MT Development Experience of Vietnam

VU Tat Thang, Ph.D.
Institute of Information Technology
Vietnamese Academy of Science and Technology
vtthang@ioit.ac.vn

Thang VU

- 2002 ~ 2005: IOIT, Vietnam
  Speech Processing Problems:
  - ⇨ Hybrid model of ANN/HMM for Speech recognition system
  - ⇨ HMM-based approach for Vietnamese LVCSR
  - ⇨ Fujisaki model in Vietnamese synthesis
- 2005 ~ 2008: JAIST, Japan
  - ⇨ The Restoration of bone-conducted speech
- 2008: ATR SLC, Japan
- 2008 ~ 2010: NICT SLC, Japan
  Speech Translation Problem
  - ⇨ Vietnamese LVCSR, Tone recognition
  - ⇨ HMM-based Vietnamese speech synthesis
  - ⇨ Machine Translation (Vietnamese – English)
- 2010 ~: IOIT, Vietnam

- Research/Development in some National Projects:
  2007-2010: *VLSP - Vietnamese language and speech processing*
  2011-2014: *S2s – English-Vietnamese and Vietnamese-English Speech translation in Specific domain*

Outline

- Vietnamese Language
- Some Results in MT from Vietnam
  - ➔ Experience with VLSP Project
  - ➔ Experience with S2s Project

Outline

- Vietnamese Language
- Some Results in MT from Vietnam
  - ➔ Experience with VLSP Project
  - ➔ Experience with S2s Project

54 ethnic groups in Vietnam

Language groups

- Mon-Khmer
- Tay-Thai
- Tibeto-Burman
- Malayo-Polysian
- Kadai
- Mong-Dao
- Han

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

Vietnamese language

- Vietnamese is an analytic language (words are composed of a single morpheme).
  - ➔ ngôn ngữ (analytic), lang-gua-ge (synthetic), 言語 (synthetic)
- Vietnamese does not use morphological marking of case, gender, number, and tense.
  - ➔ Trưa nay tôi ăn ba thằng tôm
- Syntax conforms to Subject Verb Object word order
  - ➔ Cái thằng chồng em nó chẳng ra gì.
    FOCUS CLASSIFIER
    husband I he not turn.out what
    "That husband of mine, he is good for nothing."

VLSP Project 2007-2010

- Objectives:
  1. Basic research on methods for processing Vietnamese language and speech
  2. Build and develop several typical products for VLSP for public end-users.
  3. Build and develop indispensable resources and tools for the VLSP development
- All the tools are constructed based on the same view of words, label assignment, sentences, and resources.
- Using statistical and machine learning methods to build the tools with the corpora.
- Tools and resources are to be given to the public

Typical products

Resources and tools

Computation methods

NLP groups

Group	Experience
National Center for Technology Progress	Rule-based MT -> The only MT commercial systems in Vietnam (EVTRAN3.0, VETRAN3.0)
Univ. of Natural Sciences, VNUHCM	Transfer based MT using Bitext Transfer Learning doing dictionary, bilingual corpus.
HCM Univ. of Technology, VNUHCM	Since 1989 with various trails. SMT since 2002, PBT and phrase extraction from Penn Treebank (since 2003)
JAIST	SMT since 2007, improving the rule-based MT system using statistical techniques.
UNS - VNU Hanoi	Text alignment, biText, tools: POS Tagging, Chunking, Parsing
Lexicography Center	dictionary, corpora.
ColTech–VNU Hanoi	Focus on SMT, and improve the rule-based MT system using statistical techniques.
HUT	Develop tools: POS Tagging, Chunking, Parsing
Danang Univ	Develop tools: spelling, POS Tagging, Chunking, Parsing, dictionary: French-Vietnamese-French (Papillon Project)

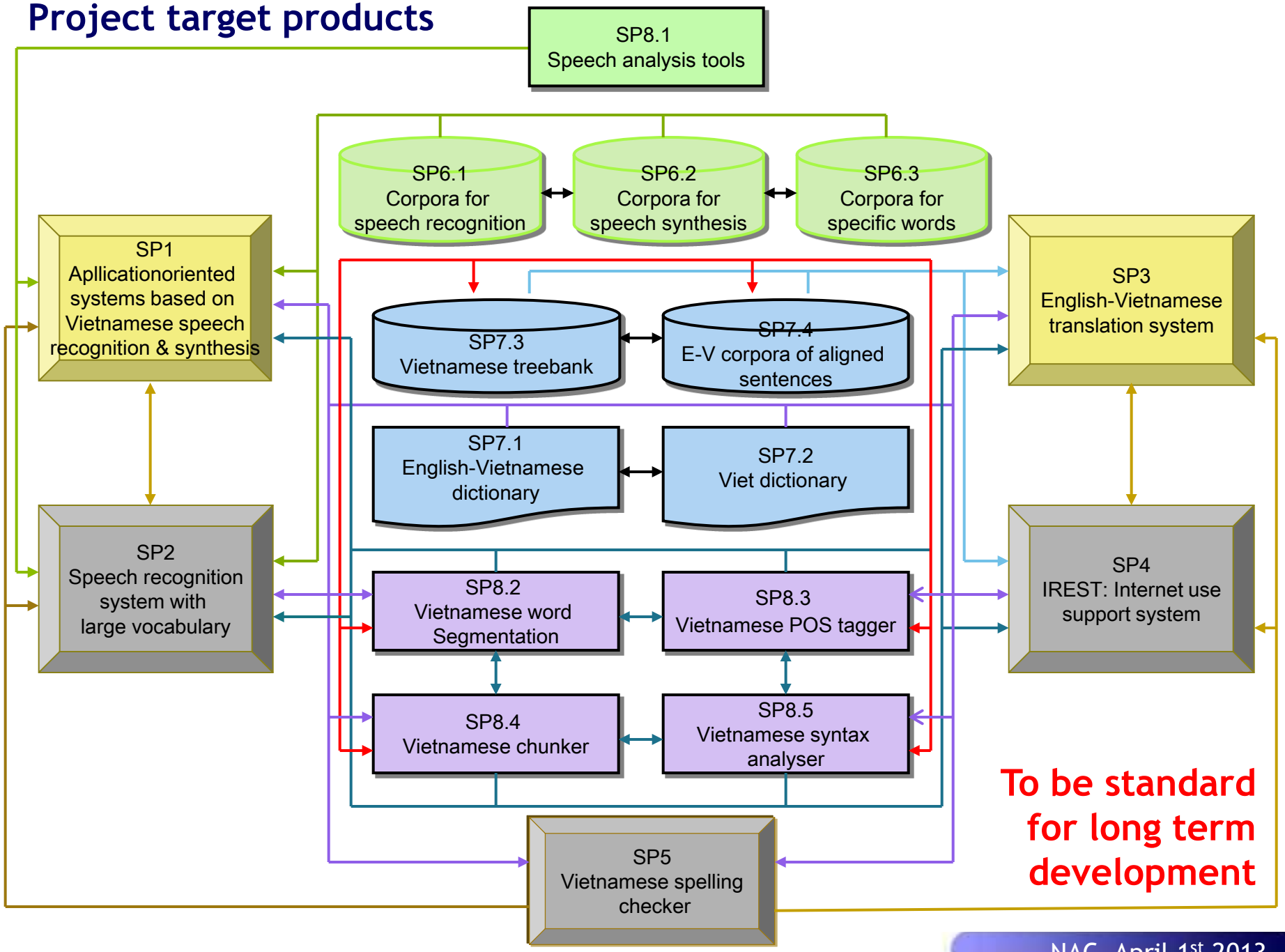
Development of Text Corpora

- VietTreeBank
  - ➔ 10.000 items of fully annotated corpus
    - ⇨ Word Boundary
    - ⇨ POS Tagging
    - ⇨ Syntax Labeling
  - ➔ Text corpus with 1 million syllables with word boundary
  - ➔ Web-based tool for access and updated sentences with POSTaging
- Billingual Corpora: English-Vitenamese
  - ➔ 100.000 sentence pair (including 60.000 parallel sentence pair and 40.000 comparable sentence pair)
- Vietnamese Machine Readable Dictionary
  - ➔ 35.000 items with fully lexical, syntax and semantic information,
  - ➔ Cover all of model Vietnamese Words

Development of NLP Tools

- Word boundary detection
  - ➔ Accuracy about 99%
  - ➔ Text corpus with XML format, boundary lebelling
- POS Tagging
  - ➔ Accuracy >90%
  - ➔ Common rule of POS Tagging with VietTreeBank
  - ➔ Training on 10.000 sentences with labelling
- Chunking
  - ➔ Accuracy >85%
- Syntax Parser
  - ➔ Accuracy >80%

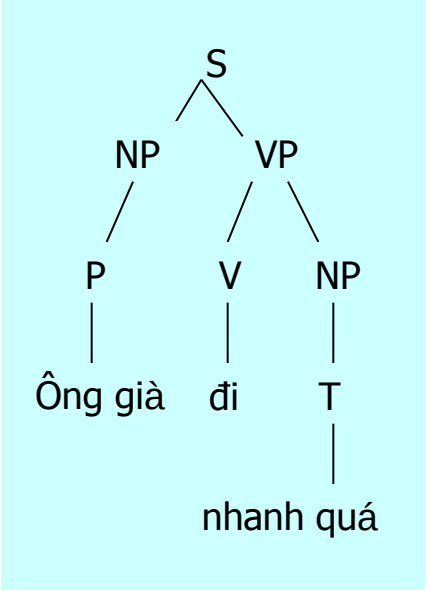
Project target products

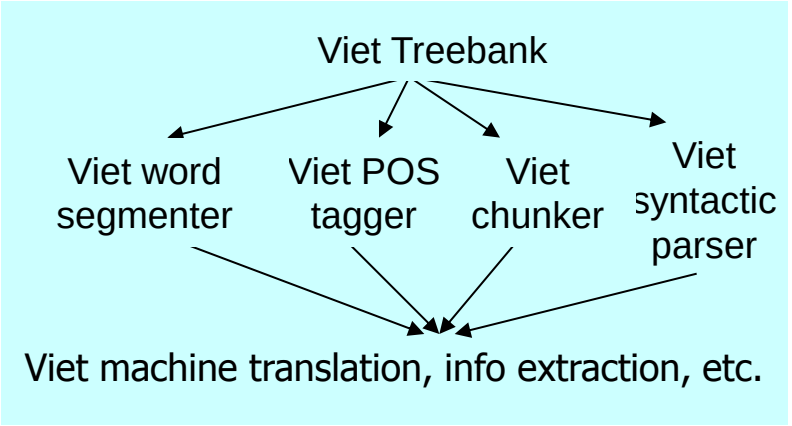


To be standard for long term development

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**:
  - ➔ 10,000 trees
  - ➔ 1,000,000 morphemes





SP7.3: Viet Treebank

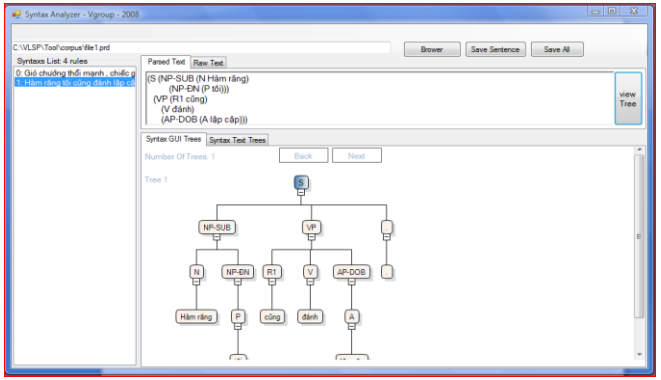
- Study various existing treebanks, modern theories for syntax and Vietnamese language
- Build guidelines for word segmentation, POS, and syntax
  - ➔ "Nhà cửa bề bộn quá" and "Ở nhà cửa ngõ chẳng đóng gì cả" ("the house is in jumble" and "at home the door is not closed")
  - ➔ "Cô ấy giữ gìn sắc đẹp" 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)



Setting up the "standards" for VLSP

- An appropriate view from different research group
- Challenge: Standards for sustainable development
- Guideline for
  - ➔ Words recognition and description: morphological, syntactic, semantic criteria
  - ➔ Label set: noun phrase, verb phrase, clause, …
  - ➔ sentence split
- i.e: 36 word labels in English, from Penn Treebank (1989)

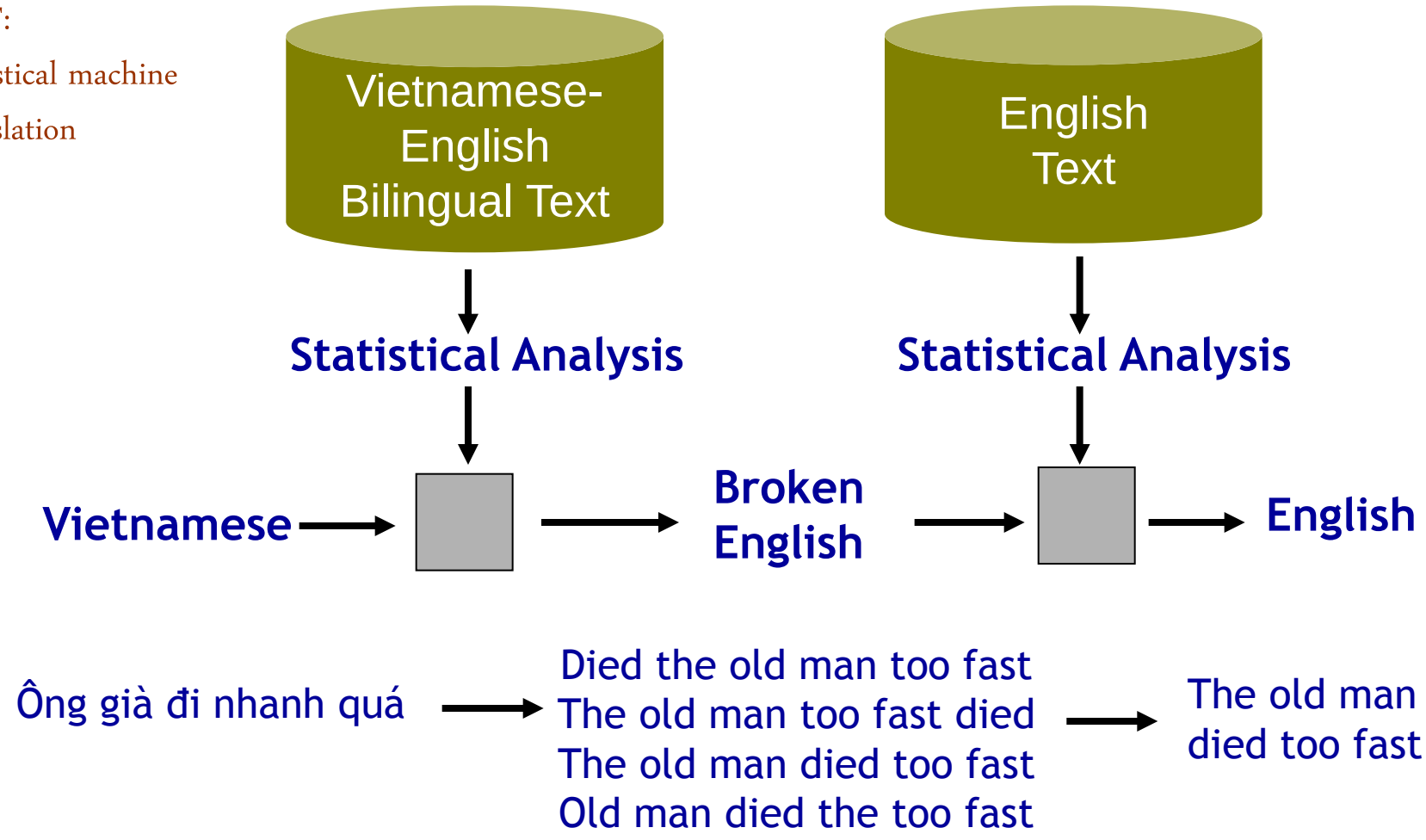
  30 word labels in Chinese, from Chinese TreeBank (1998)

  47 word labels in Thai, from Orchid corpus (1997)

  ➔ How many POS Tag for Vietnamese?

SP3: Machine translation and EVSMT1.0

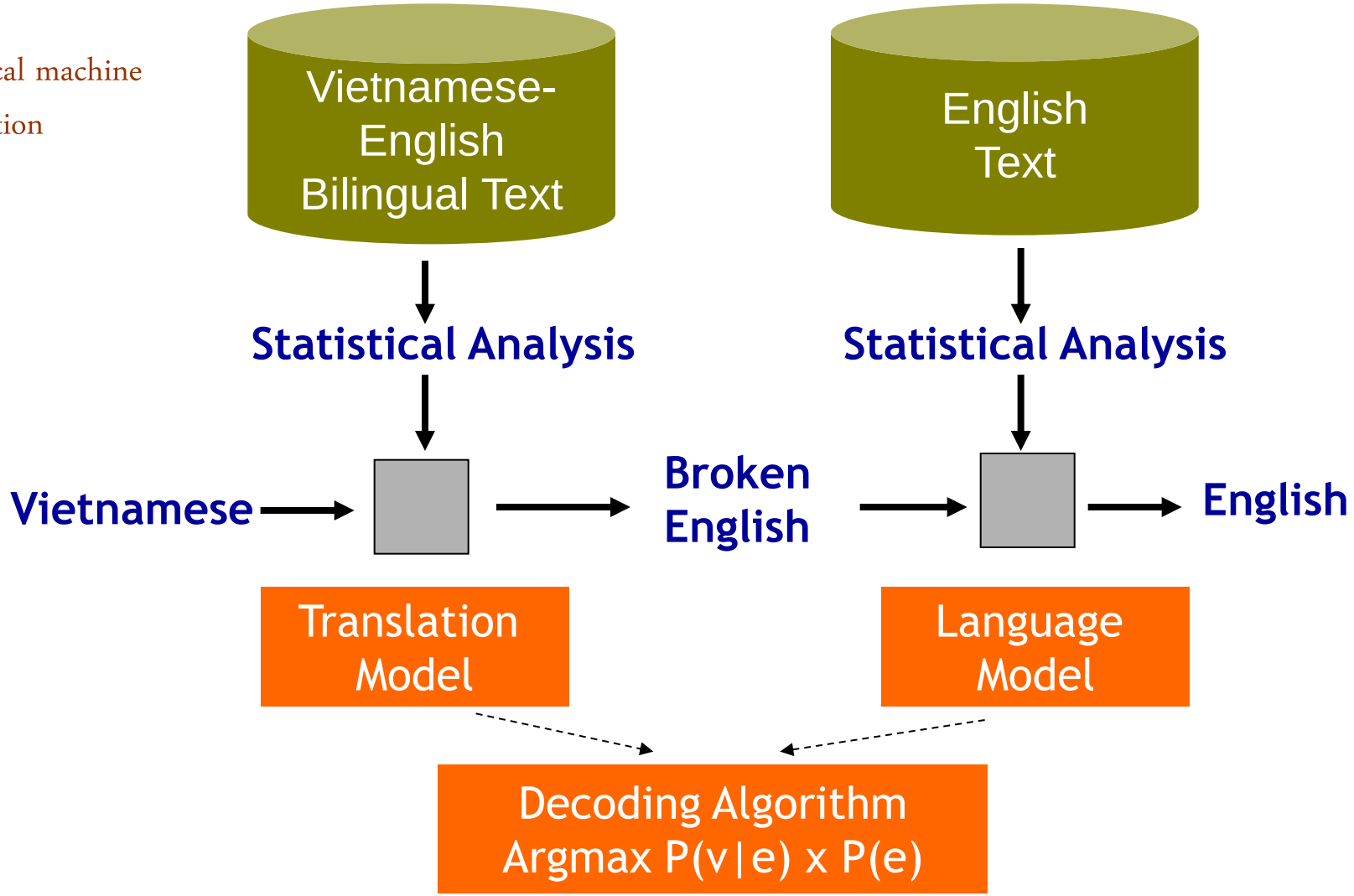
SMT:
statistical machine
translation



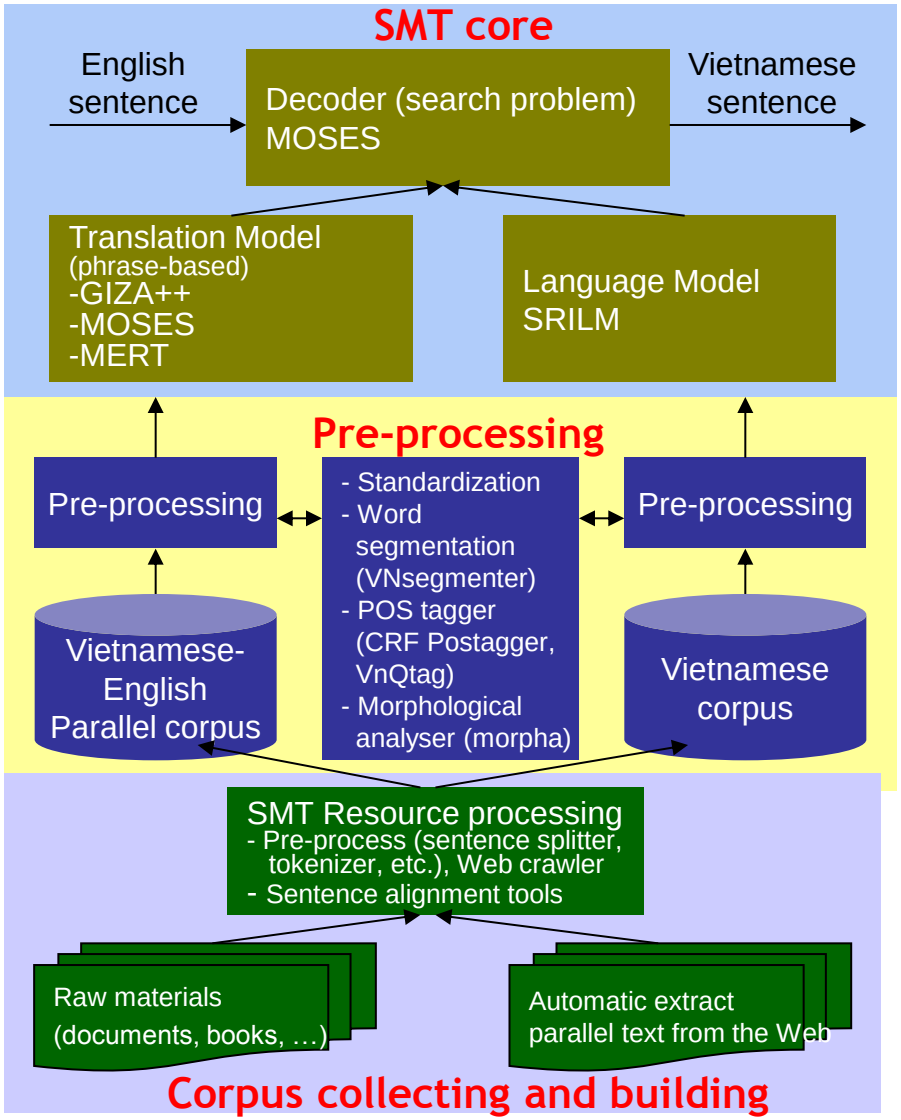
(Slides 31-32 adapted from tutorial on SMT, K. Knight and P. Koehn)

SP3: Machine translation and EVSMT1.0

SMT:
statistical machine translation



SP3: Machine translation and EVSMT1.0



Issues in Vietnamese SMT

- ➔ Corpus building
- ➔ Language Modeling
- ➔ Translation Model
- ➔ Decoder
- ➔ Others

Language	Tokens	Token types
English	342,035	8,128
Vietnamese	285,137	5805

Table 2: The parallel corpus

Models	BLEU score
Baseline phrase-based	0.5826
Words+POS	0.5964
Words+POS+Morphological	**0.6014**

Table 3: Comparison by the Bleu score

Conclusion

- Complete the first phase in VLSP infrastructure.
- Advanced technologies and experience from processing of other languages, especially statistical learning from large corpora.
- Work in collaboration and sharing
- Look for investment from the government and industry for the next phase, and for collaboration.

Thank you !