Open Collaborative Development of the Thai Linguistics Resources

Thatsanee Charoenporn[1], Virach Sornlertlamvanich[1], Sawit Kasuriya[2], Chatchawarn Hansakunbuntheung[2], and Hitoshi Isahara[1]
[1]*Thai Computational Linguistics Laboratory, CRL*
[2]*National Electronics and Computer Technology Center*
Thailand Science Park, 112 Phahon Yothin Rd., Klong 1, Klong Luang, Pathumthani 12120
{thatsanee@crl-asia.org, virach@crl-asia.org, sawitk@nectec.or.th, chatchawarnh@nectec.or.th, isahara@crl.go.jp}

ABSTRACT

This paper describes the development of Thai large corpora germinated with an open collaboration platform. We started developing a corpus-based Thai-English lexicon database (LEXiTRON) since 1994. It was originated from a dictionary designed for using in developing a machine translation system. Since then the Thai POS was designed and evaluated in several applications (word segmentation, machine translation, grapheme-to-phoneme, etc.) Extending the lexicon database to together with several tools that have been developed, POS tagged corpus (ORCHID) and speech corpus are developed under the collaboration from several universities and research organizations.

1. INTRODUCTION

A large corpus plays its very essential role when stochastic and learning approaches come to their ages. Many research units put great efforts on developing the corpus for their particular purpose. But a large and complete corpus consumes a lot of man-power, time and budget. Collaboration, therefore, is established for the prompt need of the corpus. ORCHID, a Thai POS-tagged corpus and NECTEC-ATR Speech corpus are the concrete examples of the successful collaborative projects.

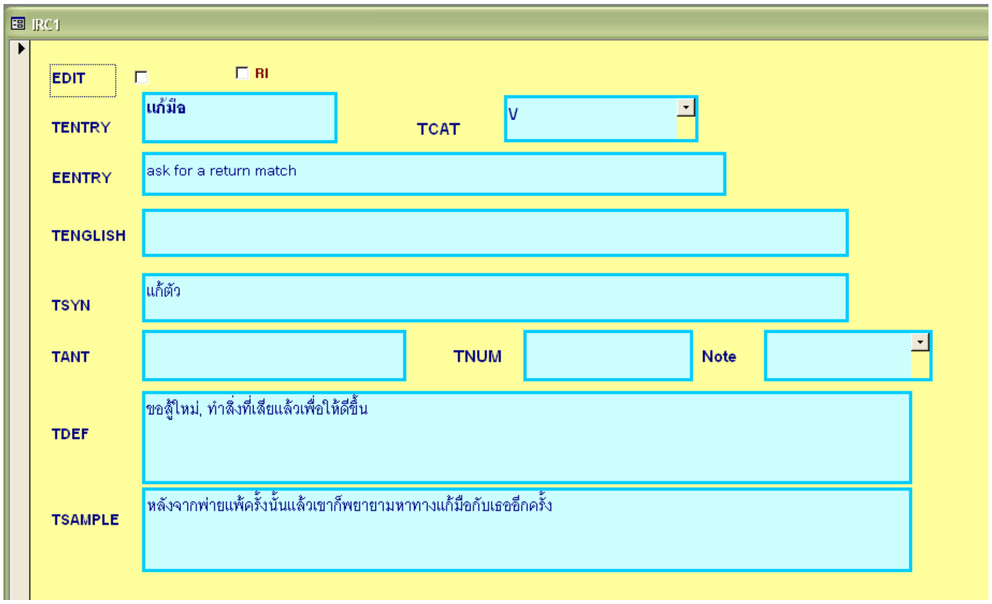
This paper is organized in 2 sections. The development of the text corpus will be described in Section 2. This section includes the steps of development, and design of the POS tagged corpus. Section 3 describes the development of speech corpus using several tools to automate the annotation process.

2. TEXT CORPUS

Text corpus is the collection of a large database in the text level, which includes lexicon database, and annotated text.

2.1 Lexicon database

We started the project of building a lexicon database, LEXiTRON, since 1994. LEXiTRON is the first Thai-English corpus-based dictionary. The words are defined by a set of sample sentences and the usages in addition to their basic information of part-of-speech, classifier, verb pattern, synonym, antonym, and pronunciation. It was aimed to be a dictionary for writing. Most of the lexicons are originated from the dictionary developed for using in the Machine Translation project (the research and development of Multi-lingual Machine Translation System for Asian countries, 1987-1997). It then includes the information and word entry that are suitable for both human and machine use. The first version of LEXiTRON was launched in 1996 as a CD-ROM dictionary for human use. Recently, after a concentrated revision, the second version was launched under the Open Source concept for the contents. It is available in both stand-alone and on-line versions in http://lexitron.nectec.or.th/.

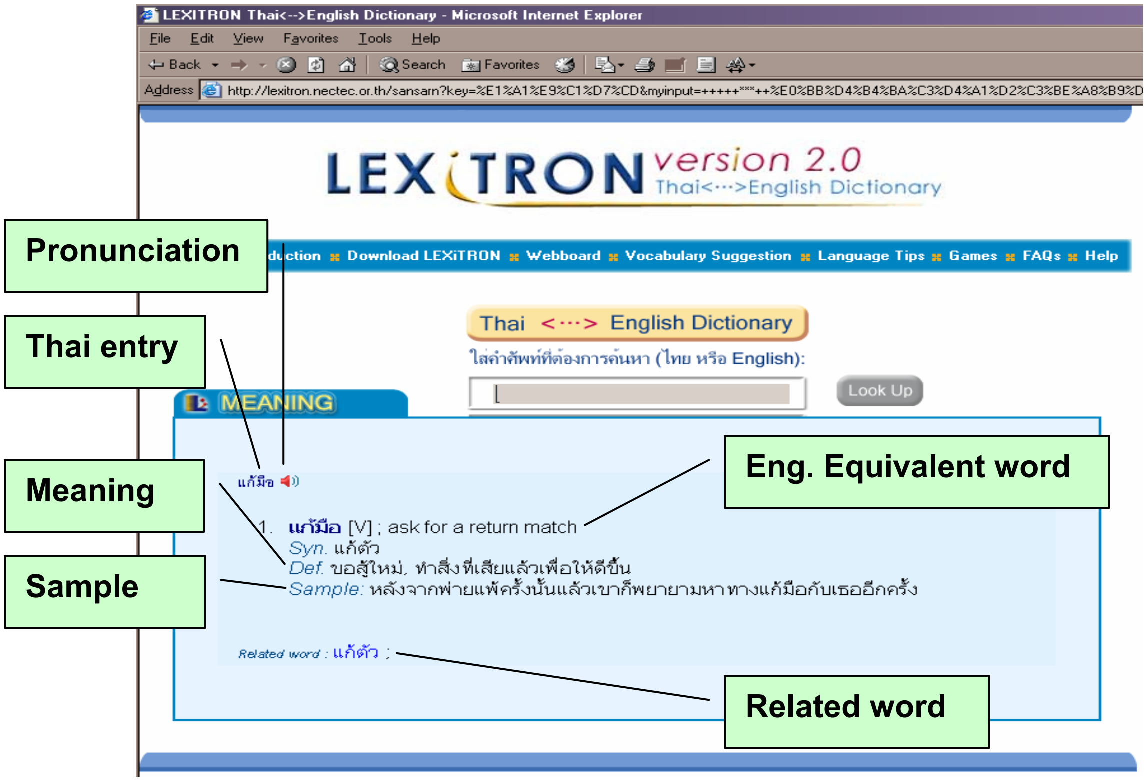


IRC1

EDIT □ □ RI

TENTRY	แก้มือ	TCAT	V		
EENTRY	ask for a return match				
TENGLISH					
TSYN	แก้ตัว				
TANT		TNUM		Note	
TDEF	ขอสู้ใหม่, ทำสิ่งที่เสียแล้วเพื่อให้ดีขึ้น				
TSAMPLE	หลังจากพ่ายแพ้ครั้งนั้นแล้วเขาก็พยายามหาทางแก้มือกับเธออีกครั้ง				

Figure 1 Screen of the coding tools for LEXiTRON database construction



LEXiTRON Thai<-->English Dictionary - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites
Address http://lexitron.nectec.or.th/sansarn?key=%E1%A1%E9%C1%D7%CD&myinput=+++++"""++%E0%BB%D4%B4%BA%C3%D4%A1%D2%C3%BE%A8%B9%D
LEXiTRON version 2.0
Thai<···>English Dictionary
duction Download LEXiTRON Webboard Vocabulary Suggestion Language Tips Games FAQs Help
Thai <···> English Dictionary
ใส่คำศัพท์ที่ต้องการค้นหา (ไทย หรือ English):
Look Up
MEANING
Pronunciation
Thai entry
Meaning
Sample
Eng. Equivalent word
Related word

แก้มือ

1. แก้มือ [V] ; ask for a return match
Syn. แก้ตัว
Def. ขอสู้ใหม่, ทำสิ่งที่เสียแล้วเพื่อให้ดีขึ้น
Sample: หลังจากพ่ายแพ้ครั้งนั้นแล้วเขาก็พยายามหาทางแก้มือกับเธออีกครั้ง

Related word : แก้ตัว ;

Figure 2 LEXiTRON on web: http://lexitron.nectec.or.th/

LEXiTRON is a Thai-English corpus-based lexicon database. The finest unit is a “word” which is extracted from large corpora according to the frequency of its occurrence. It contains 53,000 English entries and 35,000 Thai entries. Each entry or word is assigned its linguistics information as follows,

- English-to-Thai:
  English entry, Thai equivalent word, pronunciation, head word, index for searching, synonym, antonym, Thai equivalent meaning, English sample sentence.
- Thai-to-English:
  Thai entry, English equivalent word, synonym, antonym, Thai meaning, classifier, and Thai sample sentence.

It is noted that the linguistics information (fields) in English-to-Thai and Thai-to-English lexicon databases are different in terms of the language features. However, they are both stored in the same designed format and linked to share the common information.

Figure 1 shows a screen shot of the editing tools that provides a template for editing the field value. Figure 2 shows the online version of LEXiTRON which provides a free service for both English-to-Thai and Thai-to-English word lookup. The online version also provides a link to the wave file for Thai pronunciation which is generated by our Thai text to speech system.

2.2 Annotated corpus

ORCHID is a part-of-speech tagged corpus developed under the collaboration between the Communications Research Laboratory in Japan and NECTEC with the technical support from the Electrotechnical Laboratory in Japan, since 1996.

The original texts are a collection of technical papers of the proceedings of the National Electronics and Computer Technology Center (NECTEC) annual conferences. Each article is annotated in three levels, namely the level of paragraph, sentence and word. Figure 3 shows the procedure in annotating ORCHID. The process of annotation is semi-automatic. Separating paragraph into sentences (sentence segmentation), and post editing are conducted manually, while word segmentation and POS tagging are done automatically.

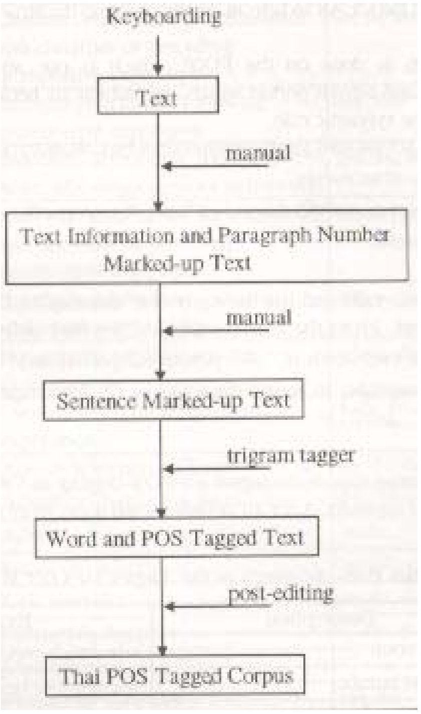


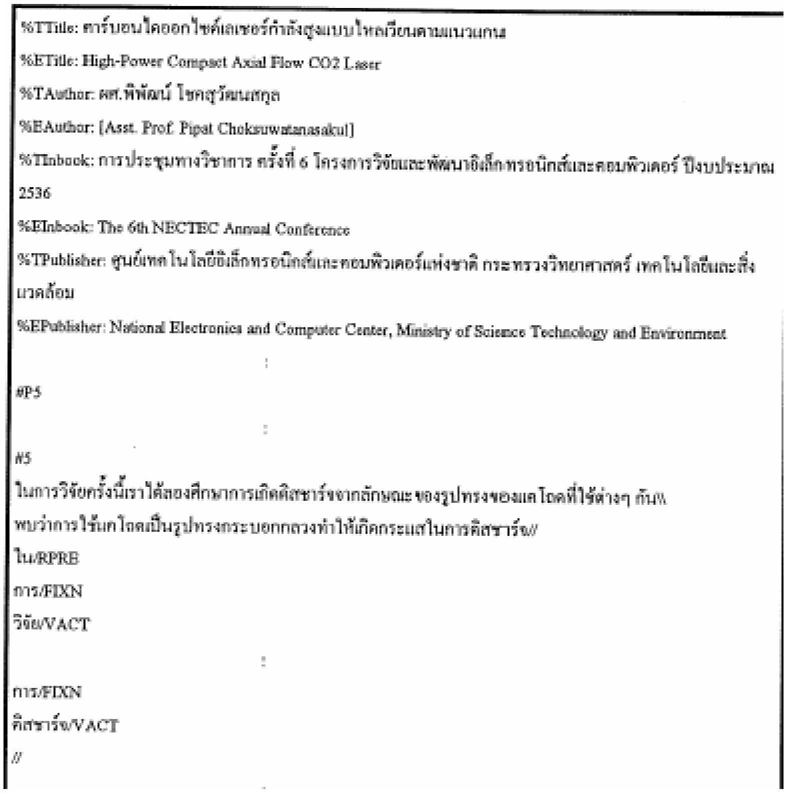
Figure 3 Construction procedure of ORCHID

Each paragraph is manually tagged, from the input text into sentences. Each sentence in a tagged paragraph is then manually tagged with a delimiter. In the word level, word segmentation and POS-tagging processes are conducted automatically by a POS trigram word segmentation, SWATH. The POS set in ORCHID is the one that was designed for developing the MMT system. It consists of 14 categories with 47 subcategories, as shown in Appendix.

Figure 4 shows a sample of the annotated text of the ORCHID corpus. ORCHID is now available on http://links.nectec.or.th/orchid. It is designed to follow the following tag description.

Table 1 Tag set in ORCHID POS tagged corpus and the description

Tag	Description
%TTitle:	Title written in Thai
%ETitle:	Title written in English
%TAuthor:	Authors' name written in Thai
%EAuthor:	Authors' name written in English
%TInbook:	Book name of the article written in Thai
%EInbook:	Book name of the article written in English
%TPublisher:	Publisher's name written in Thai
%EPublisher:	Publisher's name written in English
#Pn	Paragraph number counted from the beginning to the end of the article
#n	Sentence number counted from the beginning to the end of the article
\\	Line break within a sentence
//	Sentence end
word/POS	Word, delimiter ("/") and the corresponding POS



```
%TTitle: คาร์บอนไดออกไซด์เลเซอร์กำลังสูงแบบไหลเวียนตามแนวแกน
%ETitle: High-Power Compact Axial Flow CO2 Laser
%TAuthor: ผศ.พิพัฒน์ โชคสุวัฒนสกุล
%EAuthor: [Asst. Prof. Pipat Choksuwatanasakul]
%TInbook: การประชุมทางวิชาการ ครั้งที่ 6 โครงการวิจัยและพัฒนาอิเล็กทรอนิกส์และคอมพิวเตอร์ ปีงบประมาณ
2536
%EInbook: The 6th NECTEC Annual Conference
%TPublisher: ศูนย์เทคโนโลยีอิเล็กทรอนิกส์และคอมพิวเตอร์แห่งชาติ กระทรวงวิทยาศาสตร์ เทคโนโลยีและสิ่ง
แวดล้อม
%EPublisher: National Electronics and Computer Center, Ministry of Science Technology and Environment
                              :
#P5
                              :
#5
ในการวิจัยครั้งนี้เราได้ลองศึกษาการเกิดดิสชาร์จจากลักษณะ ของรูปทรงของแคโถดที่ใช้ต่างๆ กัน\\
พบว่าการใช้แคโถดเป็นรูปทรงกระบอกกลวงทำให้เกิดกระแสในการดิสชาร์จ//
ใน/RPRE
การ/FIXN
วิจัย/VACT
                              :
การ/FIXN
ดิสชาร์จ/VACT
//
```

Figure 4 Sample of the annotated text in the ORCHID Corpus

3. SPEECH CORPUS

To develop a corpus for using in the research of Thai speech recognition and synthesis, we called for collaboration from universities and research organizations to produce a large corpus. Language and acoustic models are needed in speech recognition while the models of prosody and continuous speech unit are needed in natural speech synthesis research. This speech corpus design is aimed to support both speech recognition and synthesis research. Currently there are 2 parallel projects running for NECTEC-ATR Thai Speech Database development and Thai Large vocabulary continuous speech Corpus (TLEC) development.

3.1 NECTEC-ATR Thai Speech Database (2001-2002)

This project is the collaboration between NECTEC and ATR to develop a Thai dialogue speech corpus based on the hotel reservation task. The database consists of three sets, namely the isolated word set (DB1), the phonetic balanced sentences (DB2), and the hotel reservation dialogues (DB3).

DB1: The isolated word set contains two subsets as follows.

- *5,000-words vocabulary (D0-D4):* The words in this subset are selected from the most frequently used 5,000 words. They are collected from articles in magazines, journals, and daily news. This subset includes four minor subsets. Each minor subset consists of 1,000 words.

- *PB words and extra-words (D5):* This subset contains two minor subsets. The first minor subset is the PB words set that are selected from the set of 5,000 words by minimizing the number of words. Those words cover all Thai phonemes occurred in the set of 5,000 words and are balanced by the occurrence of each phoneme. The second minor subset contains the extra words. In this set, there are 131 words that are selected from the vocabularies in DB3 which are not included in the set of 5,000 words, such as the name of credit card companies, currencies, the types of room in hotel, etc. Consequently, this corpus consists of 5,131 words.

DB2: The set of 390 phonetic balanced sentences that are selected to cover all possible Thai phonemic units in the minimum number of the sentences. They are limited by text corpus and the domain. In general the phonemic model is defined in three types, namely monophone models, biphone models, and triphone models. In this project, the biphone model is adopted.

DB3: The set 50 dialogues of hotel reservation are transcribed from ATR English transcription, Hotel Reservation Transcription (HRT) set. Each dialogue is uttered by two speakers, namely clerk and customer.

All utterances are recorded in quasi-quiet room. The qualities of them are around 20 dB. A number of speakers are 20 males and 20 females (18 to 40 years old).

Procedure

DB1 (Isolated word set)

This set consists of three subsets. We divided the first subset into five minor subsets (D0 to D4). Each minor subset contains 1,000 words. The other two subsets are the phonetic balanced words and the extra words. The details of each subset creation are described in the followings.

- *5,000-words vocabulary subset (D0-D4):* We counted the frequency of each vocabulary from the text corpus (Thai magazines, journals, and encyclopaedias) and the most frequently used 5,000 words were selected. Then they are randomly divided into five minor subsets.

- *PB word subset (D5):* Using the 5,000 words to select the PB words. All phonemes with at least amount of words and balanced occurrence were collected. Hence, the phoneme occurrences in this subset equal to the phoneme occurrences in the 5,000 words subset. A number of words in this subset are 640. Furthermore, this selection procedure is also used in PB sentence selection.

- *Extra word subset (D5):* The words that occurred in Hotel Reservation Transcription (HRT) set and did not occur in 5,000 vocabularies subset or PB word subset, is called "Extra words". It contains 131 words.

DB2 (Phonetic balanced sentence set)
This set is the collection of the sentences that contains the whole set of Thai biphones. A large text corpus is required in order to extract the set of biphone. However, the current text corpus is not large enough to cover all Thai biphones.

DB3 (Hotel Reservation Transcriptions, HRT)
A hotel reservation system is the one of well-known speech recognition application. Developing the corpus is the major problem in this application because there are many speaking styles, several different dialogues, many types of hotel reservation procedure. The transcriptions used in this project are translated from the set 50 dialogues in HRT of the Spoken Language Translation Research Laboratories (SLT), Advanced Telecommunication Research International Institute (ATR), Kyoto, Japan. The dialogues have been translated to more than two languages such as English and Japanese.

3.2 Thai Large vocabulary continuous speECh Corpus (TLEC) (2001-)

TLEC is the collaborative project between NECTEC, Faculty of Electrical Engineering, Mahanakorn University of Technology (MUT), and Faculty of Computer Engineering, Prince of Songkhla University. The contents of this corpus consist of two sets, namely (1) the phonetically distributed (PD) sentence set and (2) 5,000-words vocabulary sentence set.

Table 1 Summary of phonetically distributed sentence set

Attribute	PD set
No. of sentences	802
No. of vocabularies	2,269
No. of words	7,847
No. of syllables	12,702
No. of phonemes	38,106

Table 2 Summary of 5,000-words vocabulary sentence set

Attribute	TR set	DT set	ET set
No. of sentences	3,007	500	500
No. of vocabularies	5,000	1,622	1,630
No. of words	55,504	8,076	8,290
Difference from TR	0	3,378	3,370
Difference from DT	0	0	609
Difference from ET	0	617	0

The utterances are recorded in two environments: the clean speech environment (CS) and the office environment (OF). These environments are separated by the signal to noise ratio (SNR) Moreover, the SNR of CS and OF are around 30 dB and 20 dB respectively. All utterances are recorded according to reading styles. A number of speakers are 248 speakers (PSU: 100, MUT: 100, and NECTEC: 48).

Procedure:

(1) Phonetically distributed sentence (PD) set

To initial acoustic model efficiently, phonetically balanced sentences (PB) is usually used for training. PB is the smallest set of sentences covering all phonemic units in the language. In our case, the phonemic unit is biplone. PD is the extension of PB. It does not only cover all biphone, but the text distribution is also similar to the daily used context (ORCHID corpus in this case).

The PD selection process starts from PB construction. In PB construction process, the sentence containing mostly unselected biphone is chosen one by one until all biphones are included in the PB set. Before constructing PD set, the biphone distributions of ORCHID are calculated. Then, some sentences are added to PB to change the distribution as same as distributions of ORCHID. The number of adding sentences should be kept at minimum while the biphone distribution of PD set is closest to ORCHID's distribution.

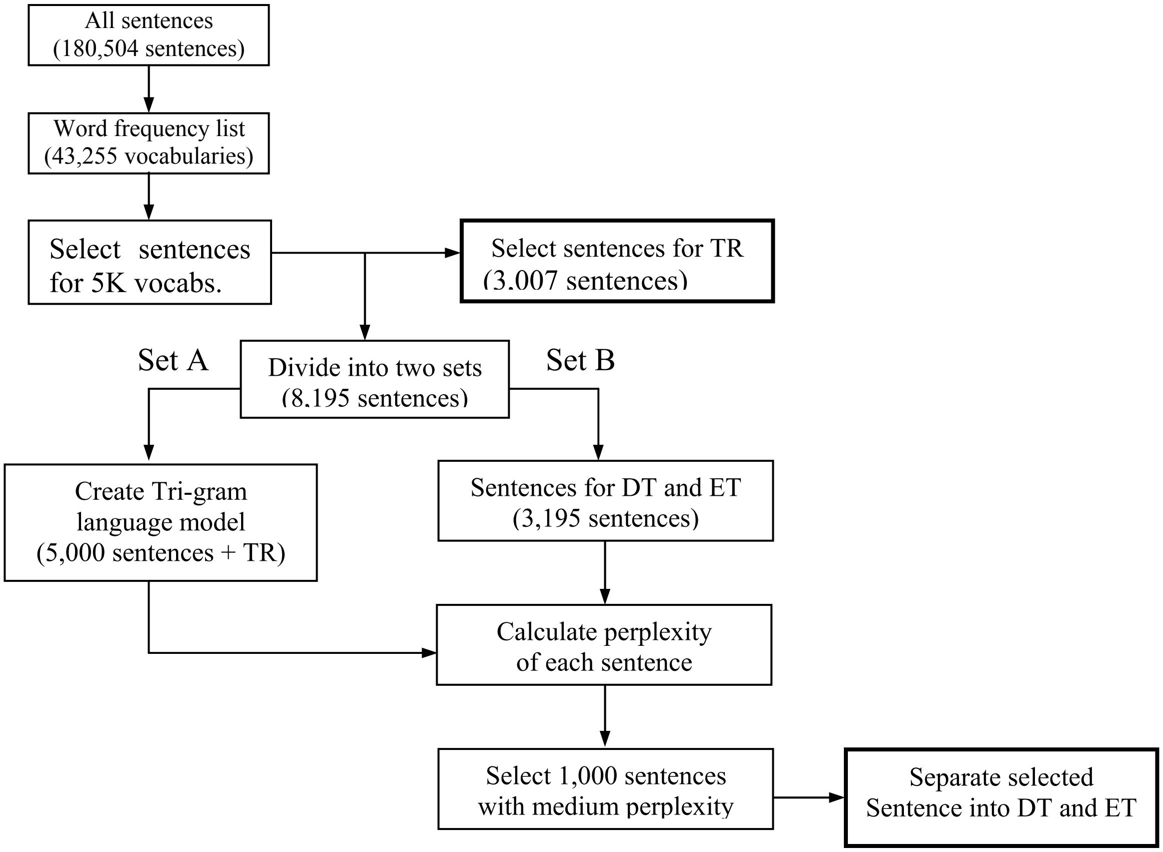


Figure 5 TR, DT and ET selection

(2) 5,000-words vocabulary set

The objective of this set is to collect the structure of Thai language for language model (LM) construction. This set is divided into three subsets: the training set (TR), the development test set (DT), and the evaluation test set (ET). The TR set is used to train language models. The DT and ET sets are used for testing in development and evaluation phases respectively.

The process of TR, DT, and ET selections are illustrated in Figure 5. Firstly, the words of all sentences are listed and sorted. There are 43,255 vocabularies. The sentences containing the first 5,000 vocabularies that most frequently occurring, are selected. These sentences (11,202 sentences) are chosen to the next step. The TR set (3,007 sentences) is selected by collecting the minimum amount of sentences that pertains 5,000 vocabularies. The remaining sentences are divided into two sets: set A and B, for language model construction (5,000 sentences) and DT, ET selection (3,195 sentences), consecutively. In addition, the set B is selected by calculating the sentence scores of each sentences and choosing the 3,195 sentences that are the highest sentence scores. On the other set, the tri-gram language model is created by 5,000 sentences and 3,007 sentences (TR set). There are 8,007 sentences that use for LM construction. And LM is used for calculating the perplexity of each sentence in set B. In the next procedure, the 1,000 sentences that have the medium perplexity (around 100 to 300), are selected and randomly divided into DT and ET sets.

3.3 NECTEC's Thai Speech Corpus for Speech Synthesis

The aims of this corpus is (1) to construct a chunk of speech unit candidates for developing a unit selection speech synthesis system and (2) to build a speech corpus with linguistic tags and acoustic information for conducting research on Thai reading-style prosodic model.

1. Specification

- 5,200 sentences collected from ORCHID Thai POS tagged corpus
- Text corpus in multilevel XML format: document, paragraph, sentence and word levels
- Linguistic tag and acoustic information: POS, Tone, Phone boundary, Syllable boundary, Word boundary, Phrase boundary, syllable position in word and phrase, Energy, F0, Duration, Voiced/unvoiced region, tone/toneless region
- Covering of tri-phone, tri-vowel and tri-tone unit
- Covering entire Thai and foreign (loaned) phones
- 14-hour speech sound of one female voice with standard Thai accent
- Recording environment:
    - 44kHz sampling rate, 16 bit per samples
    - SNR > 46 dB (silent room)
    - Using DAT

2. Development Procedure

The procedure in developing the corpus is divided into two parts as the followings:

2.1 Text selection

- Convert grapheme to phoneme transcription using Probabilistic GLR parser
- Define tri-phone, tri-vowel and tri-tone unit
- Score each units using occurring probabilities
- Select sentences that cover the defined unit using greedy algorithm

2.2 Speech Tagging

- Use automatic phone segmentation tool based on HTK
- Revise the phone marks and insert phrase marks by linguists
- Mark syllable boundaries automatically using phone labels
- Locate syllable position in words and phrases using phone labels, word marks and phrase marks
- Mark voiced/unvoiced region
- Mark tone/toneless region using voiced/unvoiced marks and phone labels
- Extract energy and F0 curve

CONCLUSION

For years, we spend a lot of efforts to develop the linguistics resources. At the time the corpora are complete and available, it will be manage to support the research in Thai NLP and speech technology research. However, many aspects relating to the corpus construction should be developed for increasing its capacity in NLP researches such as the size, the variety of the raw data, the annotated tag, tools and so on. We are planning to widen the coverage and the collaboration to fulfill the need. Should the standard for data exchange be designed and shared among the research communities in the near future.

REFERENCE

1. Chatchawarn Hansakunbuntheung, Virongrong Tesprasit, Virach Sornlertlamvanich. *Thai Tagged Speech Corpus for Speech Synthesis*, O-COCOSDA 2003. (to be published)
2. Pongthai Tarsaku, Virach Sornlertlamvanich and Rachod Thongprasirt. *Thai Grapheme-to-Phoneme Using Probabilistic GLR Parser*, Eurospeech, Vol. 2, pp. 1057-1060, 2001.
3. Pornpimon Palingoon, Pornchan Chantanapraiwan, Supranee Theerawattanasuk, Thatsanee Charoenporn and Virach Sornlertlamvanich. *Qualitative and Quantitative Approaches in Bilingual Corpus-Based Dictionary*. The Fifth Symposium on Natural Language Processing 2002 & Oriental COCOSDA Workshop 2002 (SNLP-O-COCOSDA 2002), Hua Hin, Thailand, pp 152-158, May 2002.
4. Rachod Thongprasirt, Virach Sornlertlamvanich, Patcharikra Cotsomrong, Sinaporn Subevisai and Supphanat Kanokphara. *Progress Report on Corpus Development and Speech Technology in Thailand* . The Fifth Symposium on Natural Language Processing 2002 & Oriental COCOSDA Workshop 2002 (SNLP-O-COCOSDA 2002), Hua Hin, Thailand, pp 300-306, May 2002.
5. Supphanat Kanokprara, Virongrong Testprasit, and Rachod Thongprasirt. Pronunciation Variation Speech Recognition without Dictionary Modification on Sparse Database, International Conference of Acoustic and Speech Signal Processing, 2003.
6. Sawit Kasuriya, Virach Sornlertlamvanich, Patcharika Cotsomrong, Takatoshi Jitsuhiro, Genichiro Kikui and Yoshinori Sagisaka. Thai Speech Database for Speech Recognition (NECTEC-ATR Thai Speech Database), O-COCOSDA, 2003. (to be published)
7. Thatsanee Charoenporn, Virach Sornlertlamvanich and Hitoshi Isahara. *Building A Large Thai Text Corpus---Part-Of-Speech Tagged Corpus: ORCHID---.* Proceedings of the Natural Language Processing Pacific Rim Symposium, 1997.
8. Virach Sornlertlamvanich, Naoto Takahashi and Hitoshi Isahara. *Building a Thai Part-Of-Speech Tagged Corpus (ORCHID)*. The Journal of the Acoustical Society of Japan (E), Vol.20, No.3, pp 189-140, May 1999.

9. Virach Sornlertlamvanich and Rachod Thongpresirt. *Speech Technology and Corpus Development in Thailand,* Proceedings of O-COCOSDA2001, Korea, Aug 2001.

APPENDIX

Part-of-speech and tag set of ORCHID

TYPE	ORCHID TAG	EXAMPLE
Proper noun	NPRP	โคโรน่า (Corona)
Common noun	NCMN	หนังสือ (book), อาหาร (food)
Cardinal numeral noun	NCNM	หนึ่ง (one), สอง (two)
Ordinal numeral noun	NONM	ที่หนึ่ง (first), ที่สาม (third)
Label noun	NLBL	ก, ข, …1, 2,…
Title noun	NTTL	ดร. (Dr.), นาย (Mr.)
Active verb	VACT	ทำงาน (work), ร้องเพลง (sing), กิน (eat)
Stative verb	VSTA	เห็น (see), รู้ (know), คือ (be)
Attributive verb	VATT	อ้วน (fat), ดี (good), สวย (beautiful)
Adverb with normal form	ADVN	เก่ง, เร็ว
Adverb with iterative form	ADVI	เร็วๆ, ช้าๆ, เสมอๆ
Adverb with prefixed form	ADVP	โดยเร็ว, อย่างเชื่องช้า
sentence modifier	ADVS	โดยปกติ, ตามธรรมดา
Personal pronoun	PPRS	คุณ (you), เขา (he), ฉัน (I)
Demonstrative pronoun	PDMN	นั่น (that), ทั้งหมด (all),บ้าง (some)
Interrogative pronoun	PNTR	ใคร (who), อะไร (what), อย่างไร (how)
Relative pronoun	PREL	ที่, ซึ่ง, อัน (that, which)
Pre-verb auxiliary, before negator	XVBM	ฝนเกิดไม่ตก, การบ้านเกือบเสร็จแล้ว, คนไข้กำลังหลับ.
Pre-verb auxiliary, after negator	XVAM	เขาไม่ค่อยมาที่นี่, บ้านนี้น่าอยู่, เราได้เห็นฝีมือเขาแล้ว.
Pre-verb auxiliary, before or after negator	XVMM	เธอ (ไม่) ควรไปพบเขา or เราควร (ไม่) พูดเลยวันนี้, เล็ก (ไม่) เคยเอาใจใส่เรา or เล็กเคย (ไม่) เอาใจใส่เรา, เรา (ไม่) ต้องบอกเขาก่อน or เราต้อง (ไม่) บอกเขาก่อน.
Post-verb auxiliary	XVAE	แก้วแตกไปหลายใบ or เด็กกินไปเล่นไป, ฉันเห็นมากับตา or เขาพักที่นี่มาหลายวันแล้ว, ยกมือขึ้น.
Pre-verb auxiliary in imperative mood	XVBB	กรุณา, จง, เชิญ (please), อย่า, ห้าม (don’t).
Definite determiner, after noun without classifier in between	DDAN	นี่, นั่น, โน่น, ทั้งหมด
Definite determiner, allowing classifier in between	DDAC	นี้, นั้น, โน้น, นู้น

Definite determiner, between noun and classifier or preceding quantitative expression	DDBQ	ทั้ง, อีก, เพียง
Definite determiner, following quantitative expression	DDAQ	พอดี, ถ้วน
Indefinite determiner, following noun; allowing classifier in between	DIAC	ไหน, อื่น, ต่างๆ
Indefinite determiner, between noun and classifier or preceding quantitative expression	DIBQ	บาง, ประมาณ, เกือบ
Indefinite determiner, following quantitative expression	DIAQ	กว่า, เศษ
Determiner, cardinal number expression	DCNM	หนึ่งคน, เสือ 2 ตัว
Determiner, ordinal number expression	DONM	ที่หนึ่ง, ที่สอง, ที่สุดท้าย
coordinating	JCRG	และ, หรือ, แต่
subordinating	JSBR	เพราะว่า, เนื่องจาก
comparative	JCMP	กว่า, เหมือนกับ, เท่ากับ
Unit classifier	CNIT	ตัว, คน, เล่ม
Collective classifier	CLTV	คู่, กลุ่ม, ฝูง, เชิง, ทาง, ด้าน, แบบ, รุ่น
Measurement classifier	CMTR	กิโลกรัม, แก้ว, ชั่วโมง
Frequency classifier	CFQC	ครั้ง, เที่ยว
Verbal classifier	CVBL	ม้วน, มัด
Nominal prefix	FIXN	การทำงาน, ความสนุกสนาน
Adverbial prefix	FIXV	อย่างเร็ว
Ending for affirmative sentence	EAFF	จ๊ะ, จ้ะ, ค่ะ, ครับ, นะ, น่า, เถอะ.
Ending for interrogative sentence particle	EITT	หรือ, เหรอ, ไหม, มั้ย (yes or no).
Negator	NEG	ไม่, มิได้, ไม่ได้ (not)