
- All the tools: Word segmentation, POS tagging, Chunking, Syntax analysis are constructed based on the same view of words, label assignment, sentences, Viet dictionary and Viet Treebank.
- Using statistical and machine learning methods in building such tools.
- All the tools and resources is given to the R&D community.

Vietnamese WordNet 2012-2015

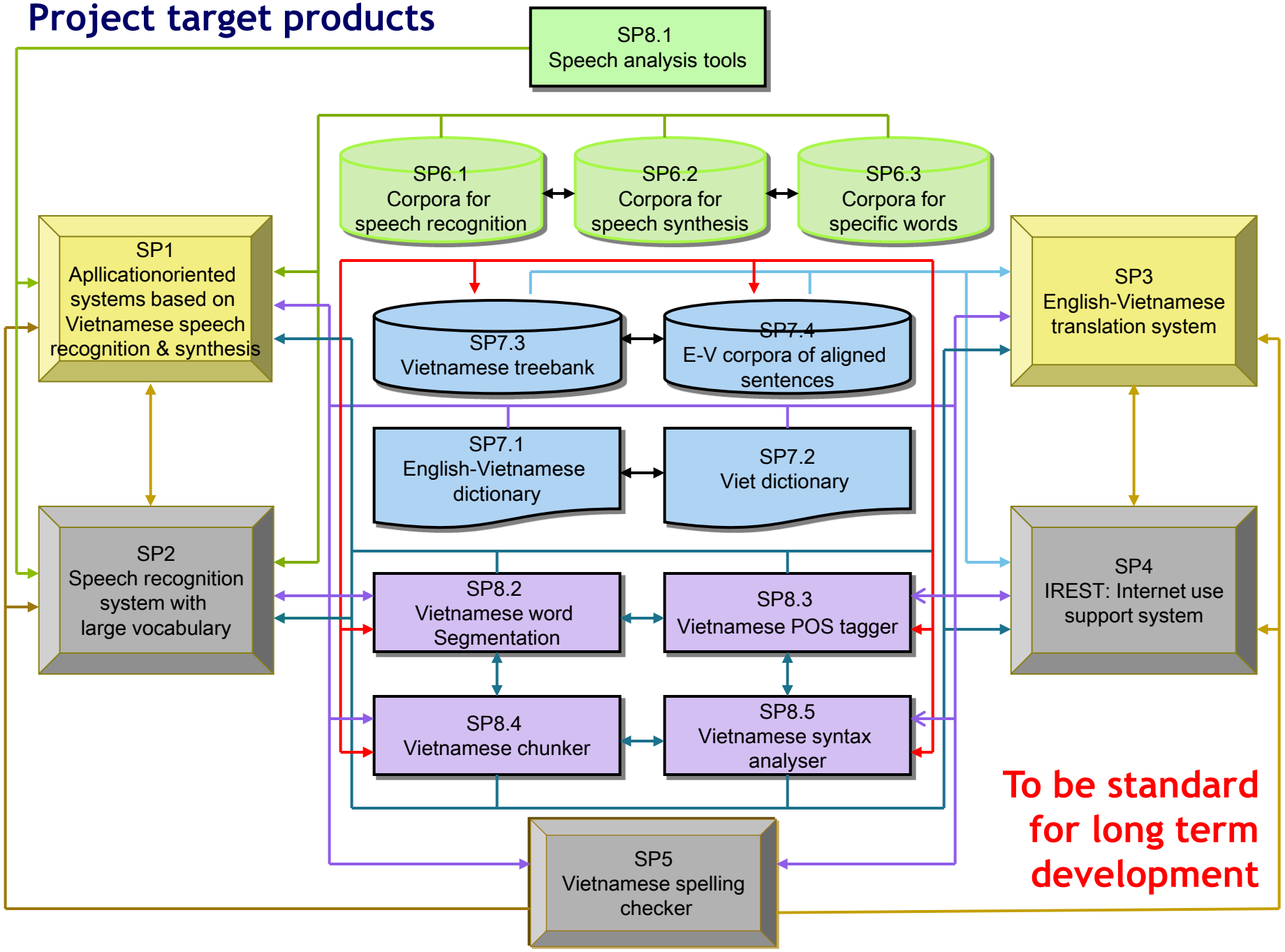
- Developing Vietnamese WordNet with the following features:
  - Vietnamese WordNet with 50.000 words (30.000 popular words and 20.000 domain-based)
  - 30.000 synset
  - Accuracy: 95% for terms in the same synset, 90% in the relationship between different synsets
  - Develop API for WordNet users
  - Develop a tool to access, verify and update
  - Propose guideline for long term WordNet development

NLP Resources

- VietTreebank
  - 10,000 trees; 1,000,000 morphemes
  - Tools: text graphical edit, log and history view, agreement check, search by words, syntactic patterns
- Vietnamese Machine Readable Dictionary
  - Model of VCL (Vietnamese Computational Lexicon) by learning from other language's MRDs with morphological, syntactic and semantic information.
  - 35,000 Vietnamese common used words in modern Vietnamese
  - Develop a tool for building VCL with XML representation

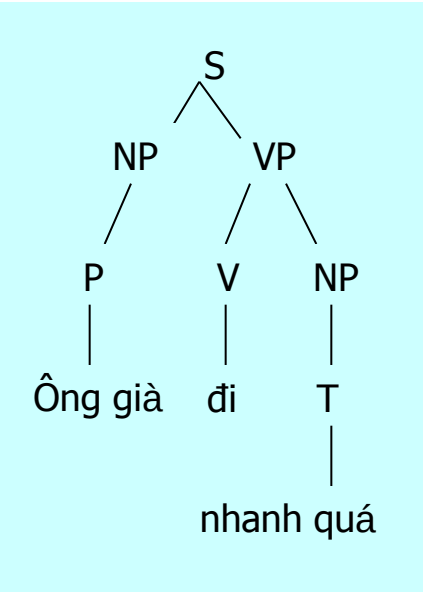


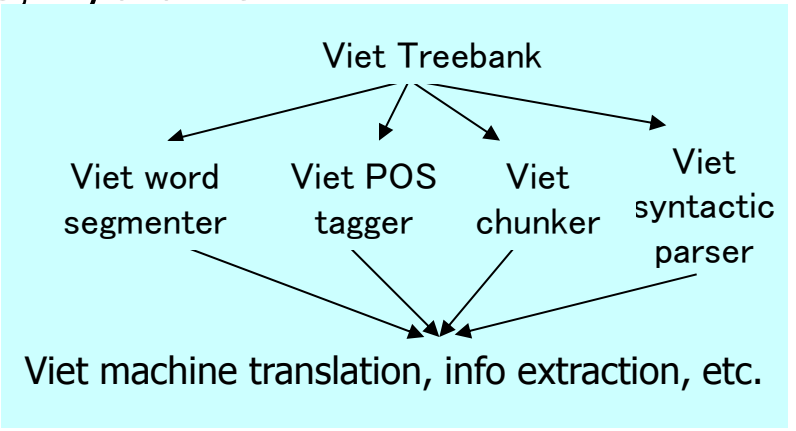
Project target products



SP7.3: Viet Treebank

- A **Treebank** or **parsed corpus** is a text corpus in which each sentence has been parsed, i.e. annotated with syntactic structure.
  - **English**: Penn Treebank (4.5M words) and many others;
  - **Chinese**: Penn Chinese Treebank (507K words), Sinica Treebank (61,087 trees, 361K words);
  - **Japanese**: ATR Dependency corpus, Kyoto Text Corpus, Verbmobil treebanks;
  - **Korean**: Korean Treebank (5078 trees, 54K words)
- **Viet Treebank (2012)**:
  - 10,000 trees
  - 1,000,000 morphemes





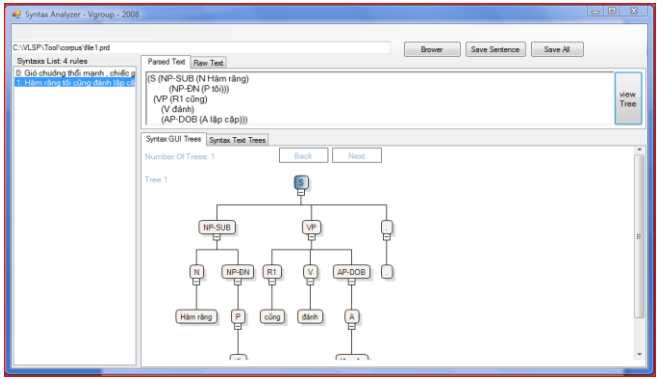
SP7.3: Viet Treebank

- Study various existing treebanks, modern theories for syntax and Vietnamese language
- Build guidelines for word segmentation, POS, and syntax
  - "Nhà cửa bề bộn quá" and "Ở nhà cửa ngõ chẳng đóng gì cả" ("the house is in jumble" and "at home the door is not closed")
  - "Cô ấy giữ gìn sắc đẹp" and "Bức này màu sắc đẹp hơn" (She keeps her beauty" and "this painting has better color")
- Build the tools
- Labeling

Agreement between labelers (95%)

Ví dụ: Hằng ngắm mưa trong công viên.

Người 1	Người 2
(S (NP (Np Hằng)) (VP (V ngắm) (NP (N mưa)) (PP (E trong) (NP (N công viên)))) (. .))	(S (NP (Np Hằng)) (VP (V ngắm) (NP (NP (N mưa)) (PP (E trong) (NP (N công viên))))) (. .))
(1,6,S); (1,1,NP); (2,5,VP); (3,3,NP); (4,5, PP); (5,5,NP)	(1,6,S); (1,1,NP); (2,5,VP); (3,3,NP); *(3,5,NP)*; (4,5, PP); (5,5,NP)



NLP Tools

- Word segmentation
  - Methods: n-gram + dictionary + regular expression
  - 97,1% based on VieTreebank with annotated 220.000 vietnamese words
  - 98,2% based on 100 sentences not included in **VieTreebank**
- POS tagger
  - Methods: MEMs, CRFs
  - Training: 20.000 sentences with POS from **VieTreebank** and VN dictionary
  - 90%
- Syntactic parser 1
  - Method: HPSG grammar
  - P = 82%, R = 74%, F-score = 78% tested on 100 sentences in **VieTreebank**
  - Syntactic parser 2
  - Method: LPCFG, Bikel's implementation
  - F-score = 78% tested on 9600 sentences in **VieTreebank**
- Chunker
  - CRF, online learning on > 9.000 sentences with POS as in **VieTreebank**
  - 94%

Content

- IOIT and International Collaborations
- Vietnamese Language
- VLSP Standard
- Current status of Vietnamese Processing
- Propose idea

We need **Asian Language Treebank**

- ALT is the key resources of most of Asian languages.

Word segmenter	Search engines,
POS tagger	Information retrieval
Chunker	machine translation
Syntactic parser	QA system
….	….

- Can constructs from **multi-lingual corpora** among all **Asian languages** with
  - The same standard of infrastructure
  - The same kind of tool
  - …
- Accelerates research of NLP for Asian languages
  - We have Treebank for English, Japanese, Vietnamese
  - How about Indonesian, Thai, Khmer, Laos, Malay, Myammar, Philippine..