ระบบป้อนข้อมูลภาษาไทยด้วยตัวอักษรโรมัน

ROMANIZED THAI INPUT METHOD EDITOR

ธิทชานนี์ รัตนเหรียญ

วิมลศิริ โหล่เกษม

ปริญญานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรวิทยาศาสตรบัณฑิต

สาขาวิชาเทคโนโลยีสารสนเทศ

คณะเทคโนโลยีสารสนเทศ

สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง

ภาคเรียนที่ 2 ปีการศึกษา 2558

ระบบป้อนข้อมูลภาษาไทยด้วยตัวอักษรโรมัน

ROMANIZED THAI INPUT METHOD EDITOR

โดย

สิทธานต์ รัตนเหลี่ยม

SITTAN RATTANALIAM

วิมลสิริ โพธิ์เกษม

WIMONSIRI POKASAME

อาจารย์ที่ปรึกษา

ดร.สุภวรรณ อันนันหนับ

อาจารย์ที่ปรึกษาร่วม

ดร.นล เปรมัษเฐียร

ปริญญานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิทยาศาสตรบัณฑิต

สาขาวิชาเทคโนโลยีสารสนเทศ

คณะเทคโนโลยีสารสนเทศ

สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง

ภาคเรียนที่ 2 ปีการศึกษา 2558

ระบบป้อนข้อมูลภาษาไทยด้วยตัวอักษรโรมัน

ROMANIZED THAI INPUT METHOD EDITOR

โดย

สิทธานต์ รัตนเหลี่ยม

วิมลสิริ โพธิ์เกษม

อาจารย์ที่ปรึกษา

ดร.สุภวรรณ อันนันหนับ

อาจารย์ที่ปรึกษาร่วม

ดร.นล เปรมัษเฐียร

ปริญญานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิทยาศาสตรบัณฑิต

สาขาวิชาเทคโนโลยีสารสนเทศ

คณะเทคโนโลยีสารสนเทศ

สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง

ภาคเรียนที่ 2 ปีการศึกษา 2558

ROMANIZED THAI INPUT METHOD EDITOR

SITTAN RATTANALIAM

WIMONSIRI POKASAME

A PROJECT SUBMITTED IN PARTIAL FULFILLMENT

OF THE REQUIREMENT FOR THE DEGREE OF

BACHELOR OF SCIENCE PROGRAM IN INFORMATION TECHNOLOGY

FACULTY OF INFORMATION TECHNOLOGY

KING MONGKUT'S INSTITUTE OF TECHNOLOGY LADKRABANG

2/2015

ใบรับรองปริญญานิพนธ์ ประจำปีการศึกษา 2558

คณะเทคโนโลยีสารสนเทศ

สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง

เรื่อง ระบบป้อนข้อมูลภาษาไทยด้วยตัวอักษรโรมัน

ROMANIZED THAI INPUT METHOD EDITOR

ผู้จัดทำ

1. นายสิทธานต์ รัตนเหลี่ยม รหัสนักศึกษา 55070130
2. นางสาววิมลสิริ โพธิ์เกษม รหัสนักศึกษา 55070109

สุภวรรณ อัน ..... อาจารย์ที่ปรึกษา

(ดร.สุภวรรณ อันนันหนับ)

นล เปรมัษเฐียร .....อาจารย์ที่ปรึกษาร่วม

(ดร.นล เปรมัษเฐียร)

หัวข้อโครงงาน	ระบบป้อนข้อมูลภาษาไทยด้วยตัวอักษรโรมัน	
นักศึกษา	นายสิทธานต์ รัตนเหลี่ยม	รหัสนักศึกษา 55070130
	นางสาววิมลสิริ โพธิ์เกษม	รหัสนักศึกษา 55070109
ปริญญา	วิทยาศาสตรบัณฑิต	
สาขาวิชา	เทคโนโลยีสารสนเทศ	
ปีการศึกษา	2558	
อาจารย์ที่ปรึกษา	ดร.สุภวรรณ อันนันหนับ	
อาจารย์ที่ปรึกษาร่วม	ดร.นล เปรมัษเฐียร	

บทคัดย่อ

ระบบป้อนข้อมูลภาษาไทยด้วยตัวอักษรโรมันเป็นวิธีการหนึ่งที่จะทำให้สามารถพิมพ์ภาษาไทยด้วยการใช้ตัวอักษรโรมันได้ เนื่องจากปัจจุบันชาวต่างชาติได้เข้ามามีบทบาทในประเทศไทยมากขึ้น มีการติดต่อสื่อสารกับชาวต่างชาติเพิ่มขึ้น ราชบัณฑิตยสถานจึงได้ประกาศกฎเกณฑ์การถอดอักษรไทยด้วยตัวอักษรโรมันขึ้น เพื่อใช้เป็นมาตรฐานในการเขียนข้อความหรือคำภาษาไทยด้วยการใช้ตัวอักษรโรมันเพื่ออำนวยความสะดวกให้กับชาวต่างชาติมากขึ้น ดังที่เห็นในแผนที่หรือป้ายบอกทางต่างๆ ในวิทยานิพนธ์นี้ผู้จัดทำจึงได้พัฒนาระบบป้อนข้อมูลภาษาไทยด้วยตัวอักษรโรมันขึ้น โดยยึดหลักเกณฑ์การถอดอักษรไทยด้วยตัวอักษรโรมันแบบถ่ายเสียงของราชบัณฑิตยสถานมาใช้ประกอบการพัฒนา วิทยานิพนธ์ฉบับนี้ได้นำเสนอขั้นตอนวิธีการถอดอักษรโดยใช้วิธีการเข้ารหัสผ่านตัวกลางโดยใช้ซาวน์เด็กซ์ ซึ่งเป็นวิธีที่จะอนุญาตให้ใช้คำภาษาไทยโดยใช้ตัวอักษรโรมันได้เพื่อค้นคืนคำไทยที่หลักการเขียนที่ตรงกันในอีกภาษาหนึ่งโดยจะมีตัวกลางที่ตรงกันสำหรับสองภาษา ขั้นตอนวิธีที่นำเสนอแบ่งออกเป็นสองส่วนคือ ขั้นตอนวิธีการเข้ารหัสตัวอักษรภาษาไทยและขั้นตอนวิธีการเข้ารหัสตัวอักษรภาษาอังกฤษ ผลการทดลองจะแสดงให้เห็นว่าขั้นตอนวิธีการเข้ารหัสภาษาไทยด้วยตัวอักษรโรมันโดยใช้วิธีซาวน์เด็กซ์สามารถแสดงผลออกมาได้สูงถึง 90 เปอร์เซ็นต์ นอกจากนั้นจะมีปัญหาในเรื่องความกำกวมในการออกเสียงของคำนั้นๆ และการสะกดคำที่แตกต่างกันในแต่ละบุคคลก็ส่งผลให้ระบบไม่สามารถค้นคืนคำไทยออกมาได้

Project Title	Thai Romanized Input Method Editor	
Student	Mr. Sittan Rattanaliam	Student ID 55070130
	Miss Wimonsiri Pokasame	Student ID 55070109
Degree	Bachelor of Science	
Program	Information Technology	
Academic Year	2015	
Advisor	Dr. Supawan Annannub	
Co-Advisor	Dr. Nol Premasathian	

ABSTRACT

Thai Romanized Input Method Editor is the way to write Thai language using Roman alphabets. Nowadays, the foreigners has come to Thailand a lots and they have to communicate with Thai people. The Royal Institute has established the standard by proposing the principle of Romanization[5] on the basis of transcription. In this study, we aim to develop such a system and we base on the rules of The Royal Institute. In this paper, we present an algorithm of transcription called 'Soundex' that allow the users input Roman alphabet and transform to Numeric codes. The algorithm has divided to two parts for Thai alphabets and Roman alphabet. The results has shown that the algorithm can perform to 90%, but some missing is a result of ambiguous of language and each people are spelling difference.

กิตติกรรมประกาศ

ผู้วิจัยขอขอบพระคุณ ดร.สุภวรรณ อันนันหนับ อาจารย์ที่ปรึกษาและ ดร.นล เปรมัษเฐียร อาจารย์ที่ปรึกษาร่วมเป็นอย่างสูง ปริญญานิพนธ์เล่มนี้มิอาจสำเร็จลุล่วงได้ หากไม่ได้รับความช่วยเหลือและคำปรึกษาต่างๆจากอาจารย์ทั้งสองท่าน

นอกจากนี้ผู้วิจัยขอขอบคุณประยุทธ สุวรรณวิสารท ที่ให้ความร่วมมือในการให้ข้อมูลการทำซาวน์เด็กซ์ภาษาไทย และได้ให้คำแนะนำต่างๆ เพื่อใช้เป็นแนวทางในการดำเนินโครงงาน ให้เป็นไปในแนวทางที่ถูกต้องและสำเร็จลุล่วงไปได้ด้วยดี

สิทธานต์ รัตนเหลี่ยม

วิมลสิริ โพธิ์เกษม

สารบัญ

หน้า

สารบัญ (ต่อ)

หน้า

สารบัญตาราง

หน้า

ตารางที่

สารบัญรูป

หน้า

รูปที่

บทที่ 1
บทนำ

1.1 ความเป็นมาและความสำคัญของปัญหา

ภาษาแต่ละภาษามีลักษณะที่แตกต่างกันออกไป แต่การออกเสียงยังคงมีลักษณะที่คล้ายกัน เช่น การใช้พยัญชนะ สระ ตัวสะกด รวมกันเป็นคำ ในรูปของตัวอักษรที่ต่างกัน หรืออาจจะคล้ายกันแล้วแต่ภูมิภาคที่ใกล้เคียงกัน มนุษย์เราใช้อักษรเหล่านี้ในการบันทึก หรือศิลปะวรรณกรรม แต่ส่วนที่สำคัญที่สุดของตัวอักษรเหล่านี้คือ การสื่อสาร ตัวอักษรสามารถที่จะเป็นเครื่องมือในการสื่อสารได้เป็นอย่างดี

การพิมพ์เป็นขั้นตอนในการสื่อสารผ่านตัวอักษรในโลกเทคโนโลยี การที่เราจะพิมพ์บนแป้นพิมพ์ในภาษาต่างๆ ก็ต้องอาศัยความรู้และความเข้าใจในไวยากรณ์และคำศัพท์ของภาษานั้นๆ เช่นเดียวกันกับภาษาไทย ที่มีหลักภาษาและตัวอักษรที่มีลักษณะเฉพาะ

เนื่องจากมีชาวต่างชาติที่สนใจที่จะศึกษาภาษาไทย ที่ต้องการพิมพ์ภาษาไทยได้อย่างถูกต้อง แต่ยังมีปัญหาเรื่องของการเขียนภาษาไทยหรือการสะกดคำต่างๆ อยากพิมพ์ภาษาไทยให้ได้ตรงตามความหมายที่จะสื่อสารได้อย่างถูกต้อง หรือต้องการค้นคว้าวิจัยแหล่งข้อมูลต่างๆที่เป็นภาษาไทย ได้อย่างมีประสิทธิภาพ สำหรับคนไทยที่ต้องใช้แป้นพิมพ์ภาษาต่างประเทศและยังจำตำแหน่งต่างๆของแป้นพิมพ์ภาษาไทยไม่ได้ ก็สามารถที่จะพิมพ์ได้ถูกต้องเช่นกัน

สำหรับปริญญานิพนธ์นี้จะมุ่งเน้นในการพัฒนา Input Method Editor(IME) ที่สามารถแปลงตัวอักษรโรมันที่ผสมกันแล้วอ่านออกเสียงคล้ายคำที่มีความหมายในภาษาไทย แล้วแปลงคำเป็นคำนั้นๆในภาษาไทย โดยสามารถนำไปใช้ได้กับระบบปฏิบัติการ Windows

1.2 ความมุ่งหมายและวัตถุประสงค์ของการศึกษา

1.2.1 เพื่อศึกษาและพัฒนาตัวแก้ไขวิธีการป้อนข้อมูล (IME) สำหรับการป้อนข้อมูลลงในคอมพิวเตอร์ด้วยการพิมพ์ตัวอักษรบน US – Standard Keyboard และแสดงผลเป็นตัวอักษรภาษาไทย

1.2.2 เพื่อจะช่วยอำนวยความสะดวกให้กับชาวต่างชาติที่ต้องการพิมพ์ภาษาไทย เพื่อการเรียนรู้ และเขียนได้อย่างถูกต้อง

1.2.3 เพื่อจะช่วยอำนวยความสะดวกแก่คนไทยที่ใช้แป้นพิมพ์ภาษาอังกฤษ (ที่ไม่มีภาษาไทย) เพื่อพิมพ์ได้อย่างถูกต้อง

1.3 สมมติฐานของการศึกษา

1.3.1 คำทุกคำในภาษาไทยสามารถนำมาพิมพ์ทับศัพท์ด้วยอักษรโรมันได้ทุกคำ (อาจจะมีคำที่เขียนเหมือนกันหรือคล้ายกัน)

1.3.2 คำบางคำในภาษาไทยนำเอามาจากภาษาอังกฤษ คำเหล่านี้ต้องใช้กฎการถอดอักษรจากงานวิจัยอื่นๆ หรือสร้างขึ้นเองโดยอ้างอิงหลักการของราชบัณฑิตยสถาน

1.3.3 โปรแกรมหรือ IME ที่สร้างขึ้นสามารถนำไปใช้เป็นเครื่องมือในการช่วยพิมพ์ได้จริง และมีความถูกต้องไม่ต่ำกว่า 90%

1.4 ทฤษฎีหรือแนวคิดที่ใช้ในโครงงาน

การ Romanization คือการแปลงจากการเขียนด้วยอักษรภาษาต้นฉบับ เป็นการเขียนด้วยอักษรโรมัน อาศัยการออกเสียงของตัวอักษรโรมันมาประสมกันให้มีเสียงอ่านที่เหมือนหรือคล้ายกันกับภาษาต้นฉบับ โดยในโครงงานนี้จะยึดหลักการการแปลงอักษรตามแบบฉบับของราชบัณฑิตยสถาน ตัวระบบเราจะใช้ภาษา C# ในการพัฒนาเป็นหลัก การหาคำต่างๆเราใช้อัลกอริทึม Metaphone เป็นตัวในการจำแนกคำโดยการสะกดคำ

1.5 ขอบเขตโครงงาน

1.5.1 คำศัพท์ในตัวอักษรภาษาอังกฤษไม่รวมถึงคำย่อหรือรัสพจน์ (Acronym)

1.5.2 คำที่ใช้ในการทดสอบขั้นตอนวิธี จะใช้หลักเกณฑ์การถอดอักษรของราชบัณฑิตยสถาน

1.6 วิธีการดำเนินงาน

1.6.1 ศึกษาขั้นตอนวิธีการค้นคืนสารสนเทศข้ามภาษา

1.6.2 ศึกษาหลักภาษาในการถอดอักษร และหลักเกณฑ์การทับศัพท์จากภาษาอังกฤษเป็นภาษาไทย และจากภาษาไทยไปเป็นภาษาอังกฤษ

1.6.3 ศึกษาขั้นตอนวิธีการใช้อัลกอริทึมซาวน์เด็กซ์ ภาษาอังกฤษและภาษาไทย

1.6.4 ออกแบบและพัฒนาขั้นตอนวิธีการเข้ารหัสคำทับศัพท์และคำไทยเพื่อการค้นคืนข้ามภาษา

1.6.5 ออกแบบวิธีการทดสอบขั้นตอนวิธีในข้อ 4

1.6.6 ทดสอบและปรับปรุงคุณภาพของขั้นตอนวิธี

1.6.7 สรุปผลการวิจัย

1.7 โครงสร้างของปริญญานิพนธ์

ปริญญานิพนธ์ฉบับนี้ประกอบไปด้วยเนื้อหา 5 บท โดยแต่ละบทจะมีเนื้อหาดังนี้

บทที่ 1 บทนำ บอกถึงความเป็นมาและความสำคัญของปัญหา วัตถุประสงค์ของการศึกษา ประโยชน์ ขอบเขตโครงงาน และวิธีการดำเนินงาน

บทที่ 2 ทฤษฎีและหลักการ บอกถึงทฤษฎีและหลักการต่างๆที่ใช้ในการดำเนินโครงงาน

บทที่ 3 วิเคราะห์และการออกแบบ แสดงถึงการวิเคราะห์ต่างเพื่อที่จะมาใช้ในการพัฒนาโครงงาน และการออกแบบโครงสร้าง ลักษณะหน้าตาหรือรูปแบบของโครงงาน

บทที่ 4 ระบบต้นแบบ นำเสนอระบบที่ได้จากการพัฒนาลักษณะการใช้งานต่างๆ

บทที่ 5 สรุปและประเมินผลโครงงาน วิเคราะห์ผลที่ได้จากการพัฒนาโครงงาน และนำเสนอ

บทที่ 2

เอกสารและงานวิจัยที่เกี่ยวข้อง

ในบทนี้จะกล่าวถึงทฤษฎีต่าง ๆ ที่เกี่ยวกับหลักภาษาไทย และ หลักการถอดอักษร ได้แก่ การถอดอักษร ประวัติความเป็นมาของการถอดอักษรในภาษาไทย การถ่ายเสียงด้วยตัวอักษรโรมัน ตามหลักเกณฑ์ของราชบัณฑิตยสถาน การเข้ารหัสซาวน์เด็กซ์ ขั้นตอนวิธีซาวน์เด็กซ์ภาษาอังกฤษ ขั้นตอนวิธีซาวน์เด็กซ์ภาษาไทย งานวิจัยการใช้หลักคำพ้องเสียง เพื่อค้นหาชุดอักขระภาษาไทยที่ออกเสียงเหมือนกัน[4] งานวิจัย A Thai Soundex System for Spelling Correction[3] งานวิจัยการเข้ารหัสคำทับศัพท์เพื่อการค้นคืนข้ามภาษาไทย-อังกฤษ[1]

2.1 การถอดอักษร

การถอดอักษร (Transliteration) หมายถึง การนำคำในภาษาหนึ่งมาเขียนด้วยตัวอักษรอีกภาษาหนึ่งแบบอักษรต่ออักษร โดยพยายามใช้หน่วยเสียงของอักษรทั้งสองภาษาใกล้เคียงกันมากที่สุด[1] เช่นคำว่า "CAMBRIDGE" ในภาษาอังกฤษถอดอักษรเป็น "เคมบริดจ์" ในภาษาไทย เป็นต้น การถอดอักษรระหว่างภาษาเป็นกระบวนการหนึ่งที่เกิดขึ้น เมื่อเราต้องการนำคำจากภาษาหนึ่งมาใช้ในอีกภาษาหนึ่ง ซึ่งในแต่ละภาษาก็จะมีวิธีการเขียนที่แตกต่างกัน เนื่องจากระบบเสียงในแต่ละภาษานั้นแตกต่างกัน จึงต้องใช้คำจากภาษาต่างประเทศมาช่วยในการถอดอักษรระหว่างภาษานี้ขึ้น และนอกจากนั้นยังช่วยให้ชาวต่างชาตินั้นอ่านและใช้คำของเราได้ การถอดอักษรระหว่างภาษานี้ทำได้หลายวิธีและมีคำศัพท์ที่ใช้แตกต่างกันอยู่ 3 คำหลักๆ ได้แก่ Transliteration, Transcription และ Romanization

Transliteration หมายถึง การเขียนตัวเขียนในภาษาหนึ่งด้วยตัวเขียนที่สอดคล้องกันในอีกภาษาหนึ่งคำว่า Transliteration นี้มักจะหมายถึง "การถอดอักษร" แต่บางคนก็ใช้คำอื่น เช่น "การถ่ายถอดอักษร" "การถ่ายทอดตัวอักษร" "การถ่ายรูปอักษร" เวลาที่ชาวตะวันตกศึกษาเกี่ยวกับเรื่องเกี่ยวกับประเทศอื่น ๆ ที่มีระบบตัวเขียนต่างออกไป ก็มักใช้วิธีการ Transliteration นี้ คือถอดอักษรภาษานั้นๆออกมาด้วยตัวอักษรที่ตนใช้[4]

Transcription หมายถึงการบันทึกข้อมูลเสียงของภาษาโดยใช้ระบบการเขียนที่กำหนด มักหมายถึง "การถ่ายเสียง" (ธีระพันธ์ 2526), การถ่ายถอดเสียง (อุไรศรี และ อรวรรณ 2545) ในที่นี้จะขอใช้คำว่า "การถ่ายเสียง" เพื่อให้แตกต่างจากคำ "การถอดอักษร" อย่างชัดเจน นักภาษาศาสตร์ได้กำหนดสัญลักษณ์สากลที่ใช้ในการถ่ายเสียงภาษาต่างๆ อักษรที่ใช้ในการถ่ายเสียงนั้นเรียกว่า สัทอักษร (Phonetic) และในการถ่ายเสียงนั้นอาจใช้ตัวอักษรโรมันมาแทนคำในภาษาหนึ่งด้วยก็ได้ ดังนั้นจึงคาบเกี่ยวกับความหมายของ Romanization ด้วย[4]

ดังนั้น Romanization จึงหมายถึงการแทนคำในภาษาหนึ่งด้วยตัวอักษรโรมัน ซึ่งอาจใช้วิธีการถอดอักษร (Transliteration) หรือ วิธีการถ่ายเสียง (Transcription) ก็ได้ แต่ความจริงมีคำไทยอีกประเภทหนึ่งซึ่งมีความหมายทำนองเดียวกับ Romanization คือคำว่า "การทับศัพท์" เพียงแต่เรามักใช้ในความหมายของการเขียนคำภาษาต่างประเทศด้วยอักษรไทย ในขณะที่ Romanization เป็นการเขียนคำภาษาไทยหรือภาษาอื่นๆด้วยอักษรโรมัน แต่จริงๆแล้วหลักการของการทับศัพท์อาจจะใช้วิธีการถอดอักษรหรือการถ่ายเสียงก็ได้ (ความหมายตามพจนานุกรมไทยฉบับราชบัณฑิตยสถาน พ.ศ. ๒๕๔๒ คือ การที่รับเอาคำของภาษาหนึ่งมาใช้ในอีกภาษาหนึ่งโดยวิธีการถ่ายเสียงและถอดอักษร)

2.2 ประวัติความเป็นมาของการถอดอักษรในภาษาไทย

กริสโวลด์ได้กล่าวถึงความเป็นมาของการถอดอักษรไทยไว้ว่า มีมาเนิ่นนานแล้ว ตัวอย่างเช่น ชาวโปรตุเกสในศตวรรษที่ 16 เรียกเมืองไทยว่า SIÃO หรือ MUANTAI มีเมืองหลวงชื่อ HUDIA (อยุธยา) ซึ่งกริสโวลด์เห็นว่าเป็นการถอดอักษรไทยที่ไม่เหมาะสม การถอดอักษรไทยโดยชาวต่างประเทศนั้นมักเป็นไปตามอำเภอใจของแต่ละคน ในสมัยรัชกาลสมเด็จพระนารายณ์มหาราช (พ.ศ. 2200-2231) กลุ่มมิชชันนารีชาวฝรั่งเศสได้คิดวิธีการถอดอักษรไทยเพื่อให้ได้เสียงใกล้เคียงกับเสียงภาษาไทย เช่น "เมื่อท่านมาเราได้กินสำเร็จแล้ว" จะเขียนว่า "MEÜA TÂN MÂ' RÂO DÁÏ KIN SAM-RED LÊOU" ซึ่งปรากฏอยู่ในหนังสือของลาลูแบร์ เอกอัครราชทูตของพระเจ้าหลุยส์ที่ 14 แต่ดูเหมือนว่าชาวต่างประเทศต่าง ๆ ก็จะมีวิธีการเขียนคำไทยโดยอิงกับระบบภาษาของตน ไม่ได้ใช้ระบบที่บาทหลวงฝรั่งเศสเสนอขึ้น จนในสมัยรัชกาลที่ 5 เมื่อเริ่มมีปัญหาเรื่องการเขียนชื่อทางภูมิศาสตร์ในแผนที่ซึ่งไทยได้ทำร่วมกับอังกฤษและฝรั่งเศส จึงมีความพยายามร่วมกันระหว่างไทยและฝรั่งเศสว่าจะสร้างระบบการเขียนคำไทยด้วยตัวอักษรโรมัน แต่ระบบนี้ก็ไม่ได้เป็นที่แพร่หลาย เพราะปัญหาเรื่องการถอดอักษรไทยเป็นโรมันนี้เป็นที่ถกเถียงกันมากขึ้นในหมู่นักวิชาการในเวลาต่อมา[9] จะเห็นได้จากบทความของพระบาทสมเด็จพระมงกุฎเกล้าเจ้าอยู่หัว[10] ที่ทรงวิตกกังวลว่า หากปล่อยให้ต่างคนต่างเขียนก็จะทำให้เกิดความสับสน จึงได้เสนอให้ใช้ระบบที่ยึดตามตัวสะกดของคำศัพท์เดิมในภาษาบาลีสันสกฤตหรือที่รู้จักกันในชื่อของระบบกราฟิก (Graphic System หรือ Hunterian System) พระองค์ทรงไม่เห็นด้วยกับการถอดอักษรโดยวิธีการถ่ายเสียง จนกระทั่งปี พ.ศ. 2474 กระทรวงกรรมาธิการ (กระทรวงศึกษาธิการในปัจจุบัน) ก็ได้ตั้งคณะกรรมการเพื่อพิจารณาเรื่องหลักการถอดอักษรไทยเป็นโรมันที่จะใช้เป็นมาตรฐานร่วมกันและได้นำเสนอผลเพื่อรับฟังความคิดเห็นจากกระทรวงกลาโหม ราชบัณฑิตยสถาน สยามสมาคม และศาสตราจารย์ยอร์ช เซเดส์ แห่งวิทยาลัยฝรั่งเศสของตะวันออกไกลและต่อมาราชบัณฑิตยสถานได้สานงานต่อจนเป็นที่มาของประกาศเกณฑ์การถอดอักษรไทยเป็นโรมันฉบับแรกในปี พ.ศ. 2482 โดยกำหนดไว้สองระบบคือ ระบบทั่วไป (General System) และ ระบบพิสดาร (Precise System)

ทั้งสองระบบจะใช้วิธีการต่างกัน โดยระบบทั่วไปจะใช้วิธีการถ่ายเสียง เช่นคำว่า กษัตริย์ ถ่ายเสียงได้เป็น KASAT แต่ระบบพิสดารจะใช้วิธีการถอดอักษร ก็จะถอดอักษรเป็น KASATRIY แต่ประกาศที่ออกมานั้นก็ไม่เป็นที่ใช้งานกันอย่างแพร่หลายนัก มีเพียงบางกลุ่ม บางหน่วยงานที่ใช้เกณฑ์ดังกล่าว[6] เช่น กรมทางหลวงใช้เขียนชื่อถนน กรมแผนที่ทหาร การรถไฟใช้เขียนชื่อสถานีรถไฟ กรุงเทพมหานครใช้เขียนชื่อถนนและซอยต่างๆ หลังจากนั้นจึงมีการพัฒนาต่อมาเรื่อยๆ กระทั่งปี พ.ศ. 2541 มาตรฐาน ISO11940:1998 ได้รับการรับรองให้ใช้เป็นแบบ Transliteration คือแต่ละตัวอักษรจะถอดแบบหนึ่งต่อหนึ่ง และได้รับการพัฒนามาเรื่อยๆ จนกระทั่งปีพ.ศ. 2542 ราชบัณฑิตยสถานได้ยกเลิกประกาศเกณฑ์การถอดอักษรไทยเป็นอักษรโรมันเดิมและยกเลิกแถลงการณ์เรื่องการเขียนชื่อจังหวัด เขตอำเภอ และกิ่งอำเภอ โดยใช้หลักการเขียนแบบทั่วไป และให้ใช้หลักเกณฑ์การถอดอักษรไทยเป็นโรมันแบบถ่ายเสียงที่ได้ปรับปรุงใหม่ให้เหมาะสมยิ่งขึ้น[5]

2.3 การถอดอักษรไทยเป็นอักษรโรมันแบบถ่ายเสียงของราชบัณฑิตยสถาน

การถอดอักษรไทยเป็นอักษรโรมันมี 2 แบบ ได้แก่ การถอดอักษรตามวิธีเขียน (Transliteration) และการถอดอักษรตามวิธีอ่านหรือการถอดแบบถ่ายเสียง (Transcription) การถอดอักษรตามวิธีเขียน เป็นการถอดตามตัวอักษรทุกตัวที่มีอยู่ในคำนั้นๆ[6] มีข้อดีคือทราบที่มาของคำ สามารถสะกดคำได้ และสามารถถอดอักษรกลับไปยังคำเดิม (retransliteration) ได้ ข้อด้อยคืออ่านออกเสียงได้ยาก แม้จะใส่เครื่องหมายออกเสียงช่วยไว้ด้วยก็ตาม และคำที่ถอดออกมาดูยุ่งยาก รุงรัง ส่วนการถอดอักษรแบบถ่ายเสียงเป็นการถอดอักษรที่ไม่จำเป็นต้องถอดครบทุกตัวอักษร เน้นการออกเสียงเป็นหลัก ข้อดีคือทำให้อ่านคำไทยที่ถอดเป็นอักษรโรมันได้เสียงใกล้เคียง คำที่ถอดออกมามีความกะทัดรัด เข้าใจง่าย ข้อด้อยคือ ไม่เอื้อให้ถอดอักษรกลับคืนสู่คำเดิมได้

ตัวอย่างการถอดอักษรตามวิธีเขียน มะม่วง ⟶ Mamwng

ตัวอย่างการถอดอักษรแบบถ่ายเสียง มะม่วง ⟶ Mamuang

ระบบการถอดอักษรด้วยวิธีการถ่ายเสียง นั้นคือการแทนตัวอักษรในภาษาหนึ่งด้วยรูปแทนเสียงของภาษานั้น เพื่อให้อ่านคำภาษาไทยที่เขียนด้วยตัวอักษรโรมันให้ได้เสียงใกล้เคียงโดยไม่คำนึงถึงการสะกดตัวการันต์และวรรณยุกต์[6] เช่น จันทร์ = chan, พระ = phra, แก้ว = kaeo การถอดอักษรไทยเป็นอักษรโรมันแบบถ่ายเสียง เป็นหลักการถอดตัวอักษรไทยเป็นอักษรโรมันอย่างเป็นทางการโดยราชบัณฑิตยสถาน ใช้สำหรับหนังสือและสิ่งพิมพ์ของรัฐบาล และป้ายชื่อถนนต่างๆในประเทศไทย โดยใช้หลักเกณฑ์การถอดอักษรไทยเป็นอักษรโรมันตั้งแต่ พ.ศ. 2482 และใน พ.ศ. 2542 ได้ปรับปรุงให้เหมาะสมยิ่งขึ้น โดยยกเลิกประกาศเดิมและให้ใช้หลักเกณฑ์การถอดอักษรไทยเป็นอักษรโรมันแบบถ่ายเสียง[5]

การถอดอักษรไทยเป็นโรมันตามแบบราชบัณฑิตยสถานนี้ เป็นที่ยอมรับและใช้กันอย่างแพร่หลาย โดยเฉพาะตามหน่วยงานราชการต่าง ๆ รวมทั้งได้รับการรองรับจากองค์การ

สหประชาชาติให้ใช้เป็นมาตรฐานในการถอดอักษรไทยเป็นโรมันสำหรับชื่อต่างๆ อย่างไรก็ตาม ก็มีผู้วิจารณ์ว่าระบบที่ใช้นี้ไม่เหมาะสมสำหรับใช้สอนภาษาไทยให้ชาวต่างชาติ เนื่องจากไม่ได้มีการกำหนดวรรณยุกต์ ความสั้นหรือยาวของสระ รวมทั้งไม่แยกความแตกต่างของเสียงสระ เช่น โอะ – โอ ฯลฯ และไม่ได้แยกความแตกต่างของเสียงระหว่าง จ กับ ช ด้วย จึงมีผลทำให้ชาวต่างประเทศออกเสียงผิดได้ แต่ทั้งนี้คงต้องเข้าใจว่า จุดมุ่งหมายหลักของระบบแบบราชบัณฑิตยสถานนั้นไม่ได้มุ่งเน้นที่การเรียนการสอนภาษาไทยให้ชาวต่างประเทศ แต่ต้องการให้ชาวต่างประเทศโดยทั่วไปที่จำเป็นต้องติดต่อสื่อสาร สามารถอ่านและเขียนคำไทยด้วยตัวอักษรโรมันได้

องค์การระหว่างประเทศว่าด้วยการมาตรฐาน (International Organization for Standardization : ISO) ได้กำหนดมาตรฐานการถอดอักษรไทยเป็นอักษรโรมันขึ้นในปี 1998 เพื่อจัดให้มีเครื่องมือสำหรับการสื่อสารสากลในการเขียนที่ชัดเจน ไม่คลุมเครือ เพื่อการถ่ายทอดอัตโนมัติที่สามารถถอดกลับได้ถูกต้องสมบูรณ์ โดยอาจจะเป็นคนหรือเครื่องจักร โดยได้กำหนดสัญลักษณ์สำหรับการออกเสียงสระ สัญลักษณ์เพื่อแยกความแตกต่างระหว่างพยัญชนะที่ออกเสียงเหมือนกัน ด้วยวิธีนี้ทำให้อักษรแต่ละตัวถูกถอดเป็นอักษรโรมันและสัญลักษณ์กำกับที่ไม่ซ้ำกันสามารถถอดเป็นอักษรโรมันและถอดกลับเป็นอักษรไทยได้อย่างถูกต้อง

ตารางที่ 2.1 ตารางเทียบเสียงพยัญชนะ และ สระ ตามหลักเกณฑ์ที่กำหนดโดยราชบัณฑิตยสถาน

พยัญชนะ			สระ	
พยัญชนะไทย	อักษรโรมัน		สระไทย	อักษรโรมัน
	ตัวต้น	ตัวสะกด		
ก	k	k	อะ, ◌ั (อะ ลดรูป), รร (มีตัวสะกด), อา	a
ข ฃ ค ฅ ฆ	kh	k	รร (ไม่มีตัวสะกด)	an
ง	ng	ng	อำ	am
จ ฉ ช ฌ	ch	t	อิ, อี	i
ซ ทร (เสียง ซ)	s	t	อึ, อื	ue
ศ ษ ส	s	t	อุ, อู	u
ญ	y	n	เอะ, เ◌็- (เอะ ลดรูป), เอ	e
ฎ ฑ (เสียง ด) ด	d	t	แอะ, แอ	ae
ฏ ต	t	t	โอะ, - (โอะ ลดรูป), โอ, เอาะ, ออ	o
ฐ ฑ ฒ ถ ท ธ	th	t	เออะ, เ◌ิ-(เออะ ลดรูป), เออ	oe
ณ น	n	n	เอียะ, เอีย	ia

ตารางที่ 2.1 (ต่อ) ตารางเทียบเสียงพยัญชนะและสระตามหลักเกณฑ์ที่กำหนดโดย ราชบัณฑิตยสถาน

พยัญชนะ			สระ	
พยัญชนะไทย	อักษรโรมัน		สระไทย	อักษรโรมัน
	ตัวต้น	ตัวสะกด		
บ	b	p	เอือะ, เอือ	uea
ป	p	p	อัวะ, อัว, -ว- (อัว ลดรูป)	ua
ผ พ ภ	ph	p	ใอ, ไอ, อัย, ไอย, อาย	ai
ฝ ฟ	f	p	เอา, อาว	ao
ม	m	m	อุย	ui
ย	y	-	โอย,ออย	oi
ร	r	n	เอย	oei
ล ฬ	l	n	เอือย	ueai
ว	w	-	อวย	uai
ห ฮ	h	-	อิว	io
			เอ็ว, เอว	eo
			แอ็ว, แอว	aeo
			เอียว	iao
			ฤ (เสียง รึ) ฤๅ	rua
			ฤ (เสียง ริ)	ri
			ฤ (เสียง เรอ)	roe
			ฦ, ฦๅ	lue

2.4 วิธีเชิงสัทลักษณ์ (Phonetic algorithm)[12]

ขั้นตอนวิธีเชิงสัทลักษณ์ (Phonetic algorithm) คือขั้นตอนวิธีสำหรับการกำหนดดัชนีของคำต่างๆ โดยใช้การออกเสียงเป็นเกณฑ์ ขั้นตอนวิธีเชิงสัทลักษณ์ส่วนใหญ่พัฒนาขึ้นเพื่อใช้กับภาษาอังกฤษ ดังนั้นการใช้กฎเกณฑ์ดังกล่าวกับคำในภาษาอื่นอาจไม่ให้ผลลัพธ์ที่มีความหมาย

ขั้นตอนวิธีเหล่านั้นมีความซับซ้อนด้วยกฎและข้อยกเว้นหลายประการในภาษานั้นๆ เนื่องจากการสะกดคำ และการออกเสียงในภาษาอังกฤษ[12]

ขั้นตอนวิธีเชิงสัทลักษณ์ที่เป็นที่รู้จักนั้นมีดังนี้

- ซาวน์เด็กซ์ ถูกพัฒนาขึ้น โดย Robert C. Russell และ Margaret K. Odell พัฒนาขึ้นเพื่อใช้เข้ารหัสสำหรับชื่อคนในภาษาอังกฤษในการทำสำมะโนประชากร ใช้แก้ปัญหาเนื่องจากบางคำนั้นสามารถสะกดได้หลายแบบอ่านออกเสียงคล้ายกันอาจจะแตกต่างกันเพราะสำเนียงหรือวัฒนธรรมซึ่งจะเป็นปัญหาหลักในการเก็บข้อมูลประวัติของชื่อบุคคล
- Metaphone และ Double Metaphone ใช้สำหรับคำในภาษาอังกฤษทั่วไป ไม่เพียงชื่อเฉพาะบุคคล ขั้นตอนวิธีนี้เป็นรากฐานของซอฟต์แวร์ตรวจการสะกดคำในหลายโปรแกรม (Spell Checking)
- New York State Identification and Intelligence System (NYSIIS) เป็นการจับคู่หน่วยเสียง (phoneme) ที่คล้ายกันให้เข้ากับอักษรตัวเดียวกัน ให้ผลลัพธ์เป็นสายอักขระซึ่งสามารถอ่านได้โดยไม่ต้องถอดรหัส
- การเข้าหาด้วยการจัดระดับคู่ (Match Rating Approach) ถูกพัฒนาขึ้นโดยสายการบินเวสเทอร์นแอร์ไลน์เมื่อปี พ.ศ. 2520 ใช้เทคนิคการเข้ารหัสและเปรียบเทียบและจัดระดับความคล้ายคลึงให้กับชื่อภาษาอังกฤษที่มีลักษณะพ้องเสียงกัน ประโยชน์ของขั้นตอนวิธีนี้คือเพิ่มความแม่นยำและรวดเร็วในการค้นหาชื่อภาษาอังกฤษไม่เพียงแค่ตรวจหาคำที่ตรงที่สุดเพียงอันเดียวแต่ครอบคลุมไปถึงชื่อที่มีลักษณะพ้องเสียงกันแม้อาจจะมีการสะกดที่แตกต่างกัน

2.5 ขั้นตอนวิธีซาวน์เด็กซ์ภาษาอังกฤษ

M.K. Odell และ R. C. Russell ได้ออกแบบขั้นตอนวิธีการเข้ารหัสชื่อในภาษาอังกฤษโดยยึดหลักของการอ่านออกเสียง เพื่อให้ชื่อที่อ่านออกเสียงคล้ายกันก็จะได้รหัสเหมือนกัน หรือที่เรียกว่า "ซาวน์เด็กซ์" ขั้นตอนวิธีดังกล่าวได้ใช้แนวคิดทางภาษาศาสตร์และชวเลขที่ว่าชื่อในภาษาอังกฤษสามารถจำแนกความแตกต่างได้โดยพิจารณาเพียงพยัญชนะเท่านั้น [13]

ขั้นตอนวิธีการเข้ารหัสของซาวน์เด็กซ์ โดยเริ่มจากการนำตัวอักษรตัวแรกของคำไปเป็นรหัส ส่วนตัวอักษรที่เหลือจะแปลงเป็นตัวเลขโดยใช้ตารางการกำหนดรหัสซาวน์เด็กซ์ ดังแสดงในตารางที่ 2.2 จากนั้นจะตัดรหัสตัวเลขศูนย์ออกไป แล้วถ้ารหัสตัวเลขที่อยู่ตำแหน่งติดกันมีค่าเท่ากัน

จะเก็บรหัสเพียงแค่หนึ่งรหัสเท่านั้น แล้วรหัสซาวน์เด็กซ์ที่ได้จะต้องมีความยาวทั้งหมด 4 หลัก กรณีที่ไม่ถึง 4 หลัก ก็จะเติมตัวเลข 0 ลงไปจนครบ 4 หลัก

ตารางที่ 2.2 ตารางการเข้ารหัส ซาวน์เด็กซ์ ของ Odell และ Russell

ตัวอักษร	รหัสตัวเลข
A E I O U H W Y	0
B F P V	1
C G J K Q S X Z	2
D T	3
L	4
M N	5
R	6

ขั้นตอนซาวน์เด็กซ์นี้เป็นขั้นตอนที่ง่ายและทำงานได้รวดเร็ว ส่วนในเรื่องของประสิทธิภาพในการค้นคืนพบว่า เปอร์เซ็นต์ของค่าเรียกคืน (Recall) มีค่าเรียกคืนสูงมาก แต่เปอร์เซ็นของค่าแม่นยำ (Precision) ต่ำ

ปัจจุบันซาวน์เด็กซ์ภาษาอังกฤษได้มีการใช้อย่างแพร่หลาย และข้อดีของการเข้ารหัสซาวน์เด็กซ์คือสามารถปรับแปลงค่าของตัวเลขได้ เพื่อให้เหมาะสมกับภาษาท้องถิ่นนั้นๆ

2.6 ขั้นตอนวิธีซาวน์เด็กซ์ภาษาไทย

ปัจจุบันมีทีมงานวิจัยหลายท่านได้นำเอาขั้นตอนวิธีของซาวน์เด็กซ์ไปใช้เพื่อแก้ปัญหาต่างๆ เช่น การสืบค้นชื่อนามสกุลที่อ่านออกเสียงเหมือนกันในทะเบียนรายชื่อ การสืบค้นคำไทยที่มักสะกดผิด และการตรวจสอบตัวสะกด เป็นต้น[1] ในการออกแบบและการพัฒนาของแต่ละท่านจึงมีความแตกต่างกันออกไป ในที่นี้จะขอกล่าวถึงงานวิจัยการใช้หลักคำพ้องเสียง เพื่อค้นหาชุดอักขระภาษาไทยที่ออกเสียงเหมือนกัน งานวิจัย A Thai Soundex System for Spelling Correction, งานวิจัยการเข้ารหัสคำทับศัพท์เพื่อการค้นคืนข้ามภาษาไทย-อังกฤษ

2.6.1 งานวิจัยการใช้หลักคำพ้องเสียง เพื่อค้นหาชุดอักขระภาษาไทยที่ออกเสียงเหมือนกัน

มีการออกแบบขั้นตอนวิธีการเข้ารหัสซาวน์เด็กซ์ภาษาไทยให้ชื่อที่อ่านออกเสียงเหมือนกันได้รหัสที่ตรงกัน เพื่อแก้ปัญหาการสืบค้นชื่อและนามสกุลที่อ่านออกเสียงเหมือนกันใน

ทะเบียนรายชื่อ หลักเกณฑ์และข้อกำหนดต่างๆ จะแปลงมาจากขั้นตอนวิธีซาวน์เด็กซ์ภาษาอังกฤษของ Odell และ Russell มีการเพิ่มเติมข้อกำหนดบางส่วนเพื่อให้เหมาะสมกับภาษาไทย โดยมีหลักเกณฑ์ดังนี้[4]

- กำหนดให้ขนาดของรหัสมีความยาวเท่ากับ 7 หลัก ซึ่งประกอบด้วยตัวอักษร 1 หลัก และตามด้วยตัวเลขอีก 6 หลัก โดยมีรูปแบบดังนี้
  รหัสซาวน์เด็กซ์ = (ตัวอักษร) (ตัวเลข) (ตัวเลข) (ตัวเลข) (ตัวเลข) (ตัวเลข) (ตัวเลข)
- เปลี่ยนพยัญชนะตัวแรกของคำเป็นรหัสตัวอักษรตามตารางที่ 2.3
- เปลี่ยนอักษรตั้งแต่ตัวที่สองเป็นต้นไปของคำเป็นรหัสดังตารางที่ 2.4
- ไม่มีการนำสระ วรรณยุกต์ และไม้ไต่คู้ มาสร้างรหัส ยกเว้นสระที่ให้เสียงสะกดเป็นพยัญชนะ ได้แก่ ำ ไ ใ
- ถ้ารหัสซาวน์เด็กซ์ที่ได้มีความยาวต่ำกว่า 7 หลัก ให้เติม 0 ต่อไปจนรหัสมีความยาวครบ 7 หลัก

ตารางที่ 2.3 การกำหนดรหัสตัวอักษรของรหัสซาวน์เด็กซ์ภาษาไทยสำหรับอักษรตัวแรก

รหัสตัวอักษร	ตัวอักษร	รหัสตัวอักษร	ตัวอักษร
ก	ก	บ	บ
ข	ข ฃ ค ฅ ฆ	ป	ป
ง	ง	พ	พ ภ ผ
จ	จ	ฟ	ฝ ฟ
ช	ช ฌ ฉ	ม	ม
ส	ซ ศ ษ ส	ย	ณ ย
ด	ด ฎ	ร	ร ล ฬ ฤ ฦ
ต	ต ฏ	ว	ว
ท	ฐ ฑ ฒ ถ ท ธ	อ	อ
น	ณ น	ฮ	ห ฮ

ตารางที่ 2.4 การกำหนดรหัสตัวอักษรของรหัสซาวน์เด็กซ์ภาษาไทยสำหรับตัวอักษรที่เหลือ

รหัสตัวเลข	ตัวอักษร
0	ม ว ◌ำ
1	ก ข ฃ ค ฅ ฆ
2	ง ย
3	ญ ณ น
4	ฎ ฏ ด ต ศ ษ ส
5	บ ป พ ภ
6	ผ ฝ พ ห อ ฮ
7	จ ฉ ช ซ ฌ
8	ฐ ฑ ฒ ถ ท ธ
9	ร ล ฬ ฤ ฦ

ส่วนข้อกำหนดเพื่อให้เหมาะสมกับภาษาไทยแบ่งเป็นกรณีต่าง ๆ ดังนี้

- กรณีพบ ใ- ไ- ไ-ย และ ◌ัย จะเปลี่ยนให้อยู่ในรูปแบบเดียวกันคือ ◌ัย ก่อนทำการเข้ารหัส เพื่อให้ได้รหัสซาวน์เด็กซ์ที่เหมือนกัน เนื่องจากสระดังกล่าวอ่านออกเสียงเหมือนกัน เช่น ใท ไท ไทย และ ทัย จึงจะเปลี่ยนเป็น ทัย
- กรณีพบ รร จะเปลี่ยนเป็น ◌ัน เมื่อไม่มีตัวสะกดตามหลัง และจะเปลี่ยนเป็น ◌ั เมื่อมีตัวสะกดตามหลัง
- กรณีพบการันต์ จะตัดการันต์และพยัญชนะที่มีตัวการันต์กำกับรวมทั้งสระและอักษรควบการันต์ทิ้ง เช่น คำว่า จันทร์ ศักดิ์ และ พันธุ์ เปลี่ยนเป็น จัน ศัก และ พัน ตามลำดับ

ตารางที่ 2.5 ตัวอย่างการเข้ารหัสซาวน์เด็กซ์ภาษาไทย

ชื่อภาษาไทย	รหัสซาวน์เด็กซ์ภาษาไทย
อัมพร	อ059000
อำกรณ์	อ059000
พรรณศักดิ์	พ341000
พันธุ์ศักดิ์	พ341000
เนืองนิจ	น6234000
เนืองนิตย์	น6234000

2.6.2 งานวิจัย A Thai Soundex System for Spelling Correction[3]

งานวิจัย A Thai Soundex System for Spelling Correction ได้นำเสนอวิธีในการเข้ารหัสสำหรับซาวน์เด็กซ์ภาษาไทย โดยมีแนวคิดว่าระบบหน่วยคำในภาษาไทยไม่มีเครื่องหมายการแบ่งพยางค์ จึงทำให้บางคำนั้น สามารถอ่านออกเสียงได้หลายแบบ เช่น คำว่า ตากลม สามารถอ่านเป็น ตา-กลม หรือ ตาก-ลม ก็ได้ ซึ่งการเข้ารหัสซาวน์เด็กซ์เป็นการเข้ารหัสแบบไม่มีการแบ่งพยางค์ จึงทำให้รหัสที่ได้เกิดความผิดพลาดไปด้วย[3] ดังนั้นเพื่อลดความผิดพลาดในการอ่านออกเสียง จึงได้ใช้เทคนิค Nondeterministic Finite Automaton with Output เพื่อสร้างรหัสซาวน์เด็กซ์สำหรับการอ่านออกเสียงในทุกกรณีของคำ และในการค้นคืนก็จะนำรหัสที่ได้ไปเปรียบเทียบกับรหัสทุกตัวของคำที่ต้องการค้นคืน ถ้าพบรหัสที่เหมือนกันเพียงหนึ่งคู่ก็จะถือว่าทั้งคู่เป็นคู่คำที่อ่านออกเสียงคล้ายกัน[3]

การกำหนดรหัสซาวน์เด็กซ์สำหรับอักขระไทยดังนี้

1. กำหนดรหัสซาวน์เด็กซ์สำหรับพยัญชนะต้น จะแบ่งตามเสียงของพยัญชนะ 21 เสียงของไทย และได้รวม ร กับ ล เป็นกลุ่มเดียวกัน โดยจะใช้พยัญชนะไทย 1 ตัวแทนกลุ่มเสียง ดังแสดงในตารางที่ 2.6
2. กำหนดรหัสซาวน์เด็กซ์สำหรับสระ จะแบ่งตามเสียงของสระและรวมสระเสียงสั้นและยาวเข้าด้วยกัน โดยจะใช้อักขระโรมันมาแทนกลุ่มเสียงสระ ดังตารางที่ 2.7
3. กำหนดรหัสซาวน์เด็กซ์สำหรับตัวสะกด จะแบ่งกลุ่มเสียงตามมาตราต่างๆ ของอักษรไทย และได้เพิ่มกลุ่ม ฮ สำหรับคำที่ไม่มีตัวสะกดดังแสดงในตารางที่ 2.8

ตารางที่ 2.6 การเข้ารหัสสำหรับพยัญชนะต้น

รหัสตัวอักษร	ตัวอักษร	รหัสตัวอักษร	ตัวอักษร
ก	ก	บ	บ
ค	ข ฃ ค ฅ ฆ	ป	ป
ง	ง	พ	พ ภ ผ
จ	จ	ฟ	ฝ ฟ
ช	ช ฌ ฉ	ม	ม
ซ	ซ ศ ษ ส	ย	ญ ย
ด	ด ฎ ฑ*	ร	ร ล ฬ
ต	ต ฏ	ว	ว
ท	ฐ ฑ* ฒ ถ ท ธ	อ	อ
น	ณ น	ฮ	ห ฮ

* ฑ สามารถอ่านออกเสียงได้สองแบบ คือ /ด/ และ /ท/ จึงสร้างรหัสทั้ง 2 แบบ

ตารางที่ 2.7 การเข้ารหัสสำหรับพยัญชนะสระ

รหัสซาวน์เด็กซ์	รูปแบบสระ
A	ะ ั รร อ ำ ใ- ไ- ไ-ย เ-า –า
I	ิ ี -ฤ- (ริ)
V	ึ ื -ฤ- (รึ)
U	ุ ู
E	เ-ะ เ็ เ-
X	แ – ะ แ็ แ -
O	โ-ะ อ + ตัวสะกด โ-
C	เ-าะ ็อ –อ- -ร- (อร)
D	เ-อะ เ-อ เ ิ เ-ย
J	เ ี ยะ เ ี ย
W	เ ื อะ เ ื อ
$	ั ว ะ ั ว -ว-

ตารางที่ 2.8 การเข้ารหัสสำหรับพยัญชนะสะกด

รหัสตัวอักษร	ตัวอักษร
ก	ก ข ค ฆ
ง	ง
ด	จ ฉ ช ซ ฌ ฎ ฏ ฐ ฒ ด ต ถ ท ธ ศ ษ ส
น	ญ ณ น ร ล ฬ
บ	บ ป พ ฟ ภ ผ ฝ
ม	ม ำ
ย	ย
ว	ว
ฮ	-

ขั้นตอนวิธีการเข้ารหัสซาวน์เด็กซ์มีดังนี้

1.สร้างโครงร่างเครื่องเข้ารหัสซาวน์เด็กซ์ จากการรวบรวมกฎเกณฑ์ต่างๆในการสร้างคำในภาษาไทย

2. ขยายเครื่องเข้ารหัสซาวน์เด็กซ์ โดยกำหนดข้อมูล Output ให้สำหรับแต่ละกฎเกณฑ์

3. นำคำที่ต้องการเข้ารหัสผ่านเครื่องเข้ารหัสซาวน์เด็กซ์ ผลลัพธ์ที่สมบูรณ์แต่ละตัวที่ได้คือรหัสซาวน์เด็กซ์

ตัวอย่างคำว่า "เสนา" ผ่านเครื่องเข้ารหัสซาวน์เด็กซ์ ได้ผลลัพธ์ที่สมบูรณ์สองค่าคือ

(1) -เ/e-> (5) -ส/ซ-> (c) -e/a-> (d) -e/ฮ-> (e) -น/น-> (6) -า/a-> (7) -e/ว-> (8)*

(1) -เ/e-> (5) -ส/ซ-> (c) -e/e-> (6) -e/e-> (4) -e/ฮ-> (8) -e/e-> (1) -น/น-> (9) -e/e-> (2) -า/a-> (4) -e/ฮ-> (8)*

เราจึงสรุปรหัสซาวน์เด็กซ์ที่ได้คือ "ซaฮนaว" และ "ซeนฮaฮ"

ตารางที่ 2.9 ตัวอย่างการเข้ารหัสซาวน์เด็กซ์ภาษาไทย

ชื่อภาษาไทย	รหัสซาวน์เด็กซ์ภาษาไทยที่ได้
บุญญา	**บนยaฮ**, บนฮยoนยaฮ, บนนยaฮยaฮ, บนฮยcฮยaฮ, บนฮยaฮยaฮ, บนนยaฮ
บุณยา	**บนยaฮ**, บนนนaฮยaฮ, บนฮนcฮยaฮ, บนฮนaฮยaฮ
สับปะรด	**ซaบปaฮลoด**, ซaบปaฮปaฮลoดดaฮ, ซaบปaฮลoดดaฮ, ซaบบaฮปaฮลcฮดcฮ, ซaบบaฮปaฮลaฮดcฮ, ซaบปaฮลcฮดcฮ, ซaบปaฮลaฮดcฮ, ซaบบaฮปaฮลcฮดaฮ, ซaบบaฮปaฮลaฮดaฮ, ซaบปaฮลcฮดaฮ, ซaบปaฮลaฮดaฮ, ซaบบaฮปaฮลoด
สับปะรด	**ซaบปaฮลoด**, ซaบปaฮลoดซaฮ, ซaบลoดซaฮ, ซaบปaฮซcฮ, ซaบซcฮ, ซaบปaฮลcฮซcฮ, ซaบปaฮลaฮซcฮ, ซaบลcฮซcฮ, ซaบลaฮซcฮ, ซaบปaฮซaฮ, ซaบซaฮ, ซaบปaฮลcฮซaฮ, ซaบปaฮลaฮซaฮ, ซaบลcฮซaฮ, ซaบลaฮซaฮ, ซaบลoด
ก้าว	**กaว**, กaววaฮ, กaฮวcฮ, กaฮวaฮ
เก้า	**กaว**

2.6.3 งานวิจัยการเข้ารหัสคำทับศัพท์เพื่อการค้นคืนข้ามภาษาไทย-อังกฤษ[1]

งานวิจัยการเข้ารหัสคำทับศัพท์เพื่อการค้นคืนข้ามภาษาไทย-อังกฤษ[1]ได้ดัดแปลงต้นแบบการเข้ารหัสคำ ดังตารางที่ 2.10 โดยมีรายละเอียดดังนี้

- ใช้รหัสตัวเลขเป็นอักษรตัวแรกของรหัส แทนการใช้ตัวอักษรของคำเป็นตัวแรก เนื่องจากพบว่ามีตัวอักษรอังกฤษหลายตัวถอดเป็นอักษรไทยเหมือนกัน เช่น V และ W ถอดเป็นตัว ว และได้มีการเสนอรหัสเพิ่มอีก 3 ตัว คือ 7 8 และ 9 ซึ่งจะใช้ในกรณีที่เป็นตัวอักษรแรกของคำเท่านั้น โดยมีรายละเอียดดังนี้
  - รหัส 7 สำหรับคำที่ขึ้นต้นด้วย A E I O U หรือ อ เนื่องจากพบว่าคำภาษาอังกฤษที่ขึ้นต้นด้วยสระมักจะใช้ อ เป็นพยัญชนะต้นในการอ่านออกเสียง เช่น ABRAHAM (อับราฮัม) EDWARD (เอ็ดเวิร์ด) OHM (โอห์ม) เป็นต้น
  - รหัส 8 สำหรับคำที่ขึ้นต้นด้วย H เนื่องจาก H ที่เป็นตัวอักษรแรกของคำมักจะเป็นพยัญชนะและอ่านออกเสียงเป็น ฮ เช่น HOPSKINS (ฮอปกินส์) ส่วน H ที่อยู่ส่วนอื่นของคำมักจะเป็นอักษรควบและไม่อ่านออกเสียง เช่น WHITE (ไวต์) JOHN (จอห์น) SHOW (โชว์) เป็นต้น
  - รหัส 9 สำหรับตัวอักษร Y เนื่องจากพบว่า Y ที่เป็นตัวอักษรแรกของคำมักจะเป็นพยัญชนะและอ่านออกเสียง เช่น YAHOO (ยาฮู) ส่วน Y ที่อยู่ส่วนอื่นของคำมักจะเป็นสระ เช่น ONYX (โอนิกซ์) PHYSICS (ฟิสิกส์) เป็นต้น
- เพิ่มรหัส 52 สำหรับตัวอักษร ง เนื่องจากอักษร ง ถอดมาจากตัวอักษรอังกฤษ NG โดยรหัส 5 สำหรับอักษร N และรหัส 2 สำหรับ G
- ขยายความยาวของรหัสคำที่ได้จากเดิมคือ 4 หลัก (ตามหลักของ Odell และ Russell) เป็น ไม่จำกัดความยาวของรหัสคำที่ได้ เนื่องจากพบว่าคำที่มีความยาวมากอาจได้รหัสตรงกันทั้งๆที่ออกเสียงไม่คล้ายกันเลย เช่น คริสต์เตียน ได้รหัส 262395 และ คริสต์โตฟเฟล ได้รหัส 262314 ซึ่งถ้าพิจารณาเพียง 4 หลัก คือ 2623 จะถือว่าคำทั้งสองออกเสียงคล้ายกัน
- เพิ่มพารามิเตอร์ K ซึ่งเป็นความยาวน้อยสุดของรหัสคำ โดยพิจารณาเฉพาะคำที่มีความยาวน้อยสุดของรหัสคำที่มากกว่า K หลัก
- วรรณยุกต์ และสระในภาษาไทยจะไม่นำมาพิจารณาในการเข้ารหัส

ตารางที่ 2.10 การกำหนดรหัสซาวน์เด็กซ์สำหรับอักษรไทยและอักษรอังกฤษ

โดยประยุทธ สุวรรณวิสารท

ตัวอักษรโรมัน	ตัวอักษรไทย	รหัส	หมายเหตุ
A E I O U H W Y	อ	0	#1
B F P V	บ ฝ ฟ ป ผ พ ภ ว	1	
C G J K Q S X Z	ข ฃ ค ฅ ฆ ฉ ช ฌ ก จ ซ ศ ษ ส	2	
D T	ฎ ด ฏ ต ฐ ฑ ฒ ถ ท ธ	3	
L	ล ฬ	4	
M N	ม ณ น	5	
R	ร	6	
A E I O U	อ	7	#2
H	ห ฮ	8	#2
W	ว	1	#2
Y	ย ญ	9	#2
	ง	52	

หมายเหตุ #1 สำหรับตัวอักษรตัวแรก (ซ้ายสุด)

#2 สำหรับตัวอักษรตัวที่สองจนตัวอักษรสุดท้าย

บทที่ 3

การวิเคราะห์และออกแบบระบบ

หลังจากที่ได้ทบทวนงานวิจัยเรื่องการถอดอักษรในภาษาต่างๆแล้วนั้น พบว่าในการถอดอักษรนั้นมีหลายขั้นตอนมีหลายแนวทาง เพื่อให้ได้มาซึ่งโปรแกรมการถอดอักษรที่สมบูรณ์ สำหรับวิทยานิพนธ์เล่มนี้จะเน้นความสำคัญไปที่การพัฒนาโปรแกรมการถอดอักษรภาษาอังกฤษไปเป็นภาษาไทยโดยใช้เกณฑ์การถอดอักษรไทยเป็นโรมันแบบถ่ายเสียงของราชบัณฑิตยสถาน ผู้วิจัยต้องการจะใช้วิธีการถอดอักษรโดยใช้ตัวกลางในการถอดอักษร หรือรูปแทนเสียง เพราะเห็นว่าหลักเกณฑ์ของราชบัณฑิตยสถานนั้นยังไม่ครอบคลุมมากพอ ยังคงมีปัญหาเรื่องการกำหนดความสัมพันธ์ของการถ่ายเสียง ซึ่งไม่เป็นแบบหนึ่งต่อหนึ่ง ตัวอักษร กล่าวคือ ตัวอักษรโรมันหนึ่งตัวสามารถถ่ายเสียงออกมาได้เป็นหลายตัวอักษรของภาษาไทย นอกจากนี้ในการถอดอักษรภาษาอังกฤษเป็นภาษาไทยโดยไม่ใช้ตัวกลางนั้น มีความจำเป็นที่จะต้องมีข้อมูลคำไทยและคำที่ถอดเป็นอักษรโรมันคู่กัน ซึ่งไม่มีหน่วยงานใดเผยแพร่ชิ้นงานนี้ ทำให้ไม่สามารถวิจัยด้วยการถอดอักษรโดยตรงโดยไม่ใช้ตัวกลางได้ ดังนั้นผู้วิจัยจึงเห็นว่าการถอดอักษรภาษาอังกฤษเป็นภาษาไทยโดยใช้ตัวกลางในการถ่ายเสียงนั้นเหมาะสมที่สุด โดยผู้วิจัยได้เก็บรวบรวมข้อมูลที่เผยแพร่โดยศูนย์เทคโนโลยีอิเล็กทรอนิกส์และคอมพิวเตอร์แห่งชาติ และคณะอักษรศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย เป็นคลังข้อมูลที่มีขนาด 7,235 คำ แล้วนำข้อมูลเหล่านี้มาประมวลผลโดยการเข้ารหัสซาวน์เด็กซ์ โดยการเข้ารหัสจะใช้ตารางที่ได้ปรับปรุงจากงานวิจัยของประยุทธ สุวรรณวิสารทเรื่องการเข้ารหัสคำทับศัพท์เพื่อการค้นคืนข้ามภาษาไทย – อังกฤษ[1] และงานวิจัยของเทพพิทักษ์ การุญบุญญานันท์เรื่อง A Thai Soundex System for Spelling Correction[3]

ในรายงานฉบับนี้จะแบ่งออกเป็นสองส่วนคือการเข้ารหัสซาวน์เด็กภาษาไทย และรหัสซาวน์เด็กซ์สำหรับคำทับศัพท์ ในส่วนของกระบวนการพัฒนาโปรแกรมนั้นมีขั้นตอนแรกเป็นการเก็บรวมรวมข้อมูลคำไทยเพื่อสร้างเป็นคลังคำศัพท์ของราชบัณฑิตยสถานขนาด 7,235 คำ ส่วนขั้นตอนต่อไปจะเป็นการเข้ารหัสซาวน์เด็กซ์เพื่อนำข้อมูลคำไทยที่ผ่านการเข้ารหัสมาใส่ลงในฐานข้อมูลเพื่อใช้เป็นแบบจำลองการเปรียบเทียบ จากนั้นจะนำรหัสที่ได้มาเปรียบเทียบกับข้อมูลตัวอักษรภาษาอังกฤษที่ได้ป้อนข้อมูลเข้ามาว่ามีข้อมูลรหัสซาวน์เด็กซ์ตัวใดที่สอดคล้องกับคำภาษาอังกฤษบ้าง จากนั้นจึงจะนำข้อมูลที่สอดคล้องมาแสดงผลในโปรแกรม ขั้นตอนสุดท้ายจะเป็นขั้นตอนการประเมินผลความถูกต้องแม่นยำของโปรแกรมที่พัฒนาขึ้น

กระบวนการพัฒนาโปรแกรมการถอดอักษรภาษาอังกฤษเป็นภาษาไทยโดยใช้ตัวกลางในการถอดอักษรแบบถ่ายเสียงสามารถแสดงได้ดังรูปที่ 3.1

รูปที่ 3.1 กระบวนการพัฒนาโปรแกรมการถอดอักษรภาษาอังกฤษเป็นภาษาไทยโดยใช้ตัวกลางในการถอดอักษรแบบถ่ายเสียง

3.1 การเก็บรวบรวมข้อมูล

ขั้นตอนแรกของการพัฒนา คือ การเก็บรวบรวมข้อมูล ข้อมูลที่ต้องการคือคำศัพท์ไทยที่พิมพ์เผยแพร่โดยศูนย์วิจัยอิเล็กทรอนิกส์และคอมพิวเตอร์แห่งชาติ และข้อมูลคำไทยใช้บ่อยที่เผยแพร่โดยโครงการคลังข้อมูลภาษาไทยแห่งชาติ ซึ่งจัดรวบรวมโดยคณะอักษรศาสตร์จุฬาลงกรณ์

มหาวิทยาลัย โดยรวบรวมจากงานประเภทงานเขียน เช่น งานวิชาการ กึ่งวิชาการ เรื่องแต่ง นิยาย โคลงกลอน เรื่องสั้น อื่นๆ รวมกันเป็นจำนวนประมาณ 7,235 คำ

ขั้นตอนการเตรียมข้อมูลเพื่อใช้สร้างฐานข้อมูล ในการดำเนินโครงงาน

1. รวบรวมข้อมูลประกอบด้วยคำไทยได้มาทั้งหมด 7,235 คำ
2. นำข้อมูลที่เตรียมไว้มาจัดเก็บลงในฐานข้อมูล Microsoft Access แบ่งเป็น 2 สดมภ์ โดย

สดมภ์ที่ 1 จัดเก็บคำไทยที่รวบรวมได้

สดมภ์ที่ 2 จัดเก็บ รหัสซาวน์เด็กซ์ของคำไทยที่ได้

3. เขียนโปรแกรมการสลับที่ตัวอักษร ให้ตรงกับการถอดอักษรโดยใช้ตัวอักษรโรมัน โดยจะกล่าวในขั้นตอนต่อไป
4. นำข้อมูลที่ได้จากข้อ 3. ไปเข้ารหัสซาวน์เด็กซ์ทีละคำ โดยจะกล่าวในขั้นตอนต่อไป
5. จัดเก็บรหัสที่ได้ลงในฐานข้อมูล สดมภ์ที่สอง

แสดงเป็นตัวอย่างได้ดังนี้

ตารางที่ 3.1 ตัวอย่างข้อมูลที่เก็บรวบรวมโดยโครงการคลังข้อมูลภาษาไทยแห่งชาติ

คำศัพท์ภาษาไทย	รหัสซาวน์เด็กซ์ที่ได้
การ	K97500
เป็น	P10150
ใน	N97105
จะ	C97000
ของ	K11152
มี	M41050
และ	L97101
ไม่	M49710
ได้	D97105

ข้อมูลที่รวบรวมได้มีทั้งหมดจำนวนประมาณ 7,235 คำ จัดเก็บอยู่ในตาราง (Table) ทั้งหมด 1 ตาราง โดยจำแนกตามที่มาของการเผยแพร่ หลังจากนั้นจะนำข้อมูลทั้งหมดที่เก็บได้มาเข้ารหัสซาวน์เด็กซ์ในขั้นตอนต่อไป

3.2 การเขียนโปรแกรมเข้ารหัสซาวน์เด็กซ์สำหรับข้อมูลคำภาษาไทย

หลังจากที่ได้ข้อมูลมาอยู่ในลักษณะของฐานข้อมูลแล้ว ก็จะเริ่มเรียงลำดับข้อมูลใหม่อยู่ในรูปของการสลับที่ตัวอักษรสระประสมให้สอดคล้องกับลำดับการเขียนในภาษาอังกฤษเนื่องจากซาวน์เด็กซ์จะไม่สามารถแยกคำไทยได้ เพราะคำไทยมีความกำกวมหลายประการ แตกต่างจากภาษาต้นฉบับ โดยเริ่มเขียนโปรแกรมแปลงอัตโนมัติเพื่อความสะดวกและรวดเร็ว เช่น 'เมื่อ' ระบบจะเรียงลำดับตัวอักษรที่เป็นสระประสมเป็น 'มเือ' จะตรงกับการเขียนทับศัพท์คือ 'muea' แล้วก็จะนำคำที่ได้เก็บไว้ลงตารางรอเข้าโปรแกรมแปลงรหัสอีกครั้ง หลังจากนั้นจะเขียนโปรแกรมเข้ารหัสซาวน์เด็กซ์ เพื่อถอดอักษรคำไทยให้อยู่ในรูปข้อมูลของรูปแทนเสียง เพื่อใช้ในการเปรียบเทียบคำภาษาอังกฤษ ด้วยวิธีการเข้ารหัสแบบอัตโนมัติ นั่นคือการนำคำภาษาไทยมาเข้ารหัสโดยใช้โปรแกรมการเข้ารหัส เหตุผลที่ใช้โปรแกรมเข้ารหัสเพราะเป็นวิธีที่สะดวกและรวดเร็ว

ในการทำโครงงานนี้ ซาวน์เด็กซ์ที่นำมาใช้เป็นตัวที่ถูกพัฒนาเพื่อใช้สำหรับภาษาไทย โดยหลักการคือ แบ่งกลุ่มพยัญชนะในภาษาไทยออกตามการออกเสียงที่เหมือนกันเป็น 21 กลุ่มด้วยกัน ดังตารางที่ 3.2

ตารางที่ 3.2 ตารางแสดงกลุ่มพยัญชนะในภาษาไทยตามการออกเสียง

ก	ฎ ด	ฝ ฟ
ข ฃ ค ฅ ฆ	ฏ ต	ม
ง	ฐ ฑ ฒ ถ ท ธ	ร
จ	ณ น	ล ฬ
ฉ ช ฌ	บ	ว
ซ ศ ษ ส	ป	ห ฮ
ญ ย	ผ พ ภ	อ

จากนั้นนำ 21 กลุ่มมาจัดให้เหลือ 7 กลุ่ม อ้างอิงจากตารางอัลกอริทึมซาวน์เด็กซ์ของ Odell and Russell (ในตาราง 2.2) โดยอาศัยการออกเสียงที่คล้ายกัน จะได้ผลลัพธ์ตามตาราง 3.3

ตารางที่ 3.3 ตารางกลุ่มตัวอักษรไทยในวิธีการเข้ารหัสซาวน์เด็กซ์ของ Odell and Russell

ตัวอักษรอังกฤษ	ตัวอักษรไทย
A E I O U H W Y	อ ห ฮ ว ญ ย
B F P V	บ ภ ฟ ป ผ พ ภ ว
C G J K Q S X Z	ฃ ข ค ฅ ฆ ฌ ช ฒ ฌ ก จ ซ ศ ษ ส
D T	ฎ ก ฏ ต ฐ ฑ ฒ ถ ท ธ
L	ล ฬ
M N	ม ณ น
R	ร

สำหรับการเขียนโปรแกรมการเข้ารหัสจะยึดหลักเกณฑ์งานวิจัยการเข้ารหัสคำทับศัพท์เพื่อค้นคืนข้ามภาษาไทย – อังกฤษ โดยประยุทธ สุวรรณวิสารท ซึ่งผู้วิจัยได้มีการปรับปรุงอัลกอริทึมเพื่อให้สามารถใช้กับคำไทยได้ดังนี้

- ใช้ตัวอักษรภาษาอังกฤษที่ตรงกับเสียงนั้นเป็นตัวแรกของรหัสซาวน์เด็กซ์ หากตัวอักษรตัวแรกของคำเป็นเสียงสระ (เสียง อ) จะใช้ตัวอักษร 'A E I O U' แทนสระนั้น (แสดงใน **ตารางที่ 3.4**)
- ใช้รหัส 52 สำหรับตัวอักษร 'ง' เมื่ออยู่ส่วนอื่นของคำ และใช้ตัวอักษร 'NG' เมื่อเป็นตัวแรกของคำ (แสดงใน**ตารางที่ 3.4 และ ตารางที่ 3.5**)

ตารางที่ 3.4 แสดงการแทนรหัสตัวอักษรตัวแรกสำหรับการเข้ารหัสตัวอักษรไทย

Thai alphabet	Code
อ	A E I O U
บ	B
ฝ ฟ	F
ป ผ พ ภ	P
ก ข ฃ ค ฅ ฆ	K
จ ฉ ช ฌ	C
ซ ศ ษ ส	S
ฎ ด	D
ฏ ต ฐ ฑ ฒ ถ ท ธ	T
ล ฬ	L
ม	M

ตารางที่ 3.4 (ต่อ) แสดงการแทนรหัสตัวอักษรตัวแรกสำหรับการเข้ารหัสตัวอักษรไทย

Thai alphabet	Code
ณ น	N
ร	R
ง	NG
ห ฮ	H
ญ ย	Y
ว	W

หมายเหตุ: รหัส A E I O U จะขึ้นอยู่กับการออกเสียงตัวแรกของสระที่ได้จากการแปลงเป็นตัวอักษรภาษาอังกฤษ

ตารางที่ 3.5 แสดงการเข้ารหัสสำหรับตัวอักษรตัวที่ไม่ใช่ตัวแรกของคำ

Thai	Code	Note
บ ฝ ฟ ป ผ พ ภ	1	
ก ข ฃ ค ฅ ฆ จ ฉ ช ฌ ซ ศ ษ ส	2	
จ ฉ ช ฌ ซ ศ ษ ส	3	#1
ฎ ด ฏ ต ฐ ฑ ฒ ถ ท ธ	3	
ล ฬ	4	
ม	@	
ณ น	5	
ร	6	
ร ญ ล ฬ	5	#1
ห ฮ	7	
ว	8	
ญ ย	9	
ง	52	

#1 : สำหรับพยัญชนะสะกด

- เพิ่มขนาดความยาวของรหัสให้มีความยาวเท่ากับ 9 ตัวอักษรจากเดิมรหัสมีความยาวเพียง 4 ตัวอักษรซึ่งไม่เหมาะสมกับคำไทย

- ใช้รหัส ASCII เป็นรหัสสำหรับสระ (แสดงในตารางที่ 3.6)
- ไม่มีการเข้ารหัสสำหรับวรรณยุกต์
- แปลงสระในตัวอักษรไทย ให้อยู่ในรูปของการเขียนในภาษาอังกฤษ เช่น 'เ-า' -> 'ao', 'ไ' -> 'ai'. (แสดงในตารางที่ 3.6)

ตารางที่ 3.6 แสดงการเข้ารหัสสำหรับสระ

ตัวอักษรไทย	ตัวอักษรโรมัน	รหัสที่ได้
อะ, อา, อั	a	97
อัวะ, อัว	ua	11797
อำ	am	9764
อิ, อี	i	105
อุ, อู	u	117
อึ, อื	ue	11769
เอย	oei	111101105
เอะ, เอ็, เอ	e	101
เออะ, เออ, เอิ	oe	111101
เอา	ao	97111
เอียะ, เอีย	ia	10597
เอือะ, เอือ	uea	11710197
แอะ, แอ็, แอ	ae	97101
โอะ, โอ	o	111
เอาะ, อ็, ออ	o	111
ใอ, ไอ, ไอย	ai	97105
เอือย	ueai	11710197105
-วย	uai	11797105
ิว	io	105111
เ็ว, เ-ว	eo	101111
เอียว	iao	10597111

- ตัวสะกด ในภาษาไทยเช่น จ ฉ ช ฌ ซ ศ ษ ส ญ ร ล ฬ เมื่อนำตัวอักษรเหล่านี้มาเขียนแบบถ่ายเสียงภาษาไทยโดยใช้ตัวอักษรโรมันตามเกณฑ์การถอดอักษรของราชบัณฑิตยสถาน

จะพบว่าตัวสะกดที่ได้ จะถูกแปลงเป็นคนละตัวกับอักษรไทย เช่น จ.จาน เมื่อใช้เป็นพยัญชนะต้น จะใช้ 'ch' แต่เมื่อเป็นตัวสะกดจะใช้ตัวอักษร 't' ดังนั้น รหัสซาวน์เด็กซ์ ที่แปลงได้ในฐานข้อมูลก็จะไม่ตรงกับคำไทยที่ใช้ตัวอักษรโรมัน ที่แปลงได้จากโปรแกรม

ตารางที่ 3.7 ตารางแสดงตัวอักษรที่ทำให้เกิดข้อผิดพลาดของโปรแกรม

รูปพยัญชนะไทย	พยัญชนะต้น	ตัวสะกด	รหัสที่ถอดได้	
			พยัญชนะต้น	ตัวสะกด
จ ฉ ช ฌ	ch	t	2	3
ซ ศ ษ ส ทร (ออกเสียง ซ)	s	t	2	3
ญ	y	n	9	5
ร	r	n	6	5
ล ฬ	l	n	4	5

- กรณีที่พบการอ่านออกเสียงเชื่อมระหว่างพยางค์ เช่น พัฒนา มีการออกเสียง /tha/ เชื่อมระหว่างพยางค์ โปรแกรมจะต้องเพิ่มตัวอักษรเสียงเดียวกันเพิ่มขึ้นมาในคำ เพื่อให้ตรงกับเสียงที่อ่าน
- คำที่มีตัวการันต์ประกอบ ให้นำไปตัดตัวการันต์ออกก่อนที่จะนำไปเข้ารหัสใช้แก้ไขปัญหากรณีที่พบตัวการันต์ ตัวอักษรก่อนหน้าจะต้องไม่ออกเสียงด้วย และในกรณีที่มี ตัวการันต์ไม่ออกเสียง 2 ตัว เช่น จันทร์ ก็จะต้องไม่นำตำแหน่ง ทร์ ไปเข้ารหัสด้วย
- หากมีอักษรนำเสียงสนิท เช่น อย-หน-หล-หม-หญ-หง-หร-หว-หย จะต้องไม่มีการเข้ารหัสตัวอักษร ห หรือ อ จะตัดตัวอักษร อ หรือ ห ทิ้ง แล้วให้อักษรตัวถัดไปมาเป็นรหัสแทน กรณีคำที่พบเป็นอักษรนำเสียงสนิท
- กรณีที่มีสระ ำ (อำ) อยู่ในคำ ให้เปลี่ยนสระ ำ (อำ) เป็น ัม (อัม) โดยแทนรหัสเป็น 5
- กรณีที่พบ รร (ร หัน) เช่น สรรค์ สรรพ มหัศจรรย์ จะแก้ไขโดยแยกเป็นอีก 2 กรณีคือ
    1. ร หัน แต่ไม่มีตัวสะกด จะเปลี่ยนให้เป็น ัน (อัน)
    2. ร หัน มีตัวสะกด จะเปลี่ยน ร หัน ให้เป็น ั (สระ อะ ลดรูป)

เมื่อเขียนโปรแกรมการเข้ารหัสแล้ว จากนั้นจะดึงคำไทยจากฐานข้อมูลมาเข้ารหัสทีละตัวแล้วจัดเก็บลงใน Microsoft Access ในสดมภ์ที่ 2 แสดงเป็นตัวอย่างได้ดังนี้

ตารางที่ 3.8 ตัวอย่างข้อมูลที่ผ่านการเข้ารหัสซาวน์เด็กซ์

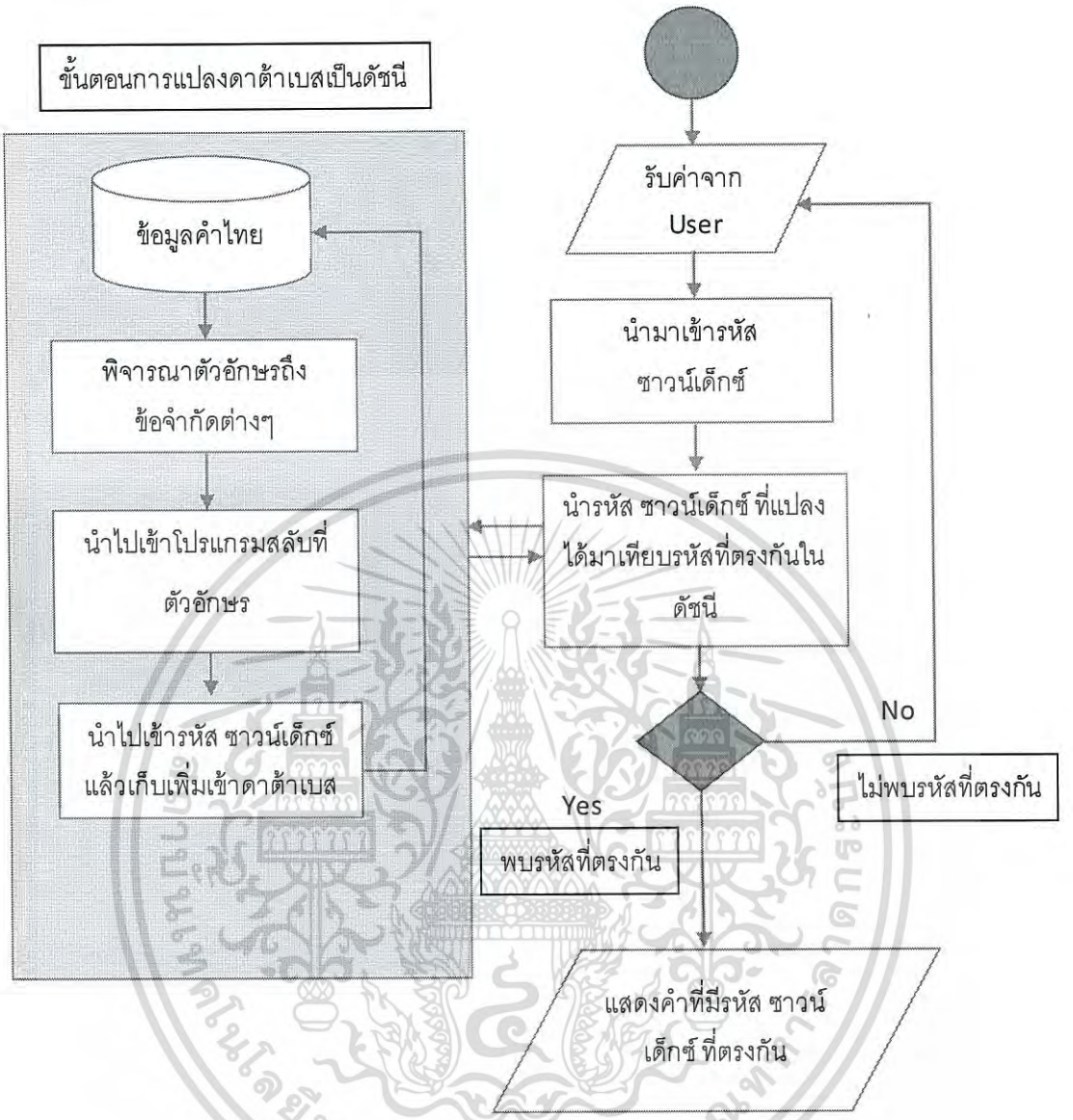
ที่	T10500
การ	K97500
เป็น	P10150
ใน	N97105
จะ	C97000
ของ	K11152

3.3 การเขียนโปรแกรมเข้ารหัสซาวน์เด็กซ์สำหรับคำภาษาอังกฤษ

หลังจากเตรียมข้อมูลรหัสซาวน์เด็กซ์คำไทยเพื่อใช้สำหรับการเปรียบเทียบแล้ว จากนั้นจะเป็นขั้นตอนการเขียนโปรแกรมการเข้ารหัสซาวน์เด็กซ์สำหรับคำภาษาอังกฤษ โดยหลักๆแล้วจะใช้เกณฑ์การเข้ารหัสโดยใช้ตารางเดียวกันกับการเข้ารหัสซาวน์เด็กซ์ภาษาไทย โดยขั้นแรกจะให้คำที่ป้อนเข้ามาในระบบเปลี่ยนให้เป็นตัวพิมพ์ใหญ่ทั้งหมด จากนั้นจึงนำมาเข้ารหัสโดยใช้ตารางที่ 3.7 ตัวอย่างคำภาษาอังกฤษที่ผ่านการเข้ารหัสดังนี้

ตารางที่ 3.9 ตัวอย่างการเข้ารหัสคำภาษาอังกฤษ

ที่	THI	T10500
การ	KAN	K97500
เป็น	PEN	P10150
ใน	NAI	N97105
จะ	CHA	C97000
ของ	KHONG	K11152
มี	MI	M41050
และ	LAE	L97101
ไม่	MAI	M49710
ได้	DAI	D97105



รูปที่ 3.2 แผนภาพแสดงการทำงานโดยรวมของระบบ

ซึ่งผู้วิจัยได้มีการปรับปรุงขั้นตอนวิธีการเข้ารหัสซาวน์เด็กซ์คำภาษาอังกฤษดังนี้

- หากตัวอักษรแรกของคำเป็นสระ A E I O U จะแทนรหัสด้วย 'A E I O U' เช่นเดียวกับขั้นตอนการเข้ารหัสในอักษรไทย แต่ตัวอักษรนั้นจะขึ้นอยู่กับตัวอักษรตัวแรกของสระที่ได้จากการแปลงเป็นตัวอักษรภาษาอังกฤษ (แสดงใน**ตารางที่ 3.6**)
- ไม่มีการเข้ารหัสสำหรับตัวอักษร 'H' ที่อยู่หลังตัวอักษร C, K, T, P
- คำภาษาอังกฤษที่ขึ้นต้นด้วย NG จะให้รหัสซาวน์เด็กซ์ตัวแรกเป็น NG เพื่อเป็นการกำหนดว่าคำที่ต้องค้นนั้นเป็น ตัวอักษร ง ไม่ใช่ตัวอักษร น หรือ ณ และเมื่ออยู่ส่วนอื่นของคำให้แทนรหัสเป็น 52 โดยที่ 5 แทนตัวอักษร N และ 2 แทนตัวอักษร G (แสดงใน**ตารางที่ 3.11**)

- สำหรับสระ A E I O U ที่ไม่ใช่ตัวแรกของคำ จะแทนด้วยรหัส ASCII เช่นเดียวกับขั้นตอนการเข้ารหัสในอักษรไทย (ดังแสดงใน**ตารางที่ 3.6**)
- รหัสที่ได้จะมีความยาวเท่ากับ 9 ตัวอักษร โดยที่ตัวแรกจะเป็นตัวอักษร และอีก 6 ตัวที่เหลือจะเป็นตัวเลขเช่นเดียวกับขั้นตอนการเข้ารหัสในอักษรไทย

จากนั้นเมื่อได้รหัสแล้วจะนำรหัสที่ได้มาเทียบกับฐานข้อมูลดังกล่าวในขั้นตอนต่อไป

ตารางที่ 3.10 แสดงการแทนรหัสตัวอักษรตัวแรกในขั้นตอนการแปลงรหัสสำหรับตัวอักษรโรมัน

Roman alphabet	Code
A E I O U	A E I O U
B	B
F V	F
PH P	P
K KH C G	K
C CH J	C
S	S
D	D
T	T
L	L
M	M
N	N
R	R
NG	NG
H	H
Y	Y
W	W

หมายเหตุ: A E I O U จะขึ้นอยู่กับตัวแรกของสระที่ได้จากการแปลงเป็นตัวอักษรภาษาอังกฤษ

ตารางที่ 3.11 แสดงการแทนรหัสสำหรับตัวอักษรตัวที่ไม่ใช่ตัวแรกของคำในขั้นตอนการเข้ารหัสตัวอักษรโรมัน

Roman alphabet	Thai Sound	Code	Note
B F P V PH	บ ฝ ฟ ป ผ พ ภ	1	
K C G J KH CH S	ก ข ฃ ค ฅ ฆ ฆ จ ฉ ช ฌ ซ ศ ษ ส	2	
	จ ฉ ช ฌ ซ ศ ษ ส	3	#1
D T	ฎ ด ฎ ต ฐ ฑ ฒ ถ ท ธ	3	
L	ล ฬ	4	
M	ม	@	
N	ณ น	5	
R	ร	6	
	ร ญ ล ฬ	5	#1
H	ห ฮ	7	
W	ว	8	
Y	ญ ย	9	
NG	ง	52	

#1 : for final consonant

3.4 การกำหนดรหัสซาวน์เด็กซ์ สำหรับสระ ในภาษาไทย

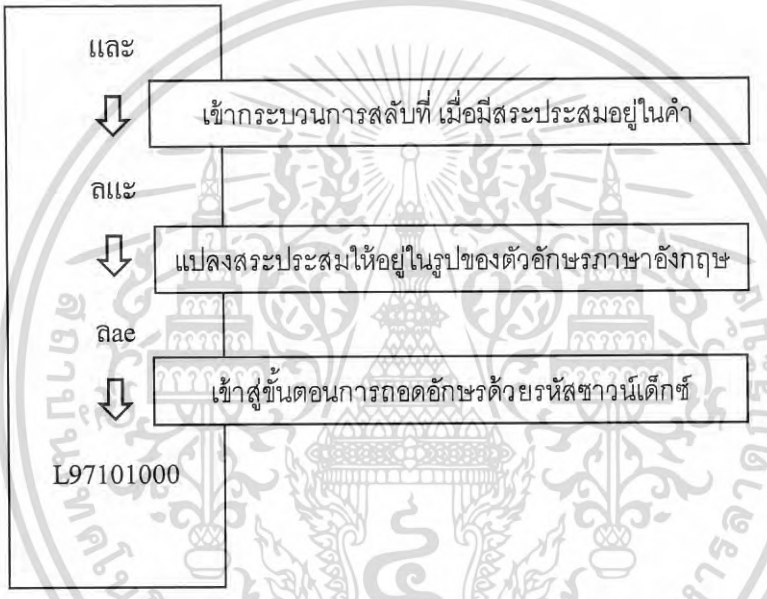
ในการศึกษานี้ผู้วิจัยได้มีวิธีการเข้ารหัสสระเพิ่มขึ้นมาแตกต่างจากต้นฉบับเนื่องจากผลลัพธ์ที่ค้นหาได้มีจำนวนที่เยอะเกินไป และคำศัพท์ส่วนใหญ่มีความเกี่ยวข้องน้อย หรือไม่ตรงกับที่ผู้ใช้ต้องการ เป็นผลมาจากเราเข้ารหัสซาวน์เด็กซ์เฉพาะตัวอักษรเท่านั้น เราจึงเพิ่มการเข้ารหัสให้กับสระด้วย เพื่อให้การค้นคำมีขอบเขตที่แคบและ ใกล้เคียงกับคำที่ต้องการมากขึ้น โดยจะใช้รหัส ASCII ของตัวอักษรสระในภาษาอังกฤษ มาเป็นรหัสซาวน์เด็กซ์แทน

ยกตัวอย่างเช่น เข้ารหัสซาวน์เด็กซ์คำว่า “และ”

ขั้นตอนแรก ก่อนนำมาเขารหัสซาวน์เด็กซ์จะต้องทำการสลับตำแหน่งเมื่อมีสระประสม หรือสระที่มีตำแหน่งนำข้างหน้าพยัญชนะต้น จากคำว่า "และ" จะถูกสลับเป็น "ลและ"

ขั้นตอนที่สอง แปลงสระนั้นๆให้อยู่ในรูปของตัวอักษรภาษาอังกฤษ ซึ่งในตัวอย่างนี้สระคือ "และ" (แอะ) จะถูกแปลงเป็น "ae" (ดูได้จาก**ตารางที่3.6**) เมื่อแปลงเสร็จจะได้เป็น "ลae"

ขั้นตอนสุดท้าย นำผลลัพธ์ที่ได้จากขั้นตอนที่สอง มาแปลงเป็นรหัสซาวน์เด็กซ์ จากตัวอย่างในขั้นตอนที่สอง นำ "ลae" มาแปลงจะได้เป็น L97101000 (L คือตัว 'ล' จาก**ตารางที่ 3.4** เลขรหัส 97101 คือ "ae" จาก**ตารางที่ 3.6** ส่วนเลขศูนย์อีกสามตัวท้ายถูกเพิ่มให้รหัสครบ 9 ตัว)



รูปที่ 3.3 แสดงตัวอย่างขั้นตอนการเข้ารหัสซาวน์เด็กซ์ของคำว่า "และ"

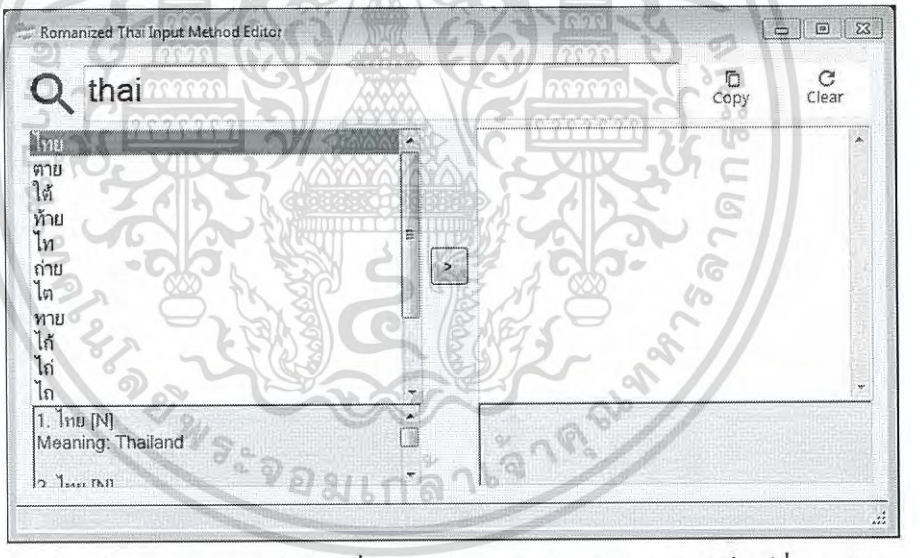
บทที่ 4

ผลการดำเนินงาน

4.1 การเปรียบเทียบความสอดคล้องของรหัสซาวน์เด็กซ์จากอักษรภาษาอังกฤษและจากอักษรภาษาไทย

เมื่อได้ข้อมูลรหัสซาวน์เด็กซ์ที่ใช้เป็นตัวกลางในการถอดอักษรแล้ว ก็จะนำรหัสซาวน์เด็กซ์ภาษาอังกฤษที่ได้มาเทียบกับข้อมูลซาวน์เด็กซ์ภาษาไทยที่ได้จัดเก็บไว้ในขั้นตอนที่ 3.2 วิธีการเปรียบเทียบข้อมูลจะใช้คำสั่ง SQL ในการเปรียบเทียบว่า ให้เลือกคำไทยในสดมภ์ที่ 1 เมื่อรหัสซาวน์เด็กซ์ภาษาอังกฤษตรงกับรหัสซาวน์เด็กซ์ในสดมภ์ที่ 2 จากนั้นจะนำไปแสดงผลลงในโปรแกรมออกมาเป็นรายการคำที่พบเจอ ดังรูปที่ 3.3

เมื่อพิมพ์คำว่า "thai" ลงในช่อง INPUT ซึ่งคำที่ต้องการคือคำว่า "ไทย" ระบบจะคืนค่าคำไทยที่มีรหัสซาวน์เด็กตรงกับคำว่า "thai" ขึ้นมา



รูปที่ 4.1 ตัวอย่างการแสดงผลคำไทยที่พบเจอโดยการเทียบรหัสซาวน์เด็กซ์ที่ตรงกัน

ตารางที่ 4.1 แสดงรายการคำที่มีรหัสตรงกัน เมื่อค้นหาคำว่า thai

คำไทยที่ได้	รหัสซาวน์เด็กซ์
ไทย	T9710500
ตาย	T9710500
ใต้	T9710500
ท้าย	T9710500
ไท	T9710500

ตารางที่ 4.1 (ต่อ) แสดงรายการคำที่มีรหัสตรงกัน เมื่อค้นหาคำว่า thai

คำไทยที่ได้	รหัสซาวน์เด็กซ์
ถ่าย	T9710500
ไต	T9710500
ทาย	T9710500
ใต	T9710500
ไต่	T9710500
ไต้	T9710500
ไต๋	T9710500
ไถ	T9710500
ไถ่	T9710500
ไถ้	T9710500

จะเห็นได้ว่าข้อมูลทั้งหมดที่พบนั้น มีรูปแบบการเขียนที่สอดคล้องกับคำว่า Thai ทั้งสิ้น ส่วนคำที่มีหลายพยางค์ระบบก็นำข้อมูลนั้นออกมาให้เพื่อช่วยให้ผู้ใช้สามารถใช้งานได้ง่ายขึ้น โดยไม่จำเป็นต้องพิมพ์ให้ครบทั้งคำ

4.2 การประเมินผลความถูกต้องแม่นยำของโปรแกรมการถอดอักษร

ในการทดสอบว่าโปรแกรมการถอดอักษรที่พัฒนาขึ้นมานี้ทำงานได้ดีเพียงใด ก็จะต้องมีการตรวจสอบความถูกต้องของคำที่ถอดอักษรได้จากโปรแกรมเปรียบเทียบกับคำที่ถอดอักษรโดยราชบัณฑิตยสถาน ซึ่งผู้วิจัยได้มีการเตรียมข้อมูลเพื่อใช้ทดสอบนี้เป็นข้อมูลคำไทยที่ถอดอักษรเป็นโรมันโดยใช้เกณฑ์การถอดอักษรของราชบัณฑิตยสถาน[3] เป็นข้อมูลจำนวน 4,808 คู่คำ โดยจะเปรียบเทียบระหว่างสดมภ์ของ SOUNDEX_CODE (รหัสที่ได้เมื่อผ่านกระบวนการเข้ารหัส) กับ COMPARE_CODE (รหัสที่ควรได้) แล้วนำค่า Boolean ที่ได้มาจัดเก็บในสดมภ์ RESULT ดังแสดงตารางใน Microsoft Excel ดังนี้

ตารางที่ 4.2 แสดงการเปรียบเทียบรหัสซาวน์เด็กซ์คำภาษาอังกฤษที่ได้กับรหัสซาวน์เด็กซ์คำภาษาไทย

THAI_WORD	SOUNDEX_CODE	COMPARE_CODE	RESULT
กักขัง	K9729752	K9729752	TRUE
กัง	K9752000	K9752000	TRUE
กังวล	K9752811	K9752811	TRUE
กัด	K9730000	K9730000	TRUE
กัน	K9750000	K9750000	TRUE
กั้น	K9750000	K9750000	TRUE
กันเอง	K9751015	K9751015	TRUE
กันและกัน	K9754971	K9754971	TRUE
กันยายน	K9759979	K9759979	TRUE
กับ	K9710000	K9710000	TRUE
กับข้าว	K9712971	K9712971	TRUE
กัมพูชา	K9764111	K9764111	TRUE
กา	K9700000	K9700000	TRUE

ขั้นตอนทั้งหมดที่กล่าวไปเป็นวิธีการดำเนินการวิจัยโดยละเอียด มีการทดสอบและเปรียบเทียบข้อมูลหลายๆครั้ง เนื่องจากข้อมูลมีปริมาณมาก บางครั้งโปรแกรมก็คำนวณออกมาผิดพลาด ดังนั้นจึงต้องทดสอบให้ได้ค่าที่คงที่ที่สุด

จากนั้นจะทำการนับข้อมูลที่มีค่า RESULT เป็น TRUE นั่นหมายถึงข้อมูลที่เปรียบเทียบกันมีความถูกต้องได้รหัสที่ตรงกัน จากนั้นก็จะนำมาหาค่าเฉลี่ยของข้อมูลทั้งหมดที่ได้จัดเก็บว่ามีความถูกต้องคิดเป็นกี่เปอร์เซ็นต์ และผลความถูกต้องแม่นยำของโปรแกรมที่พัฒนาในบทต่อไปจะกล่าวถึงผลการถอดอักษรโดยใช้โปรแกรมคอมพิวเตอร์กับตัวกลางในการถอดอักษร

4.3 การแก้ไขอัลกอริทึมเพื่อเพิ่มความแม่นยำในการทำงาน

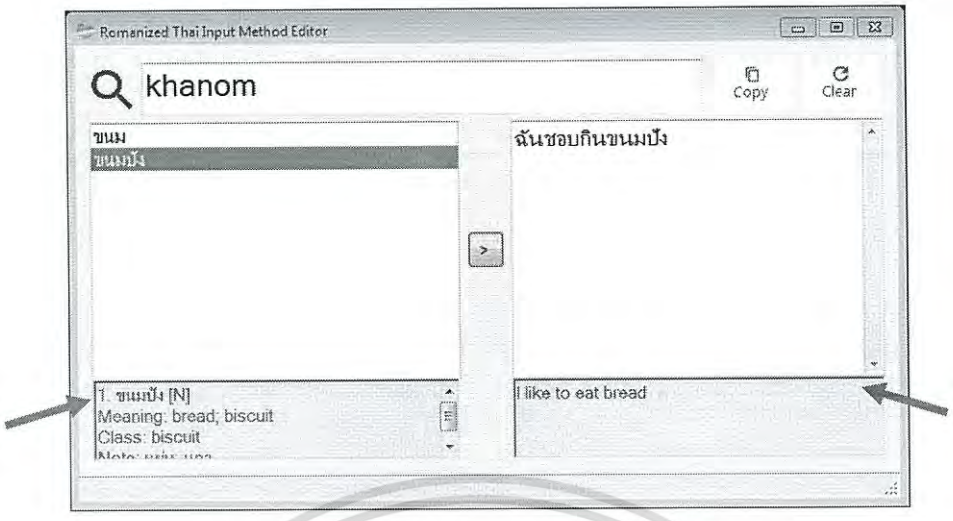
หลังจากที่เราได้ทดสอบการทำงานของโปรแกรม จะพบว่าผลจากการค้นคำมีจำนวนมาก และคำบางคำที่ได้ค่อนข้างที่จะไม่มีความสอดคล้องกับคำที่ป้อนลงไปในโปรแกรมเช่น คำที่มีตัวสะกดเป็น ม จะแสดงพร้อมกับคำที่มีตัวสะกดเป็นตัว น เนื่องจากพบว่ารหัสซาวน์เด็กซ์ที่นำมาใช้กำหนดให้ พยัญชนะเสียง ม และ พยัญชนะเสียง น มีการเข้ารหัสเป็นรหัสตัวเดียวกันคือ 5 ดังนั้นจึงกำหนดรหัสใหม่ให้พยัญชนะเสียง ม มีรหัสเป็น @ เพื่อลดความผิดพลาดในการค้นคำ

หรือกรณีที่ผู้ใช้ไม่ได้ใช้หลักการถอดอักษรในการพิมพ์คำไทยด้วยตัวอักษรภาษาอังกฤษ เช่น ใช้ตัวอักษร J แทน จ ซึ่งตามมาตรฐานจะต้องใช้ CH ผู้วิจัยได้มีการแก้ไขอัลกอริทึมให้รองรับความเคยชินของผู้ใช้งานมากขึ้น ผู้ใช้สามารถใช้ตัวอักษรที่ไม่ตรงตามมาตรฐานได้ แต่ทำได้เพียงบางตัวเท่านั้น เพราะยังมีข้อจำกัดอยู่บางประการ

รูปที่ 4.2 ตัวอย่างการแสดงผลข้อมูลด้วยการใช้ตัวอักษรที่ไม่ได้ระบุในกฎการถอดอักษรโดยราชบัณฑิตยสถาน

4.4 กำกับความหมายและประเภทของคำศัพท์ในคลังข้อมูล

เพื่อให้ผู้ใช้สามารถเลือกใช้คำได้อย่างถูกต้อง เนื่องจากคำในภาษาไทยมีคำพ้องรูปและคำพ้องเสียงมากมาย จึงต้องมีการกำกับคำศัพท์ด้วยความหมายและประเภทของคำเพิ่มเติม โดยได้ใช้ Google api มาเป็นตัวช่วยในการแปลความหมายของคำ ความหมายที่ได้จะใช้เป็นภาษาอังกฤษซึ่งสามารถอ่านเข้าใจได้ในระดับสากล ทั้งชาวไทยและชาวต่างชาติ



รูปที่ 4.3 ตัวอย่างการแสดงความหมายประกอบของคำที่ค้นหา

4.5 สรุปผลจากการศึกษาแนวทางที่เหมาะสมในการพัฒนาระบบการถอดอักษรอังกฤษเป็นไทยแบบถ่ายเสียง

ผู้จัดทำได้ทำการทดสอบโดยใช้คู่คำไทย – อังกฤษ ทั้งหมด 4,807 คู่คำ หลังจากที่ได้มีการปรับปรุงอัลกอริทึม ทำให้ค่าความถูกต้องของอัลกอริทึมรหัสซาวน์เด็กซ์ของคำไทยและอังกฤษตรงกันเพิ่มขึ้น

จากผลการทดสอบพบว่ามีบางคำที่ระบบไม่สามารถดึงข้อมูลออกมาแสดงผลได้ เนื่องจากบางคำอัลกอริทึมไม่สามารถแยกเสียงของตัวอักษรนั้นออกมาได้อย่างถูกต้อง โดยตัวอักษรที่เป็นปัญหามากที่สุดคือ อักษร ย และ อ เนื่องจากความหลากหลายในการอ่านออกเสียงสองตัวนี้ในคำต่างๆ ในขั้นตอนของการแปลงสระให้เป็นตัวอักษรโรมัน เช่น คำว่า ทายาท ระบบแยกคำออกมาได้เป็น ทาย –าท ซึ่งไม่ถูกต้อง เมื่อนำคำไปสืบค้นจึงไม่ปรากฏคำนี้ขึ้น

ประเด็นสำคัญที่ทำให้คำระบบไม่สามารถค้นคำที่ต้องการออกมาแสดงผลได้คือ ขั้นตอนของการเรียงลำดับตัวอักษรใหม่ให้สอดคล้องกับลำดับของการเขียนในรูปของคำทับศัพท์ เพราะถ้าหากมีการเรียงลำดับตัวอักษรผิด รหัสที่ได้ก็จะผิดไปด้วยเพราะระบบจะต้องนำเอาลำดับของตัวอักษรไปพิจารณาว่าสอดคล้องกับพยัญชนะ หรือสระตัวใดในภาษาไทย และอีกประเด็นหนึ่งที่สำคัญไม่แพ้กัน คือความหลากหลายของการออกเสียง โดยปกติแล้วชาวต่างชาติมักจะถอดเสียงคำไทยโดยใช้เสียงของภาษาดั้งเดิมของตนเอง ซึ่งจะแตกต่างกับภาษาไทย เช่น คำว่า เป็น ชาวต่างชาติ

ส่วนใหญ่มักถอดเสียงออกมาเป็น ben , bin หรือเสียง C , G ชาวต่างชาติส่วนใหญ่มักออกเสียงเป็นเสียง ก หรือ จ ทำให้ระบบไม่สามารถค้นเจอคำที่การออกมาอย่างหลีกเลี่ยงไม่ได้

จากปัญหาที่กล่าวมาข้างต้น ผู้วิจัยจึงได้เพิ่มส่วนของการแสดงความหมายของคำขึ้นมา หากผู้ใช้ออกเสียงผิดไปจากกฎการถอดอักษรหรือใช้ความคุ้นชินของเสียงมาเป็นหลัก เพื่อช่วยให้เกิดความแม่นยำ และคำที่แสดงผลออกมานั้นได้เรียงลำดับตามความถี่ในการใช้งานในปัจจุบัน

บทที่ 5

สรุปผลโครงงาน

5.1 สรุปผลการทำโครงงานและข้อเสนอแนะ

โครงงานนี้เกี่ยวกับการสร้างวิธีขั้นตอนในการพิมพ์ภาษาไทยด้วยตัวอักษรภาษาอังกฤษ เพื่อใช้เป็นเครื่องมือในการศึกษาภาษาไทยและเป็นการช่วยในการพิมพ์ภาษาไทยด้วยตัวอักษรภาษาอังกฤษโดยคำนึงจากการอ่านออกเสียงของตัวอักษรและการประสมกันเป็นพยางค์หรือคำนั้นๆ โดยได้ใช้หลักการจับคู่ตัวอักษรในการแยกสระกับพยัญชนะ และใช้ซาวน์เด็กซ์อัลกอริทึมในการเข้ารหัสเพื่อจำแนกคำด้วยคำอ่านของคำศัพท์

และด้วยข้อจำกัดในการใช้ฐานข้อมูลคำศัพท์ ซึ่งเป็นวิธีที่ไม่เหมาะสมสำหรับการทำ Romanization เพราะอาจจะมีคำใหม่ๆเกิดขึ้นได้ในอนาคตอยู่ตลอดเวลา จึงไม่เหมาะสมที่จะเผยแพร่โปรแกรมที่พัฒนาเพื่อใช้เป็นเครื่องมือการเขียนคำทับศัพท์สำหรับผู้ใช้โดยทั่วไป ผู้วิจัยจึงคาดหวังไว้ว่าต้องการให้รายงานฉบับนี้เป็นข้อมูลอ้างอิง หรือเป็นโปรแกรมต้นแบบเพื่อนำไปศึกษาและพัฒนาให้มีประสิทธิภาพต่อไปในอนาคต

5.2 ปัญหาที่พบระหว่างดำเนินโครงงาน

1. หลักการถอดอักษรไทยเป็นอักษรโรมันแบบถ่ายเสียงของราชบัณฑิตยสถานยังมีความไม่ชัดเจนในเรื่องของรายละเอียดในภาษาไทย โดยที่พบคือ
   - สระเสียงสั้นเสียงยาว ใช้ตัวอักษรเดียวกัน เช่น อะ กับ อา ใช้ตัว 'A' เหมือนกัน
   - ตัวอักษรภาษาไทย 1 เสียงบางตัวอักษรมีหลายรูป อีกทั้งมีตัวที่เสียงไม่เหมือนกันใช้ตัวเดียวกัน เช่น 'ส ศ ษ' กับ 'ซ' แทนด้วย 'S'

   โดยปัญหาเหล่านี้ทำให้การพิมพ์ไม่สามารถเจาะจงคำ หรือทำให้ผลลัพธ์ที่ได้มีขอบเขตกว้างเกินไป ส่งผลให้ผู้ใช้

2. พยัญชนะต้น หรือคำควบกล้ำ ควบกล้ำทั้งแท้และควบกล้ำไม่แท้ในภาษาไทยทำให้ระบบถอดตัวอักษรทำงานได้ไม่ดีนัก เช่น จรวด จรด ทราบ

3. ตัวสะกดบางตัวทำให้ระบบไม่สามารถถอดออกมาให้ตรงกับคำนั้นได้ เช่น มาตรฐาน สามารถ การใช้ตัวสะกด ตร หรือ รถ เป็นการยืมคำมาจากภาษาอื่น

4. อัลกอริทึมซาวน์เด็กซ์ที่นำมาใช้ยังมีข้อบกพร่องอยู่มาก จึงต้องมีการปรับเปลี่ยนเพื่อให้เกิดความถูกต้องแม่นยำมากขึ้น

5.3 ประโยชน์ที่ได้รับจากการดำเนินโครงงาน

จากการจัดทำโครงงาน 1 ผู้จัดทำได้รับประโยชน์ดังนี้

1. ได้ศึกษาทฤษฎีความรู้ถึงความคล้ายคลึงกันของภาษาต่างๆ
2. ฝึกการแก้ไขปัญหาการลดความบกพร่องของระบบ

5.4 แนวทางการพัฒนาในอนาคต

- พัฒนาให้ระบบสามารถถอดอักษรได้แม่นยำมากขึ้นที่ระดับ 100%
- พัฒนาขั้นตอนวิธีที่ครอบคลุมในส่วนของคำอังกฤษทับศัพท์ภาษาไทย และภาษาไทยทับศัพท์อังกฤษให้เป็นขั้นตอนวิธีเดียว เพื่อให้เกิดความสะดวกในการใช้งาน
- มีระบบแนะนำความน่าจะเป็นของคำต่อไป โดยใช้หลัก n-gram มาพิจารณาความเป็นไปได้ของคำถัดไป
- พัฒนาให้รองรับการใช้งานบนมือถือ

บรรณานุกรม

[1] ประยุทธ สุวรรณวิสารท. **การเข้ารหัสคำทับศัพท์เพื่อการค้นคืนข้ามภาษาไทย-อังกฤษ.** วิทยานิพนธ์ปริญญามหาบัณฑิต ภาควิชาวิศวกรรมคอมพิวเตอร์ คณะวิทยาศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย, 2541.

[2] วิโรจน์ อรุณมานะกุล. **อักขรวิธีไทยและการถอดอักษรระหว่างภาษาไทยและภาษาอังกฤษ.** พิมพ์ครั้งที่ 1 กรุงเทพมหานคร : โครงการเผยแพร่ผลงานวิชาการ คณะอักษรศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย, 2551

[3] T. Karoonboonyanan, V. Sornlertlamvanich and S. Meknavin. **A Thai Soundex System for Spelling Correction. Proceedings of the Natural Language Processing Pacific Rim Symposium (1997)** : 633-636.

[4] วรรณี อุดมพานิชย์. **การใช้หลักคำพ้องเสียง เพื่อค้นหาชุดอักขระภาษาไทยที่ออกเสียงเหมือนกัน.** วิทยานิพนธ์ปริญญามหาบัณฑิต ภาควิชาวิศวกรรมคอมพิวเตอร์ บัณฑิตวิทยาลัย จุฬาลงกรณ์มหาวิทยาลัย, 2526.

[5] ราชบัณฑิตยสถาน. **ประกาศราชบัณฑิตยสถาน เรื่อง การถอดอักษรโรมัน**, 2482.

[6] วัลย์วรา ไชยฤกษ์, **การพัฒนาโปรแกรมการถอดอักษรภาษาอังกฤษเป็นไทยโดยใช้คลังคำทับศัพท์ของราชบัณฑิตยสถาน**, วิทยานิพนธ์ปริญญามหาบัณฑิต ภาควิชาภาษาศาสตร์คณะอักษรศาสตร์ บัณฑิตวิทยาลัย จุฬาลงกรณ์มหาวิทยาลัย, 2547.

[7] Suwanvisat, Prayut and Prasitjutrakul, Somchai. 1998. **Thai-English Cross-Language Transliterated Word Retrieval using Soundex Technique**. The National Computer Science and Engineering Conference, Kasetsart University, Bangkok, Thailand.

[8] Suwanvisat, Prayut and Prasitjutrakul, Somchai, 1999. **Transliterated Word Encoding and Retrieval Algorithms for Thai-English Cross-Language Retrieval**. The National Computer Science and Engineering Conference, Bangkok, Thailand.

บรรณานุกรม (ต่อ)

[9] A.B. Grisworld. **Thought on the Romanization of Siamese.** Printed for distribution at the obsequies of Colonel Kasem Nandhakij, February 25, 1969

[10] Vajiravudh, King. 1912. **The Romanization of Siamese Words. In Journal of the Siam Society**, Vol.9, Part 4, 1-10.

[11] Travis Leithead, Takayoshi Kochi, Kenji Baheux, Hironori Bono. **"Input Method Editor PI"**[Online]. Available: http://www.w3.org/TR/ime-api. 2015.

[12] **"ขั้นตอนวิธีเชิงสัทลักษณ์"** [Online]. Available: https://th.wikipedia.org/wiki/ขั้นตอนวิธีเชิงสัทลักษณ์.

[13] "Soundex" [online]. Available: https://en.wikipedia.org/wiki/Soundex

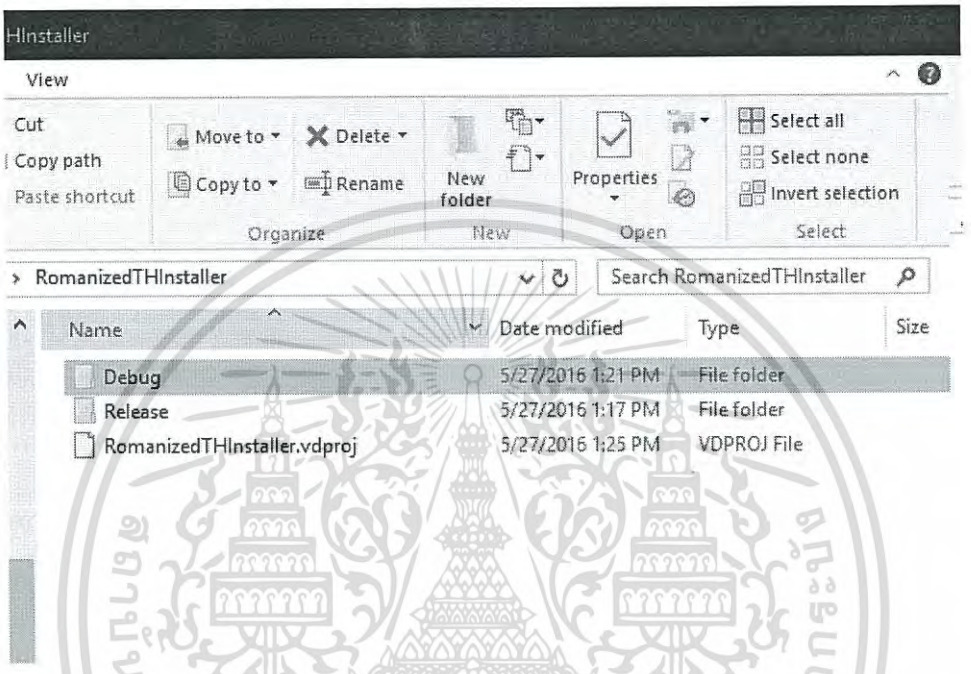
ภาคผนวก

คู่มือการใช้โปรแกรม

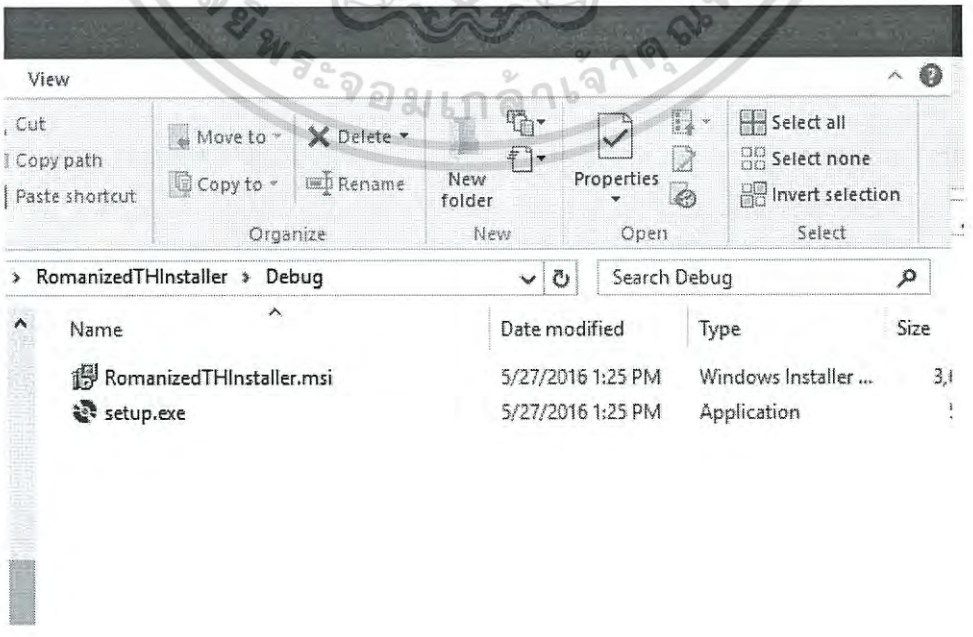
คู่มือการติดตั้งโปรแกรม

Romanized Thai Input Method Editor

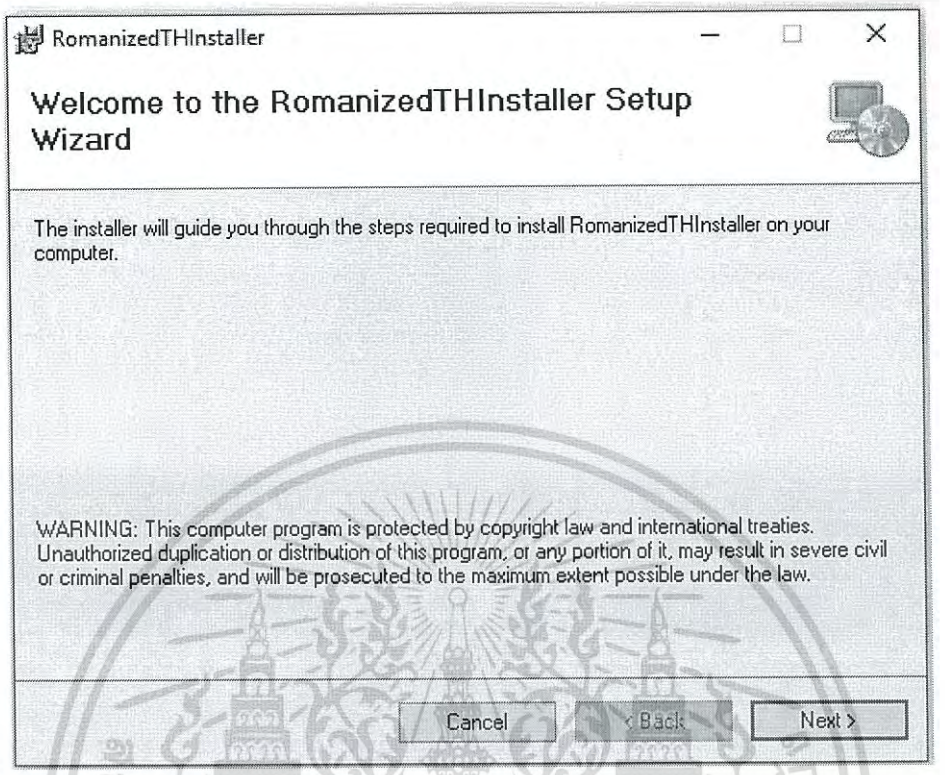
1. เข้าไปใน folder ของโปรแกรมจากนั้นเข้าไปที่ folder ชื่อ Debug



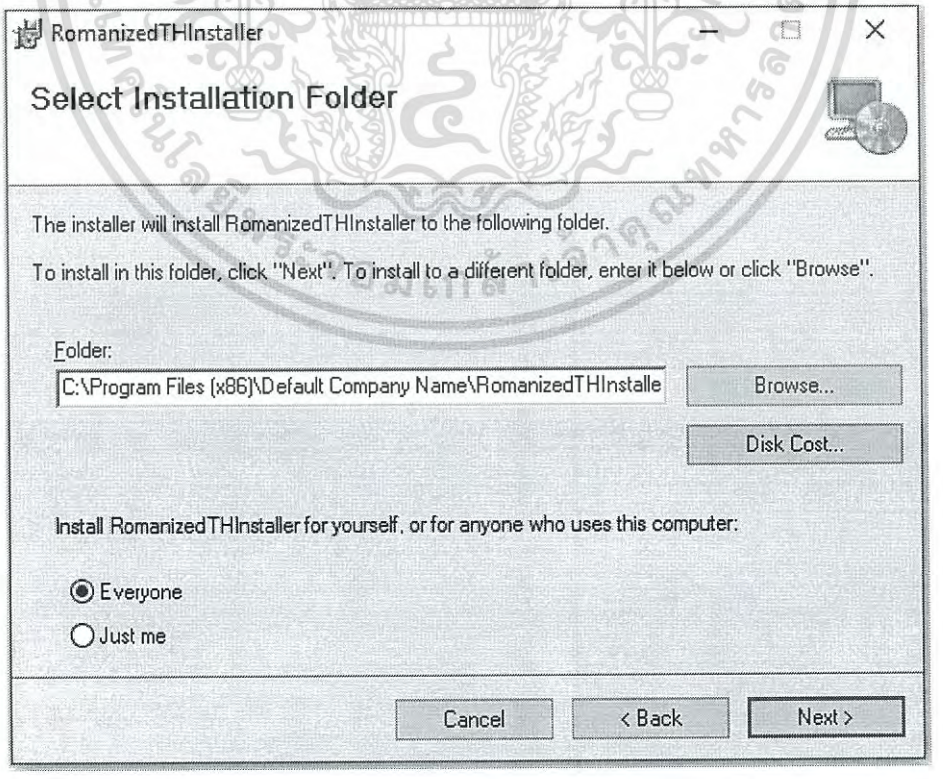
2. ดับเบิ้ลคลิกที่ RomanizedTHInstaller.msi



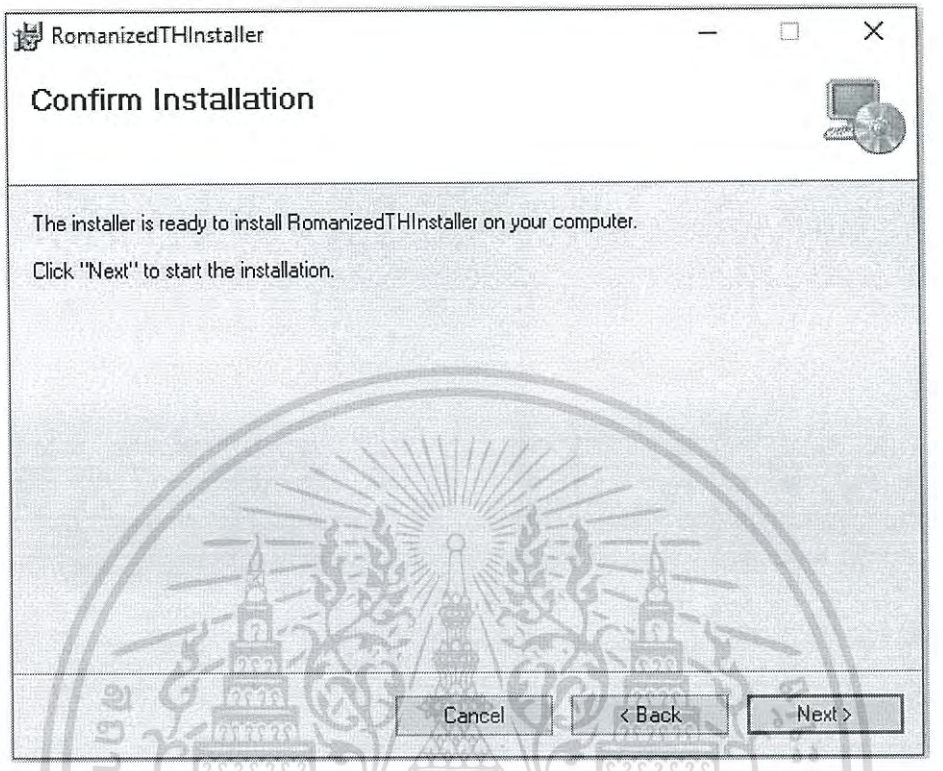
3. คลิก Next



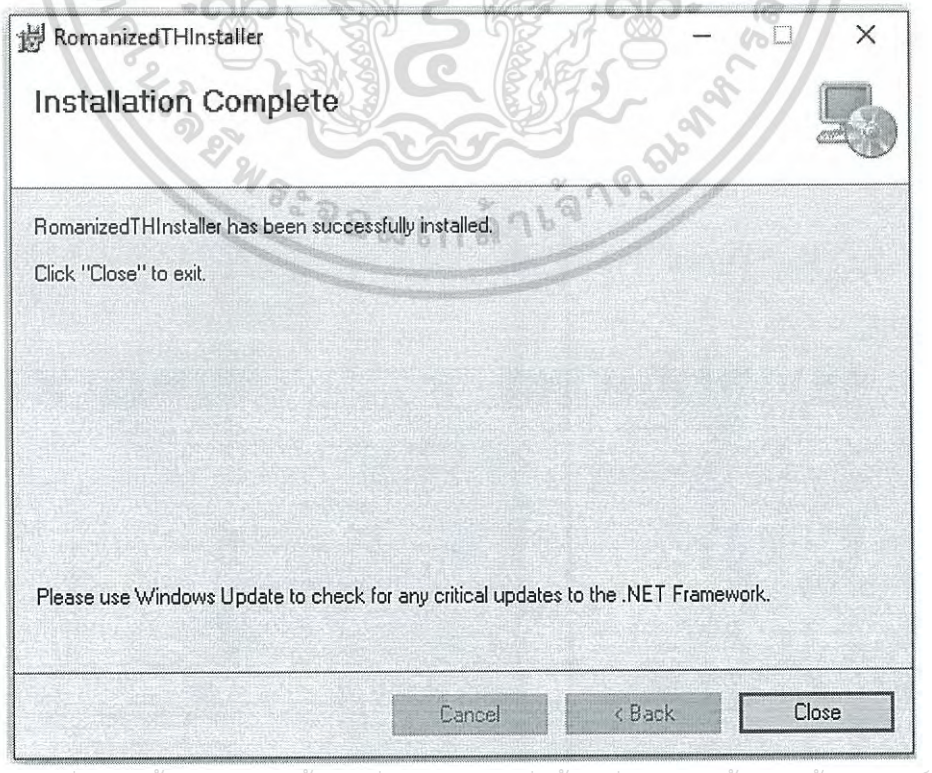
4. คลิก Browse เพื่อเลือกพื้นที่ ที่ต้องการติดตั้งจากนั้น คลิก Next



5. หลังจากขั้นตอน Select Installation Folder เสร็จแล้วให้คลิก Next อีกครั้ง เพื่อเป็นการยืนยันการ Install จากนั้นรอจนกว่าจะ ดำเนินการเสร็จ



6. หลังจากตัว Install ดำเนินการเสร็จแล้วให้คลิก Close



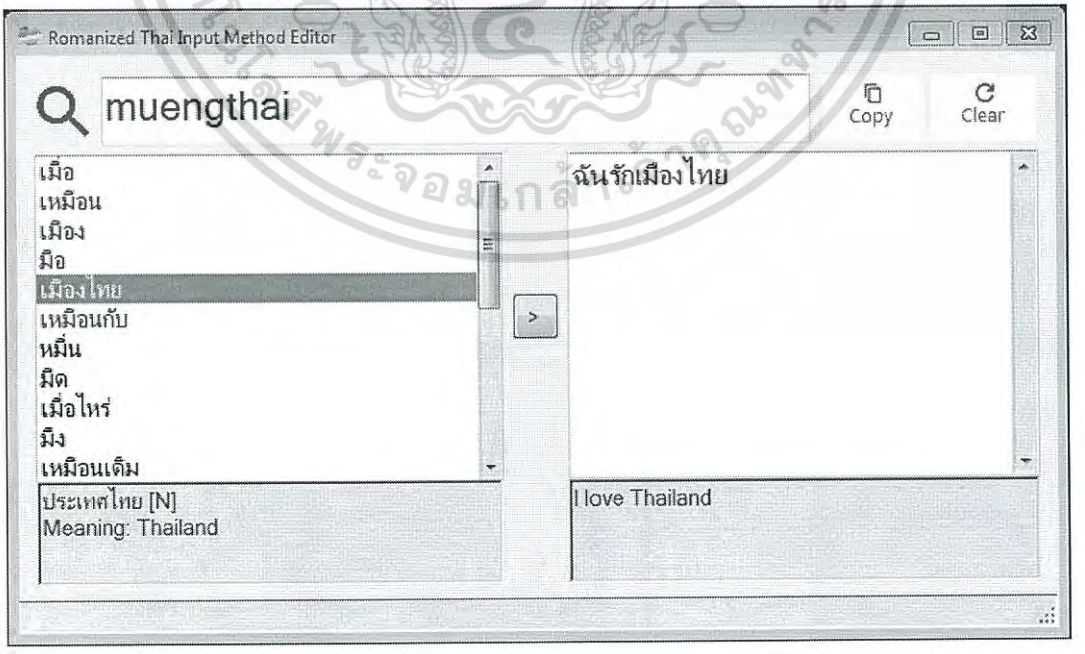
คู่มือการใช้งานโปรแกรม

Romanized Thai Input Method Editor

โปรแกรม Romanized Thai Input Method Editor เป็นโปรแกรมที่ช่วยแปลงคำไทยที่พิมพ์ด้วยตัวอักษรโรมัน ให้เป็นคำไทยที่พิมพ์ด้วยตัวอักษรไทย จุดมุ่งหมายของโปรแกรมคือต้องการช่วยเหลือชาวต่างชาติที่ต้องการศึกษาภาษาไทย หรือต้องการพิมพ์ภาษาไทยให้ได้อย่างถูกต้อง แหล่งที่มาของคำไทยในโปรแกรมนั้น ได้มาจากคณะอักษรศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย และศูนย์เทคโนโลยีอิเล็กทรอนิกส์และคอมพิวเตอร์แห่งชาติ (NECTEC) ซึ่งได้จากการรวบรวมข่าวสาร บทความ นวนิยายต่างๆมารวมกัน เพื่อให้มีประโยชน์ในการใช้งานในชีวิตประจำวันมากที่สุด โดยคำศัพท์ที่ค้นเจอจะเป็นคำศัพท์ที่ถูกเรียงตามความถี่ของการใช้งานในปัจจุบัน นอกจากนี้ยังช่วยเหลือให้ชาวต่างชาติที่ไม่แน่ใจในคำศัพท์ไทยด้วยการแสดงความหมายของรายการคำไทยที่ค้นเจอเป็นภาษาอังกฤษ และแสดงความหมายรวมทั้งประโยคเมื่อผู้ใช้ได้เลือกรายการคำเหล่านั้นแล้วนำมาสร้างเป็นประโยคใหม่ขึ้น

ขั้นตอนการใช้งาน

1. เปิดโปรแกรม Romanzied Thai Input Method Editor ขึ้นมา
2. ช่อง  ใช้พิมพ์คำไทยที่ต้องการค้นหาด้วยตัวอักษรโรมัน
3. เมื่อพิมพ์คำไทยในข้อที่ 2 แล้วระบบจะทำการค้นหาคำไทยที่ตรงกันในฐานข้อมูล มาแสดงในช่องแสดงรายการคำศัพท์ พร้อมทั้งแสดงความหมายของแต่ละรายการคำศัพท์ที่ค้นเจอ
4. กดปุ่ม  เพื่อเลือกรายการคำที่ต้องการจากช่องแสดงรายการคำศัพท์ที่ค้นเจอ หรือผู้ใช้สามารถใช้ปุ่ม enter บนแป้นพิมพ์เพื่อเลือกคำศัพท์ก็ได้ ระบบจะนำมาแสดงในช่องรายการคำศัพท์ที่เลือกทางด้านขวา และพร้อมแสดงความหมายเมื่อนำคำศัพท์ที่เลือกมารวมกันเป็นประโยค
5. ปุ่ม  จะทำการคัดลอกรายการคำที่ช่องแสดงรายการคำศัพท์ที่เลือกทางด้านขวา
6. ปุ่ม  จะทำการล้างข้อความทั้งหมดที่อยู่ในช่องป้อนข้อความ และช่องแสดงรายการคำศัพท์ที่เลือก



ตัวอย่างการใช้งาน เมื่อพิมพ์ chan rak muengthai ตามลำดับ

ระบบป้อนข้อมูลภาษาไทยด้วยตัวอักษรโรมัน

วิมลสิริ โพธิ์เกษม และ สิทธานต์ รัตนเหลี่ยม

[1]คณะเทคโนโลยีสารสนเทศ สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง

Emails: wimonsiri.pks@gmail.com, sittan.rattanaliam@gmail.com

บทคัดย่อ

ระบบป้อนข้อมูลภาษาไทยด้วยตัวอักษรโรมันเป็นวิธีการหนึ่งที่จะช่วยให้พิมพ์ภาษาไทยด้วยการใช้ตัวอักษรโรมันได้ เนื่องจากปัจจุบันชาวต่างชาติได้เข้ามามีบทบาทในประเทศไทยมากขึ้น มีการติดต่อสื่อสารกับชาวต่างชาติเพิ่มขึ้น ราชบัณฑิตยสถานจึงได้ประกาศกฎเกณฑ์การถอดอักษรไทยด้วยตัวอักษรโรมันขึ้น เพื่อใช้เป็นมาตรฐานในการเขียนข้อความหรือคำภาษาไทยด้วยการใช้ตัวอักษรโรมันเพื่ออำนวยความสะดวกให้กับชาวต่างชาติมากขึ้น ดังที่เห็นในแผนที่ หรือป้ายบอกทางต่างๆ ในวิทยานิพนธ์นี้ผู้จัดทำจึงได้พัฒนาระบบป้อนข้อมูลภาษาไทยด้วยตัวอักษรโรมันขึ้น โดยยึดหลักเกณฑ์การถอดอักษรไทยด้วยตัวอักษรโรมันแบบถ่ายเสียงของราชบัณฑิตยสถานมาใช้ประกอบการพัฒนา วิทยานิพนธ์ฉบับนี้ได้นำเสนอขั้นตอนวิธีการถอดอักษรโดยใช้วิธีการเข้ารหัสผ่านตัวกลางโดยใช้ ซาวน์เด็กซ์ ซึ่งเป็นวิธีที่จะอนุญาตให้ใช้คำภาษาไทยโดยใช้ตัวอักษรโรมันได้เพื่อค้นคืนคำไทยที่หลักการเขียนที่ตรงกันในอีกภาษาหนึ่งโดยจะมีตัวกลางที่ตรงกันสำหรับสองภาษา ขั้นตอนวิธีที่นำเสนอแบ่งออกเป็นสองส่วนคือ ขั้นตอนวิธีการเข้ารหัสตัวอักษรภาษาไทยและขั้นตอนวิธีการเข้ารหัสตัวอักษรภาษาอังกฤษ ผลการทดลองจะแสดงให้เห็นว่าขั้นตอนวิธีการเข้ารหัสภาษาไทยด้วยตัวอักษรโรมันโดยใช้วิธีซาวน์เด็กซ์สามารถแสดงผลออกมาได้สูงถึง 90 เปอร์เซ็นต์ นอกจากนั้นจะมีปัญหาในเรื่องความกำกวมในการออกเสียงของคำนั้นๆ และการสะกดคำที่แตกต่างกันในแต่ละบุคคลก็ส่งผลให้ระบบไม่สามารถค้นคืนคำไทยออกมาได้

คำสำคัญ – Romanization; Thai; Automatic Thai Romanization; Input method editor

1. บทนำ

การถอดอักษรไทย-โรมันเริ่มใช้ในช่วงทศวรรษที่ 17 คิดค้นขึ้นโดยกลุ่มมิชชันนารี [4] อย่างไรก็ตามยังไม่มีระบบที่สามารถทำให้ชาวต่างชาติออกเสียงได้อย่างถูกต้อง ในปัจจุบันได้มีการทำระบบการถอดอักษรโรมันอยู่มากรวมถึงการใช้คำทับศัพท์ ต่อมาในปี ค.ศ. 1998 ได้มีการกำหนดมาตรฐาน ISO-11940 และถูกแก้ไขพัฒนาเพิ่มเติมในปี ค.ศ. 2003 โดยระหว่างการพัฒนานี้ ทางราชบัณฑิตยสถานได้ตีพิมพ์หลักเกณฑ์การถอดอักษรไทยเป็นโรมันแบบถ่ายเสียงในปี ค.ศ. 1999[2] แต่มาตรฐานนี้ยังคงต้องการการตรวจสอบความถูกต้อง ซึ่งปัจจุบันบุคคลทั่วไปจึงนิยมถอดอักษรตามแบบของตัวเองมากกว่าใช้หลักของราชบัณฑิตยสถานโดยใช้อัลกอริทึม

ซาวน์เด็กซ์(เราจะอธิบายในตอนที่ 3) คลังข้อมูลคำศัพท์ที่ใช้ถูกรวบรวมมาจากศูนย์เทคโนโลยีอิเล็กทรอนิกส์และคอมพิวเตอร์แห่งชาติ (NECTEC) ก่อนทำการถอดอักษรไทย-โรมัน คำศัพท์ภาษาไทยจะถูกนำมาเรียงใหม่ตามตัวอักษรที่ใช้แทนในเสียงภาษาอังกฤษ และรูปวรรณยุกต์จะถูกตัดออก โดยจะถูกเรียงโดยขั้นตอนวิธีด้วยคอมพิวเตอร์ หลังจากนั้นจะตรวจสอบและแก้ไขให้อีกถูกต้องโดยบุคคลอีกรอบ เป็นไปได้ที่จะมีตัวอักษรภาษาอังกฤษมากกว่าหนึ่งตัวเพื่อใช้ในการแทนตัวอักษรภาษาไทย เช่น ‘th’ แทน ‘ท’, ‘ia’ แทน ‘เ- ีย’ ยกตัวอย่างขั้นตอนการเรียงตัวอักษรใหม่ได้ดังต่อไปนี้

เสีย	→	สเ ีย	→	Sia
ร่างกาย	→	รางกาย	→	Rangkai
และ	→	ลแะ	→	Lae
ทำให้	→	ทำหใ	→	Tamhai
เสียง	→	สเ ียง	→	Siang

โดยในบทความนี้เราจะแบ่งออกเป็น 5 ส่วน ส่วนที่ 2 คือ การสลับตำแหน่งสระประสม ส่วนที่ 3 คือ อัลกอริทึมซาวน์เด็กซ์ ส่วนที่ 4 คือ การเข้ารหัสตัวอักษร และส่วนที่ 5 คือ สรุปผล

การอ่านออกเสียงซึ่งจะทำให้ชื่อที่อ่านออกเสียงเหมือนหรือคล้ายกันมีรหัสเสียงที่เหมือนกัน อัลกอริทึมนี้จึงมีพื้นฐานมาจากชื่อที่เป็นภาษาอังกฤษและยังเข้ารหัสเพียงพยัญชนะเท่านั้น การเข้ารหัสเสียงนี้จะแปลงรหัสตาม ตารางที่ 1.จากนั้นรหัสที่เป็นเลข 0 จะถูกตัดออกทั้งหมด และรหัสที่ซ้ำกันและอยู่ติดกันจะถูกลดให้เหลือเพียงตัวเดียว ขั้นตอนสุดท้ายคือให้เรียงโดยใช้ตัวอักษรแรกเป็นรหัสตามด้วยรหัสตัวเลขสามตัวเลขที่แปลงได้ ตัวอย่างเช่น ALEXANDER จะถูกแปลงเป็น A425

2. การสลับตำแหน่งสระประสม

ก่อนการถอดอักษรไทย-โรมัน เราต้องเตรียมข้อมูลเพื่อสร้างเป็นคลังคำศัพท์ที่รวบรวมการจับคู่กันระหว่างคำศัพท์ภาษาไทยและคำอ่านภาษาอังกฤษ วิธีการนี้เราเรียกว่าการจับคู่ข้อมูล วิธีนี้จะจับคู่คำศัพท์ภาษาไทยกับคำภาษาอังกฤษที่มีคำอ่านเหมือนหรือคล้ายกับคำศัพท์ภาษาไทยนั้นๆ ในขั้นตอนนี้ตัวอักษรจะถูกแยกและนำมาเรียงเป็นพยางค์โดยอัลกอริทึมบนคอมพิวเตอร์ [5] เพื่อการนำมาเรียงได้อย่างถูกต้อง เราต้องกำหนดคู่ตัวอักษรของภาษาไทยและภาษาอังกฤษก่อน แต่ตัวอักษรบางกลุ่มยังไม่สามารถหาคู่ที่ถูกต้องได้เนื่องจากการผันเสียงหรือการงดออกเสียงของคำ เช่น ‘ทร’สามารถออกเสียงได้ 2 แบบ ทั้งเสียง ‘ทร’ และเสียง ‘ซ’ ในขั้นตอนนี้จึงมีความสำคัญมากต่อการนำไปเข้ารหัสในขั้นตอนต่อไป

3. ซาวน์เด็กซ์อัลกอริทึม

ซาวน์เด็กซ์ถูกใช้ในการตรวจสอบการสะกดคำ [1]และการสืบค้นข้ามภาษา[3] ในขั้นตอนนี้เราพบว่าการจับคู่เสียงของตัวอักษรระหว่างภาษาไทยกับภาษาอังกฤษไม่เป็นความสัมพันธ์แบบหนึ่งต่อหนึ่ง เนื่องจากตัวอักษรไทยหนึ่งเสียงสามารถเขียนได้ในหลายรูป เพื่อการแก้ปัญหานี้เราได้ใช้อัลกอริทึมซาวน์เด็กซ์เริ่มแรกถูกพัฒนาขึ้นโดย M.K. Odell and R.C.

Russell ถูกใช้ในการเข้ารหัสเสียงของชื่อตามเสียงพูดของพยัญชนะ(ไม่สนใจตัวอักษรทางซ้ายสุด)เป็นตัวเลขตาม

หลักการเข้ารหัสเหมือนกับของ M.K. Odell and R.C. Russell คือใช้ตัวแรกเป็นตัวอักษรและตามด้วยตัวเลขโดยไม่สนใจวรรณยุกต์และสระ

ตารางที่ 1. การจับคู่รหัสซาวน์เด็กซ์กับตัวอักษรภาษาอังกฤษของ M.K. Odell and R.C. Russell

ตัวอักษร	รหัสตัวเลข
A E I O U H W Y	0
B F P V	1
C G J K Q S X Z	2
D T	3
L	4
M N	5
R	6

ซาวน์เด็กซ์อัลกอริทึมเป็นวิธีที่เร็วและค่อนข้างมีความแม่นยำ แต่ก็อาจจะพบปัญหาได้หากตัวอักษรนั้นออกเสียงไม่ตรงกับที่กำหนดไว้

4. การเข้ารหัสตัวอักษร

สำหรับการเข้ารหัสซาวน์เด็กซ์ในภาษาไทยโดยส่วนใหญ่จะมีต้นแบบมาจากการเข้ารหัสซาวน์เด็กซ์ของ M.K. Odell and R.C. Russell เช่น ของคุณประยุทธ สุวรรณวิสารท ได้ใช้

- การใช้ตัวอักษรตัวแรกเป็นรหัส ถ้าเป็นสระให้ใช้อักษร 'A' เป็นตัวแทนสำหรับสระนั้นๆ
- ขยายความยาวของรหัสซาวน์เด็กซ์ที่ได้เป็น 9 ตัว จาก 4 ตัว เพื่อรหัสที่มีความเจาะจงมากขึ้น
- ใช้รหัส ASCII แทนสระ (ตามตารางที่ 4) และแปลงสระเป็นรูปของอักษรภาษาอังกฤษ เช่น 'เ-า' เป็น 'ao', 'ไ' เป็น 'ai' (แสดงในตารางที่ 4)
- ไม่คำนึงถึงเสียงสูงต่ำในขั้นตอนการเข้ารหัสซาวน์เด็กซ์

ในโครงงานนี้ได้นำอัลกอริทึมซาวน์เด็กซ์ภาษาไทยที่พัฒนาโดยคุณประยุทธ สุวรรณวิสารท มาดัดแปลงเพิ่มเติม โดยจะเข้ารหัสตัวเลขทั้งพยัญชนะและสระ

ตารางที่ 2. แสดงการจับคู่รหัสสำหรับตัวอักษรตัวหลังจากตัวแรก(รหัสตัวเลข)

ตัวอักษรอังกฤษ	ตัวอักษรไทย	รหัส	หมายเหตุ
B F P V Ph	บ ฝ ฟ ป ผ พ ภ	1	
K C G Kh Ch S	ก ข ฃ ค ฅ ฆ จ ฉ ช ฌ ซ ศ ษ ส	2	
	จ ฉ ช ฌ ซ ศ ษ ส	3	#1
D T	ฎ ด ฏ ต ฐ ฑ ฒ ณ ถ ท ธ	3	
L	ล ฬ	4	
M	ม	@	
N	ณ น	5	
R	ร	6	
	ร ญ ล ฬ	5	#1
H	ห ฮ	7	
W	ว	8	
Y	ญ ย	9	
Ng	ง	52	

#1 : สำหรับตัวอักษรตัวสุดท้าย

จากตารางที่ 2. ได้กำหนดหลักเกณฑ์และแก้ไขกฎในการเข้ารหัสดังนี้

- เพิ่มรหัส 'NG' ตัวอักษร 'ง' เมื่อเป็นตัวอักษรตัวแรกของคำ ซึ่งเกี่ยวข้องกับรหัสเลขของ 'ง' คือ 52 (N มีรหัส 5, G มีรหัส 2)

ตารางที่ 3. แสดงการจับคู่รหัสสำหรับตัวอักษรตัวแรก

ตัวอักษรอังกฤษ	ตัวอักษรไทย	รหัส
A E I O U	อ	A E I O U
B	บ	B
F V	ฝ ฟ	F
Ph P	ป ผ พ ภ	P
K Kh C G	ก ข ฃ ค ฅ ฆ	K
C Ch	จ ฉ ช ฌ	C
S	ซ ศ ษ ส	S
D	ฎ ด	D
T	ฏ ต ฐ ฑ ฒ ฌ ท ธ	T
L	ล ฬ	L
M	ม	M
N	ณ น	N
R	ร	R
Ng	ง	NG
H	ห ฮ	H
Y	ญ ย	Y
W	ว	W

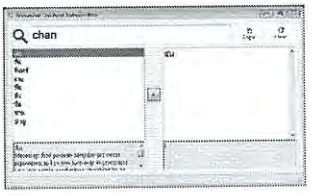
หมายเหตุ: A E I O U จะขึ้นอยู่กับอักษรตัวแรกของสระที่ได้จากการแปลงเป็นตัวอักษรภาษาอังกฤษ (แสดงในตารางที่ 4.)

ตารางที่ 4. แสดงการเข้ารหัสสำหรับสระ

ตัวอักษรไทย	ตัวอักษรโรมัน	รหัส(ASCII)
อะ, อา, อั	a	97
อัวะ, อัว	ua	11797
อำ	am	9764
อิ, อี	i	105
อุ, อู	u	117
อึ, อือ	ue	11769
เอย	oei	111101105
เอะ,เอ็, เอ	e	101
เออะ, เออ, เอิ	oe	111101
เอา	ao	97111
เอียะ,เอีย	ia	10597
เอือะ, เอือ	uea	11710197
แอะ, แอ็, แอ	ae	97101
โอะ, โอ	o	111
เอาะ, อ็, ออ	o	111
ไอ, ไอ, ไอย	ai	97105
อิว	io	105111
เอ็ว, เอว	eo	101111
เอียว	iao	10597111
เอือย	ueai	11710197105
อวย	uai	11797105

5. สรุปผล

ผลที่ได้จากการปรับปรุงแก้ไขซาวน์เด็กซ์ทำให้ได้ระบบการถอดอักษรโรมันที่มีความถูกต้องมากขึ้น สามารถแยกแยะคำได้อย่างถูกต้อง เราจำจัดทำโปรแกรมที่สามารถนำกฎเกณฑ์เหล่านี้มาใช้แสดงผลได้



รูปที่ 1. แสดงตัวอย่างโปรแกรมที่ได้

เอกสารอ้างอิง

[1] T. Karoonboonyanan, V. Sornlertlamvanich and S. Meknavin. **A Thai Soundex System for Spelling Correction. Proceedings of the Natural Language Processing Pacific Rim Symposium (1997)**: 633-636.

[2] The Royal Institute. The principle of Romanization on the basis of transcription, 2482.

[3] Suwanvisat, Prayut and Prasitjutrakul, Somchai. 1998. **Thai-English Cross-Language Transliterated Word Retrieval using Soundex Technique.** The National Computer Science and Engineering Conference, Kasetsart University, Bangkok, Thailand.

[4] A.B. Grisworld. **Thought on the Romanization of Siamese.** Printed for distribution at the obsequies of Colonel Kasem Nandhakij, February 25, 1969

[5] Wanwara Chaiyapruek, The Development of an English Thai Transliteration Program Base on Transloterated-word Corpus of the Royal Institute, Department of Linguistics, Faculty of Arts Chulalongkorn University, 2004.

Thai alphabet (อักษรไทย)

Created by Simon Ager, Omniglot.com – the guide to writing systems and languages

Consonants (พยัญชนะ)

ก	ข	ฃ	ค	ฅ	ฆ	ง	จ
ก ไก่	ข ไข่	ฃ ขวด	ค ควาย	ฅ คน	ฆ ระฆัง	ง งู	จ จาน
ko khai	kho khai	kho khuat	kho khwai	kho khon	kho rakhang	ngo ngu	cho chan
k/k	kh/k	kh/k	kh/k	kh/k	kh/k	ng/ng	ch/t
mid	high	high	low	low	low	low	mid

ฉ	ช	ซ	ฌ	ญ	ฎ	ฏ	ฐ
ฉ ฉิ่ง	ช ช้าง	ซ โซ่	ฌ เฌอ	ญ หญิง	ฎ ชฎา	ฏ ปฏัก	ฐ ฐาน
cho ching	cho chang	so so	cho choe	yo ying	do cha-da	to pa-tak	tho than
ch/-	ch/t	s/t	ch/-	y/n	d/t	t/t	th/t
high	low	low	low	low	mid	mid	high

ฑ	ฒ	ณ	ด	ต	ถ	ท	ธ
ฑ มณโฑ	ฒ ผู้เฒ่า	ณ เณร	ด เด็ก	ต เต่า	ถ ถุง	ท ทหาร	ธ ธง
tho montho	tho phuthao	no nen	do dek	to tao	tho thung	tho thahan	tho thong
th/t	th/t	n/n	d/t	t/t	th/t	th/t	th/t
low	low	low	mid	mid	high	low	low

น	บ	ป	ผ	ฝ	พ	ฟ	ภ
น หนู	บ ใบไม้	ป ปลา	ผ ผึ้ง	ฝ ฝา	พ พาน	ฟ ฟัน	ภ สำเภา
no nu	bo baimai	po pla	pho phueng	fo fa	pho phan	fo fan	pho samphao
n/n	b/p	p/p	ph/-	f/-	ph/p	f/p	ph/p
low	mid	mid	high	high	low	low	low

ม	ย	ร	ล	ว	ศ	ษ	ส
ม ม้า	ย ยักษ์	ร เรือ	ล ลิง	ว แหวน	ศ ศาลา	ษ ฤาษี	ส เสือ
mo ma	yo yak	ro ruea	lo ling	wo waen	so sala	so rue-si	so suea
m/m	y/y	r/n	l/n	w/w	s/t	s/t	s/t
low	low	low	low	low	high	high	high

Thai alphabet (อักษรไทย)

Created by Simon Ager, Omniglot.com – the guide to writing systems and languages

ห	ฟ	อ	ฮ
ห หีบ	ฟ ฉุฟา	อ อ่าง	ฮ นกฮูก
ho hip	lo chu-la	o ang	ho nok-huk
h/-	l/n	-	h/-
high	low	mid	low

Notes

· The sounds represented by some consonants change when they are used at the end of a syllable (indicated by the letters on the right of the slash). Some consonants can only be used at the beginning of a syllable.

· Duplicate consonants represent different Sanskrit and Pali consonants sounds which are pronounced identically in Thai.

Vowel diacritics (สระ)

อะ	อั	อา	อิ	อี	อึ	อื
a	a	a	i	i	ue	ue
a	a	a:	i	i:	ɯ	ɯ:
อือ	อุ	อู	เอะ	เอ็	เอ	แอะ
ue	u	u	e	e	e	ae
ɯ:	u	u:	e	e	e:	ɛ
แอ็	แอ	โอะ	โอ	เอาะ	อ็อ	ออ
ae	ae	o	o	o	o	o
ɛ	ɛ:	o	o:	ɔ	ɔ	ɔ:
อ็	เออะ	เออ	เอิ			
o	oe	oe	oe			
ɔ:	ɤʔ	ɤ:	ɤ:			

Thai alphabet (อักษรไทย)

Created by Simon Ager, Omniglot.com – the guide to writing systems and languages

Diphthongs

เอียะ	เอีย	เอือะ	เอือ	อัวะ	อัว	อว
ia	ia	uea	uea	ua	ua	ua
iaʔ	ia	ɯaʔ	ɯa	uaʔ	ua	ua

อิว	เอ็อย	เอว	แอว	เอา	อาว	เอียว
io	eo	eo	aeo	ao	ao	iao
iu/iw	eu/ew	eːu/eːw	ɛːu/ɛːw	au/aw	aːu	iau/iaw

อัย	ไอ	ไอ	ไอย	อาย	อ็อย	ออย
ai	ai	ai	ai	ai	oi	oi
ai/aj	ai/aj	ai/aj	ai/aj	aːi/aːj	ɔi/ɔj	ɔːi/ɔːj

โอย	อุย	เอย	อวย	เอือย	อำ	ฤ
oi	ui	oei	uai	ueai	am	rue
oːi/oːj	ui/uj	ɤːi/ɤːj	uai/uaj	ɯai/ɯaj	am	rɯ/ri

ฤๅ	ฦ	ฦๅ	อๆ	อ์
rue	lue	lue	ๆ	silences
rɯː	lɯ	lɯː	~	consonant

The vowel diacritics are shown with the letter o ang (อ), which acts as a silent vowel carrier at the beginning of words that start with a vowel.

Thai alphabet (อักษรไทย)

Created by Simon Ager, Omniglot.com – the guide to writing systems and languages

Punctuation

◉	ตาไก่	dtaa gai beginning of a new paragraph	๛	โคมูตร	khoh muut marks the end of a story

Tone markers

Tone markers

ก่	ไม้เอก mai ehk	low tone	ก๊	ไม้โท mai toh	falling tone
ก๋	ไม้ตรี maai dtree	high tone	ก๎	ไม้จัตวา maai jat dta waa	rising tone

Tone indication

	Open syllables			Closed syllables *	
	unmarked	อ่	อ้	short vowel	long vowel
Class 1	mid	low	falling	low	low
Class 2	rising	low	falling	low	low
Class 3	mid	falling	high	high	falling

* Closed syllables are those ending with p, t or k

Numerals (เลขไทย)

๐	๑	๒	๓	๔	๕	๖	๗	๘	๙	๑๐
ศูนย์	หนึ่ง	สอง	สาม	สี่	ห้า	หก	เจ็ด	แปด	เก้า	สิบ
sun	neung	song	sam	si	ha	hok	chet	paet	kao	sip
0	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**