ระบบตัดพยางค์และแปลงหน่วยเสียงสำหรับตรวจกลอนสุภาพ
Thai Phonemes Transformetrics System for Klon Supab Poetry

อรทัย คงธรรม, ณัฐโชติ พรหมฤทธิ์ และ สัจจาภรณ์ ไวจรรยา

ภาควิชาคอมพิวเตอร์ คณะวิทยาศาสตร์ มหาวิทยาลัยศิลปากร นครปฐม
Emails: khongtum_o.su.ac.th, promrit_n@silpakorn.edu, waijanya_s@silpakorn.edu

บทคัดย่อ

งานวิจัยนี้นำเสนอวิธีการตัดพยางค์ และแปลงพยางค์เป็นหน่วยเสียงสำหรับตรวจกลอนสุภาพ โดยการใช้กฎไวยากรณ์ทางภาษา (Rule Based) ซึ่งรูปแบบกฎเกิดจากการวิเคราะห์โครงสร้างพยางค์ของคำในภาษาไทย ร่วมกับหลักการอ่านภาษาไทย และใช้นิพจน์ปกติ (Regular expression) เพื่อตัดพยางค์ตามรูปแบบที่กำหนด ซึ่งครอบคลุมการตัดพยางค์คำไทย สมาส-สนธิ และคำเฉพาะ เพื่อตรวจสอบความถูกต้องของการตัดพยางค์ งานวิจัยนี้จึงนำกลอนสุภาพ จำนวน 300 สำนวน ซึ่งมี ถุงคำ (Bag-of-word) จำนวน 3,800 คำ จากสมาคมกวีร่วมสมัย และได้ค่าความถูกต้องของการตัดพยางค์เท่ากับร้อยละ 95.78

ABSTRACT

This research proposes Thai syllable separation and transformation to be a phoneme for Thai poem evaluation system which used Thai grammar rule-based methodology. A pattern rule has defined by analytical structure of Thai syllable and Thai reading principle and to use the regular expression in order to separate the syllables by specific pattern, the ability of system to separate the syllable of general Thai words, Thai compound words and Thai specific words. Three hundred Thai poems (klon-supab) from Thai Contemporary Poets Association which bag of word size 3800 words are used to evaluate performance of system. The result of evaluation shows the accuracy 95.78%.

คำสำคัญ— ตัดพยางค์; พยางค์; หน่วยเสียง; syllable segment, syllable; phonemes

1. บทนำ

ลักษณะของประโยคภาษาไทยประกอบไปด้วยการนำคำหลายๆ คำมาเรียงต่อกันโดยไม่มีการใช้สัญลักษณ์ใดเพื่อบ่งบอกขอบเขตของคำ แต่ละคำเป็นได้ทั้งคำไทย คำสมาสและคำสนธิ ซึ่งในพจนานุกรมราชบัณฑิตยสถานได้ให้ความหมายของคำไว้ว่า คือ “เสียงที่เปล่งออกมาครั้งหนึ่ง ๆ เสียงพูด หรือลายลักษณ์อักษรที่เขียนหรือพิมพ์เพื่อแสดงความคิด โดยปกติถือว่าเป็นหน่วยที่เล็กที่สุดซึ่งมีความหมายในตัว” การอ่านออกเสียงคำมีทั้งคำที่ออกเสียงพยางค์เดียว และออกเสียงหลายพยางค์ ยกตัวอย่างคำว่า “ถม” มี 1 พยางค์ “ถนน” มี 2 พยางค์ และ “พัฒนา” มี 3 พยางค์ องค์ประกอบของการเกิดเสียงพยางค์ประกอบด้วย เสียงพยัญชนะต้น เสียงสระ และเสียงวรรณยุกต์ องค์ประกอบของเสียงพยางค์เป็นสิ่งที่ทำให้เกิดจังหวะ และโทนสูง-ต่ำ ในภาษาไทย ซึ่งมีความสำคัญในการประพันธ์บทร้อยกรอง โดยการประพันธ์ร้อยกรองเป็นงานเขียนที่มีการกำกับรูปแบบด้วยฉันทลักษณ์ และฉันทลักษณ์ที่สำคัญอย่างหนึ่ง คือ จำนวนพยางค์ในแต่ละวรรค รวมทั้งเสียงของพยางค์ ซึ่งเมื่อเราต้องการตรวจสอบการประพันธ์ทางคอมพิวเตอร์ จึงมีความจำเป็นต้องใช้กระบวนการตัดคำ กระบวนการตัดพยางค์ กระบวนการแปลงคำเป็นคำอ่าน และกระบวนการแปลงหน่วยเสียงที่ถูกต้องเพื่อตรวจสอบฉันทลักษณ์และความไพเราะของกลอน

ผู้วิจัยพบว่าที่ผ่านมามีโครงงานที่จัดทำระบบตัดคำ ตัดพยางค์ แปลงคำอ่าน และแปลงหน่วยเสียง [1, 2] แต่โครงงานเหล่านั้นยังมีข้อจำกัดคือ ระบบยังไม่สามารถตัดพยางค์ของคำสมาส คำสนธิได้ หากการตัดพยางค์ไม่มีประสิทธิภาพมากพอจะทำให้ไม่สามารถตรวจสอบฉันทลักษณ์และความไพเราะของกลอนได้อย่างถูกต้อง

งานวิจัยนี้จึงนำเสนอวิธีการตัดพยางค์ และแปลงพยางค์เป็นหน่วยเสียงสำหรับตรวจกลอนสุภาพ โดยการใช้กฎไวยากรณ์ทางภาษา (Rule Based) ซึ่งรูปแบบกฎเกิดจากการวิเคราะห์โครงสร้างพยางค์ของคำในภาษาไทย ร่วมกับหลักการอ่านภาษาไทย[3] และใช้นิพจน์ปกติ (Regular expression)

เพื่อตัดพยางค์ตามรูปแบบที่กำหนด โดยในเนื้อหาที่เหลือจะประกอบไปด้วย ทฤษฎีและงานวิจัยที่เกี่ยวข้อง วิธีดำเนินงานวิจัย การทดลองและผลการทดลอง และบทสรุป

2. ทฤษฎีและงานวิจัยที่เกี่ยวข้อง

ในการศึกษาวิธีการตัดพยางค์และแปลงหน่วยเสียงมีทฤษฎีที่ผู้วิจัยได้รวบรวมขึ้นมาเพื่อศึกษาดังนี้

2.1 พยางค์ในภาษาไทย

พยางค์ คือเสียงที่เปล่งออกมาครั้งหนึ่ง ๆ ซึ่งมีเสียงสระเป็นเสียงที่ดังเด่น 1 เสียง และเสียงที่อยู่ข้างเคียงอย่างน้อย 2 เสียง ได้แก่ เสียงพยัญชนะและเสียงวรรณยุกต์ องค์ประกอบของพยางค์ในภาษาไทยมีองค์ประกอบสำคัญอย่างน้อย 3 ส่วน คือ เสียงพยัญชนะต้น + เสียงสระ + เสียงวรรณยุกต์ โครงสร้างพยางค์ มี 4 แบบ [3] ดังนี้

1) การประสม 3 ส่วนคือ พยัญชนะต้น + สระ + วรรณยุกต์ เช่น สี เป็นต้น

2) การประสม 4 ส่วนคือ พยัญชนะต้น + สระ + พยัญชนะตัวสะกด + วรรณยุกต์ เช่น ตาม เป็นต้น

3) การประสม 4 ส่วนพิเศษคือ พยัญชนะต้น + สระ + วรรณยุกต์+การันต์ เช่น เล่ห์ เป็นต้น

4) การประสม 5 ส่วนคือ พยัญชนะต้น + สระ + พยัญชนะตัวสะกด + วรรณยุกต์+การันต์ เช่น ขันธ์ เป็นต้น

2.2. ลักษณะภาษาไทย

ลักษณะภาษาไทยประกอบด้วย รูปพยัญชนะไทย สระ วรรณยุกต์ และมาตราตัวสะกด ซึ่งมีความหมายดังนี้

2.2.1. รูปพยัญชนะไทย

รูปพยัญชนะไทยมี 44 ตัวคือ “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” แบ่งออกเป็น 3 กลุ่ม [3] คือ พยัญชนะสูงมี 11 ตัวได้แก่ “ข ฃ ฉ ฐ ถ ผ ฝ ศ ส ษ ห” พยัญชนะกลางมี 9 ตัวได้แก่ “ก จ ด ต ฎ ฏ บ ป อ” พยัญชนะต่ำ มี 24 ตัว แบ่งเป็นพยัญชนะต่ำคู่ มี 14 ตัวได้แก่ “ค ฅ ฆ ช ซ ฌ ฑ ฒ ท ธ พ ฟ ภ ฮ” พยัญชนะต่ำเดี่ยว 10 ตัวได้แก่ “ง ญ ณ น ม ย ร ล ว ฬ”

พยัญชนะต้น คือ เสียงพยัญชนะที่อยู่ต้นพยางค์และนำหน้าสระ แบ่งออกเป็น 2 ประเภทได้แก่ 1) เสียงพยัญชนะต้นเดี่ยวมี 21 เสียงดังตารางที่ 1.

ตาราง 1. เสียง และรูปพยัญชนะต้นเดี่ยว

พยัญชนะ 21 เสียง	พยัญชนะ 44 รูป
ก	ก
ค	ข ฃ ค ฅ ฆ
ง	ง
จ	จ
ช	ช ฌ ฉ
ซ	ซ ศ ษ ส
ด	ด ฎ
ต	ต ฏ
ท	ท ธ ฑ ฒ ถ ฐ
น	น ณ
บ	บ
ป	ป
พ	พ ภ ผ
ฟ	ฟ ฝ
ม	ม
ย	ย ญ
ร	ร
ล	ล ฬ
ว	ว
ฮ	อ ฮ
อ	อ

2) พยัญชนะต้นประสมคือ พยัญชนะสองตัวอักษรที่ประสมกันแบ่งเป็น 2 กลุ่มคือ อักษรควบกล้ำ และอักษรนำ

อักษรควบกล้ำ คือพยัญชนะซึ่งควบกับ ร ล ว ซึ่งแบ่งเป็นอักษรควบแท้ คืออักษรควบซึ่งออกเสียงพยัญชนะตัวหน้ากับพยัญชนะตัวหลังควบกล้ำพร้อมกัน มีรูปได้แก่ กร กล กว คร ขร คล ขล คว ขว ตร ปร ปล พร พล ผล และอักษรควบไม่แท้ คืออักษรที่ควบกล้ำกันแต่ออกเสียงเฉพาะพยัญชนะตัวหน้า เช่น เศร้า ทราย

อักษรนำ คือพยัญชนะสองตัวที่ประสมกันแต่มีวิธีการออกเสียงต่างกับอักษรควบกล้ำมีลักษณะคือไม่ออกเสียงตัวนำได้แก่ อ นำ ย เช่น อย่า อยู่ เป็นตัว หรือ ห นำอักษรต่ำเดี่ยวและออกเสียงตัวนำ ได้แก่อักษรสูงนำอักษรต่ำเดี่ยว อักษรกลางนำอักษรต่ำเดี่ยว หรืออักษรสูงนำอักษรต่ำคู่

2.2.2. สระ

รูปสระ และเสียงสระในภาษาไทยมี 21 รูป 32 เสียง ดังตารางที่ 2.

ตาราง 2. รูปสระ

รูป	เรียกว่า
ะ	วิสรรชนีย์
ั	ไม้หันอากาศ
็	ไม้ไต่คู้
า	ลากข้าง
ิ	พินท์อิ

รูป	เรียกว่า
◌ั	ฝนทอง
◌ุ	นฤคหิต
◌ึ	ฟันหนู
◌ิ	ตีนเหยียด
◌ู	ตีนคู้
เ-	ไม้หน้า
ใ	ไม้ม้วน
ไ	ไม้มลาย
โ	ไม้โอ
-อ	ตัวออ
-ว	ตัววอ
-ย	ตัวยอ
ฤ	ตัวรึ
ฤๅ	ตัวรือ
ฦ	ตัวลึ
ฦๅ	ตัวลือ

2.2.3. วรรณยุกต์

วรรณยุกต์ในภาษาไทยจัดเป็น 5 เสียง และมีรูปต่างกัน 4 รูปดังตารางที่ 3.

ตาราง 3. วรรณยุกต์

เสียงวรรณยุกต์	สามัญ	เอก	โท	ตรี	จัตวา
รูปวรรณยุกต์	-	◌่	◌้	◌๊	◌๋

2.2.4. มาตราตัวสะกด

มาตราตัวสะกดมีทั้งหมด 8 แม่ [3] คือ กก กด กบ กม เกย เกอว กง และกน แบ่งได้ดังนี้

1) แม่กง ใช้ ง สะกด เช่น หาง ปลิง เป็นต้น
2) แม่กม ใช้ ม สะกด เช่น ลม แต้ม โสม มุม เป็นต้น
3) แม่เกย ใช้ ย สะกด เช่น สาย ลอย โปรย เป็นต้น
4) แม่เกอว ใช้ ว สะกด เช่น แห้ว กาว เปรี้ยว เป็นต้น
5) แม่กน ใช้ น ญ ณ ร ล ฬ สะกด เช่น นาน วิญญาณ วานร เป็นต้น
6) แม่กก ใช้ ก ข ค ฆ สะกด เช่น ปัก เลข เป็นต้น
7) แม่กด ใช้ ด จ ช ซ ฎ ฏ ฐ ฑ ฒ ต ถ ท ธ ศ ษ ส สะกด เช่น แปด ตรวจ เป็นต้น
8) แม่กบ ใช้ บ ป ภ พ ฟ สะกด เช่น กลับ บาป ลาภ นพรัตน์ กราฟ เป็นต้น

2.3. หลักการอ่านภาษาไทย

การอ่านออกเสียงคำให้ถูกต้องในภาษาไทย มีหลักการอ่านภาษาไทย [4] ดังนี้

2.3.1. การอ่านอักษรควบกล้ำมี 2 ชนิดดังนี้

1) อักษรควบแท้ คือพยัญชนะอื่นที่ควบกล้ำกับพยัญชนะ ร ล หรือ ว แล้วอ่านออกเสียงพยัญชนะทั้งสองตัวนั้นเป็นเสียงเดียวกัน เช่น คลาด คำอ่าน คลาด
2) อักษรควบไม่แท้ คือพยัญชนะอื่นที่ควบหรือกล้ำกับตัว ร แล้วออกเสียงเฉพาะเสียงเฉพาะพยัญชนะต้นตัวหน้าเพียงตัวเดียว ไม่ออกเสียง ร และเมื่อ ทร ควบกันจะออกเป็นเสียง ซ เช่น จริง คำอ่าน จิง

2.3.2. การอ่านอักษรนำ

1) อักษรสูงนำอักษรเดี่ยว หรือ อักษรกลางนำอักษรเดี่ยว ให้อ่านออกเสียง 2 พยางค์ พยางค์แรกออกเสียง อะ เพียงครึ่งหนึ่งหรือกึ่งเสียง พยางค์หลังให้อ่านแบบมี ห นำ เช่น ขณ คำอ่าน ขะ - หนะ
2) อักษรสูงนำอักษรกลาง เวลาอ่านให้ออกเสียงเป็น 2 พยางค์แล้วผันตามอักษรกลางไม่ผันตามอักษรสูงดังต่อไปนี้ เช่น ขจิต คำอ่าน ขะ - จิด

2.3.3. การอ่านตัว ฤ

1) ฤ อ่าน รึ เมื่อตามหลังพยัญชนะ ค น พ ม ห หรือ ฤ เป็นพยัญชนะต้น เช่น คฤหาสน์ คำอ่าน คะ - รึ - หาด
2) ฤ อ่าน ริ เมื่อตามหลังตัว ก ต ท บ ป ศ ส ห เช่น กฤช คำอ่าน กริด
3) ฤ อ่าน เริก เมื่อเป็นพยัญชนะต้นและตามด้วย ก เช่น ฤกษ์ คำอ่าน เริก

2.3.4. การอ่านคำสมาส

คำสมาสที่มาจากบาลีสันสกฤต ต้องอ่านออกเสียงต่อเนื่องกันไป โดยมีหลักเกณฑ์ดังต่อไปนี้

1) อ่านออกเสียงสระอะ ที่พยางค์ท้าย เมื่อพยางค์ท้ายไม่มี สระ อ เช่น คณบดี อ่านว่า คะ - นะ - บอ - ดี
2) อ่านออกเสียง อิ ที่พยางค์ท้าย เมื่อพยางค์ท้ายมีสระ อิ เช่น เกียรติคุณ อ่านว่า เกียด - ติ - คุน
3) อ่านออกเสียง สระอุ ที่พยางค์ท้าย ถ้าพยางค์ท้ายมีสระอุ เช่น ประทุษร้าย อ่านว่า ประ - ทุด - สะ - ร้าย
4) อ่านออกเสียงต่อเนื่องเหมือนคำสมาส เช่น ดาษดา อ่านว่า ดาด - สะ - ดา
5) อ่านออกเสียงไม่ต่อเนื่องเหมือนคำสมาส เช่น กาลเวลา อ่านว่า กาน - เว - ลา

2.2.5. การอ่านตัว รร

1) พยัญชนะที่ไม่มีรูปสระแล้วมี “รร” อยู่หลัง ให้อ่านออกเสียงเป็น อัน โดยให้ “ร” ตัวแรกเป็นไม้หันอากาศ แล้วให้ “ร” ตัวหลังเป็น น ทำหน้าที่เป็นตัวสะกด เช่น ขรรค์ อ่านว่า ขัน

2) พยัญชนะที่มี “รร” ตามหลัง แล้วมีตัวสะกดด้วยให้ตัว “รร” เป็นเท่ากับไม้หันอากาศ เช่น กรรมการ อ่านว่า กำ - มะ – กาน

2.4. คำสมาสและคำสนธิ

คำสมาส [5] หมายถึง คำที่เกิดจากการนำคำตั้งแต่ 2 คำขึ้นไปมาต่อกันเป็นคำเดียวตามหลักของภาษาบาลี – สันสกฤต โดยคำที่นำมาประกอบกันนั้นเป็นคำภาษาบาลี – สันสกฤต ที่แปลความหมายจากหลังมาหน้า ตัวอย่างเช่น จิตรกรรม อ่านว่า จิด – ตฺระ – กำ

คำสนธิ [5] หมายถึง คำที่เกิดจากการที่หน่วยเสียง 2 หน่วย มาอยู่ประชิดกันแล้ว หน่วยเสียงใดหน่วยเสียงหนึ่งหรือทั้งสองแปรไป หรือหน่วยเสียง 2 หน่วย รวมเข้าเป็นหน่วยเสียงเดียวกัน คำที่นำมาประกอบกันจะเป็นคำภาษาบาลี – สันสกฤต และแปลความหมายจากหลังมาหน้า ตัวอย่างเช่น ราชูปโภค อ่านว่า รา – ชู – ปะ – โพก

ผู้วิจัยได้สังเกตคำสมาส-สนธิ และนำไปสร้างกฎ พบว่าเป็นคำที่ลงท้ายไม่ตรงตามมาตราตัวสะกด พยัญชนะที่ตรงตามมาตราตัวสะกดได้แก่ “ง ม ย ก น บ ด ว”

2.5. นิพจน์ปกติ (Regular Expression)

นิพจน์ปกติ คือ การกำหนดรูปแบบเพื่อค้นหาข้อความตามโครงสร้างรูปแบบที่กำหนด[6] โดยมีสัญลักษณ์ที่ใช้อธิบายดังตารางที่ 4.

ตาราง 4. สัญลักษณ์ที่ใช้อธิบายนิพจน์ปกติ

เครื่องหมาย	ความหมาย	เครื่องหมาย	ความหมาย
.	ใช้แทนตัวอักษรใดๆ	+	ใช้แทนหนึ่งหรือมากกว่าของนิพจน์ก่อนหน้า เช่น A+ มีความหมายว่า A AA หรือ AAA..
^	ใช้แทนรูปแบบในช่วงเริ่มต้นของข้อความ เช่น “^abc” มีความหมายว่า ข้อความที่ขึ้นต้นตรงกับ “abc”	?	ใช้แทนศูนย์หรือหรือหนึ่งของนิพจน์ก่อนหน้า เช่น a? หมายความว่าจะมี a หรือ ไม่มี a ก็ได้
$	ใช้แทนส่วนท้ายของข้อความ เช่น “abc$” มีความหมายว่า ข้อความที่ส่วนท้ายตรงกับ “abc”	{n}	มีนิพจน์ซ้ำกันจำนวน n นิพจน์โดยที่ n ไม่สามารถเป็นตัวเลขติดลบ เช่น “go{5}gle” มีความหมายว่า gooooogle
[ ]	ใช้แทนกลุ่มของตัวอักษรตัวใดตัวหนึ่ง เช่น “[ABC]” มีความหมายว่า ข้อความที่มี “A” หรือ “B” หรือ “C”	{n,}	มีนิพจน์ซ้ำกันอย่างน้อย n นิพจน์ เช่น “a{2,}” มีความหมายว่า aa หรือ aaa..
[A-Z0-9]	ใช่แทนข้อความที่ตรงกับหนึ่งในช่วงตัวอักษร	{,n}	มีนิพจน์ซ้ำกันไม่เกิน n นิพจน์ เช่น “b{,2}” มีความหมายว่า b หรือ bb
|	ใช้สำหรับสร้างทางเลือก เช่น “ed|ing|s” มีความหมายว่า “ed” หรือ “ing” หรือ “s”	{m,n}	มีนิพจน์ซ้ำกันอย่างน้อย m นิพจน์และไม่เกิน n นิพจน์ เช่น “go{2,4}gle” มีความหมายว่า google gooogle และ gooooogle
*	ใช้แทนศูนย์หรือมากกว่าของนิพจน์ก่อนหน้า	()	ใช้บ่งบอกขอบเขตของนิพจน์ เช่น “ex(pres|pan)sion” มีความหมาย ว่า

เครื่องหมาย	ความหมาย	เครื่องหมาย	ความหมาย
	เช่น A* มีความหมายว่าไม่มี A หรือมี A ก็ได้		expression หรือ expansion

2.6. JSON

JSON ย่อมากจาก Java Script Object Notation เป็นรูปแบบมาตรฐานการจัดเก็บข้อมูลและการแลกเปลี่ยนข้อมูลที่เข้าใจง่าย ซึ่ง JSON ถูกกำหนดด้วยภาษา Java Script โดย JSON จะแลกเปลี่ยนข้อมูลผ่านเบราว์เซอร์และเซิร์ฟเวอร์ ตัวอย่างโครงสร้างของ JSON ดังรูปที่ 1.

```
{
    "employees": [{
        "firstname": "Orathai"
    }, {
        "firstname": "Oraphan"
    }, {
        "firstname": "Janjira"
    }]
}
```

รูปที่ 1. โครงสร้างของ JSON

โครงสร้างของ JSON [7] มีลักษณะดังนี้

1) ข้อมูลของ JSON จะต้องกำหนดเป็นคู่ โดยจะมีชื่อข้อมูล(Name) และเนื้อข้อมูล(Values) เช่น “firstName”:“” โดยที่เนื้อข้อมูลสามารถเป็นค่าจำนวนตัวเลข ข้อมูลตัวอักษร ค่าบูลีนได้ หากเนื้อข้อมูลเป็นอาเรย์ (Arrays) ให้ใช้เครื่องหมาย Square brackets ([ ]) ครอบข้อมูลเหล่านั้น หรือใช้เครื่องหมาย { } ใช้ในการครอบวัตถุข้อมูล
2) ข้อมูลแต่ละคู่ จะคั่นด้วยเครื่องหมาย Comma (,)
3) เครื่องหมาย { } ใช้ในการครอบวัตถุข้อมูลหนึ่งๆ หรือคั่นข้อมูลแต่ละระเบียน (Record)
4) เครื่องหมาย [ ] ใช้ในการสนับสนุนการทำงานแบบอาร์เรย์ (Arrays)

2.7. งานวิจัยที่เกี่ยวข้อง

ในการศึกษาวิธีการตัดพยางค์และแปลงหน่วยเสียงมีงานวิจัยที่เกี่ยวข้อง ที่ผู้วิจัยได้รวบรวมขึ้นมาเพื่อศึกษาดังนี้

2.7.1 พจนานุกรมคำอ่านไทย

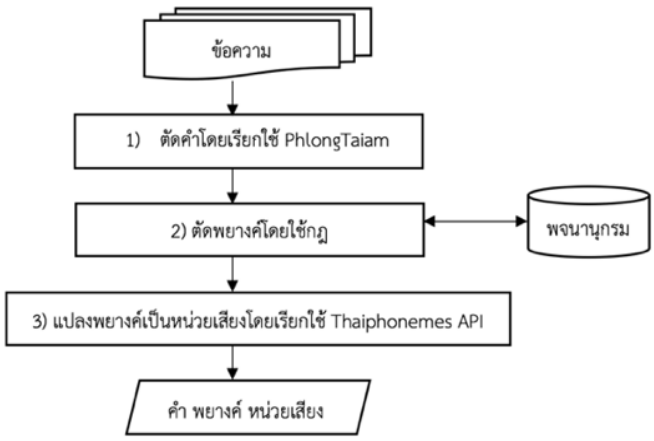
ในโครงงานนี้ได้พัฒนาโครงงานที่สามารถเปลี่ยนคำในภาษาไทยให้เป็นรูปแบบคำอ่านที่ถูกต้อง โดยวิธีการที่ใช้คือหลักการ Thai Minimum Clusters (TMC) [1] ซึ่ง TMC คือ กลุ่มตัวอักษรที่น้อยที่สุดที่สามารถออกเสียงเป็นหนึ่งพยางค์ได้ โดยโครงงานนี้มีข้อจำกัดคือ โครงงานไม่สามารถตัดคำสมาส-สนธิ และคำเฉพาะได้ครอบคลุมทุกคำ ยกตัวอย่างเช่น ศิลปากร โครงงานแบ่งพยางค์ได้เป็น ศิล-ปาก-ร

2.7.2 ตัวตรวจทางฉันทลักษณ์ และคุณภาพของกลอนสุภาพ

โครงงานนี้นำเสนอไวยากรณ์ BNF [7] สำหรับกลอนสุภาพ และใช้กฎการเขียนพยางค์ไทยแปลงพยางค์เป็นสัทอักษร สำหรับในส่วนแปลงเสียงอักษรไทย ทำหน้าที่แปลงอักษรภาษาไทย เป็นหน่วยเสียงสากล โดยโครงงานนี้มีข้อจำกัดคือ โครงงานนี้ จะนำเข้าข้อมูลในรูปแบบพยางค์ที่คั่นด้วยเครื่องหมาย “–” ถ้าใส่เป็นประโยคจะไม่สามารถแปลงเป็นหน่วยเสียงได้ ขณะที่โครงงานนี้พัฒนากระบวนการแปลงประโยคเป็นพยางค์โดยอัตโนมัติ

3. วิธีการดำเนินงานวิจัย

งานวิจัยนี้แบ่งการทำงานออกเป็น 3 ส่วนดังนี้ ส่วนที่ 1) กระบวนการตัดคำ ส่วนที่ 2) กระบวนการตัดพยางค์ และส่วนที่ 3) กระบวนการแปลงเป็นหน่วยเสียง โดยภาพรวมของระบบดังรูปที่ 2.

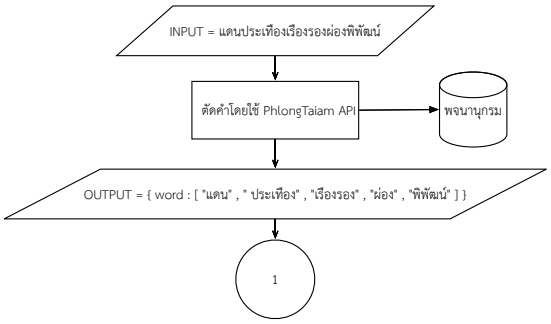


รูปที่ 2. กระบวนการแปลงคำเป็นพยางค์

3.1. กระบวนการตัดคำ

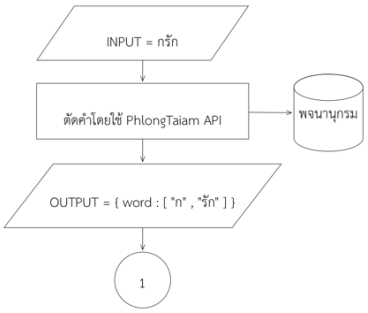
งานวิจัยนี้เรียกใช้ PhlongTalam API ในการตัดคำ เป็นวิธีการตัดคำโดยใช้พจนานุกรม ซึ่งใช้วิธีการตัดคำแบบเลือกคำที่ยาวที่สุด (Longest Matching) โดย PhlongTalam API จะรับข้อมูลเข้าในรูปแบบข้อความ และคืนค่าข้อมูลเป็น 2 ลักษณะดังนี้ 1) ข้อความถูกแบ่งคำได้ถูกต้องโดยเมื่อพบคำในข้อความปรากฏในพจนานุกรม 2) ข้อความถูกแบ่งเป็นลักษณะอักขระเพียง

ตัวอักษรตัวเดียวเมื่อไม่พบคำในพจนานุกรม ข้อความที่ถูกแบ่งคำได้ถูกต้องจะคืนค่าคำที่ถูกต้องดังรูปที่ 4.



รูปที่ 4. กระบวนการตัดคำเมื่อพบคำในพจนานุกรม

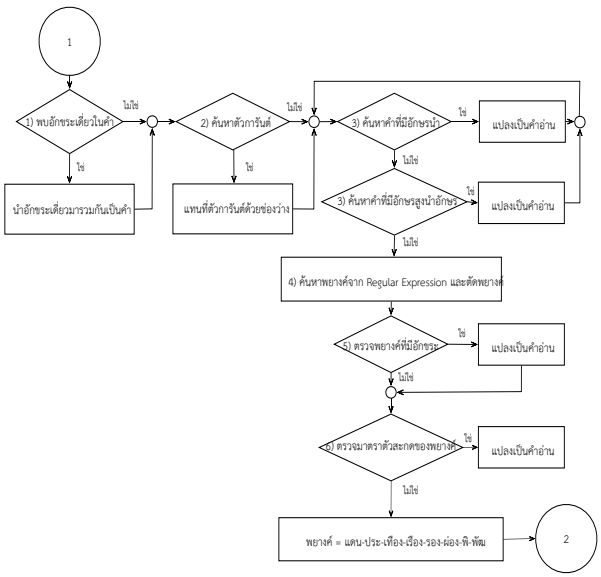
ข้อความถูกแบ่งเป็นอักขระเพียงตัวอักษรตัวเดียว เมื่อไม่พบคำนั้นในพจนานุกรม ดังรูปที่ 5.



รูปที่ 5. กระบวนการตัดคำเมื่อไม่พบคำในพจนานุกรม

3.2. กระบวนการตัดพยางค์

การตัดพยางค์ในโครงงานนี้จะใช้กฎที่ผู้วิจัยออกแบบตามโครงสร้างพยางค์ของคำในภาษาไทย ร่วมกับหลักการอ่านภาษาไทยมีขั้นตอนการตัดพยางค์ดังรูปที่ 6.



รูปที่ 6. ขั้นตอนการตัดพยางค์

การทำงานของโครงงานมีขั้นตอนตัดพยางค์ดังนี้

1) ตรวจสอบคำหากพบคำตัดได้อักขระตัวเดียว จะนำอักขระตัวเดียวมารวมกันเป็นคำ ยกตัวอย่างเช่น “ก / รัก” เมื่อนำมารวมกันจะได้ “กรัก”
2) ตรวจสอบคำกับกฎตัวการันต์ ค้นหาโดยใช้ Regular expression มีกฎดังนี้ “([ก-ฮ]|ทร์|ตร์| ธุ์|ดิ์|ธิ์|ทน์|ษณ์|คุ์)” ถ้าค้นหาพบให้แทนที่การันต์ด้วยช่องว่าง ยกตัวอย่างเช่น “การันต์” จะถูกเปลี่ยนเป็น “การัน”
3) ค้นหาคำโดยใช้สัญลักษณ์นิพจน์ปกติ (Regular expression) ดังตารางที่ 5.

ตาราง 5. สัญลักษณ์ที่ใช้ในกฎ

สัญลักษณ์	ความหมาย	พยัญชนะ
C	พยัญชนะต้นเดี่ยว	ก-ฮ
CC	พยัญชนะต้นควบ	กร กล กว ขร ขล ขว คร คล คว ชร ตร ทร ปร ปล ผล พร พล
V	สระตามหลังพยัญชนะ	ิ ี ึ ื ุ ู
S	พยัญชนะตัวสะกด	ก ข ค ฆ จ ด ต ถ ท ธ ฎ ฏ ฑ ฒ ช ซ ศ ษ ส บ ป พ ภ ฟ น ณ ร ล ฬ ง ม ย ว ญ
N	อักษรนำ	ขณ ขน ขม ขย ฉง ฉน ฉม ฉล ฉว ถง ถน ถม ถล ถว ผง ผน ผย ผล ศย ศล ศว ศล สง สน สม สย สร สล สว กน กล จม จร จว ตม ตล ตว ตน ปร ปล อง อน อร อล
HM	อักษรสูงนำอักษรกลาง	ขจ ขบ ฉก ผก ผจ ผด ศก ศด สก สด สต สบ สป
T	วรรณยุกต์	่ ้ ๊ ๋ ◌

นำคำมาค้นหาคำที่มีอักษรนำ และคำที่มีอักษรสูงนำอักษรกลางจากกฎดังตารางที่ 6.

ตาราง 6. กฎค้นหาอักษรนำ

ข้อ	โครงสร้างกฎ	ข้อ	โครงสร้างกฎ
1.	(N)+(T)?ะ	19.	(เ|แ|ใ)(N)+(T)?+(S)
2.	แ(N)+(T)?ะ	20.	(เ|แ)(N)+(T)?

ข้อ	โครงสร้างกฎ	ข้อ	โครงสร้างกฎ
3.	เ(N)+(T)?าะ	21.	(N)+(V)+(T)?+(S)
4.	(N)+(T)?ำ	22.	(N)+(V)+(T)?
5.	เ(N)+(T)?า	23.	(N)+(T)?+(S)
6.	ไ(N)+(T)?	24.	(HM)+(T)?า+(S)
7.	(N◌ั)+(T)?	25.	(HM)+(T)?า
8.	เ(N)+(T)?อ	26.	(HM)+(V)+(T)?+(S)
9.	(เ)(N◌็)+(S)	27.	(HM)+(V)+(T)?
10.	(N)+(T)?(อ|ว)+(S)	28.	(เ|แ|ใ)+(HM)+(T)?+(S)
11.	(N)+(T)?อ	29.	(เ|แ|ใ)+(HM)+(T)?
12.	(N◌ั)+(T)?+(S)	30.	เ(HM◌ี)+(T)?ย+(S)
13.	เ(N◌็)+(T)?+(S)	31.	(HM)+(T)?(อ|ว)+(S)
14.	เ(N◌ี)+(T)?ย+(S)	32.	(HM◌ั)+(T)?+(S)
15.	เ(N◌็)+(T)?ย	33.	(เ)(HM◌์)+(S)
16.	เ(N◌ี)+(T)?อ+(S)	34.	เ(HM◌็)+(T)?+(S)
17.	(N)+(T)?า+(S)	35.	(HM)+(T)?+(S)
18.	(N)+(T)?า		

ถ้าพบคำที่มีอักษรนำ หรือ อักษรสูงนำอักษรกลางให้แปลงเป็นคำจากหลักภาษาไทยในหัวข้อที่ 2.3.2. โดยวิธีการแปลงเป็น 2 วิธีดังนี้ 1) พบคำที่มีอักษรนำ ให้แปลงเป็นคำอ่านโดยให้เติมสระอะที่พยางค์แรกและพยางค์หลังให้ออกเสียงเหมือน “ห” นำ ยกตัวอย่างเช่น ขยาย แปลงเป็น ขะ-หยาย 2) พบคำที่มีอักษรสูงนำอักษรกลาง ให้เติมสระอะที่พยางค์แรกและพยางค์หลังผันตามอักษรกลาง ยกตัวอย่างเช่น ผกา แปลงเป็น ผะ-กา

4) ตัดพยางค์จากกฎตัดพยางค์ที่ออกแบบตามโครงสร้างพยางค์ไทย โดยใช้สัญลักษณ์นิพจน์ปกติ (Regular expression) ดังตารางที่ 7.

ตารางที่ 7. กฎการตัดพยางค์

ข้อ	โครงสร้างกฎ	ข้อ	โครงสร้างกฎ
1.	(CC|C)+(T)?ะ	19.	(CC|C)+(T) ?อ
2.	(C◌ั)+(T)?วะ	20.	เ(CC|C)+(T) ?อ
3.	เ(C)+(T)?ะ	21.	(CC|C)+(V)+(T)? +(S)
4.	แ(C)+(T)?ะ	22.	(เ|แ|ใ)(CC|C)+(T)?+(S)
5.	เ(CC|C◌ี)+(T) ?ยะ	23.	(C◌์)
6.	เ(CC|C◌ี)+(T)?อะ	24.	(เ|แ)(CC|C◌็)+(S)

ข้อ	โครงสร้างกฎ	ข้อ	โครงสร้างกฎ
7.	โ(C)+(T) ?ะ	25.	(CC|C◌็)อ+(S)
8.	เ(CC|C)+(T) ?าะ	26.	(CC|C)+(V)+(T) ?
9.	เ(CC|C)+(T) ?อะ	27.	(เ|แ|ใ)(CC|C)+(T)?
10.	(CC|C)+(T)?ำ	28.	เ(CC|C◌ี)+(T)?ย
11.	เ(CC|C)+(T) ?า	29.	เ(CC|C◌ี)+(T)?อ
12.	ไ(CC|C)+(T) ?	30.	(CC|C)+(T)?า
13.	ใ(CC|C)+(T) ?	31.	เCC|C◌ี)+(T)?+(S)
14.	(C◌ั)+(T)?	32.	(CC|C)+(T)?(อ|ว)+(S)
15.	เ(CC|C◌ี)+(T) ?อ+(S)	33.	(CC|C◌ั)+(T)?+(S)
16.	(CC|C)+(T) ?+(S)	34.	(CC|C)+(T)?+(S)
17.	เ(C◌ี)+(T) ?ย+(S)	35.	C
18.	เ(CC|C◌ี)+(T) ?อ		

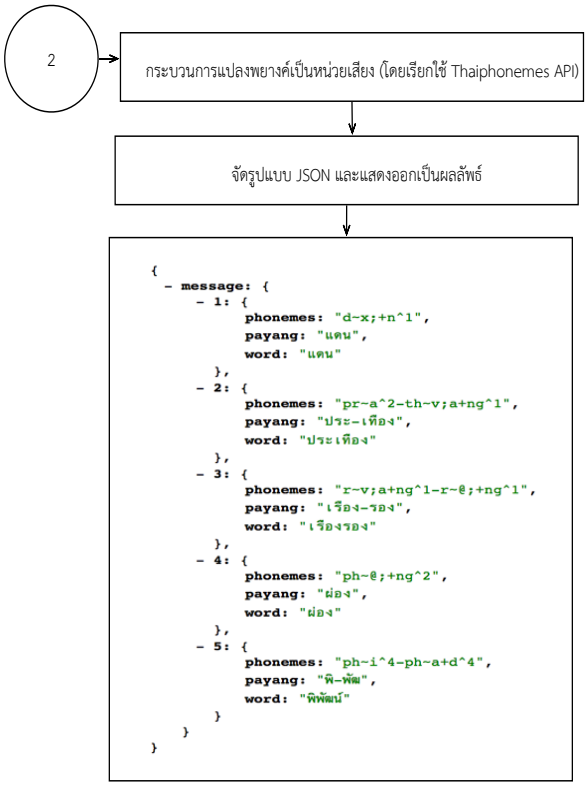
5) ตรวจพยางค์ที่ถูกแบ่งได้อักขระเดียว ถ้าพบให้เติม “สระอะ” หลังพยัญชนะ ยกตัวอย่างเช่น วลี แบ่งพยางค์ได้เป็น ว-ลี แปลงเป็น วะ-ลี

6) ตรวจมาตราตัวสะกดเพื่อตรวจสอบว่าพยางค์นั้นควรอ่านแบบคำสมาส-สนธิในหลักการอ่านหัวข้อ 2.4. ผู้วิจัยจึงได้สังเกตคำสมาส-สนธิ และนำไปสร้างกฎ โดยนำพยางค์มาตรวจสอบมาตราตัวสะกด หากพบพยางค์ที่ไม่ได้ลงตามมาตราตัวสะกด ให้เพิ่มพยางค์ด้วยวิธีนำพยัญชนะท้าย+สระอะ ยกตัวอย่างเช่น “นิจศีล” ในระดับพยางค์จะถูกแบ่งได้เป็น “นิจ-ศีล” เมื่อนำพยางค์แรกมาตรวจสอบพยัญชนะท้ายไม่ตรงตามมาตราตัวสะกด จึงเติมพยางค์เพิ่ม ได้พยางค์ใหม่เป็น “ นิจ-จะ-ศีล ”

7) เมื่อได้พยางค์ทั้งหมด จะส่งพยางค์ที่มีคั่นด้วยเครื่องหมาย “-” ไปเรียกใช้ ThaiPhoonemes API เพื่อแปลงพยางค์เป็นหน่วยเสียง ยกตัวอย่างเช่น แดน-ประ-เทือง-เรือง-รอง-ผ่อง-พิ-พัฒ

3.3. กระบวนการแปลงพยางค์เป็นหน่วยเสียง

งานวิจัยนี้เรียกใช้ ThaiPhonemes API ในการแปลงพยางค์เป็นหน่วยเสียง [8] โดย ThaiPhonemes API จะรับข้อมูลเข้าเป็นพยางค์ที่คั่นด้วยเครื่องหมาย “ - ” ดังนั้นเมื่อได้คำที่แบ่งพยางค์แล้ว จะต้องทำการใส่เครื่องหมาย “ - ” ระหว่างพยางค์เพื่อเป็นข้อมูลเข้าในส่วนการแปลงเป็นหน่วยเสียง เช่น “แดน-ประ-เทือง-เรือง-รอง-ผ่อง-พิ-พัฒ” โดยผลลัพธ์ ThaiPhonemes API จะคืนค่าแต่ละพยางค์เป็นหน่วยเสียง และนำมาจัดรูปแบบ JSON ให้คืนค่า คำ พยางค์และหน่วยเสียงดังรูปที่ 8.



รูปที่ 8. หน่วยเสียงในรูปแบบ JSON

4. การทดสอบ และผลการทดสอบ

เพื่อทดสอบความถูกต้องของระบบที่พัฒนาขึ้น ผู้พัฒนาได้นำคำภาษาไทยจากกลอนจำนวน 300 สำนวนจากสมาคมกวีร่วมสมัย [9] ซึ่งมีจำนวน ถุงคำ (Bag-of-word) จำนวน 3800 คำ มาใช้ในการทดสอบและวัดผลความถูกต้องของการแปลงคำเป็นพยางค์ของระบบโดยใช้ตัวชี้วัดความถูกต้องคือ คำที่ตัดพยางค์ถูกต้องต่อจำนวนคำทั้งหมดคิดเป็นร้อยละดังสมการนี้

$$ความถูกต้อง\ (\%) = \frac{W}{N} * 100 \qquad (1)$$

เมื่อ W คือ คำที่ตัดพยางค์ถูกต้อง
และ N คือ จำนวนคำทั้งหมด

ผลการทำสอบความถูกต้องของการแปลงคำเป็นพยางค์ดังแสดงตารางที่ 6.

ตาราง 6. ผลการทดสอบความถูกต้องของการแปลงคำเป็นพยางค์

ประเภทเอกสาร	จำนวนคำ	ตัดพยางค์ถูก	คิดเป็น (%)
กลอนสุภาพ 300 สำนวน	3,800	3,640	95.78

5. บทสรุป

การทดลองนี้ได้ใช้วิธีการตัดพยางค์จากกฎไวยากรณ์โครงสร้างพยางค์ไทย ร่วมกับหลักการอ่านภาษาไทย จากผลการทดลองตัดพยางค์จากกลอน 300 สำนวน ได้ความถูกต้องในการตัดพยางค์ร้อยละ 95.78 ซึ่งปัญหาที่พบเกิดจากพยางค์ที่ไม่สามารถแยกได้ว่าเป็นพยัญชนะตัวสะกดของพยางค์แรก หรือเป็นพยัญชนะต้นของพยางค์ถัดไป เช่น จิตรา แบ่งพยางค์ได้เป็น จิต-รา และยังพบปัญหาที่เกิดจากคำเฉพาะ เนื่องจากส่วนของการตัดคำไม่สามารถตัดคำเฉพาะได้ถูกต้อง เช่น เหลืองสมานกุล ถูกแยกคำได้เป็น เหลือง-สมา-นกุล ทำให้ในส่วนของการตัดพยางค์ถูกแยกจากกัน ได้เป็น เหลือง-สะ-มา-นะ-กุล และเนื่องจากในหลักภาษาไทยยังมีคำอ่านคำเฉพาะที่ไม่เป็นไปตามกฎ ทำให้ไม่สามารถนำมาเขียนกฎได้ครอบคลุมคำภาษาไทยทั้งหมด สำหรับแนวทางในการพัฒนาระบบต่อในอนาคตคือสร้างกฎเพิ่มเติมเพื่อให้ผลดียิ่งขึ้น

6.เอกสารอ้างอิง

[1] นพพล น้อยโต, 2555, พจนานุกรมคำอ่านไทย, ปริญญามหาวิทยาลัยศิลปากร.

[2] ณัฐชา จามรธรรมกุล และพฤกษาพรรณ มีหมก, 2557, เครื่องมือแปลงหน่วยอักขระไปสู่หน่วยเสียงสำหรับภาษาไทย, ปริญญานิพนธ์วิทยาศาสตร์บัณฑิต สาขาวิทยาการคอมพิวเตอร์ มหาวิทยาลัยศิลปากร.

[3] พระยาอุปกิตศิลปสาร, 2511, หลักภาษาไทย ,กรุงเทพฯ: ไทยวัฒนาพาณิชย์

[4] เอกรัตน์ อุดมพร, 2547, หลักการอ่านภาษาไทย ,กรุงเทพฯ: พัฒนาศึกษา

[5] 2551, “เอกสารแบ่งคำไทย” ,[ออนไลน์] เข้าถึงได้: http://lexitron.nectec.or.th/KM_HL5001/file_HL5001/Document/krrn_14625.doc สืบค้นวันที่ 16 มกราคม 2560

[6] 2558, “regular expressions” ,[ออนไลน์] เข้าถึงได้: http://www.nltk.org/book/ch03.html สืบค้นวันที่ 16 มกราคม 2560

[7] ณัฐภัทร แก้วรัตนภัทร, 2558, “คู่มือโครงสร้างเจสัน” ,[ออนไลน์] เข้าถึงได้:http://www.teacher.ssru.ac.th/nutthapat_ke/file.php/1/IntroJSON3_new.pdf สืบค้นวันที่ 16 มกราคม 2560

[8] สัจจาภรณ์ ไวจรรยา, 2550, “ตัวตรวจทานฉันทลักษณ์และคุณภาพของกลอนสุภาพ” ,[ออนไลน์] เข้าถึงได้ : http://www.gits.kmutnb.ac.th/ethesis/data/4740583382.pdf/ สืบค้นวันที่ 10 ตุลาคม 2559

[9] 2550, “สมาคมกวีร่วมสมัย” ,[ออนไลน์] เข้าถึงได้: http://www.kawethai.com/ สืบค้นวันที่ 16 มกราคม 2560