การถอดคำแบบถ่ายเสียงสำหรับชื่อบุคคลภาษาไทยที่เขียนด้วยอักษรโรมัน

นางสาวชุลีกร กิตติกูล

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิทยาศาสตรมหาบัณฑิต
สาขาวิชาวิทยาศาสตร์คอมพิวเตอร์ ภาควิชาวิศวกรรมคอมพิวเตอร์
คณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย
ปีการศึกษา 2554

GENERATING TRANSCRIPTIONS FOR ROMANIZED THAI PERSON NAMES

Ms. Chuleekorn Kittikool

A Thesis Submitted in Partial Fulfillment of the Requirements

for the Degree of Master of Science Program in Computer Science

Department of Computer Engineering

Faculty of Engineering

Chulalongkorn University

Academic Year 2011

หัวข้อวิทยานิพนธ์	การถอดคำแบบถ่ายเสียงสำหรับชื่อบุคคลภาษาไทยที่เขียนด้วยอักษรโรมัน
โดย	นางสาวชุลีกร กิตติกูล
สาขาวิชา	วิทยาศาสตร์คอมพิวเตอร์
อาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก	ผู้ช่วยศาสตราจารย์ ดร. โปรดปราน บุณยพุกกณะ
อาจารย์ที่ปรึกษาวิทยานิพนธ์ร่วม	ผู้ช่วยศาสตราจารย์ ดร.อติวงศ์ สุชาโต

คณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย อนุมัติให้นับวิทยานิพนธ์ฉบับนี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญามหาบัณฑิต

…….. คณบดีคณะวิศวกรรมศาสตร์
(รองศาสตราจารย์ ดร.บุญสม เลิศหิรัญวงศ์)

คณะกรรมการสอบวิทยานิพนธ์

…… ประธานกรรมการ
(ผู้ช่วยศาสตราจารย์ ดร. เศรษฐา ปานงาม)

…….. อาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก
(ผู้ช่วยศาสตราจารย์ ดร. โปรดปราน บุณยพุกกณะ)

…… อาจารย์ที่ปรึกษาวิทยานิพนธ์ร่วม
(ผู้ช่วยศาสตราจารย์ ดร.อติวงศ์ สุชาโต)

…… กรรมการ
(รองศาสตราจารย์ ดร.วิโรจน์ อรุณมานะกุล)

…… กรรมการภายนอกมหาวิทยาลัย
(ดร.ชัย วุฒิวิวัฒน์ชัย)

ชุลีกร กิตติกูล : การถอดคำแบบถ่ายเสียงสำหรับชื่อบุคคลภาษาไทยที่เขียนด้วยอักษรโรมัน. (GENERATING TRANSCRIPTIONS FOR ROMANIZED THAI PERSON NAMES) อ.ที่ปรึกษาวิทยานิพนธ์หลัก : ผศ.ดร. โปรดปราน บุณยพุกกณะ, อ.ที่ปรึกษาวิทยานิพนธ์ร่วม : ผศ.ดร.อติวงศ์ สุชาโต, 36 หน้า.

การถอดคำแบบถ่ายเสียงสำหรับแต่ละคำสามารถสร้างได้จากกฎ หรือใช้แบบจำลองทางสถิติ หรือค้นจากพจนานุกรม อย่างไรก็ตามการขาดมาตรฐานและความหลายหลากของการแปลงชื่อบุคคลไทยให้เป็นชื่อที่เขียนด้วยอักษรโรมันเป็นงานที่ท้าทาย และแม้ว่าวิธีที่ใช้พจนานุกรมเหมือนจะให้ผลที่ค่อนข้างถูกต้องที่สุด แต่ส่วนของการแปลงตัวอักษรเป็นเสียงก็ยังมีความจำเป็นสำหรับคำที่ไม่พบในพจนานุกรม งานวิจัยนี้เสนอวิธีการถอดคำแบบถ่ายเสียงสำหรับชื่อบุคคลภาษาไทยที่เขียนด้วยอักษรโรมันให้เป็นเสียงภาษาไทย โดยคำนึงถึงความนิยมในการใช้งาน ชื่อบุคคลภาษาไทยที่เขียนด้วยอักษรโรมันจะถูกแบ่งให้เป็นสายลำดับของแกรมโดยใช้พจนานุกรมแกรมสะสมซึ่งถูกสร้างจากชื่อมากกว่า 130,000 ชื่อ ผลการศึกษาพบว่าวิธีนี้ให้ความถูกต้องของผลลัพธ์ที่ 93 % และ 95 % โดยวัดจากคะแนนความเห็นของการยอมรับได้ เมื่อคำที่ถอดคำแบบถ่ายเสียงถูกสร้างจากสายลำดับที่เป็นไปได้ทั้งหมด ด้วยการไม่ถ่วงน้ำหนักแกรมภาษาไทย และด้วยการถ่วงน้ำหนักแกรมภาษาไทยตามลำดับ และเมื่อใช้การจับคู่คำแบบยาวที่สุด จะได้ความถูกต้องที่ 73% และ 77% เมื่อใช้การไม่ถ่วงน้ำหนักแกรมภาษาไทยและการถ่วงน้ำหนักแกรมภาษาไทยตามลำดับ

ภาควิชา.....วิศวกรรมคอมพิวเตอร์.....ลายมือชื่อนิสิต.....
สาขาวิชา....วิทยาศาสตร์คอมพิวเตอร์.ลายมือชื่อ อ.ที่ปรึกษาวิทยานิพนธ์หลัก.....
ปีการศึกษา ..2554.....ลายมือชื่อ อ.ที่ปรึกษาวิทยานิพนธ์ร่วม.....

5271417321 : MAJOR COMPUTER SCIENCE

KEYWORDS: TRANSCRIPTION/ROMANIZATION/THAI NAMES

CHULEEKORN KITTIKOOL : GENERATING TRANSCRIPTIONS FOR ROMANIZED THAI PERSON NAMES ADVISOR : ASST. PROF. PROADPRAN P. PUNYABUKKANA, Ph.D., CO-ADVISOR : ASST. PROF. ATIWONG SUCHATO, Ph.D., 36 pp.

A transcription of each word can either be produced by rules, statistical models, or retrieved from dictionary. However, the lack of standards and the variation of how a Thai person romanizes his or her name pose transcription a challenging task. Although the dictionary-based approach seems to produce the most accurate result, a letter-to-sound conversion module is necessary for unknown names. We propose an approach to transcribe romanized Thai person names into Thai sounds which considers the popularity of usage. The romanized Thai names are parsed into sequences of grams, utilizing the Gram lexicon, built from a corpus of more than 130,000 names. The results show 93 and 95% mean opinion score of acceptability when the transcriptions are generated from all possible sequences with unweighted and weighted Thai grams respectively. When longest-match model is used, the acceptability levels are 73 and 77% for unweighted and weighted Thai grams.

Department .....Computer Engineering.. Student's Signature .....
Field of Study ...Computer Science..... Advisor's Signature .....
Academic Year .2011..... Co-advisor's Signature .....

กิตติกรรมประกาศ

วิทยานิพนธ์ฉบับนี้สำเร็จลุล่วงไปได้ด้วยความอนุเคราะห์อย่างยิ่งของผศ.ดร.โปรดปราน บุณยพุกกณะ และ ผศ.ดร.อติวงศ์ สุชาโต อาจารย์ที่ปรึกษาทั้งสองท่านซึ่งได้ให้ความรู้ประสิทธิ์ประสาทวิชา แนะนำแนวทางการวิจัย ให้กำลังใจ และให้การสนับสนุนเป็นอย่างดี จนทำให้การวิจัยในครั้งนี้สำเร็จออกมาด้วยดี

ขอขอบพระคุณ ผู้ช่วยศาสตราจารย์ ดร.เศรษฐา ปานงาม,ผู้ช่วยศาสตราจารย์ ดร.วิโรจน์ อรุณมานะกุล และ ดร.ชัย วุฒิวิวัฒน์ชัย กรรมการสอบวิทยานิพนธ์ ที่กรุณาเสียสละเวลา ให้คำแนะนำ ตรวจสอบ และแก้ไขวิทยานิพนธ์ฉบับนี้

ท้ายที่สุด ผู้วิจัยขอขอบคุณเพื่อนๆ ทุกๆ คน รวมทั้งครอบครัว เพื่อนร่วมงาน และผู้บังคับบัญชาในสายงาน ที่คอยติดตาม ให้กำลังใจและสนับสนุน รวมถึงท่านอื่นๆ ที่มิได้กล่าวชื่อไว้ ณ ที่นี้ที่มีส่วนช่วยให้วิทยานิพนธ์สำเร็จได้ด้วยดี

สารบัญ

หน้า

หน้า

สารบัญตาราง

หน้า

สารบัญภาพ

หน้า

บทที่ 1

บทนำ

1.1 ความเป็นมาและความสำคัญของปัญหา

การถอดอักษรไทยเป็นอักษรโรมัน ใช้เพื่อให้ชาวต่างชาติสามารถอ่าน และออกเสียงได้ใกล้เคียงกับการออกเสียงที่คนไทยใช้ เช่น ใช้ในการสอนพูดภาษาไทยให้ชาวต่างชาติ ใช้เขียนแทนชื่อสถานที่ต่างๆ หรือชื่อบุคคล เนื่องจากเป็นคำศัพท์เฉพาะ ไม่สามารถแปลให้เป็นภาษาต่างชาติได้ การถอดอักษรปัจจุบันมีการประกาศหลักเกณฑ์ในการถอดอักษรไทยเป็นอักษรโรมันตามประกาศราชบัณฑิตสถาน ฉบับวันที่ 11 มกราคม พ.ศ. 2542 แต่สำหรับการถอดอักษรไทยในชื่อของคนไทยนั้น มีความหลากหลายในการถอด ตามความชอบและความนิยมส่วนตัวของบุคคล นอกจากนี้ การแปลงตัวอักษรให้เป็นเสียง(Text-To-Speech, TTS) หากมีข้อมูลนำเข้าเป็นอักษรโรมัน จะใช้หน่วยเสียงภาษาอังกฤษในการแปลง ซึ่งจะทำให้คำอ่านที่ได้มีความไม่ถูกต้อง เนื่องจากหน่วยเสียงภาษาอังกฤษมีความแตกต่างจากหน่วยเสียงภาษาไทย ทั้งในเรื่องของเสียงวรรณยุกต์ ที่ไม่ปรากฏในหน่วยเสียงภาษาอังกฤษ และเสียงพยัญชนะที่แตกต่างกัน

การถอดคำแบบถ่ายเสียง (Transcription) คือการแปลงคำที่เขียนด้วยอักษรของระบบเขียนหนึ่ง เป็นคำที่เขียนด้วยอักษรของอีกระบบหนึ่ง เพื่อให้ได้เสียงที่ใกล้เคียงกันของคำหนึ่งๆ เช่นการถอดคำแบบถ่ายเสียงจากคำที่เขียนด้วยอักษรโรมันให้เป็นคำที่เขียนด้วยอักษรไทย โดยมากการถอดคำแบบถ่ายเสียง จะต้องใช้พจนานุกรม เพื่อให้ได้ความถูกต้องของการถอดคำ แต่เนื่องจากชื่อบุคคลมีความหลากหลายในการเขียนแม้ว่าจะเป็นชื่อเดียวกัน ตามความนิยมของแต่ละบุคคล ทำให้การใช้พจนานุกรมกับการถอดคำแบบถ่ายเสียงสำหรับชื่อบุคคล จะให้ความถูกต้องไม่สูงมากนัก

จากงานวิจัยที่ผ่านมา การถอดคำแบบถ่ายเสียงมีอยู่หลายวิธี ทั้งการใช้พจนานุกรม การใช้กฎ หรือการใช้ค่าสถิติในการสร้างกฎ แต่เนื่องจาก ชื่อบุคคลมีความหลากหลาย จึงทำให้ยากต่อการสร้างกฎที่ครอบคลุมทุกคำ ดังนั้นการใช้หลักทางสถิติเข้ามาช่วย และการใช้หน่วยเสียงภาษาไทยสำหรับคำอ่าน น่าจะเป็นทางแก้ไขที่เหมาะสมในการแก้ไขปัญหาเรื่องความแตกต่างอันเกิดจากความนิยมในการถอดชื่อบุคคลที่เขียนด้วยอักษรโรมัน และความแตกต่างระหว่างหน่วยเสียงได้

1.2 วัตถุประสงค์ของการวิจัย

เพื่อศึกษาและพัฒนาอัลกอริทึมการถอดคำแบบถ่ายเสียงสำหรับชื่อบุคคลที่เขียนด้วยอักษรโรมันโดยอาศัยความนิยมในการใช้เป็นฐาน วิธีการนี้จะเป็นแนวทางในการพัฒนาโปรแกรมคอมพิวเตอร์ในการแปลงรูปเขียนเป็นคำอ่านสำหรับชื่อคนไทยที่เขียนด้วยอักษรโรมัน เพื่อการอ่านที่ถูกต้องมากขึ้น

1.3 ขอบเขตของการวิจัย

1. ข้อมูลนำเข้าเป็นชื่อ และนามสกุลในภาษาไทยที่เขียนด้วยอักษรโรมันเท่านั้น มีวรรคตรงกลางระหว่างชื่อและนามสกุล โดยไม่รวมไปถึงชื่อของบุคคลต่างชาติ
2. ไม่พิจารณาความถูกต้องของตัวสะกดของคำที่ได้หลังจากผ่านการถอดคำแบบถ่ายเสียง
3. พิจารณาความถูกต้องเฉพาะเสียงอ่านหลังจากถอดคำแบบถ่ายเสียงแล้ว
4. ชื่อที่นำมาประเมินผล เป็นชื่อที่สามารถหาได้ในฐานข้อมูล

1.4 ขั้นตอนและวิธีดำเนินการวิจัย

1. ศึกษาทฤษฎีและงานวิจัยที่เกี่ยวข้อง
2. เตรียมฐานข้อมูลชื่อบุคคลที่สะกดด้วยอักษรไทย และอักษรโรมัน
3. สุ่มแยกชุดข้อมูลทดสอบออกจากชุดข้อมูลฝึก
4. แบ่งชื่อในชุดข้อมูลฝึกที่สะกดทั้งจากอักษรไทยและอักษรโรมันออกเป็นสายลำดับของแกรม ด้วยการสร้างโปรแกรมช่วยในการแบ่งแกรม
5. สร้างแบบจำลองภาษาจากชื่อภาษาอังกฤษในชุดข้อมูลฝึกที่แบ่งแกรมแล้ว
6. สร้างแบบจำลองการถอดคำแบบถ่ายเสียงจากชื่อทั้งไทยและอักกฤษในชุดฝึกที่แบ่งแกรมแล้ว
7. ทดสอบผลความถูกต้องในการถอดคำแบบถ่ายเสียงชื่อบุคคลที่เขียนด้วยอักษรโรมัน
8. สรุปและวิจารณ์ผลที่ได้
9. จัดทำวิทยานิพนธ์

1.5 ประโยชน์ที่คาดว่าจะได้รับ

1. วิธีการถอดคำแบบถ่ายเสียงสำหรับชื่อบุคคลที่เขียนด้วยอักษรโรมันโดยอาศัยความนิยมในการใช้เป็นฐาน
2. เป็นแนวทางในการสร้างโปรแกรมคอมพิวเตอร์ในการถอดชื่อบุคคลไทยที่สะกดด้วยอักษรโรมันเป็นเสียงอ่าน

1.6 ลำดับการจัดเรียงเนื้อหาในวิทยานิพนธ์

วิทยานิพนธ์นี้แบ่งเนื้อหาออกเป็น 5 บทดังต่อไปนี้ บทที่ 1 เป็นบทนำซึ่งกล่าวถึง ความเป็นมาและความสำคัญของปัญหา รวมถึงวัตถุประสงค์ของการวิจัย บทที่ 2 กล่าวถึงทฤษฎีพื้นฐานและงานวิจัยที่เกี่ยวข้องในงานวิจัยนี้ บทที่ 3 กล่าวถึงการดำเนินงานวิจัย บทที่ 4 เป็นการทดลองและผลที่ได้จากการทดลองตามชุดการทดลองต่างๆ และท้ายสุดคือบทที่ 5 กล่าวถึงสรุปผลการวิจัยและข้อเสนอแนะ

1.7 ผลงานที่ตีพิมพ์จากวิทยานิพนธ์

ส่วนหนึ่งของวิทยานิพนธ์นี้ได้รับการตอบรับให้ตีพิมพ์เป็นบทความทางวิชาการในหัวข้อเรื่อง “Generating Transcriptions For Romanized Thai Persons’ Names” โดย Chuleekorn Kittikool, Atiwong Suchato, Proadpran Punyabukkana นำเสนอในงานประชุมวิชาการ “The 9th Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology (ECTI-CON2012)” ณ โรงแรม โนโวเทล หัวหิน เพชรบุรี ประเทศไทย ระหว่างวันที่ 16-18 พฤษภาคม 2555

บทที่ 2

ทฤษฎีและงานวิจัยที่เกี่ยวข้อง

2.1 อักษรในภาษาไทย

ในภาษาไทยมีตัวอักษรที่ใช้แทนเสียง 3 ชนิด[1] ได้แก่

1. พยัญชนะ มีทั้งหมด 44 รูปได้แก่ “ก ข ฃ ค ฅ ฆ ง จ ฉ ช ซ ฌ ญ ฎ ฏ ฐ ฑ ฒ ณ ด ต ถ ท ธ น บ ป ผ ฝ พ ฟ ภ ม ย ร ล ว ศ ษ ส ห ฬ อ ฮ” มีลักษณะเป็นอักษรไตรยางศ์ ได้แก่ อักษรกลาง อักษรสูง อักษรต่ำคู่ อักษรต่ำเดี่ยว
2. สระ มีทั้งหมด 32 รูป แบ่งได้ 3 ประเภทคือ สระเสียงเดี่ยว, สระประสม, สระเกิน
3. วรรณยุกต์ มี 4 รูป 5 เสียง ซึ่งมีความแตกต่างจากภาษาอังกฤษที่ไม่มีเสียงวรรณยุกต์

2.2 หน่วยเสียงภาษาไทย

หน่วยเสียงพื้นฐานสำหรับภาษาไทยคือ พยางค์ ซึ่งสามารถเขียนในรูป C_i V (C_f) (T) โดย C_i แทน เสียงพยัญชนะต้น, V แทนเสียงสระ, C_f แทน เสียงพยัญชนะสะกด และ T แทนเสียงวรรณยุกต์ รายละเอียดของแต่ละหน่วยย่อย [2] มีดังนี้

1. เสียงพยัญชนะต้น

อักษรไทยทั้ง 44 ตัว สามารถใช้เป็นเสียงพยัญชนะต้นได้ แต่ตัวอักษร “ฃ” และ “ฅ” ไม่ปรากฏในคำที่ใช้ในปัจจุบัน อย่างไรก็ตาม เสียงเพียง 21 เสียงก็สามารถใช้แทนตัวอักษรทั้งหมดได้ ดังที่แสดงในตารางที่ 1

ตารางที่ 1 สัญลักษณ์แทนหน่วยเสียงสำหรับพยัญชนะไทย

หน่วยเสียง	รูปพยัญชนะ	หน่วยเสียง	รูปพยัญชนะ
/p/	ป	/m/	ม
/p^h/	พ ภ ผ	/n/	น ณ
/b/	บ	/ŋ/	ง
/t/	ต ฏ	/f/	ฟ ฝ
/t^h/	ท ธ ฒ ฑ ถ ฐ	/s/	ซ ศ ษ ส
/d/	ด ฎ ฑ	/h/	ฮ ห

/tɕ/	จ	/r/	ร ฤ
/tɕ^{h}/	ช ฌ ฉ	/l/	ล ฬ
/k/	ก	/w/	ว
/k^{h}/	ค ฆ ข	/j/	ย ญ
/ʔ/	อ		

พยัญชนะควบกล้ำในภาษาไทยมาตรฐานมีหน่วยเสียง 12 หน่วยเสียง ดังแสดงในตารางที่ 3 โดยมีอักษร “ร” “ล” “ว” เป็นอักษรควบกล้ำกับพยัญชนะ “ก” “ค” “ต” “ท” “ป” และ “พ” นับเป็นพยัญชนะควบกล้ำแท้ เช่น “เกลือ” (และแบบที่ควบกล้ำไม่แท้ เช่น “ทราบ”)

พยัญชนะต้นบางคู่ สามารถเขียนและใช้เป็นพยัญชนะต้นร่วมกัน เรียนกว่า พยัญชนะคู่ โดยมีอยู่ทั้งหมด 4 ประเภท

- ควบกล้ำแท้ : (“ปร” , /pr/), (“ตร”, /tr/), (“กร”, /kr/), (“กล”, /kr/), (“กว”, /kw/), (“พร”, /phr/), (“ทร”, /thr/), (“คร”, /khr/), (“ขร”, /khr/), (“พล”, /phl/), (“ผล”, /phl/), (“คล”, /khl/), (“ขล”, /khl/), (“คว”, /khw/), (“ขว”, /khw/)

- ควบกล้ำไม่แท้ : (“ทร”, /s/), (“จร”, /tɕ/), (“ซร”, /s/), (“สร”, /s/), (“ศร”, /s/)

- พยัญชนะขนาน (parallel consonant) : เช่น (“กล”, /k-a-l/l-/), (“ปร”, /p-a-l/l-/) เป็นต้น

- อักษรนำ : (“อย”, /j/)

2. เสียงสระ

สระในภาษาไทย มีทั้งหมด 28 เสียง ประกอบด้วยสระเสียงเดี่ยว 18 เสียง, สระประสม 6 เสียง และ สระเกิน 4 เสียง ดังแสดงในตารางที่ 2

ตารางที่ 2 สัญลักษณ์แทนหน่วยเสียงสำหรับสระ

ประเภท	เสียงสั้น		เสียงยาว	
	รูปสระ	หน่วยเสียง	รูปสระ	หน่วยเสียง
สระเสียงเดี่ยว	-ะ	/a/	-า	/a:/
	ิ	/i/	ี	/i:/
	ึ	/v/	ื	/v:/
	ุ	/u/	ู	/u:/
	เ-ะ	/e/	เ-	/e:/
	แ-ะ	/x/	แ-	/x:/
	โ-ะ	/o/	โ-	/o:/

	เ-าะ	/@/	-อ	/@:/
	เ-อะ	/#/	เ-อ	/#:/
สระประสม	เ◌ียะ	/ia/	เ◌ีย	/i:a/
	เ◌ือะ	/va/	เ◌ือ	/v:a/
	◌ัวะ	/ua/	◌ัว	/u:a/
สระเกิน	◌ำ	/am/	-	-
	ไ- , ใ-	/aj/	-	-
	เ-า	/aw/	-	-

3. เสียงพยัญชนะสะกด

มีตัวอักษรไทยบางตัวเท่านั้นที่สามารถเป็นตัวสะกดได้ เช่น “ห” และ “ฮ” ไม่สามารถใช้เป็นตัวสะกดได้ โดยเสียงพยัญชนะสะกด จะมีทั้งสิ้น 9 เสียง ดังแสดงในตารางที่ 1 และเช่นเดียวกับเสียงพยัญชนะต้นที่สามารถมีเสียงควบได้เช่น “กร” จะแทนด้วยเสียง /k/ “คร” จะแทนด้วยเสียง /k/ “ตร” จะแทนด้วยเสียง /t/

4. เสียงวรรณยุกต์

ในภาษาไทย มีทั้งหมด 5 เสียงวรรณยุกต์ เอก, โท, ตรี, จัตวา โดยเสียงวรรณยุกต์จะถูกกำหนดโดยโครงสร้างของพยางค์, พยัญชนะต้น และรูปวรรณยุกต์

2.3 การถอดอักษรเป็นโรมัน

ในพจนานุกรมได้ให้คำจำกัดความของคำว่า การถอดอักษรเป็นโรมัน (Romanization) [3] ไว้ดังนี้

การถอดอักษรไทยเป็นโรมัน (Romanization) คือ การเปลี่ยนคำในรูปเขียน หรือในรูปคำพูด ให้อยู่ในรูปตัวอักษรโรมัน โดยคำต้นฉบับใช้ระบบการเขียนที่ต่างกัน ซึ่งการใช้ตัวอักษรโรมันเพื่อการถ่ายเสียงหรือเขียนนั้น มี 2 วิธี

1. การถ่ายเสียง (Transcription) คือการแทนเสียงที่พูดให้อยู่ในรูปของสัญลักษณ์เขียน สำหรับการถอดอักษรไทยเป็นอักษณโรมันจะกล่าวถึงในหัวข้อ 2.4 ต่อไป
2. การถ่ายตัวอักษร (Transliteration) คือแทนระบบการเขียนในภาษาหนึ่งด้วยอีกระบบภาษาหนึ่ง

2.4 การถอดคำแบบถ่ายเสียง

การถอดอักษรไทยเป็นอักษรโรมันแบบถ่ายเสียง (Transcription) ปัจจุบันมีหลักเกณฑ์ที่ใช้ตามประกาศราชบัณฑิตยสถาน ฉบับวันที่ 11 มกราคม พ.ศ. 2542 [4] โดยหลักเกณฑ์การถอดอักษรไทยเป็นอักษรโรมันนี้ เป็นการถอดโดยวิธีถ่ายทอดเสียง เพื่อให้อ่านคำภาษาไทยที่เขียนด้วยอักษรโรมันให้ได้เสียงใกล้เคียง โดยไม่คำนึงถึงการสะกดการันต์ และวรรณยุกต์ เช่น จันทร์ = chan, พระ = phra, แก้ว = kaeo

ตารางที่ 3 การเทียบพยัญชนะไทยกับอักษรโรมัน

พยัญชนะไทย	อักษรโรมัน		ตัวอย่าง
	ตัวต้น	ตัวสะกด	
ก	k	k	กา = ka , นก = nok
ข ฃ ค ฅ ฆ	kh	k	ขอ = kho , สุข = suk โค = kho ยุค = yuk ฆ้อง = khong เมฆ mek
ง	ng	ng	งาม = ngam สงฆ์ song
จ ฉ ช ฌ	ch	t	จีน = chin อำนาจ amnat ฉิ่ง = ching ชิน = chin คช = khot เฌอ = choe
ซ ทร (เสียง ซ) ศ ษ ส	s	t	ซา =sa ก๊าซ kat ทราย = sai ศาล =san ทศ thot รักษา =raksa กฤษณ์ krit สี = si รส = rot
ญ	y	n	ญาติ = yat ชาญ = chan
ฎ ฑ (เสียง ด) ด	d	t	ฎีกา = deak กฎ = kot บัณฑิต = bandit ด้าย = dai เป็ด = pet
ฏ ต	t	t	ปฏิมา = patima ปรากฏ prakot ตา = ta จิต =chit

ฐ ฑ ฒ ถ ท ธ	th	t	ฐาน = than รัฐ rat มณฑล = monthon เฒ่า = thao วัฒน์ = wat ถ่าน = than นาถ= nat
ณ น	n	n	ประณีต = pranit ปราณ = pran น้อย = noi จน = chon
บ	b	p	ใบ = bai กาบ = kap
ป	p	p	ไป = pai บาป = bap
ผ พ ภ	ph	p	ผา = pha พงศ์ = phong ลัพธ์ lap สำเภา = samphao ลาภ = lap
ฝ ฟ	f	p	ฝั่ง = fang ฟ้า = fa เสิร์ฟ = soep
ม	m	m	ม้าม = mam
ย	y	-	ยาย = yai
ร	r	n	ร้อน = ron พร = phon
ล ฬ	l	n	ลาน = lan ศาล = san กีฬา = kila กาฬ = kan
ว	w	-	วาย = wai
ห ฮ	h	-	หา = ha ฮา = ha

ตารางที่ 4 การเทียบสระไทยกับอักษรโรมัน

สระไทย	อักษรโรมัน	ตัวอย่าง
อะ, ัว (อะ ลดรูป), รร (มีตัวสะกด) , อา	a	ปะ = pa, วัน = wan, สรรพ = sap, มา = ma
รร (ไม่มีตัวสะกด)	an	สรรหา = sanha, สวรรค์ = sawan
อำ	am	รำ = ram
อิ , อี	i	มิ = mi, มีด = mit

อึ, อื	ue	นึก = nuek, หรือ = rue
อุ, อู	u	ลุ = lu , หรู = ru
เอะ, เ็ (เอะ ลดรูป), เอ	e	เละ= le , เล็ง=leng, เลน=len
แอะ, แอ	ae	และ=lae, แสง=saeng
โอะ, - (โอะ ลดรูป), โอ เอาะ ,ออ	o	โละ = lo, ลม=lom, โล้=lo เลาะ=lo, ลอม=lom
เออะ, เ ิ (เออะ ลดรูป),เออ	oe	เลอะ=loe เหลิง=loeng เธอ=thoe
เอียะ,เอีย	ia	เผียะ=phia เลียน=lian
เอือะ, เอือ	uea	เลือก = lueak
อัวะ,อัว,-ว- (อัวลดรูป)	ua	ผัวะ=phua,มัว=mua, รวม=ruam
ใอ,ไอ,อัย,ไอย,อาย	ai	ใย=yai, ไล่=lai, วัย=wai, ไทย=thai, สาย=sai
เอา,อาว	ao	เมา=mao, น้าว=nao
อุย	ui	ลุย=lui
โอย, ออย	oi	โรย=roi, ลอย=loi
เอย	oei	เลย=loei
เอือย	ueai	เลื้อย=lueai
อวย	uai	มวย=muai
อิว	io	ลิ่ว=lio
เอ็ว,เอว	eo	เร็ว=reo, เลว=leo
แอ็ว,แอว	aeo	แผล็ว=phlaeo ,แมว=meao
เอียว	iao	เลี้ยว=liao
ฤ (เสียง รึ) ฤๅ	rue	ฤษี ฤๅษี =ruesi
ฤ (เสียง ริ)	ri	ฤทธิ์ = rit
ฤ (เสียง เรอ)	roe	ฤกษ์ = roek
ฦ, ฦๅ	lue	ฦๅสาย = luesai

2.5 การแปลงรูปเขียนเป็นรูปอ่านสำหรับภาษาไทย (Thai grapheme to phoneme, G2P)

Grapheme-to-Phoneme(G2P) คือการแปลงตัวอักษรภาษาไทยเป็นเสียงอ่าน ซึ่งเป็นส่วนสำคัญสำหรับการสร้างระบบการสังเคราะห์เสียงพูด (Text-to-Speech) ปัจจุบันมีหลายวิธีในการแปลง เช่น การใช้กฎ , การใช้ต้นไม้การตัดสินใจ และการใช้สถิติ [5]

2.6 แบบจำลองภาษา (Language Model)

แบบจำลองภาษา [6] คือ แบบจำลองที่จะสามารถบอกเราได้ว่าประโยคหรือสายลำดับของคำใดๆ มีความเป็นไปได้ที่จะเกิดขึ้นในภาษา หรือไม่ อาทิเช่น สายลำดับ "จะ ไป" มีโอกาสเกิดขึ้นได้แต่ สายลำดับ "ไป จะ" ไม่สามารถเกิดขึ้นได้ เป็นต้น แบบจำลองภาษาอาจจะบอกเป็นค่าความน่าจะเป็นที่จะเกิดประโยค W เรียกความน่าจะเป็นนี้ว่า $P(W)$ เช่นให้ค่า P(จะ ไป) = 0.8 และ P(ไป จะ) = 0.01 การสร้างแบบจำลองภาษาที่สามารถบอกค่าความน่าจะเป็นได้นี้สร้างมาจากแบบจำลองเอ็นแกรม (n-gram)

2.6.1 การนับคำในฐานข้อมูล

เมื่อกล่าวถึงความน่าจะเป็น เราจำเป็นที่จะต้องระบุถึงสิ่งที่เราจะนับและตำแหน่งที่เราจะพบมัน ในภาษาไทยเราไม่มีการเปลี่ยนรูปแบบคำ (Word Form) ออกมาเป็นเลมมา (Lemma) เหมือนในภาษาอังกฤษ ในงานวิจัยนี้แบ่งชื่อบุคคลออกเป็นแกรม ดังนั้นประเภทของคำ (Type) และจำนวนของคำ (Tokens) ในงานวิจัยจะมีจำนวนเท่ากัน คือ นับตามจำนวนแกรม

2.6.2 เอ็นแกรมสามัญ (Simple N-Gram)

เอ็นแกรมทำงานโดยเมื่อมีประโยค $W = w_1 ... w_n$ (ในงานวิจัยนี้คือชื่อภาษาอังกฤษที่แบ่งแล้ว) เราสามารถคำนวนความน่าจะเป็นที่จะเกิดปะโยค W นี้ได้จาก $P(w_1, w_2 ..., w_{n-1}, w_n)$ ด้วยการใช้กฎลูกโซ่ของความน่าจะเป็นทำให้แยกคำนวณค่าได้โดย

$$P(w_1^n) = P(w_1)P(w_2 \mid w_1)P(w_3 \mid w_1^2)...P(w_n \mid w_1^{n-1})$$
$$= \prod_{k=1}^{n} P(w_k \mid w_1^{k-1})$$

โดยที่ $P(w_1^n)$ คือ ความน่าจะเป็นของสายลำดับ w_1 ถึง w_n

แต่เนื่องจากความยากในการหา $P(w_k \mid w_1^{k-1})$ ทำให้เราประมาณค่าความน่าจะเป็นของประโยคไม่ได้ ดังนั้นแบบจำลองเอ็นแกรมจะคำนวณค่าความน่าจะเป็นของการเกิดคำใดๆ โดยพิจารณาจาก N-1 คำก่อนหน้า อาทิเช่น N = 2 ซึ่งเรียกว่า ไบแกรม (Bigram หรือ 2-gram) นั้น จะ

ให้ค่าความน่าจะเป็นของ คำใดๆ โดยดูจากคำก่อนหน้าเพียงคำเดียว เรามักจะเขียนค่าความน่าจะเป็นนี้ในรูปของ $P(w_2 \mid w_1)$ ซึ่งหมายถึงความน่าจะเป็นที่จะพบคำ w_2 จะตามหลังคำ w_1 เมื่อรวมทั้งประโยคเราจะสามารถคำนวณความน่าจะเป็นได้โดย

$$P(w_1^n) \approx \prod_{k=1}^{n} P(w_k \mid w_{k-1})$$

ตัวอย่างเช่น กำหนดให้

ตารางที่ 5 ตัวอย่างค่าความน่าจะเป็นของไบแกรม

P(ไป|จะ) = 0.8	P(จะ|ไป) = 0.01	P(จะ|ผม) = 0.7
P(ตลาด|ไป) = 0.5	P(โรงเรียน|ไป) = 0.6	P(ผม|ไป) = 0.02

เมื่อมีประโยคยาวๆ W = (w_1 ... w_4) ค่าความน่าจะเป็นที่จะเกิดประโยคดังกล่าวก็สามารถคำนวณได้โดยการคูณต่อๆ กันไปดังนี้

P(ผม จะ ไป โรงเรียน) = P(จะ|ผม) * P(ไป|จะ) * P(โรงเรียน|ไป)

= (0.7)(0.8)(0.6) = 0.336

P(จะ ไป ผม จะ) = P(ไป|จะ) * P(ผม|ไป) * P(จะ|ผม)

= (0.8)(0.02)(0.7) = 0.0112

จะพบว่า "จะ ไป ผม จะ" มีโอกาสเกิดต่ำมากเมื่อเทียบกับ "ผม จะ ไป โรงเรียน"

2.6.3 **การทำให้ราบเรียบ** (Smoothing)

ปัญหาสำคัญของแบบจำลองเอ็นแกรมมาตรฐานนั้น คือ ต้องได้รับการฝึกจากฐานข้อมูลซึ่งมีขนาดจำกัด บางประโยคสามารถเกิดขึ้นได้ในภาษาที่สมบูรณ์แบบแต่อาจจะไม่มีอยู่ในฐานข้อมูล ในความเป็นจริงแทบจะเป็นไปไม่ได้ที่ชุดข้อมูลฝึกจะมีคำเกิดขึ้นครบทุกคู่เพื่อใช้คำนวณค่าไบแกรมได้ และถ้าคำคู่ไหนไม่เกิดขึ้นในชุดข้อมูลฝึก เช่น C(ไป,โรงเรียน) = 0 ก็จะทำให้ P(โรงเรียน|ไป) = 0 และ P(ผม จะ ไป โรงเรียน) ก็จะเท่ากับ 0 ทำให้เกิดปัญหาขึ้นเนื่องจากไม่มีข้อมูลในชุดข้อมูลฝึกซึ่งไม่ได้หมายความว่าจะไม่มีโอกาสเกิดขึ้น ในการแก้ปัญหานี้เราเรียกว่าการปรับเรียบ (Smoothing) ซึ่งมีด้วยกันหลายวิธี

การทำให้ราบเรียบโดยการเพิ่ม 1 (Add-One Smoothing) คือ ในช่วงที่เราสร้างตารางนับไบแกรมนั้น ก่อนที่เราจะปรับค่าบรรทัดฐาน (normalization) ให้เป็นความน่าจะเป็น เราจะเพิ่มหน่วยนับขึ้นอีก 1 ให้กับทุกหน่วยซึ่งจะทำให้เราหาความน่าจะเป็นเกิดขึ้นกับทุกคู่ของไบแกรมได้

แม้ว่าวิธีการนี้จะให้ผลที่ไม่ดีนักและก็ไม่เป็นที่นิยมแต่ก็ช่วยให้พอเห็นมุมมองของการปรับเรียบ การคำนวณค่าความน่าจะเป็นของคำแต่ละคู่ในไบแกรมเมื่อปรับเรียบด้วยวิธีนี้จะกลายเป็น

$$P(w_n \mid w_{n-1}) = \frac{C(w_{n-1}, w_n) + 1}{C(w_{n-1}) + V}$$ เมื่อ V คือประเภทของคำ (type)

การทำให้ราบเรียบยังมีอีกหลายวิธีการที่นิยม เช่น การทำให้ราบเรียบโดยการลดแบบวิทเท็นเบล (Witten-Bell Discounting, Witten & Bell-1991) มีแนวคิดในการใช้จำนวนของสิ่งที่เคยพบครั้งหนึ่งในการประมาณจำนวนของสิ่งที่ยังไม่เคยพบ หรือ การทำให้ราบเรียบด้วยวิธีการลดแบบกูดทูริง (Good-Turing Discounting, Good-1953) มีแนวคิดในการปรับเรียบค่าหน่วยนับที่เป็น 0 หรือมีจำนวนน้อยๆ จากการสังเกตุจำนวนหน่วยนับที่มีค่ามากกว่า

2.6.4 **การทำให้ราบเรียบย้อน** (Back-off Smoothing)

ด้วยวิธีการปรับเรียบสามารถช่วยเราแก้ปัญหาความถี่ 0 ในเอ็นแกรมได้แต่เรายังมีอีกวิธีการที่เข้ามาช่วยได้อีก ตัวอย่างเช่นในกรณีที่เราไม่มีตัวอย่างของบางไตรแกรม $w_{n-2}w_{n-1}w_n$ ในการคำนวณค่า $p(w_n \mid w_{n-1}w_{n-2})$ ดังนั้นเราจึงพยายามประมาณด้วยความน่าจะเป็นของไบแกรม $p(w_n \mid w_{n-1})$ และเช่นกันเมื่อเรายังไม่สามารถหาได้เราก็ประมาณด้วยยูนิแกรม (unigram) $p(w_n)$ ดังนั้นในการคำนวนความน่าจะเป็นของแบบจำลองไตรแกรมจะมีลักษณะ

$$P(w_i \mid w_{i-2}w_{i-1}) = \begin{cases} P(w_i \mid w_{i-2}w_{i-1}) & ,if C(w_{i-2}w_{i-1}w_i) > 0 \\ \alpha_1 P(w_i \mid w_{i-1}) & ,if C(w_{i-2}w_{i-1}w_i) = 0 and C(w_{i-1}w_i) > 0 \\ \alpha_2 P(w) & ,other \end{cases}$$

α_1 และ α_2 คือ ค่าน้ำหนัก (Back-off weight) ซึ่งขึ้นกับอัลกอริทึมที่เลือกใช้ในการทำให้ราบเรียบย้อน (Back-off Smoothing) อาทิเช่นวิธีกูดทูริง (Good-Turing)

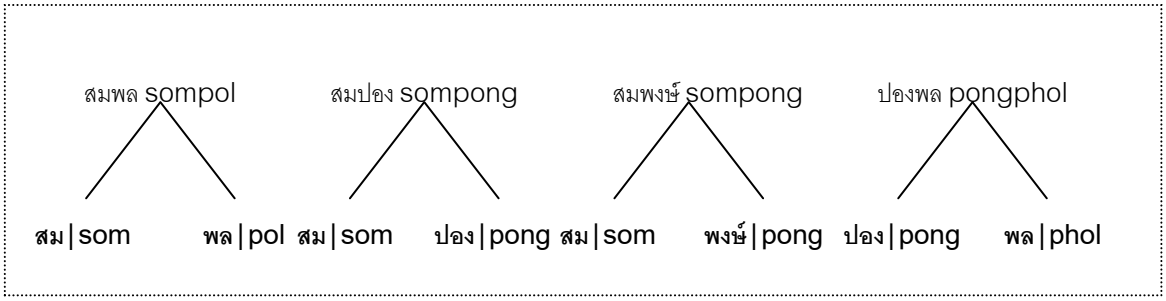
2.7 **งานวิจัยที่เกี่ยวข้อง**

งานวิจัยที่เกี่ยวข้องมี 3 ส่วน คือ การถอดชื่อบุคคลให้เป็นคำที่เขียนด้วยอักษรโรมัน, การถอดคำแบบถ่ายเสียง และการแปลงรูปเขียนเป็นคำอ่าน

2.7.1 **งานวิจัยที่เกี่ยวข้องกับการถอดชื่อบุคคลให้เป็นคำที่เขียนด้วยอักษรโรมัน**

ในการถอดชื่อบุคคลจากอักษรไทยเป็นอักษรโรมัน A. Tangverapong [7] ใช้วิธี แบ่งชื่อออกเป็นแกรม แล้วนำแต่ละแกรมมาหาสายลำดับที่ออกเสียงคล้ายกัน ที่เป็นไปได้มากที่สุดโดยการจัดกลุ่มอักษร ก:ข เป็นแกรมที่สะกดในภาษาไทยเป็นกลุ่ม ก และสะกดในภาษาอังกฤษเป็น

กลุ่ม ข ดังตัวอย่างในรูปที่ 1 แสดงการแตกองค์ประกอบของชื่อออกเป็นแกรม โดยชื่อทางด้านซ้ายสุดที่สะกดด้วยตัวอักษรภาษาไทย “สมพล” และสะกดด้วยตัวอักษรภาษาอังกฤษ “sompol” สามารถแบ่งเป็นสายลำดับของ 2 แกรม “สม|som” และ “พล|pol” เรียงต่อกัน เนื่องจากคุณสมบัติของแกรมถูกบังคับจากการสะกดของทั้ง 2 ภาษาดังนั้น แกรม 2 แกรมจะแตกต่างกันเมื่อการสะกดในภาษาไทยหรือการสะกดในภาษาอังกฤษอย่างใดอย่างหนึ่งหรือทั้งสองอย่างแตกต่างกัน จากการแตกองค์ประกอบของตัวอย่างชื่อทั้ง 4 ชื่อในภาพจะทำให้ได้แกรมที่แตกต่างกันทั้งสิ้น 5 แกรม ได้แก่ “สม|som” “พล|pol” “พล|phol” “ปอง|pong” และ “พงษ์|pong”



ภาพที่ 1 ตัวอย่างการแบ่งชื่อออกเป็นสายลำดับของแกรม

ผลจากการวิจัยนี้ให้ประสิทธิภาพสูงในการถอดจากอักษรไทยเป็นอักษรโรมัน โดยมีความถูกต้อง 75 %

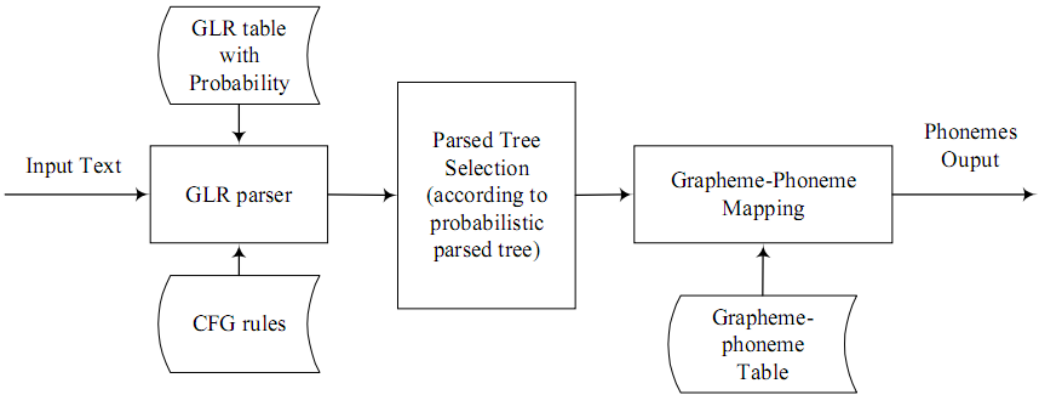
2.7.2 งานวิจัยที่เกี่ยวข้องกับการถอดคำแบบถ่ายเสียง

การถอดคำแบบถ่ายเสียงสำหรับคำที่เขียนด้วยอักษรโรมัน W. Aroonmanakun [8] ได้เสนอการใช้กฎการแปลงโดยการปรับกฎของ Bosch and Daelemans’s-1993 ให้เหมาะสมกับเสียงในภาษาไทย จาก 35,000 กฎเหลือ 440 กฎ และมีการกำหนดเสียงวรรณยุกต์ ให้กับคำอ่านที่ถอดคำมาแล้ว โดยใช้สถิติเพื่อสร้างกฎการใส่วรรณยุกต์ให้คำ จากผลการทดลอง ยังพบคำอ่านที่ไม่มีการใช้จริงถึง 56% ซึ่งอาจเป็นผลมาจากการปรับกฎยังไม่เหมาะสม

2.7.3 งานวิจัยที่เกี่ยวข้องกับการแปลงรูปเขียนเป็นคำอ่าน

P. Tarsaku [5] ได้เสนอการใช้สถิติเพื่อช่วยในการแปลงรูปเขียนเป็นคำอ่านสำหรับภาษาไทย โดยการนำข้อมูลเข้าผ่านตัวกรองข้อมูล โดยใช้ตาราง GLR(Generalized LR Parsing) [10] ซึ่งมีการใช้สถิติร่วมด้วย และใช้กฎเพื่อตรวจสอบโครงสร้างพยางค์ ข้อมูลนำออกที่ได้จะเป็นต้นไม้ของพยางค์ แล้วนำแต่ละพยางค์ไปจับคู่กับเสียงอ่าน โดยใช้ตารางการจับคู่ กระบวนการทั้งหมดที่ใช้แสดงดังภาพที่ 2 ซึ่งผลที่ได้จากการวิจัย ผลว่ามีความถูกต้องมากกว่าการใช้

พจนานุกรม หรือการใช้กฎ เนื่องจาก สามารถแปลงคำที่เกิดขึ้นใหม่หรือคำที่ไม่อยู่ในพจนานุกรมได้ โดยให้ความถูกต้องของการแปลง 90.44% เมื่อไม่คำนึงถึงเสียงสระ และให้ความถูกต้อง 72.87% เมื่อจับคู่ได้ตรง

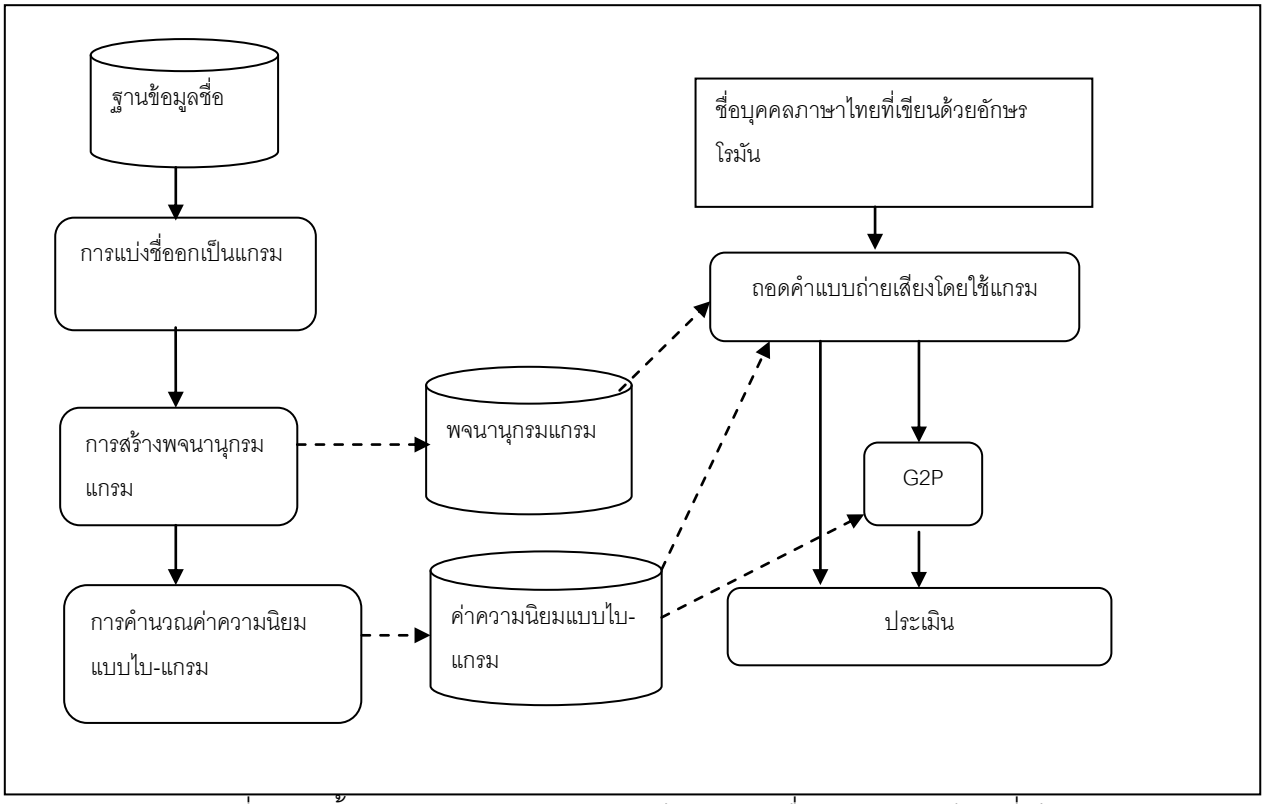


ภาพที่ 2 การแปลงรูปเขียนเป็นคำอ่านภาษาไทย โดยใช้สถิติ (PGLR)

บทที่ 3

ขั้นตอนการดำเนินงานวิจัย

วิธีการถอดคำแบบถ่ายเสียงสำหรับชื่อบุคคลภาษาไทยที่เขียนด้วยอักษรโรมันที่เสนอในงานวิจัยนี้ ประกอบด้วยขั้นตอนทั้งหมด 5 ขั้นตอน ดังแสดงในภาพที่ 3



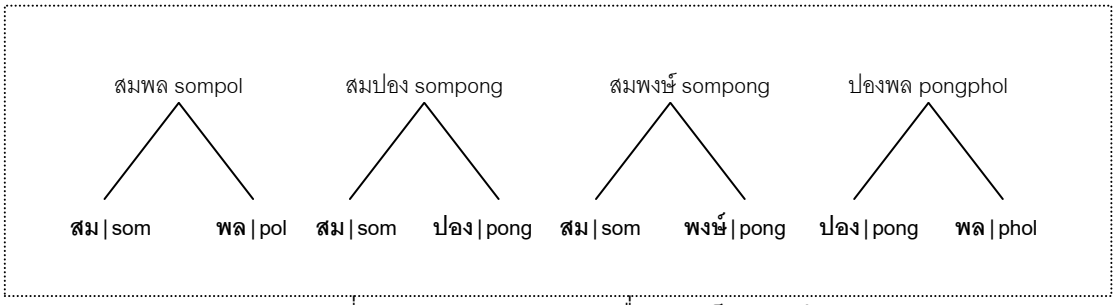
ภาพที่ 3 ขั้นตอนการถอดคำแบบถ่ายเสียงสำหรับชื่อบุคคลภาษาไทยที่เขียนด้วยอักษรโรมัน

โดยแต่ละขั้นตอนมีรายละเอียด ดังนี้

3.1 การแบ่งชื่อออกเป็นแกรม

สำหรับงานวิจัยนี้ชื่อแต่ละชื่อจะมองเป็นสายลำดับของแกรม ตามมุมมองของวิธีที่เสนอโดย A. Tangverapong [7] โดยให้กลุ่มอักษร $G^T|G^E$ เป็นแกรมที่สะกดในภาษาไทยเป็นกลุ่ม G^T และสะกดในภาษาอังกฤษเป็นกลุ่ม G^E ดังตัวอย่างในรูปที่ 10 แสดงการแตกองค์ประกอบของชื่อออกเป็นแกรม โดยชื่อทางที่สะกดด้วยตัวอักษรไทย “สมปอง” และสะกดด้วยตัวอักษรอังกฤษ “sompong” สามารถแบ่งเป็นสายลำดับของ 2 แกรม “สม|som” และ “ปอง|pong” เรียงต่อกัน

เนื่องจากคุณสมบัติของแกรมถูกบังคบจากการสะกดของทั้ง 2 ภาษาดังนั้น แกรม 2 แกรมจะแตกต่างกันเมื่อการะสะกดในภาษาไทยหรือการสะกดในภาษาอังกฤษอย่างใดอย่างหนึ่งหรือทั้งสองอย่างแตกต่างกัน จากการแตกองค์ประกอบของตัวอย่างชื่อทั้ง 4 ชื่อในภาพจะทำให้ได้ แกรมที่แตกต่างกันทั้งสิ้น 5 แกรม ได้แก่ “สม|som” “พล|pol” “พล|phol” “ปอง|pong” และ “พงษ์|pong”



ภาพที่ 4 ตัวอย่างการแบ่งชื่อออกเป็นสายลำดับของแกรม

จะสังเกตได้ว่าความสัมพันธ์ของสายลำดับอักษรไทยกับกับสายลำดับอักษรอังกฤษมีลักษณะเป็นแบบ n ต่อ n เช่น สายลำดับอักษรไทย “พล” อาจถอดได้เป็น pon หรือ phol และในทางกลับกัน สายลำดับอักษรอังกฤษ “pong” อาจถอดมาจากสายลำดับอักษร ปอง หรือ พงษ์ ดังนั้นการนิยามแกรมที่ต้องถูกบังคับจากการสะกดทั้ง 2 ภาษาจะช่วยรองรับความหลากหลายในการถอดอักษรของชื่อบุคคล

ในการแบ่งชื่อออกเป็นสายลำดับของแกรม ขั้นแรกชื่อที่เขียนด้วยอักษรไทยจะถูกแบ่งออกให้อยู่ในรูปของสายลำดับของพยางค์โดยใช้วิธีที่เสนอโดย W. Aroonmanakun [9] จากนั้นพจนานุกรมแกรมสะสมจะถูกฝึกโดยใช้วิธีที่เสนอโดย A. Tangerapong [7]

3.2 การสร้างพจนานุกรมแกรมสะสม

เมื่อแบ่งชื่อออกเป็นแกรมได้แล้ว แกรม G^T จากกลุ่มอักษร $G^T|G^E$ จะถูกค้นในพจนานุกรมแกรมสะสม หากพบในพจนานุกรมแกรม ก็จะคำนวณค่าความนิยมของแกรมนั้นใหม่ แต่หากแกรม ก ไม่พบในพจนานุกรมแกรม แกรม ก ก็จะถูกเพิ่มเข้าไปในพจนานุกรมแกรมสะสม

3.3 การคำนวณค่าความนิยมแบบไบแกรม

ในระหว่างการสร้างพจนานุกรมแกรมสะสม จะคำนวณค่าความนิยมของแต่ละคู่แกรมไปพร้อมกัน โดยใช้แบบจำลองภาษาแบบไบแกรม ในงานวิจัยนี้มีการใช้ค่าความนิยมจาก

1. ค่าความนิยมของคู่แกรมที่เขียนด้วยอักษรโรมัน เพื่อใช้ระหว่างการแบ่งสายลำดับแกรม และคำนวณคะแนนของสายลำดับแกรมที่เขียนด้วยอักษรโรมัน

2. ค่าความนิยมของคู่แกรมที่เขียนด้วยอักษรไทย เพื่อใช้ระหว่างการสร้างคำถ่ายเสียงของสายลำดับแกรมแต่ละสาย และคำนวณคะแนนของคำถ่ายเสียง
3. ค่าความนิยมของคู่เสียงที่ผ่านการแปลงจากรูปเขียนเป็นรูปอ่าน เพื่อใช้ระหว่างการสร้างรูปอ่านของคำถ่ายเสียง และคำนวณคะแนนของรูปอ่าน

3.4 การถอดคำแบบถ่ายเสียงโดยใช้แกรม

ในการถอดคำแบบถ่ายเสียงมีขั้นตอน 2 ขั้นตอนคือ

3.4.1 การแบ่งชื่อที่เขียนด้วยอักษรโรมันเป็นสายลำดับแกรม

ข้อมูลนำเข้าสำหรับการถอดคำแบบถ่ายเสียงคือ ชื่อที่เขียนด้วยอักษรโรมัน ซึ่งในงานวิจัยนี้มีวิธีการแบ่งชื่อออกเป็นสายลำดับแกรม 2 วิธี คือ

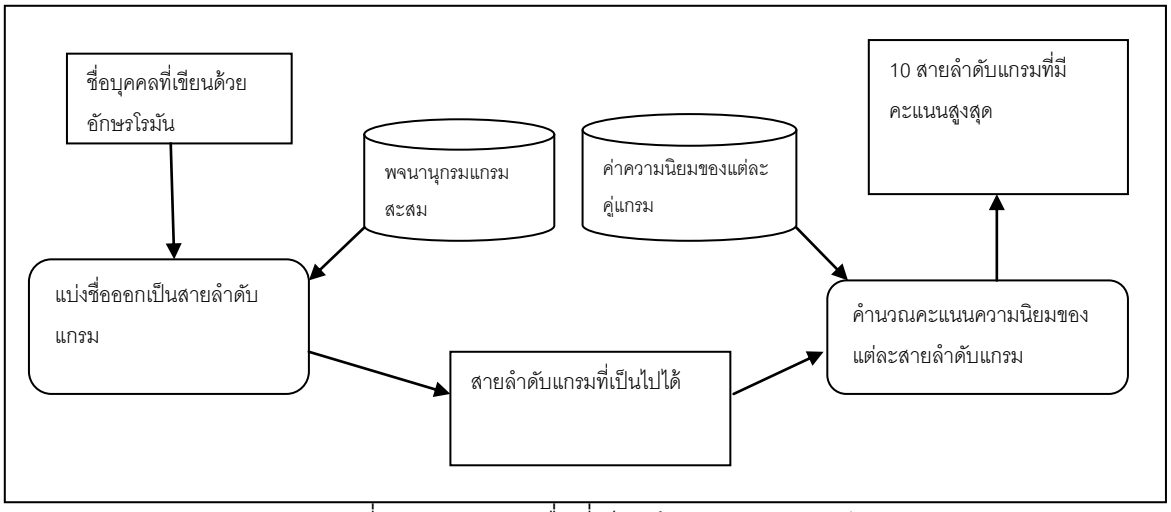
1) การค้นหาแกรมในพจนานุกรมแกรมแบบยาวที่สุด (Longest–matching) วิธีนี้จะใช้การค้นหาแกรมในพจนานุกรม จากแกรมที่ยาวที่สุดที่เป็นไปได้ในชื่อนั้น และสร้างเป็นสายลำดับแกรม โดยจะได้สายลำดับแกรมเพียง 1 สายลำดับต่อ 1 ชื่อเท่านั้น

2) การค้นหาแกรมในพจนานุกรมแกรมที่เป็นไปได้ทั้งหมด วิธีนี้จะใช้การค้นหาแกรมในพจนานุกรมแกรมทุกแกรมที่เป็นไปได้ มาจัดเรียงเป็นสายลำดับแกรมสำหรับชื่อนั้น โดยวิธีนี้จะมีสายลำดับแกรมมากกว่า 1 สายลำดับต่อชื่อ 1 ชื่อ สายลำดับแกรมที่ได้อาจมีมากถึง 100 สายลำดับ ในงานวิจัยนี้จะเลือกมาเพียง 10 สายลำดับแกรมแรกที่มีคะแนนความนิยมมากที่สุด

ในการคำนวณคะแนนความนิยมของแต่ละสายลำดับแกรม จะคำนวณโดยใช้แบบจำลองภาษา ตามสมการที่ (1) โดยให้ ชื่อที่เขียนด้วยอักษรโรมันแทนด้วย T= $w_1w_2w_3...w_n$ โดยที่ w_i แทนแกรมที่เขียนด้วยอักษรโรมัน, $P(w_i)$ คือ ค่าความเป็นไปได้ของแกรม w_i ในตำแหน่งที่ i ในสายลำดับแกรม และ $P(w_i|w_{i-1})$ คือค่าความเป็นไปได้ที่จะเกิด w_i เมื่อเกิด w_{i-1}

$$\begin{aligned} Score &= \log P(w_1) + \log P(w_2 \mid w_1) + ... + \log P(w_n \mid w_{n-1}) \\ &= \log P(w_1) + \sum_{k=2}^{n} \log P(w_k \mid w_{k-1}) \end{aligned} \quad (1)$$

การแบ่งชื่อออกเป็นสายลำดับแกรมสามารถเขียนเป็นแผนภาพได้ดังภาพที่ 5



ภาพที่ 5 การแบ่งชื่อที่เขียนด้วยอักษรโรมันเป็นสายลำดับแกรม

3.4.2 การสร้างคำถ่ายเสียงสำหรับแต่ละสายลำดับแกรม

คำถ่ายเสียงสำหรับแต่ละสายลำดับแกรม สร้างจากการนำแกรมที่เขียนด้วยภาษาไทยทั้งหมดที่สัมพันธ์กับแกรมที่เขียนด้วยอักษรโรมันมาสร้างเป็นคำ โดยจะมีการคำนวณคะแนนความนิยมของแต่ละคำ ด้วยแบบจำลองภาษาแบบไบแกรม ตามสมการที่ (1) เช่นเดียวกับการคำนวณคะแนนของสายลำดับแกรมที่เขียนด้วยอักษรโรมัน

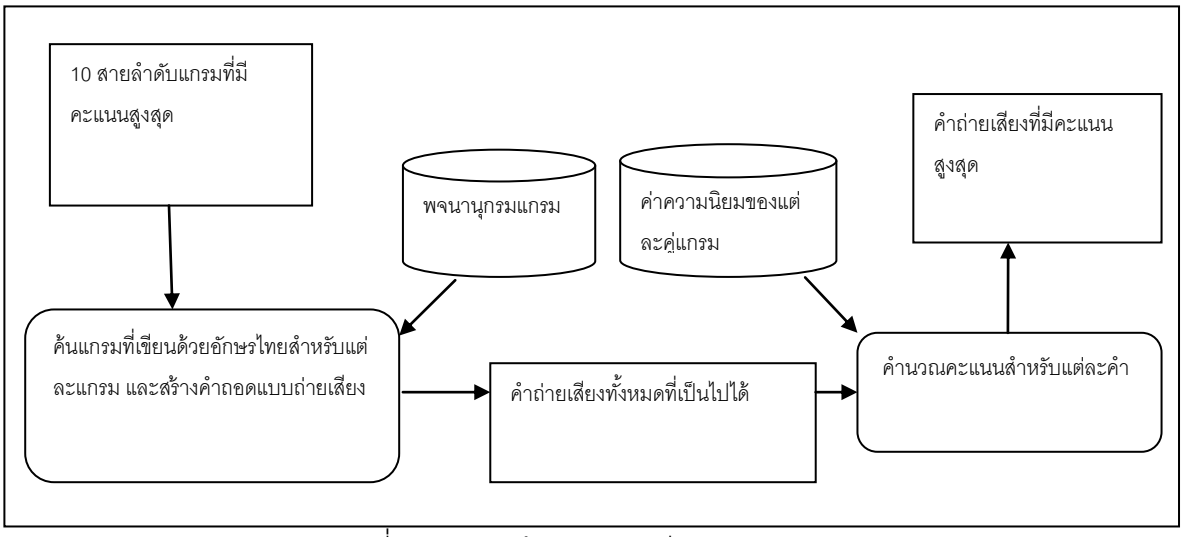
เนื่องจากแกรมที่เขียนด้วยอักษรโรมัน สามารถแปลงเป็นแกรมที่เขียนด้วยอักษรไทยได้มากกว่า 1 แกรม ทำให้ต้องมีการถ่วงน้ำหนักของแกรมแต่ละแกรม โดยยึดแกรมที่เขียนด้วยอักษรโรมันเดียวกัน เช่น แกรม lee สามารถแปลงเป็นคำว่า รี ลี ลีย์ รีย์ เป็นต้น แต่ความนิยมของแต่ละแกรม ก็แตกต่างกันไป เช่น รี อาจมีความนิยมมากกว่า ลีย์ ทำให้การคำนวณคะแนนจึงต้องมีการถ่วงน้ำหนักตามความนิยมของแต่ละแกรมด้วย ตามสมการที่ (2) โดยที่ W_n คือ ค่าถ่วงน้ำหนักของแต่ละแกรมภาษาไทย โดยคำนวณจากค่าความเป็นไปได้ของแต่ละแกรม

$$Score = W_1 W_2 ... W_n (\log P(w_1) + \log P(w_2 \mid w_1)) ... + \log P(w_n \mid w_{n-1}))$$

$$= \prod_{m=1}^{n} W_m (\log P(w_1) + \sum_{k=2}^{n} \log P(w_k \mid w_{k-1})) \quad (2)$$

เมื่อคำนวณได้คะแนนความนิยมของทุกคำแล้ว จะเลือกคำที่มีค่าคะแนนสูงสุดเป็นคำถ่ายเสียงที่สัมพันธ์กับชื่อบุคคลที่เขียนด้วยอักษรโรมันนั้นๆ

การสร้างคำถ่ายเสียงสำหรับสายลำดับแกรม สามารถเขียนเป็นแผนภาพได้ตามภาพที่ 6



ภาพที่ 6 การสร้างคำถ่ายเสียงสำหรับแต่ละสายลำดับแกรม

3.5 การแปลงจากรูปเขียนเป็นรูปอ่าน

เมื่อได้คำที่ได้จากการถอดคำแบบถ่ายเสียงโดยใช้แกรมแล้ว นำไปแปลงจากรูปเขียนเป็นรูปอ่านโดยการใช้สถิติในการแปลง ซึ่งเสนอโดย P.Tarasaku [5] อันมีวิธีการทำดังได้กล่าวมาแล้วในหัวข้อ 2.7.3

บทที่ 4

การทดลองและผลการทดลอง

4.1 การทดลอง

4.1.1 ฐานข้อมูลชื่อ

ในงานวิจัยนี้ได้นำฐานข้อมูลชื่อและนามสกุลของคนไทยมาใช้ในการทดลอง เพื่อเป็นทั้งชุดข้อมูลฝึกให้กับแบบจำลองทางสถิติและชุดข้อมูลสำหรับทดสอบประเมินผล ฐานข้อมูลชื่อบุคคลนี้ได้มาจากฐานข้อมูลของนักศึกษาที่ลงทะเบียนในมหาวิทยาลัยแห่งหนึ่งในช่วงพ.ศ. 2541-2551 โดยตัดชื่อของนักศึกษาต่างชาติออกในระหว่างการประมวลผล ฐานข้อมูลนี้มีชื่อทั้งสิ้น 178,612 ชื่อซึ่งแต่ละชื่อมีชื่อที่สะกดด้วยตัวอักษรอังกฤษกำกับด้วย

ในการทดลองจะแบ่งชุดข้อมูลโดยการสุ่ม ออกเป็นสามส่วน โดยแต่ละส่วนไม่ซ้ำกัน ดังนี้

1. ชุดข้อมูลฝึกสำหรับการสร้างพจนานุกรมแกรม 80% เพื่อใช้ในการสร้างพจนานุกรมแกรม
2. ชุดข้อมูลฝึกระหว่างการพัฒนา 10% เพื่อใช้ระหว่างการพัฒนาแบบจำลอง
3. ชุดข้อมูลสำหรับทดสอบ 10% เพื่อใช้ในการทดสอบ

4.1.2 การประเมินผล

ในการประเมินผล จะใช้กลุ่มตัวอย่าง 50 คนเพื่อประเมิน การรับได้ของเสียงที่ได้จากการถอดคำแบบถ่ายเสียง โดยผลการประเมินจะแบ่งเป็น ยอมรับ หากกลุ่มตัวอย่างเห็นว่าเสียงที่ได้ตรงกับชื่อที่เขียนด้วยอักษรโรมัน และ ไม่ยอมรับ หากกลุ่มตัวอย่างเห็นว่าเสียงที่ได้ไม่ตรงกับชื่อที่เขียนด้วยอักษรโรมัน โดยใช้ชื่อทั้งหมด 690 ชื่อในการประเมินผล ซึ่งมีการสุ่มออกมาจากฐานข้อมูล และการประเมินผลของแต่ละชื่อ จะต้องมีการยอมรับอย่างน้อย 60% ของกลุ่มประเมินจึงจะถือว่าชื่อนั้นยอมรับ

4.1.3 ชุดการทดทดลอง

ในงานวิจัยนี้ได้แบ่งการทดลองออกเป็นทั้งหมด 8 การทดลองเพื่อหาวิธีที่จะให้ผลลัพธ์ที่ถูกต้องมากที่สุด ดังนี้

1. สร้างคำถ่ายเสียงจากสายลำดับแกรมที่ค้นจากพจนานุกรมแบบยาวที่สุด และไม่มีการถ่วงน้ำหนักแกรมภาษาไทย
2. สร้างคำถ่ายเสียงจากสายลำดับแกรมที่ค้นจากพจนานุกรมแบบยาวที่สุด และถ่วงน้ำหนักแกรมภาษาไทย
3. สร้างคำถ่ายเสียงจากสายลำดับแกรมที่ค้นจากพจนานุกรมที่เป็นไปได้ทั้งหมด และไม่มีการถ่วงน้ำหนักแกรมภาษาไทย
4. สร้างคำถ่ายเสียงจากสายลำดับแกรมที่ค้นจากพจนานุกรมที่เป็นไปได้ทั้งหมด และถ่วงน้ำหนักแกรมภาษาไทย
5. สร้างรูปอ่านจากสายลำดับแกรมที่ค้นจากพจนานุกรมแบบยาวที่สุด และไม่มีการถ่วงน้ำหนักแกรมภาษาไทย
6. สร้างรูปอ่านจากสายลำดับแกรมที่ค้นจากพจนานุกรมแบบยาวที่สุด และถ่วงน้ำหนักแกรมภาษาไทย
7. สร้างรูปอ่านจากสายลำดับแกรมที่ค้นจากพจนานุกรมที่เป็นไปได้ทั้งหมด และไม่มีการถ่วงน้ำหนักแกรมภาษาไทย
8. สร้างรูปอ่านจากสายลำดับแกรมที่ค้นจากพจนานุกรมที่เป็นไปได้ทั้งหมด และถ่วงน้ำหนักแกรมภาษาไทย

นอกจากนี้ มีการทดสอบเพื่อประเมินความครอบคลุมของพจนานุกรมแกรม โดยการนับจำนวนชื่อในชุดข้อมูลทดสอบ ที่ไม่พบในพจนานุกรมแกรม และไม่สามารถถอดคำแบบถ่ายเสียงได้

4.2 ผลการทดลอง

4.2.1 ผลการถอดอักษร

ผลการประเมินการยอมรับของผลลัพธ์ที่ได้จากวิธีที่นำเสนอ แสดงดังตารางที่ 6 และตารางที่ 7 จะพบว่า การสร้างคำถ่ายเสียง และรูปอ่านที่ได้จากการค้นพจนานุกรมที่เป็นไปได้ทั้งหมดโดยมีการถ่วงน้ำหนักแกรมภาษาไทย ให้ผลการยอมรับ 95% ซึ่งเป็นผลการยอมรับสูงสุดจากชุดการทดลองทั้ง 8 ชุดการทดลอง และ การสร้างคำถ่ายเสียงและรูปอ่านจากการค้นพจนานุกรมแบบยาวที่สุด โดยไม่มีการถ่วงน้ำหนักแกรมภาษาไทยจะให้ผลการยอมรับน้อยที่สุดที่ 73% ทั้งนี้ จะเห็นว่า ผลการประเมินสำหรับรูปอ่านได้เท่ากับการประเมินสำหรับคำถ่ายเสียง เนื่องจากค่าความนิยมแบบไบแกรม ของรูปอ่านและแกรมมีค่าเกือบเท่ากัน

ตารางที่ 6 ผลการประเมินสำหรับคำถ่ายเสียง

Result	Proposed method	
	Unweighted Thai Grams	Weighted Thai Grams
Longest match	73%	77%
All possible sequences of grams	92%	95%

ตารางที่ 7 ผลการประเมินสำหรับรูปอ่าน

Result	Proposed method	
	Unweighted Thai Grams	Weighted Thai Grams
Longest match	73%	77%
All possible sequences of grams	92%	95%

ในงานวิจัยนี้เสนอวิธีการถอดชื่อบุคคลแบบถ่ายเสียงโดยอาศัยความนิยมในการใช้เป็นฐาน ให้ผลการยอมรับของการถอดชื่อบุคคล 92 - 95% โดยข้อผิดพลาดของผลลัพธ์ที่ได้ส่วนใหญ่จะเกิดจากสาเหตุต่อไปนี้

1) ความยาวของเสียงสระ เช่น “Anon” คำถ่ายเสียงที่ได้คือ อานนท์ กลุ่มตัวอย่างส่วนใหญ่จะเห็นว่าชื่อนี้ควรจะเป็น อนนท์ ซึ่งเป็นสระเสียงสั้น

2) เสียงสระ เช่น “Mati” คำถ่ายเสียงที่ได้คือ “เมธิ” กลุ่มตัวอย่างส่วนใหญ่เห็นว่าชื่อนี้ควรจะเป็น มติ ซึ่งความผิดพลาดอาจจะเกิดได้จากค่าความนิยมสำหรับฐานข้อมูลนี้อาจจะมีชื่อ เมธิ มากกว่าชื่อ มติ

3) เสียงวรรณยุกต์ เช่น “Duangdao” คำถ่ายเสียงที่ได้คือ ด้วงดาว ซึ่งจะเห็นว่าเสียงวรรณยุกต์ผิดไป จากเสียงสามัญเป็นเสียงโท

4) เสียงพยัญชนะต้น เช่น “Kachaporn” ” คำถ่ายเสียงที่ได้คือ คชาพร กลุ่มตัวอย่างส่วนใหญ่เห็นว่าชื่อนี้ควรเป็น เกชาภรณ์ ซึ่งเสียงของพยัญชนะต้นมีความผิดเพี้ยนไปเนื่องจาก K สามารถอ่านออกเสียงได้ทั้ง ค หรือ ก

ทั้งนี้ ความผิดพลาดจาก ความยาวของเสียงสระ และความผิดพลาดของพยัญชนะต้นเกิดมากที่สุด ประมาณ 80% ของความผิดพลาดที่เกิดขึ้น ส่วนความผิดพลาดของเสียงสระ และเสียงวรรณยุกต์เกิดเป็นส่วนน้อย

การแบ่งสายลำดับแกรมแบบยาวที่สุดให้ผลการประเมินเพียง 73 – 77 % เนื่องจาก การค้นแกรมในพจนานุกรมแกรมแบบยาวที่สุดอาจจะให้ผลไม่ครอบคลุมทุกชื่อที่มี และอาจจะทำให้เกิดความผิดพลาดในการแบ่งแกรม ทำให้ไม่พบแกรมในพจนานุกรม

สำหรับการทดสอบเพื่อประเมินความครบคลุมของพจนานุกรมแกรม พบว่าพจนานุกรมแกรม ให้ความครบคลุมของชื่อประมาณ 65%

บทที่ 5

สรุปผลการวิจัยและข้อเสนอแนะ

5.1 สรุปผลการวิจัย

งานวิจัยนี้เสนอวิธีการถอดคำแบบถ่ายเสียง (Transcription) ที่มีความยืดหยุ่นและเหมาะสมกับการถอดอักษรกับชื่อคนไทยด้วยวิธีการใช้ความนิยมเป็นฐานในการถอด ในงานวิจัยนี้ชื่อบุคคลจะถูกมองเป็นสายลำดับของแกรมซึ่งเป็นหน่วยย่อยที่ประกอบด้วยการสะกดด้วยตัวอักษรไทยและสะกดด้วยอักษรอังกฤษที่ออกเสียงคล้ายกัน แบบจำลองความน่าจะเป็นถูกสร้างขึ้นบนพื้นฐานของชุดของแกรม วิธีการนี้ให้ผลที่น่าพอใจในการถอดคำแบบถ่ายเสียง โดยให้ผลการประเมินการยอมรับ 95% เมื่อใช้การแบ่งสายลำดับแกรมแบบเป็นไปได้ทั้งหมดและถ่วงน้ำหนักแกรมภาษาไทย

5.2 อภิปรายผลการวิจัย

ผลจากงานวิจัยนี้สามารถนำไปเป็นแนวทางในการพัฒนาโปรแกรมคอมพิวเตอร์ สำหรับการแปลงชื่อคนไทยที่เขียนด้วยอักษรโรมันเป็นคำอ่านโดยใช้หน่วยเสียงภาษาไทยได้ โดยใช้เป็นส่วนหนึ่งของระบบการแปลงรูปเขียนเป็นรูปอ่าน (Text-To-Speech,TTS) เพื่อให้ได้รูปอ่านที่ถูกต้องเมื่อพบชื่อบุคคลที่เขียนด้วยอักษรโรมัน

5.3 ข้อเสนอแนะ

ในการวิจัยนี้ยังมีความผิดพลาดส่วนหนึ่งที่เกิดขึ้นจากแกรมที่ไม่มีอยู่ในพจนานุกรม การเพิ่มฐานข้อมูลชื่ออาจเป็นทางแก้หนึ่งที่ช่วยลดความผิดพลาดได้เมื่อพจนานุกรมแกรมสะสมมีขนาดใหญ่ระดับหนึ่ง แต่ไม่สามารถแก้ปัญหาชื่อใหม่ๆ ที่เกิดขึ้นได้โดยเฉพาะชื่อที่มีรากศัพท์มาจากภาษาบาลีและสันสกฤตซึ่งเป็นที่นิยมของคนไทย การสร้างแกรมขึ้นเลียนแบบแกรมในพจนานุกรมที่มีความใกล้เคียงกันน่าจะเป็นแนวทางหนึ่งที่ช่วยแก้ไขปัญหาได้

รายการอ้างอิง

[1] ภาทิพ ศรีสุทธิ์. อักษรไทยและการผันวรรณยุกต์. [ออนไลน์]. 2006. แหล่งที่มา: http://www.st.ac.th/bhatips/gramma3.htm [2006,April]

[2] Luksaneeyananwin, S. Speech computing and Technology in Thailand, NLP in Thailand, pp.276-321, 1993.

[3] Pie, A.M., and Gaynor, F. Dictionary of Linguishes. London : Peter Owen, 1958.

[4] ราชบัณฑิตยสถาน. การถอดอักษรไทยเป็นอักษรโรมัน. [online] 1999 แหล่งที่มา: http://www.royin.go.th/th/profile/index.php?SystemModuleKey=131&SystemMenuID=1&SystemMenuIDS= [1999,January]

[5] P. Tarasaku, V. Sornlertlamvanich, R. Thongprasirt. Thai Grapheme-to-Phoneme using Probabilistic GLR Parser, Proceedings of Eurospeech 2(2001): 1057-1060.

[6] Junrafsky, D., and Marin, J.H. An introduction to Natural Language Processing, Computational Lingu. Upper Saddle River. Speech and language processing :, N.J. : Prentice Hall, 2002.

[7] A. Tangverapong, A. Suchato, and P. Punyabukkana. Romanization of Thai Proper Names Based on Popularity of Usages. Proceedings of 13th Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD2009), Bangkok, Thailand, 2009.

[8] W. Aroonmanakun, N. Thapthong, P. Wattuya, B. Kasisopa, and S. Luksaneeyanawin. Generating Thai Transcriptions for English Words. In Wilaiwan Khanittanan and Paul Sidwell (eds.), The 14th annual meeting of the Southeast Asian Linguistics Society 2004, pp.13-22, Bangkok, 2004.

[9] W. Aroonmanakun. 2006. A Chunk-based n-gram English to Thai Transliteration. Transactions on Computer and Information Technology Vol.2 No.2, pp 121-125. 2006,

[10] Tomita, M. Generalized LR Parsing. สถานที่พิมพ์: Kluwer Academic, 1991.

ภาคผนวก

ภาคผนวก ก.

การถอดอักษรไทยเป็นอักษรโรมันแบบถ่ายเสียงราชบัณฑิตยสถาน

ตารางที่ 8 ตารางการถอดอักษรไทยเป็นภาษาโรมันแบบถ่ายเสียงราชบัณฑิตยสถาน

พยัญชนะ	รูปโรมัน		รูปสระ	รูปโรมัน
	พยัญชนะต้น	ตัวสะกด		
ก	k	k	อะ, –ั (reduced form of อะ), รร (with final consonant), อา	a
ข ฃ ค ฅ ฆ	kh	k	รร (without final consonant)	an
ง	ng	ng	อำ	am
ซ ทร (pronounced ซ) ศ ษ ส	s	t	อิ, อี	i
ญ	y	t	อึ, อื	ue
ฎ ฑ (pronouncedด) ด	d	t	อุ, อู	u
ฏ ต	t	t	เอะ, เ– ็ (reduced form of เอะ), เอ	e
ฐ ฑ ฒ ถ ท ธ	th	t	แอะ, แอ	ae
ณ น	n	n	โอะ, –(reduced form of โอะ), โอ, เอาะ, ออ	o
บ	b	p	เออะ, เ– ิ (reduced form of เออะ), เออ	oe
ป	p	p	เอียะ, เอีย	ia
ผ พ ภ	ph	p	เอือะ, เอือ	uea
ฝ ฟ	f	p	อัวะ, อัว, –ว– (reduced form of อัว)	ua
ม	m	m	ใอ, ไอ, อัย, ไอย, อาย	ai
ย	y		เอา, อาว	ao
ร	r	n	อุย	ui
ล ฬ	l	n	โอย, ออย	oi
ว	w		เอย	oei
ห ฮ	h		เอือย	ueai
			อวย	uai
			อิว	io
			เอ็ว, เอว	eo
			แอ็ว, แอว	aeo
			เอียว	iao
			ฤ (pronounced รึ), ฤๅ	rue
			ฤ (pronounced ริ)	ri
			ฤ (pronounced เรอ)	roe
			ฦ, ฦๅ	lue

ภาคผนวก ข

ตัวอย่าง 500 รายชื่อในชุดทดสอบที่ให้กลุ่มตัวอย่างประเมิน

ชื่อที่เขียนด้วยอักษรโรมัน	ชื่อไทย	ชื่อที่เขียนด้วยอักษรโรมัน	ชื่อไทย
anon	อานนท์	narumon	นฤมล
araya	อารยา	naruemon	นฤมล
areeya	อารียา	natchamon	ณัชมน
ayuwat	อายุวัฒน์	natsarun	ณัฐสรัญ
adirek	อดิเรก	natthi	ณัฐธิ
adsadakorn	อัษฎากร	natakankoon	ณัฐกานต์กุล
ajchararat	อัจฉรารัตน์	natechanok	เนตรชนก
akaraphol	อัครพล	nathaporn	ณัฐพร
akkarawat	อัครวัฒน์	nattakit	ณัฐกฤต
alisara	อลิสรา	nattapol	ณัฐพล
amnuay	อำนวย	nattapong	ณัฐพงศ์
amarisa	อมริสา	nattaporn	ณัฐพร
amornmal	อมรมาลย์	nattawut	ณัฐวุฒิ
amornthep	อมรเทพ	natthakorn	ณัฐกร
anchalee	อัญชลี	natthaporn	ณัฐพร
ansaya	อัณศยา	navaporn	นวพร
ananchai	อนันต์ชัย	nawaporn	นวพร
angkana	อังคณา	neeranart	นีรนาท
aniwat	อานิวัฒน์	nhung	นูง
anongphorn	อนงค์พร	nida	นิดา
anuphong	อนุพงษ์	nipaporn	นิภาพร
anuthida	อนุธิดา	niphon	นีพร
aornrudee	อรฤดี	nirun	นิรันดร์
apichart	อภิชาติ	nisana	นิษณา
apinai	อภินัย	nithiwadee	นิธิวดี
apiradee	อภิรดี	nitipoj	นิติพจน์
apisit	อภิสิทธิ์	niyom	นิยม
apinya	อภิญญา	nippita	นิพพิธา
arjaree	อาจารี	nittha	นิตา

arpawalee	อาภาวลี	nlin	นลิน
artinee	อาทินี	nongluk	นงลักษณ์
araya	อารยา	nont	นนท์
arin	เอกรินทร์	noppol	นพพล
art	อรรถ	nopasuk	นพศักดิ์
arunroj	อรุณโรจน์	noppakune	นพคุณ
atchana	อัจนา	noppawan	นพวรรณ
athaneeporn	อัฒณีพร	norathep	นรเทพ
atika	อธิกา	nutiyaporn	นุติยาพร
atsadang	อัษฎางค์	nuch	นุช
atthakorn	อรรถกร	numfon	น้ำฝน
aumara	อมรา	nuntiya	นันทิยา
autcharee	อัจฉรีย์	nuntawatt	นันทวัฒน์
bancha	บัญชา	nutmanee	ณัฐมณี
benja	เบญจ	nuttita	ณัฐธิตา
benjamas	เบญจมาศ	nuttakarn	ณัฐกานต์
benjawan	เบญจวรรณ	nuttapon	ณัฐพร
bongkodrut	บงกชรัตน์	nuttavee	ณัฐวี
boonlert	บุญเลิศ	nutthapong	ณัฐพงศ์
boonta	บุญตา	oran	โอฬาร
boonyarat	บุญรัตน์	onniya	อรณิชยา
burimrapee	บุริมรพี	onnjira	อรจิรา
bundid	บัณฑิต	oranuch	อรนุช
bungonsiri	บังอรศิริ	orapin	อรพิน
busaraporn	บุษราพร	orathai	อรทัย
chadanan	ชาดานันท์	orawee	อรวี
chanon	ชานนท์	ornusa	อรอุษา
chatree	ชาตรี	pajareeporn	ปาจรีย์พร
chaianun	ชัยอนันต์	paploen	พาเพลิน
chairat	ชัยรัตน์	parichat	ปาริชาติ
chaiwat	ชัยวัฒน์	pacharawan	พัชรวรรณ
chaiyan	ชัยยันต์	paiboon	ไพบูลย์
chaiyapruek	ชัยพฤกษ์	pairoj	ไพโรจน์

chakree	จักรี	pajeeluck	พจีลักษณ์
chalermkiat	เฉลิมเกียรติ	pakorn	ปกรณ์
chalida	ชลิดา	panjit	พรรณจิ
chalonglarp	ฉลองลาภ	pansa	พรรษา
chanchai	ชาญชัย	panwalai	พรรณวลัย
chansak	ชาญศักดิ์	panayu	ปนายุ
chanwit	ชาญวิทย์	panida	พนิดา
chanapat	ชนาภัทร	panita	ปณิตา
chanatda	ชนัดดา	panit	พนิต
chanikan	ชนิกานต์	panote	ปณต
chanin	ชนินทร์	papichaya	ปพิชญา
chantajit	จันทจิ	parisara	ปริสรา
chaowanee	ชวณี	parinya	ปริญ
charinthip	จรินทร์ทิพย์	passa	พรรษ
chart	ชาติ	patsawon	พรรษวรรณ
chatchai	ฉัตรชัย	patchanat	พัชนัฐ
chatnaree	ฉัตรนรี	patcharang	พัชรางศุ์
chatchaval	ชัชวาล	patcharin	พัชรินทร์
chatuphon	จตุพร	patchara	พัชร
chawalit	ชวลิต	pathom	ปฐม
chaya	ชญา	patinya	ปฏิญญา
chayapa	ชญาภา	pattamakorn	ปฐมกร
cheera	จีร	pattamaprapa	ปัทมประภา
chinapa	จิมาภา	paveena	ปวีณา
chienchai	เชียรชัย	paweena	ปวีณา
chirapa	จิรภา	pawittra	ปวิตรา
chitra	จิตรา	pechanika	พีชณิกา
chotiros	โชติรส	peerakit	พีรกิต
cholarit	ชลฤทธิ์	peerawut	พีรวุฒิ
chompoonuth	ชมพูนุช	penpaktr	เพ็ญพักตร์
chonnigarn	ชนนิกานต์	peonpon	เพียรพร
chonlada	ชลดา	petchareeporn	เพชรีพร
chorchai	ช่อชัย	phanlert	พันธุ์เลิศ

chulalak	จุฬาลักษณ์	pharnit	ภานิช
chureewan	จุรีวรรณ	phengphian	เพ่งเพียร
chutima	ชุติมา	phichayada	พิชญดา
chutimon	ชุติมน	phongcharoen	พงศ์เจริญ
chulanee	จุฬนี	photchamarnphagee	พจมานพจี
chyudh	ชยุตม์	pichamon	พิชามน
darica	ดาริกา	pichet	พิเชษฐ์
danchai	เด่นชัย	pimolchaya	พิมลชญา
danupon	ดนุพล	pipat	พิพัฒน์
decha	เดชา	piriya	พิริยา
dhanadham	ธนธรรม	pisit	พิสิทธิ์
dissaya	ดิษยา	pitha	พิกา
doungjai	ดวงใจ	pitiporn	ปิติพร
duangdao	ดวงดาว	piyachart	ปิยฉัตร
duangkamol	ดวงกมล	piyanuch	ปิยนุช
duangporn	ดวงพร	piyaporn	ปิยพร
duenchai	เดือนฉาย	piyawan	ปิยวรรณ
etaya	เอธยา	pichaya	พิชญา
eitsariya	อิสริยา	pimdaw	พิมพ์ดาว
ekapong	เอกพงษ์	pimpen	พิมพ์เพ็ญ
ekkaluck	เอกลักษณ์	pimruetai	พิมพ์ฤทัย
fahprapai	ฟ้าประไพ	pinna	พิณ
gedgaew	เกศแก้ว	pissamai	พิศมัย
guntima	กันติมา	pitchaya	พิชญา
hataichanok	หทัยชนก	piyada	ปิยดา
hathairat	หทัยรัตน์	piyanuch	ปิยนุช
irin	ไอริณ	piyarat	ปิยรัตน์
intira	อินทิรา	piyawan	ปิยวรรณ
issariya	อิสริยา	ploychanok	พลอยชนก
jakhrit	จักรกริช	pojanee	พจนีย์
jarupat	จารุภัทร	ponsuk	พรศักดิ์
jaruek	จารึก	pongpan	พงศ์พันธ์
jakkrit	จักรกฤษณ์	pongsaree	ผ่องศรี

jakrapong	จักรพงษ์	pongsathorn	พงศธร
jantawee	จันทร์วิ	pongwisute	พงศ์วิสุทธิ์
janewit	เจนวิทย์	poonsuk	พูนศักดิ์
jariya	จริยา	poramet	ปรเมศวร์
jarus	ยารัตน์	pornchan	พรจันทร์
jaturong	จตุรงค์	pornpaktra	พรภัคตรา
jeeranun	จีรนันท์	pornphan	พรพรรณ
jessada	เจษฎา	pornpit	พรพิ
jirapa	จิรภา	pornrudee	พรฤดี
jiraporn	จิราพร	porntep	พรเทพ
jirat	จิรัตน์	pornthip	พรทิพย์
jindarat	จินดารัตน์	pornusa	พรอุษา
jintawat	จินตวัฒน์	prachanart	ประชานาถ
jiraphat	จิรภัทร	prakit	ประกิต
jirawat	จิรวัฒน์	pranida	ประนิดา
jitsupa	จิตสุภา	prapaporn	ประภาพร
jittiya	จิตติยา	prapat	ประพัฒน์
jom	จอม	prapos	ประภส
juraporn	จุฬาพร	prasit	ประสิทธิ์
jutamas	จุฑามาศ	pratichaya	ประติชญา
juthamas	จุฑามาศ	prayoon	ประยูร
jugkarin	จักรินทร์	praewlada	แพรวลดา
juntrarut	จันทรรัตน์	prangtip	ปรางทิพย์
kachaporn	เกชาพร	pratthana	ปรารถนา
kajohnsak	ขจรศักดิ์	preeda	ปรีดา
kampol	กำพล	preechaya	ปรีชญา
kamolmet	กมลเมตต์	priyanuch	ปรียนุช
kamon	กมล	promphan	พรหมพรรณ
kamontip	กมลทิพย์	puchita	ปูชิตา
kanchana	กาญจนา	punnee	พรรณี
kanjana	กาญจนา	punjaporn	ปัญจพร
kannika	กรรณิกา	puttarat	พุทธรัตน์
kantaya	กันตยา	rapeepun	รพีพรรณ

kanyarat	กัญญารัตน์	rachanee	รัชนี
kangsadan	กังสดาล	rajchwanlop	ราชวัลลภ
kanit	ขนิษ	rangsima	รังสิมา
kanlaya	กัลยา	rasana	รสนา
kanokkon	กนกกร	ratana	รัตน์
kanokporn	กนกพร	ratchanan	รัชนันท์
kanokwan	กนกวรรณ	ratchata	รัชต์
kant	กานต์	rattana	รัตน์
karinya	เกศรินยา	rattapong	รัฐพงศ์
kassuda	เกศสุดา	rattiya	รัตติยา
kasidi	กษิดิ	rawiwan	รวีวรรณ
katesaraporn	เกศราพร	rindhamma	รินธรรม
kecha	เกชา	rojcharek	รจเรข
keerati	กีรติ	roongaroon	รุ่งอรุณ
kesaraporn	เกศราพร	rossana	รสนา
khajonsak	ขจรศักดิ์	rujira	รุจิรา
khanittha	ขนิษฐา	ruangyuth	เรืองยุทธ
khomsan	คมสัน	rummaneeya	รัมณีย์า
kidakarn	กิดากานต์	rungraung	รุ่งเรือง
kitisark	กิตติศักดิ์	rungtiwa	รุ่งทิวา
kiattiyot	เกียรติยศ	ruxsux	รักษ์ศักดิ์
kitja	กิจ	sarinee	สาริณี
kittichote	กิตติโชติ	satinee	สาธินี
kittipat	กิตติพัฒน์	sawitri	สาวิตรี
kittipong	กิตติพงษ์	saharut	สุดารัตน์
kittisak	กิตติศักดิ์	saitan	สายธาร
kittiya	กิตติยา	sakdapol	ศักดิ์ดาพล
komin	โกมินทร์	sakol	สกล
kobkiat	กอบเกียรติ	saleela	ศลีลา
komkrit	คมกฤช	samrit	สัมฤทธิ์
kong	ก้อง	sanchaya	สัญชญา
korkiat	กรเกียรติ	santi	สันติ
korb	กอบ	saneeya	ษณีย์า

kornsajee	กรศจี	sanitphong	สนิทพงษ์
kraikul	ไกรกุล	saowaluk	เสาวลักษณ์
kriangkrai	เกรียงไกร	saracha	ราชา
kriengkrai	เกรียงไกร	sarawoot	สราวุฒิ
krisana	กฤษณ์	sarayut	ศรายุทธ
krit	กฤษ	saranporn	ศรัณย์พร
kritsada	กฤษดา	sareewan	ศรีวรรณ
krittapong	กฤตพงศ์	sarintorn	รินธร
krongtham	ครองธรรม	sarocha	สโรชา
kuakul	เกื้อกูล	sarunya	ศรัณยา
kulvadee	กุลวดี	sasichol	ศศิชล
kullanun	กุลนันท์	sasiphim	ศศิพิมพ์
kunniti	กุลนิธิ	sasithorn	ศศิธร
kuntarat	กันตรัตน์	sasiwimon	ศศิวิมล
kwanruthai	ขวัญฤทัย	sathaporn	สถาพร
ladda	ลดา	sawanan	เสาวนันท์
lalida	ลลิดา	sayphin	สายพิณ
laliew	ลลิ่ว	seksan	เสกสรรค์
levi	เลวี	setthabut	เศรษฐบุตร
lisa	ริสา	silaporn	ศิลาพร
lukana	ลักขณา	siramas	ศิริมาศ
maliwan	มาริวรรณ	sirichai	ศิริชัย
manoch	มาโนช	sirikan	ศิริกานต์
marissa	มาริสา	sirilada	ศิริลดา
mahisorn	เมธิสร	sirima	ศิริมา
mananya	มนัญญา	sirinat	ศิรินาถ
maneerat	มณีรัตน์	siripat	ศิริภัทร
manussanit	มนัสนิตย์	siriporn	ศิริพร
mati	เมธิ	siriporn	ศิริพร
maywadee	เมวดี	sirirat	ศิริรัตน์
methawee	เมธวี	sirirut	ศิริรัตน์
melanie	มาลานี	siriwan	ศิริวรรณ
molvipa	มลวิภา	siriwat	ศิริวัฒน์

monruedee	มลฤดี	sirinporn	สิรินพร
monthira	มนทิรา	siros	ศิรส
montree	มนตรี	siwat	ศิวัช
mongkol	มงคล	sindha	ศิลป์ดา
muthita	มุทิตา	sirasa	ศิริสา
nachon	ณัฐชล	sitthipan	สิทธิพันธ์
narat	ณารัตน์	sittipon	สิทธิพร
nathakan	ณัฐกานต์	siwaporn	ศิวพร
nadda	ณัฐดา	sojiwachana	โสจิวัจน์
nakkarin	นัครินทร์	sopon	โสพล
nalintip	นลินทิพย์	sombat	สมบัติ
namwan	น้ำหวาน	somchai	สมชัย
nantiya	นันทิยา	somjed	สมเจตน์
nantarat	นันทรัตน์	sommai	สมหมาย
napachai	นภาชัย	sompoch	สมพจน์
naparatn	นภารัตน์	somporn	สมพร
naphawadee	นภาวดี	somsak	สมศักดิ์
naratip	นราทิพย์	somsuda	สมสุดา
narisara	นริศรา	sonti	สันติ
narisara	นริศรา	songpon	ทรงพล
narong	ณรงค์	soonthorn	สุนทร
narttapong	นาถพงศ์	sorapop	สรภพ

ประวัติผู้เขียนวิทยานิพนธ์

นางสาวชุลีกร กิตติกูล เกิดเมื่อวันที่ 28 มกราคม พ.ศ. 2527 ที่จังหวัดกรุงเทพมหานคร สำเร็จการศึกษาหลักสูตรวิศวกรรมศาตรบัณฑิต สาขาวิชาวิศวกรรมคอมพิวเตอร์ จากภาควิชาวิศวกรรมคอมพิวเตอร์ คณะวิศวกรรมศาสตร์ มหาวิทยาลัยเกษตรศาสตร์ ในปีการศึกษา 2549 และเข้าศึกษาต่อในหลักสูตรวิทยาศาสตรมหาบัณฑิต สาขาวิชาวิทยาศาสตร์คอมพิวเตอร์ ที่ภาควิชาวิศวกรรมคอมพิวเตอร์ คณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย ในปีการศึกษา 2552