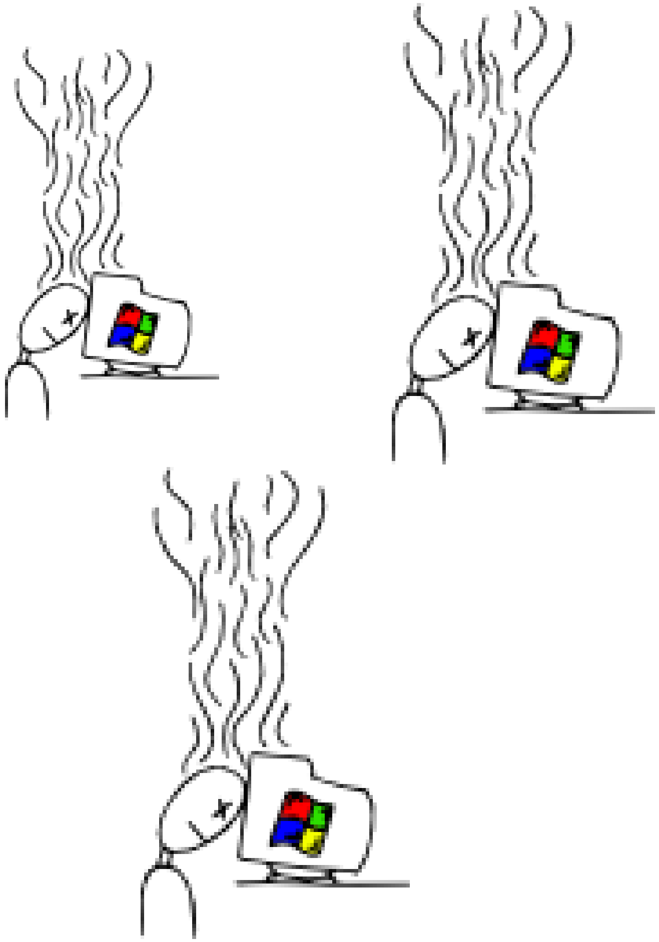
Plangsarn: Thai Romanization Library Application

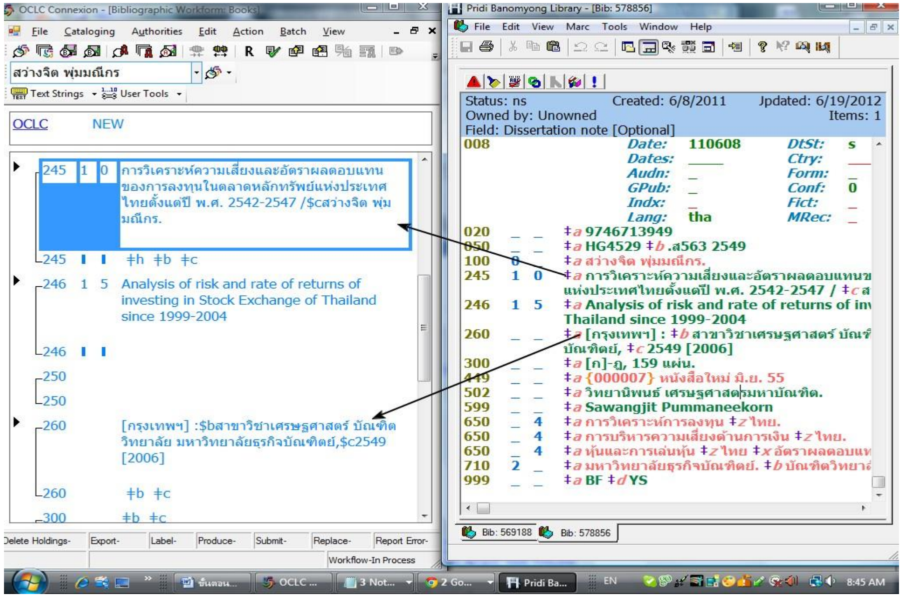
Srichan Chancheewa, THAMMASAT UNIVERSITYDr. Choochart Haruechaiyasak, NECTEC, THAILAND

Why do we need this application?

before



after



OCLC Connexion - [Bibliographic Workform: Books]

สว่างจิต พุ่มมณีกร

OCLC NEW

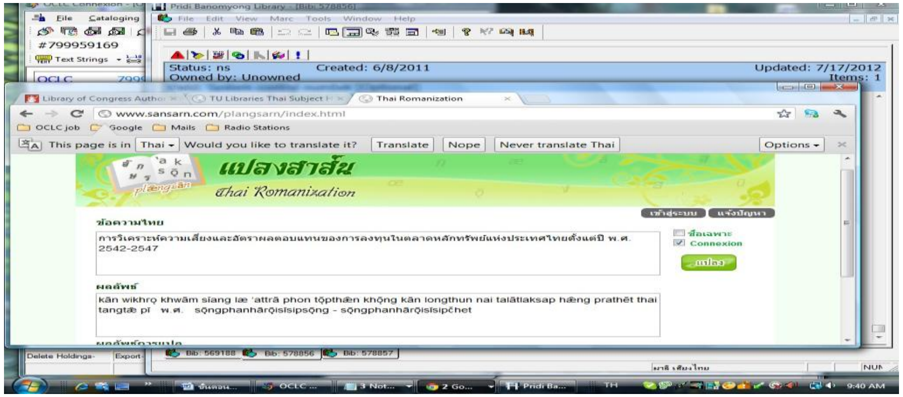
```
245 1 0 การวิเคราะห์ความเสี่ยงและอัตราผลตอบแทน
        ของการลงทุนในตลาดหลักทรัพย์แห่งประเทศ
        ไทยตั้งแต่ปี พ.ศ. 2542-2547 /$cสว่างจิต พุ่ม
        มณีกร.
245        ‡h ‡b ‡c
246 1 5 Analysis of risk and rate of returns of
        investing in Stock Exchange of Thailand
        since 1999-2004
246
250
250
260     [กรุงเทพฯ] :$bสาขาวิชาเศรษฐศาสตร์ บัณฑิต
        วิทยาลัย มหาวิทยาลัยธุรกิจบัณฑิตย์,$c2549
        [2006]
260      ‡b ‡c
300      ‡b ‡c
```

Workflow-In Process

Pridi Banomyong Library - [Bib: 578856]

```
Status: ns            Created: 6/8/2011      Jpdated: 6/19/2012
Owned by: Unowned                                      Items: 1
Field: Dissertation note [Optional]
008               Date:   110608          DtSt:  s
                  Dates:  _____           Ctry:  ___
                  Audn:   _               Form:  _
                  GPub:   _               Conf:  0
                  Indx:   _               Fict:  _
                  Lang:   tha             MRec:  _
020  _ _  ‡a 9746713949
050  _ _  ‡a HG4529 ‡b .ส563 2549
100  0    ‡a สว่างจิต พุ่มมณีกร.
245  1 0  ‡a การวิเคราะห์ความเสี่ยงและอัตราผลตอบแทนข
          แห่งประเทศไทยตั้งแต่ปี พ.ศ. 2542-2547 / ‡c ส
246  1 5  ‡a Analysis of risk and rate of returns of inv
          Thailand since 1999-2004
260  _ _  ‡a [กรุงเทพฯ] : ‡b สาขาวิชาเศรษฐศาสตร์ บัณฑิ
          บัณฑิตย์, ‡c 2549 [2006]
300  _ _  ‡a [ก]-ฎ, 159 แผ่น.
449  _ _  ‡a {000007} หนังสือใหม่ มิ.ย. 55
502  _ _  ‡a วิทยานิพนธ์ เศรษฐศาสตรมหาบัณฑิต.
599  _ _  ‡a Sawangjit Pummaneekorn
650  _ 4  ‡a การวิเคราะห์การลงทุน ‡z ไทย.
650  _ 4  ‡a การบริหารความเสี่ยงด้านการเงิน ‡z ไทย.
650  _ 4  ‡a หุ้นและการเล่นหุ้น ‡z ไทย ‡x อัตราผลตอบแท
710  2 _  ‡a มหาวิทยาลัยธุรกิจบัณฑิตย์. ‡b บัณฑิตวิทยาล
999  _ _  ‡a BF ‡d YS
```

Bib: 569188 Bib: 578856



Pridi Banomyong Library - [Bib: 578856]

#799959169

```
Status: ns            Created: 6/8/2011             Updated: 7/17/2012
Owned by: Unowned                                             Items: 1
```

www.sansarn.com/plangsarn/index.html

This page is in Thai Would you like to translate it?

แปลงสาส์น

Thai Romanization

ข้อความไทย

การวิเคราะห์ความเสี่ยงและอัตราผลตอบแทนของการลงทุนในตลาดหลักทรัพย์แห่งประเทศไทยตั้งแต่ปี พ.ศ. 2542-2547

☐ ชื่อเฉพาะ

☑ Connexion

ผลลัพธ์

kān wikhrǫ khwām sīang læ ʻattrā phon tǫpthæn khǫng kān longthun nai talātlaksap hæng prathēt thai tangtæ pī พ.ศ. sǫngphanhārǫisīsipsǫng - sǫngphanhārǫisīsipčhet

Background and Motivation

- *Romanization is the representation of a written word or spoken speech with the Roman (Latin) script.*
- *Two main methods of romanization are*

  (1) Transliteration: for representing written text,

  (2) Transcription: for representing the spoken word,

  (2.1) phonemic transcription: records the phonemes or units of semantic meaning in speech

  (2.2) phonetic transcription: records speech sounds with precision.

Background and Motivation

- *Phonetic transcription attempts to depict all phones in the source language.*
- *We adopt the International Phonetic Alphabet (IPA).*
- *IPA is an alphabetic system of phonetic notation based primarily on the Latin alphabet.*
- *IPA was devised by the International Phonetic Association as a standardized representation of the sounds of oral language.*
- *IPA symbols are composed of one or more elements of two basic types, letters and diacritics.*

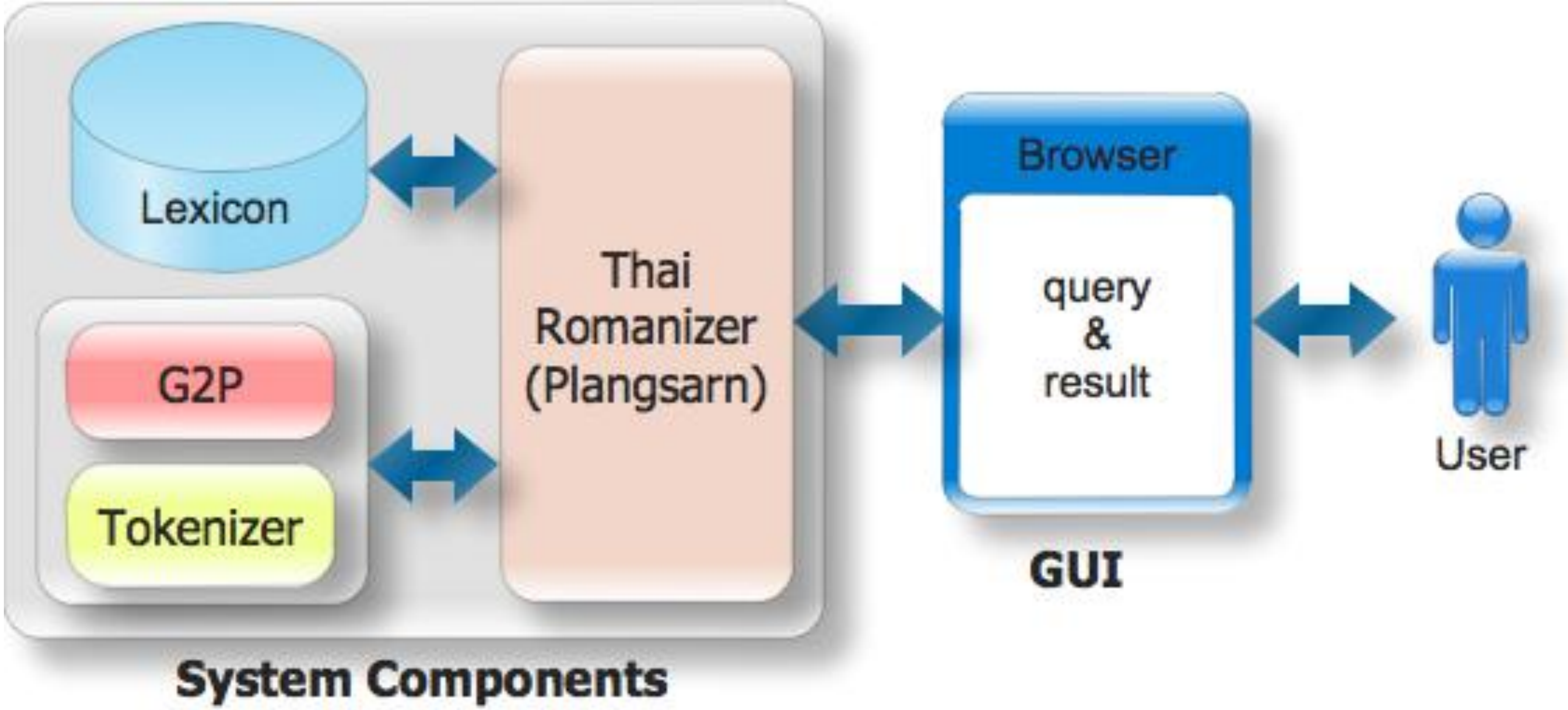
Background and Motivation

- *For library cataloging, there is a guideline for preparing romanization:*
  - *ALA-LC Romanization Tables: Transliteration Schemes for Non-Roman Scripts, which is approved by the Library of Congress and the American Library Association.*
  - *ALA-LC is a set of standards for romanization, or the representation of text in other writing systems using the Latin alphabet.*

Objective

- Design and implement a system for automatically romanizing Thai texts with the focus on bibliographic data.
- The system is intended to support librarians who perform the cataloguing task.
- The system is called Plangsarn (แปลงสาส์น).
  - plang (แปลง) = transform
  - sarn (สาส์น) = text, message

Plangsarn’s System Architecture

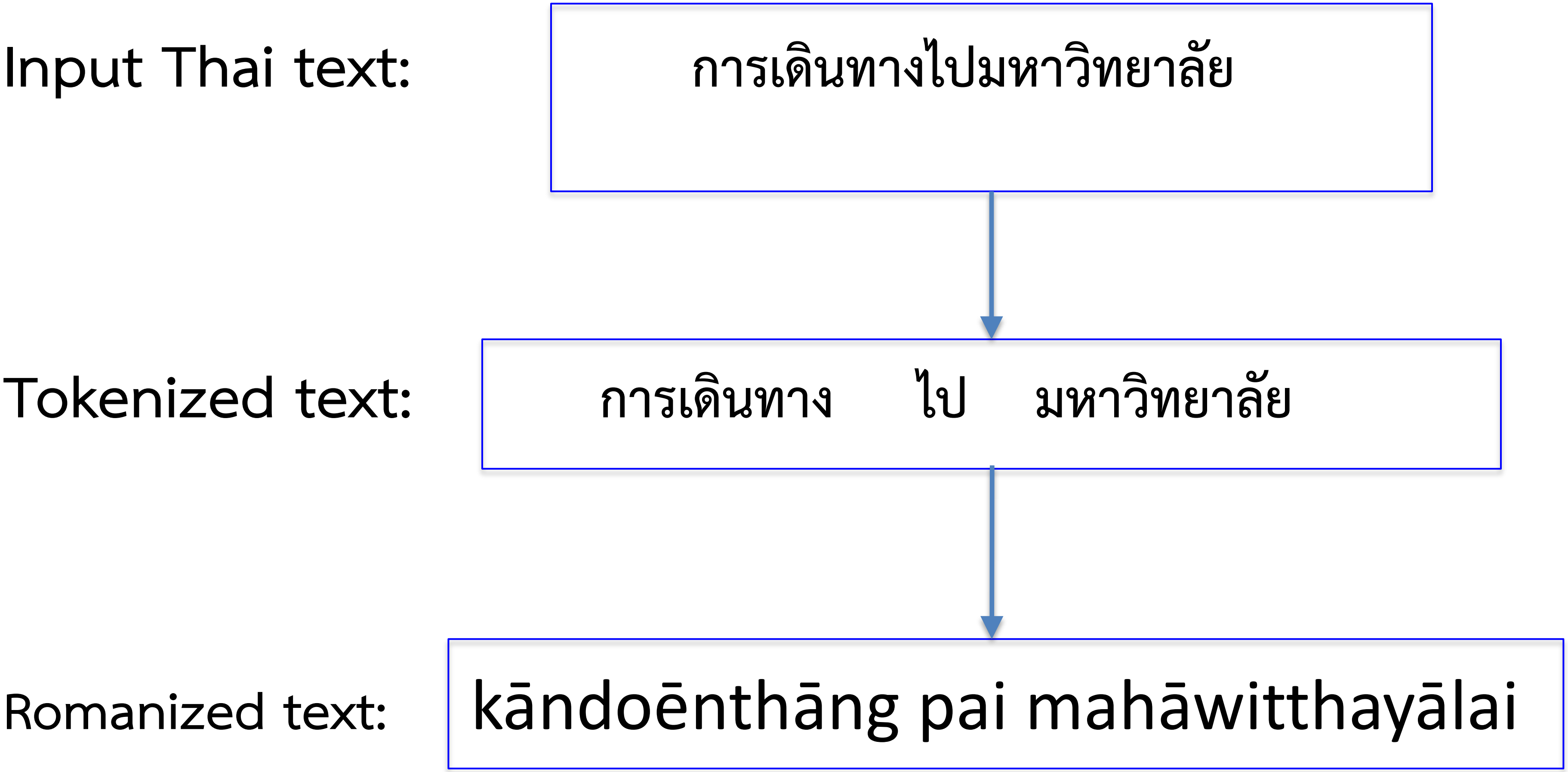


- Lexicon contains Thai word list with romanization.
- G2P (Grapheme-to-Phoneme) converts grapheme unit into phoneme unit.
- Tokenizer performs the word segmentation on Thai written texts.

Plangsarn's process flow

1. Get Thai text input from the user via browser interface.
2. Perform word segmentation on the input text.
3. For each word token, look up in the lexicon for existing romanized representation.
4. If the romanized representation already exists, then use it. Otherwise send the word to G2P module.
5. Concatenate all romanized representations and send the final result to the user via the browser interface.

Example of Plangsarn process



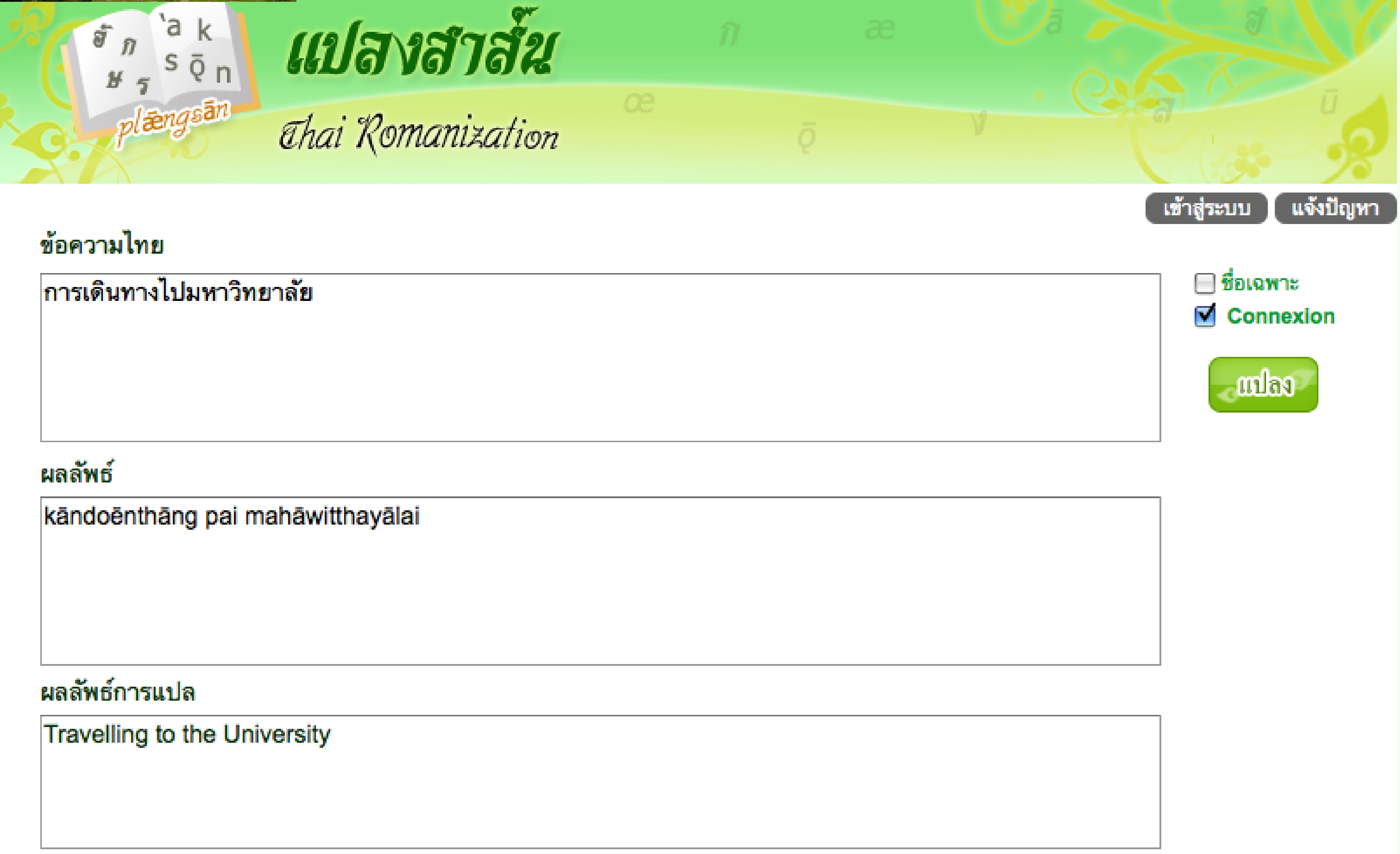
Input Thai text: การเดินทางไปมหาวิทยาลัย

Tokenized text: การเดินทาง ไป มหาวิทยาลัย

Romanized text: kāndoēnthāng pai mahāwitthayālai

Plangsarn’s Web interface:

http://164.115.23.167/plangsarn/



ข้อความไทย

การเดินทางไปมหาวิทยาลัย

☐ ชื่อเฉพาะ

☑ Connexion

ผลลัพธ์

kāndoēnthāng pai mahāwitthayālai

ผลลัพธ์การแปล

Travelling to the University

User Evaluation

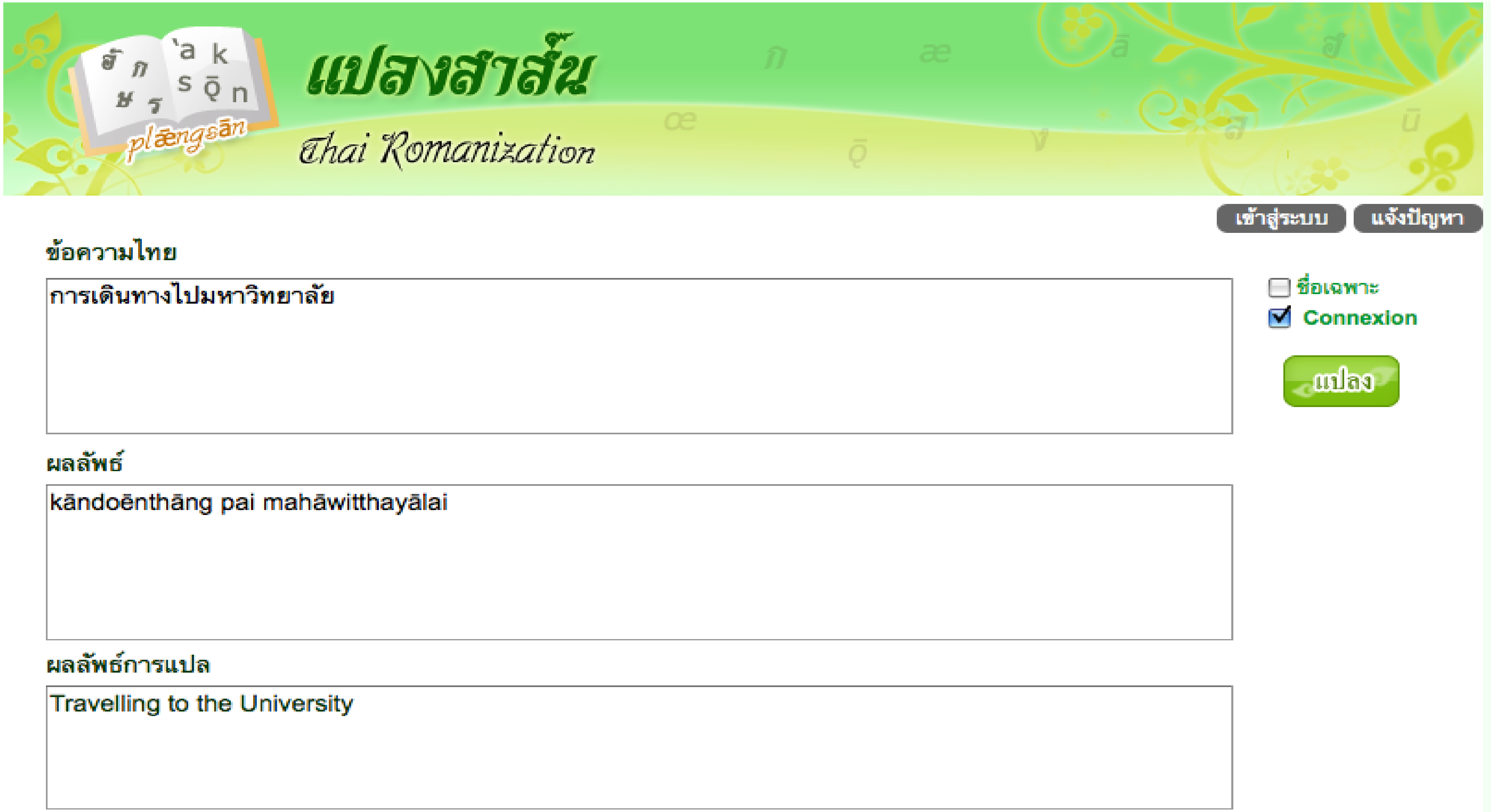
- Plangsarn is under the process of user testing and evaluation.
- Two main reported problems so far are
  - Incorrect word segmented unit.
    - ห้องครัว (Kitchen) => **hǫng khrūa (should be hǫngkhrūa)**
  - Unable to recognize and capitalize named entities such as the names of person, organization, place
    - กรุงเทพ (Bangkok) => **krung thēp (should be Krungthēp)**
- To solve the problem, the system allows authorized users to manually include words with its correct romanization into the lexicon.

Plangsarn’s Summarized Features

- Automatic romanization of Thai text using the G2P (Grapheme-to-Phoneme) technique
- Use the lexicon-based Thai word segmentation to tokenize the text
- Allow authorized users to manually add and edit the list of romanized lexicon
- Support both IPA and OCLC Connexion
- Automatic capitalization for named entities (e.g., names of person, organization, place)
- Provide the machine translation of Thai to English

PLANGSARN IS NOW AVAILABLE WITHIN THAI GOVERNMENT CLOUD SERVICE

HTTP://164.115.23.167/PLANGSARN/



ข้อความไทย

การเดินทางไปมหาวิทยาลัย

□ ชื่อเฉพาะ
☑ Connexion

ผลลัพธ์

kāndoēnthāng pai mahāwitthayālai

ผลลัพธ์การแปล

Travelling to the University

SPECIAL THANKS TO

Dr. Choochart Haruechaiyasak

Senior Researcher

Speech and Audio Technology Laboratory (SPT)

National Electronics and Computer Technology Center (NECTEC)

112 Phahonyothin Rd., Klong 1, Klong Luang,

Pathumthani 12120

THAILAND

email: choochart.haruechaiyasak@nectec.or.th

Tel. (+66) 2 564-6900 Ext. 2240

Fax. (+66) 2 564-6873