Open Collaborative Development of the Thai Linguistics Resources

Thatsanee Charoenporn¹, Virach Sornlertlamvanich¹, Sawit Kasuriya²,
Chatchawarn Hansakunbuntheung², Hitoshi Isahara¹
Thai Computational Linguistics Laboratory,
Communications Research Laboratory¹
National Electronics and Computer Technology Center, Thailand²

International Roadmap for Language Resources
ENABLER/ELSNET Workshop
Aug 28-29, 2003, Paris
Open Collaborative Development

- **Collaboration**
  - Thai Computational Linguistics Laboratory, Communications Research Laboratory (CRL)
  - National Electronics and Computer Technology Center (NECTEC), Thailand
  - Communications Research Laboratory (CRL), Japan
  - Advanced Telecommunications Research Institute International, Japan
  - Universities, Thailand:
    Prince of Songkhla Univ., Mahanakorn Univ, Sirindhorn International Institute of Technology.

- **Linguistics Resources**
  - Text Corpora:
    lexicon database, annotated corpus (POS tagged and IR corpus)
  - Speech Corpora
What do we need and what do we have?

Linguistic Resources
- Lexicon/ Dictionary
- Tagged Text/ Speech Corpora
- Language Model

Fundamental Linguistic Tools
- Word extraction
- Word Segmentation/ POS tagger
- Sentence Segmentation
- Grapheme-to-Phoneme Conversion
- Word Sense Disambiguation
- Corpus/ UW Editor

Applications
- Machine Translation (http://come.to/parsit)
- Text Summarization
- Speech Recognition/ Synthesis
Open Linguistics Resources: Lexicon#1

- **Thai Royal Institute Dictionary 1982** (Thai-Thai Dictionary)
  - Basic Term 32,000 entries
  - Technical Term 15,339 entries
  - [http://www.royin.go.th](http://www.royin.go.th)

- **LEXiTRON** (since 1994)
  - Eng-Thai 53,000 entries (Eng entry, Thai equivalent word, pronunciation, head word, index, synonym, antonym, Thai equivalent meaning, Eng sample sentence)
  - Thai-Eng 35,000 entries (Thai entry, Eng equivalent word, pronunciation, synonym, antonym, Thai meaning, Classifier, Thai sample sentence)
  - [http://www.lexitron.nectec.or.th](http://www.lexitron.nectec.or.th)
Open Linguistics Resources: Lexicon#2

LEXiTRON (since 1994)

- Entries
  - 53,000 Eng-Thai
  - 35,000 Thai-Eng

- Information (Field)
  - Equivalent word
  - Pronunciation (phonetics)
  - Root word (for English)
  - POS
  - Synonym – Antonym
  - Sample sentence

- [http://www.lexitron.nectec.or.th](http://www.lexitron.nectec.or.th)
Open Linguistics Resources: Corpus#2

ORCHID (since 1997)

ORCHID POS-Tagged Corpus (supported by CRL)

- 160 documents; 2 MB text; 400K words
- XML tagged for paragraph, sentence, word, POS (14 categories, 47 subcat.)
- http://www.links.nectec.or.th/orchid
Open Linguistics Resources: Speech Corpus

- Word Segmentation
- Sentence Extraction
- POS Tagging
- Grapheme-to-Phoneme

- Raw Text

- Corpus Editor

- XML Tagged Corpus

- Sentence Selection Process

- Speech Recording and Tagging

- Tagged Speech Corpus

- Phonetically-balanced
- Vocabulary coverage
- Prosody-balanced
Speech Corpus
1. NECTEC-ATR Thai Speech Database (2001-2002)

- Thai dialogue speech corpus, the hotel reservation task
  - Isolated word set (DB1)
    - 5,000 words vocabulary (D0-D4)
    - PB words and extra words
      - PB, selected from D0-D4 by minimizing the number of words which cover all Thai phonemes
      - Extra words = 131 words selected from DB3 excluding from D0-D4
  - Phonetic balanced sentences (DB2)
    - 390 phonetics balanced sentences
  - Hotel reservation dialogue (DB3)
    - 50 dialogues of hotel reservation transcription
1. NECTEC-ATR Thai Speech Database (2001-2002)

- Record Environment
  - Quasi-quiet room
- Qualities
  - 20 dB
- 40 speakers (18-40 years old)
  - 20 males and 20 females
Speech Corpus

2. Thai Large Vocabulary continuous speECh Corpus (TLEC) (2001-)

- Text: ORCHID
- Phonetically distributed sentence set
- 5,000 words vocabulary sentence set
- Record Environment
  - Clean speech environment
  - Office environment
- Qualities = 30 dB, 20 dB
- Reading style
- 248 speakers
Speech Corpus
3. Thai Speech Corpus for Speech Synthesis

- A chunk of speech unit candidates for developing a unit selection speech synthesis system
- Build a speech corpus with linguistics tag and acoustic info. For conducting research on Thai reading-style prosodic model.
Future Plan

- TCL-NECTEC IR Corpus
  - Thai for TREC

- TCL Parallel Corpus
  - English-Japanese-Thai Parallel Corpus
  - [http://gutenberg.net](http://gutenberg.net) (Project Gutenberg)
It's Time to Talk About Free Software Again / Bruce Perens

今こそフリーソフトウェアについて再び語るべきときだ！

I'm Bruce Perens.
You may know me as the primary author of the Debian Free Software Guidelines and the Open Source Definition.

I wrote the Electric Fence malloc() debugger, and some pieces of Debian.
And you may remember me for having brought the TIGER map database to free software.

If you want copies of that, you can get them through Dale Schectz <dwarf@polarts.net>.

About a year ago, I sent out a message announcing "Open Source".
Eric Raymond and I founded the Open Source Initiative as a way of introducing the non-hacker world to Free Software.

Well, thanks to Eric, the world noticed.
And now it's time for the second stage:
Discussion

- Language difficulties; Tai-family languages, other families
- Text sources (copyright free texts)
- Common tagset
- Resource Center
- Institutional collaboration