



รายงานวิจัยฉบับสมบูรณ์

การสร้างคลังเสียงพูดเพื่อการรู้จำเสียงพูดอัตโนมัติและการสังเคราะห์เสียงพูด
ภาษาไทยโดยการเลือกข้อความออนไลน์โดยอัตโนมัติ

Constructing Speech Corpus for Thai ASR and TTS with
Automatic online Text Selection

โดย

สุรพล วรภัทราร

ศิริกานต์ ชูเชิด

ลักษณ เจริญวัฒนา

งานวิจัยนี้ได้รับเงินอุดหนุนการวิจัยจากมหาวิทยาลัยแม่ฟ้าหลวง

ประจำปีงบประมาณ พ.ศ. 2556

บทสรุปผู้บริหาร

เทคโนโลยีการโต้ตอบกันระหว่างคอมพิวเตอร์กับมนุษย์กำลังเป็นที่สนใจของบรรดานักวิจัยและหน่วยงานต่าง ๆ ทำให้เกิดนวัตกรรมด้านเสียงพูดออกมามากขึ้น เช่น ระบบรู้จำเสียงพูดที่ใช้งานเพื่อใช้ในการโต้ตอบกันระหว่างมนุษย์กับหุ่นยนต์ โปรแกรมช่วยอ่านหน้าจอคอมพิวเตอร์ ระบบรายงานสภาพการจราจรด้วยเสียงอัตโนมัติหรือการสังเคราะห์เสียงพูดบนอุปกรณ์พกพา สิ่งที่เป็นสำหรับเทคโนโลยีการรู้จำเสียงพูดอัตโนมัติและการสังเคราะห์เสียงพูดคือ “คลังข้อความ” ซึ่งประโยคจากคลังข้อความนี้จะถูกนำไปอ่านบันทึกเสียงโดยอาสาสมัครผู้ให้เสียง เพื่อสร้างระบบการรู้จำเสียงพูดอัตโนมัติหรือระบบสังเคราะห์เสียงพูดดังกล่าว อย่างไรก็ตามประสิทธิภาพระบบรู้จำเสียงพูดและคุณภาพของเสียงพูดสังเคราะห์ขึ้นอยู่กับจำนวนหน่วยเสียงที่ครอบคลุมในฐานข้อมูล จึงจำเป็นต้องใช้เวลาในการบันทึกเสียงค่อนข้างมากเนื่องจากต้องการให้เกิดความครอบคลุมของหน่วยเสียงที่มากที่สุด ทั้งนี้คำศัพท์และสำนวนใหม่ ๆ นั้นเกิดขึ้นทุกวัน เราจึงต้องการคลังข้อความที่ทันสมัย เพื่อที่จะนำมาสังเคราะห์เสียงพูดหรือรู้จำเสียงพูดในปัจจุบันได้

ในงานวิจัยนี้เราจึงเสนอ การสร้างคลังข้อความจากการกระจายตัวของหน่วยเสียงที่กำหนด และใช้วิธีคัดเลือกประโยคจากอินเทอร์เน็ตอย่างต่อเนื่อง การกระจายตัวทางหน่วยเสียงที่กำหนด ในที่นี้หมายถึง คลังข้อความที่ได้ มีจำนวนรูปแบบของแต่ละหน่วยเสียง เป็นไปตามอัตราส่วนเป้าหมาย และยังสามารถกำหนดเกณฑ์ของจำนวนขั้นต่ำของหน่วยเสียง การคัดเลือกประโยคจะใช้วิธีการประยุกต์ขั้นตอนเชิงละโมบ (Greedy Algorithm) ซึ่งเป็นอัลกอริทึมใช้ในการหาประโยคที่ทำให้การกระจายตัวทางหน่วยเสียงเป็นไปตามที่ต้องการและเกิดความครอบคลุมทางหน่วยเสียง ใช้ค่าความต้องการของรูปแบบหน่วยเสียงที่เกิดขึ้นในประโยคในการให้คะแนนเพื่อทำให้เกิดการกระจายตัวตามที่กำหนดได้จากรูปแบบหน่วยเสียงที่อยู่ในประโยคที่ถูกเลือก ทั้งนี้ยังสามารถกำหนดลักษณะของประโยคในคลังข้อความได้ คือ สามารถกำหนดจำนวนประโยคในคลังข้อความ กำหนดลักษณะการกระจายตัวทางหน่วยเสียง กำหนดรูปแบบหน่วยเสียง กำหนดขอบเขตความยาวของประโยค เพื่อให้ได้คลังข้อความตามที่ต้องการใช้ต้องการได้ สามารถลดเวลาในขั้นตอนการบันทึกเสียงเพื่อการสร้างคลังเสียงพูด และได้รูปแบบหน่วยเสียงที่ครอบคลุมและเกิดการกระจายตัวของหน่วยเสียงที่ดี สามารถคาดเดาคุณภาพของคลังข้อความได้

บทคัดย่อ

ประสิทธิภาพของ ระบบรู้จำเสียงพูดอัตโนมัติและ ระบบสังเคราะห์เสียงพูด ขึ้นอยู่กับ ความครอบคลุมของหน่วยเสียงจากคลังข้อความที่เหมาะสม งานวิจัยนี้เสนอการสร้างคลังข้อความ อัตโนมัติ จากการกระจายตัวของหน่วยเสียงตามที่กำหนดการกระจายตัวของหน่วยตามที่กำหนด นั้น สามารถกำหนดได้จากชนิดของหน่วยเสียง ขนาดของคลังข้อความ เกณฑ์ขั้นต่ำของจำนวน หน่วยเสียง และรูปแบบของการกระจายตัวเป้าหมาย ได้คัดเลือกข้อความมาจากข้อมูลจาก อินเทอร์เน็ต โดยข้อความนั้นจะถูก จัดเก็บอย่างต่อเนื่องโดย กระบวนการดึงบทความจากหน้า เว็บบนอินเทอร์เน็ต จนกระทั่งได้คลังข้อความที่เหมาะสม ในวิทยานิพนธ์นี้ยังได้ประยุกต์ใช้วิธีการ เชิงละโมบ เพื่อเลือกประโยคที่เหมาะสมที่จะทำให้เกิดการกระจายตัวของหน่วยเสียงตาม เป้าหมาย ในการทดลองได้ใช้ข้อความจากฐานข้อมูล Large Vocabulary Continuous Speech Recognition (LVCSR) corpus for Thai language ในการสร้างเป้าหมายของการกระจายตัว หน่วยเสียง ผลการทดลองที่ได้คือ จำนวนของข้อมูลข้อความที่ดึงมาจากอินเทอร์เน็ตที่เพิ่มขึ้น สามารถทำให้การกระจายตัวของหน่วยเสียงเป็นไปตามเป้าหมายได้ และเกิดความครอบคลุมทาง หน่วยเสียงคู่ ถึง 99.13% คลังข้อความที่ถูกสร้างขึ้นนี้ จึงสามารถนำไปใช้ในการสร้างคลัง เสียงพูดได้อย่างมีประสิทธิภาพ

ABSTRACT

Performance of Automatic Speech Recognition (ASR) and Text-to-Speech (TTS) systems depend on appropriate text corpus. This article explains about the automated text corpus generating method using custom phonetic distribution. This distribution is defined by phonemes type, corpus size, minimum criterion number of phonemes, and target phonetic distribution. Generally, the system selects text data from the internet by continuously downloading them using web crawler. The greedy algorithm is applied to extract the proper sentences, in order to fit with the target phonetic distribution until the appropriate text corpus is established. The experiment is done by using the text from Large Vocabulary Continuous Speech Recognition (LVCSR) corpus for Thai language to generate target phonetic distribution. The result shown that, the increased number of data drawn from the internet is able to accomplish target phonetic distribution and generate diphone coverage for 99.13%. This text corpus then generate speech corpus efficiently.



สารบัญ

	หน้า
บทสรุปผู้บริหาร	ก
บทคัดย่อภาษาไทย	ข
บทคัดย่อภาษาอังกฤษ	ค
สารบัญ	ง
สารบัญตาราง	ช
สารบัญภาพ	ซ
บทที่ 1	บทนำ
1.1	ความเป็นมาและความสำคัญของปัญหา
1.2	วัตถุประสงค์ของการวิจัย
1.3	ขอบเขตของการวิจัย
1.4	ลำดับขั้นตอนในการเสนอผลการวิจัย
1.5	ประโยชน์ที่คาดว่าจะได้รับ
บทที่ 2	แนวคิดทฤษฎีและงานวิจัยที่เกี่ยวข้อง
2.1	แนวคิดและทฤษฎี
2.1.1	ระบบการเขียนภาษาไทย
2.1.2	ระบบการออกเสียงภาษาไทย
2.1.3	รูปแบบหน่วยเสียง
2.1.4	การแปลงข้อความภาษาไทยเป็นสัทอักษร
2.1.5	การค้นหาสารสนเทศบนอินเทอร์เน็ต
2.1.6	โครงสร้างคำสั่งในภาษา HTML
2.1.7	ขั้นตอนวิธีเชิงละโมภ: Greedy Algorithm
2.1.8	การวัดระยะทางแบบยุคลิด: Euclidean distance
2.2	เอกสารและงานวิจัยที่เกี่ยวข้อง
บทที่ 3	ระเบียบวิธีวิจัย
3.1	เครื่องมือที่ใช้ในการวิจัยขั้นตอนการดำเนินงานวิจัย
3.2	ขั้นตอนการดำเนินงานวิจัย
3.2.1	ขั้นตอนการสร้างการกระจายตัวทางหน่วยเสียงเป้าหมาย

3.2.2	ขั้นตอนการสร้างประโยคออนไลน์	23
3.2.3	ขั้นตอนการเลือกประโยคและจัดเก็บคลังข้อความ	31
บทที่ 4	การทดลอง และอภิปรายผล	42
4.1	การทดลอง	42
4.1.1	การสร้างคลังข้อความจากรูปแบบการกระจายตัวทางหน่วยเสียงจากการกระจายตัวเป้าหมาย	42
4.1.2	การสร้างคลังข้อความโดยใช้การกำหนดรูปแบบการกระจายตัวทางหน่วยเสียงแบบกำหนดเอง	50
4.1.3	ทดสอบผลของการสร้างคลังข้อความจากการปรับพารามิเตอร์ค่าคงที่ความสำคัญรูปแบบหน่วยเสียงหายาก	52
4.1.4	เปรียบเทียบผลจากวิธีการให้คะแนนในวิทยานิพนธ์กับการให้คะแนนของงานวิจัย Automatic Construction for a TTS Corpus with Limited Text (W Zhang, 2010)	55
บทที่ 5	บทสรุปผลการวิจัย และข้อเสนอแนะ	59
5.1	สรุปผลการวิจัย	59
5.2	ข้อเสนอแนะ	61
บรรณานุกรม		62
ภาคผนวก		65
	ภาคผนวก ก ตารางเทียบสัทอักษรสากลกับหน่วยเสียงไทยที่ใช้	66
	ภาคผนวก ข ตัวอย่างประโยคที่ได้จากคลังข้อความที่สร้างขึ้น	69
ประวัติคณะผู้ทำวิจัย		84

สารบัญตาราง

	หน้า
ตารางที่ 2-1	การออกเสียงของพยัญชนะภาษาไทย
ตารางที่ 2-2	การออกเสียงของพยัญชนะควบกล้ำไทย
ตารางที่ 2-3	การออกเสียงของพยัญชนะหน่วยเสียงยืมภาษาต่างประเทศ
ตารางที่ 2-4	การออกเสียงสระไทย
ตารางที่ 2-5	ตัวอย่างรูปแบบหน่วยเสียง
ตารางที่ 2-6	คำสั่งหัวเรื่องโครงสร้าง HTML
ตารางที่ 3-1	ตัวอย่างการแปลงรูปเขียนเป็นรูปอ่านจากฐานข้อมูล LOTUS
ตารางที่ 3-2	ตัวอย่างจำนวนของแต่ละรูปแบบหน่วยเสียงคู่ที่เกิดขึ้นจากฐานข้อมูล LOTUS
ตารางที่ 3-3	ตัวอย่างผลลัพธ์จากการแปลงข้อความรูปแบบบทความเป็นรูปแบบลำดับพยางค์
ตารางที่ 3-4	ตัวอย่างผลการคัดแยกประโยคตามอักขระนิพจน์ปรกติที่ต้องการ
ตารางที่ 3-5	ตัวอย่างผลการแปลงรูปเขียนเป็นรูปอ่านของประโยคออนไลน์
ตารางที่ 3-6	ข้อมูลทางสถิติการกระจายตัวทางหน่วยเสียงที่ใช้ในการคำนวณจากฐานข้อมูล
ตารางที่ 3-7	ผลคะแนนประโยคจากขั้นตอนการให้คะแนน
ตารางที่ 4-1	คุณลักษณะของคลังข้อความที่สร้างขึ้น
ตารางที่ 4-2	รูปแบบหน่วยเสียงที่ไม่ปรากฏจากการแปลงรูปเขียนเป็นรูปอ่านที่ผิด
ตารางที่ 4-3	รูปแบบหน่วยเสียงที่ไม่ปรากฏจากข้อเฉพาะ
ตารางที่ 5-1	ผลของการเปลี่ยนแปลงค่าพารามิเตอร์

สารบัญภาพ

	หน้า
ภาพที่ 2-1	ลักษณะของเสียงวรรณยุกต์ในภาษาไทย
ภาพที่ 2-2	ระบบการแปลงข้อความเป็นสัทอักษรภาษาไทย
ภาพที่ 3-1	แผนภาพความสัมพันธ์ของเครื่องมือวิจัย
ภาพที่ 3-2	ภาพรวมการทำงานหลักของระบบ
ภาพที่ 3-3	การสร้างการกระจายตัวทางหน่วยเสียงเป้าหมาย
ภาพที่ 3-4	แผนภาพการกระจายตัวทางสถิติทางหน่วยเสียงของฐานข้อมูล LOTUS
ภาพที่ 3-5	การดึงข้อความออนไลน์ด้วยตัวดึงหน้าเว็บแยกวิเคราะห์ HTML
ภาพที่ 3-6	ตัวอย่างหน้าเว็บจากอินเทอร์เน็ตที่มีเนื้อความที่เหมาะสมกับการดึงข้อความ
ภาพที่ 3-7	ตัวอย่างข้อความออนไลน์จากอินเทอร์เน็ตที่เก็บไว้ในฐานข้อมูล
ภาพที่ 3-8	การแปลงข้อความในรูปแบบบทความเป็นรูปแบบประโยค
ภาพที่ 3-9	การแปลงประโยคจากรูปเขียนเป็นรูปอ่าน
ภาพที่ 3-10	ตัวอย่างประโยคออนไลน์ในรูปเขียนและรูปอ่านที่เก็บไว้ในฐานข้อมูล
ภาพที่ 3-11	หน้าต่างโปรแกรมขั้นตอนการสร้างประโยคออนไลน์
ภาพที่ 3-12	แผนภาพขั้นตอนการเลือกประโยคและจัดเก็บคลังข้อความ
ภาพที่ 3-13	ใช้ค่าความต้องการรูปแบบหน่วยเสียงในการให้คะแนนประโยค
ภาพที่ 4-1	แผนภาพการกระจายตัวทางสถิติทางหน่วยเสียงของฐานข้อมูล LOTUS
ภาพที่ 4-2	ผลของค่าเฉลี่ยผลต่างของการกระจายตัวทางหน่วยเสียงเป้าหมายเทียบกับจำนวนประโยคออนไลน์ที่เข้ามาใหม่
ภาพที่ 4-3	ความครอบคลุมทางหน่วยเสียงเทียบกับจำนวนประโยคออนไลน์ที่เข้ามาใหม่
ภาพที่ 4-4	แผนภาพการกระจายตัวของหน่วยหลังจากการคัดเลือกประโยค
ภาพที่ 4-5	แผนภาพการกระจายตัวของหน่วยหลังจากการคัดเลือกประโยคใช้การกระจายตัวหน่วยเสียงเป้าหมายกำหนดเอง
ภาพที่ 4-6	ความครอบคลุมทางหน่วยเสียงเมื่อกำหนดให้ค่า α เท่ากับ 00
ภาพที่ 4-7	ผลของค่าเฉลี่ยผลต่างของการกระจายตัวทางหน่วยเสียงเป้าหมายเมื่อกำหนดให้ค่า α เท่ากับ 00
ภาพที่ 4-8	ความครอบคลุมทางหน่วยเสียงเมื่อกำหนดให้ค่า α เท่ากับ 15
ภาพที่ 4-9	ผลของค่าเฉลี่ยผลต่างของการกระจายตัวทางหน่วยเสียงเป้าหมายเมื่อ

	กำหนดให้ค่า α เท่ากับ 15	54
ภาพที่ 4-10	ความครอบคลุมทางหน่วยเสียงเมื่อกำหนดให้ค่า α เท่ากับ 30	54
ภาพที่ 4-11	ผลของค่าเฉลี่ยผลต่างของกระจายตัวทางหน่วยเสียงเป้าหมายเมื่อ กำหนดให้ค่า α เท่ากับ 30	55
ภาพที่ 4-12	ความครอบคลุมทางหน่วยเสียงเมื่อเปรียบเทียบวิธีการเลือกประโยคของ วิทยานิพนธ์กับวิธีการเลือกประโยคแบบเก่า	56
ภาพที่ 4-13	ผลของค่าเฉลี่ยผลต่างของกระจายตัวทางหน่วยเสียงเป้าหมายเมื่อ เปรียบเทียบวิธีการเลือกประโยคของวิทยานิพนธ์กับวิธีการเลือกประโยค แบบเก่า	56
ภาพที่ 4-14	แผนภาพการกระจายตัวของหน่วยหลังจากการคัดเลือกประโยคของ วิทยานิพนธ์เปรียบเทียบกับวิธีของ Zhang	58



บทที่ 1

บทนำ

1.1 ความเป็นมาและความสำคัญของปัญหา

เทคโนโลยีการโต้ตอบกันระหว่างคอมพิวเตอร์กับมนุษย์กำลังเป็นที่สนใจของบรรดานักวิจัยและหน่วยงานต่าง ๆ ทำให้เกิดนวัตกรรมด้านเสียงพูดออกมามากขึ้น เช่น ระบบรู้จำเสียงพูดที่ใช้งานเพื่อใช้ในการโต้ตอบกันระหว่างมนุษย์กับหุ่นยนต์ [1] โปรแกรมช่วยอ่านหน้าจอคอมพิวเตอร์ [2] ระบบรายงานสภาพการจราจรด้วยเสียงอัตโนมัติ: Traffic Voice Information Services, TVIS [3] หรือการสังเคราะห์เสียงพูดบนอุปกรณ์พกพา [4] สิ่งที่เป็นสำหรับเทคโนโลยีการรู้จำเสียงพูดอัตโนมัติและการสังเคราะห์เสียงพูดนี้คือ “คลังข้อความ” ซึ่งประโยคจากคลังข้อความนี้จะถูกนำไปอ่านบันทึกเสียงโดยอาสาสมัครผู้ให้เสียง เพื่อสร้างระบบการรู้จำเสียงพูดอัตโนมัติหรือระบบสังเคราะห์เสียงพูดดังกล่าว อย่างไรก็ตามประสิทธิภาพระบบรู้จำเสียงพูดและคุณภาพของเสียงพูดสังเคราะห์ขึ้นอยู่กับจำนวนหน่วยเสียงที่ครอบคลุมในฐานข้อมูล [5] จึงจำเป็นต้องใช้เวลาในการบันทึกเสียงค่อนข้างมากเนื่องจากต้องการให้เกิดความครอบคลุมของหน่วยเสียงที่มากที่สุด ทั้งนี้คำศัพท์และสำนวนใหม่ ๆ เกิดขึ้นทุกวัน เราจึงต้องการคลังข้อความที่ทันสมัย เพื่อที่จะนำมาสังเคราะห์เสียงพูดหรือรู้จำเสียงพูดในปัจจุบันได้ การเลือกประโยคที่เหมาะสม สามารถลดเวลาในขั้นตอนการบันทึกเสียงเพื่อการสร้างคลังเสียงพูด และได้รูปแบบหน่วยเสียงที่ครอบคลุมและเกิดการกระจายตัวของหน่วยเสียงที่ดี สามารถคาดเดาคุณภาพของคลังข้อความได้ ทำให้เกิดงานวิจัยด้านการเลือกประโยค เพื่อสร้างคลังข้อความที่เหมาะสม

มีงานวิจัยที่ใช้การสร้างประโยค ด้วยวิธีการสุ่มประโยคเพื่อสร้างคลังข้อความ [6] คือ การเลือกประโยคจากหลาย ๆ บทความ มาบรรจุในคลังข้อความแบบสุ่ม เป็นวิธีที่นิยมให้กันอย่างแพร่หลายในสมัยก่อน ปัญหาที่เกิดขึ้นจากวิธีนี้คือ การขาดความหลากหลายของรูปแบบหน่วยเสียง และใช้เวลาในการอัดเสียงที่มากและไม่สามารถคาดการณ์คุณภาพของคลังเสียงได้ ส่งผลให้เกิดงานวิจัยด้านการเลือกประโยคเพื่อสร้างคลังข้อความ คำนึงถึงมีการกระจายตัวของหน่วยเสียงสมดุลโดยใช้ขั้นตอนเชิงละโมภ: Greedy algorithm [7] ในการเลือกประโยคที่เหมาะสม [8, 9,10] โดยการเลือกประโยคที่ประกอบด้วยหน่วยเสียงที่ทำให้เกิดความครอบคลุมทางหน่วยเสียงมากที่สุดจากคลังข้อความแม่ คลังข้อความแม่หมายถึงแหล่งข้อมูลขนาดใหญ่ที่ประกอบไปด้วยประโยคจำนวนมาก แต่ปัญหาของวิธีนี้คือ ไม่สามารถควบคุมการกระจายตัวของหน่วยได้ และด้วยขนาดของข้อมูลที่ใช้ มีขนาดคงที่ ทำให้ประสิทธิภาพของการเลือกประโยคถูกจำกัด จึงมีบางงานวิจัยที่ใช้ข้อมูลอินเทอร์เน็ตมาสร้างคลังข้อความ [11] แต่งานวิจัยนี้ได้เก็บ

ข้อมูลจากอินเทอร์เน็ตมาจำนวนหนึ่งโดยไม่มีการเก็บเพิ่มอีก ทำให้การเลือกประโยคถูกจำกัดด้วยจำนวนข้อมูล

ในงานวิจัยนี้เราจึงเสนอ การสร้างคลังข้อความจากการกระจายตัวของหน่วยเสียงที่กำหนด และใช้วิธีคัดเลือกประโยคจากอินเทอร์เน็ตอย่างต่อเนื่อง การกระจายตัวทางหน่วยเสียงที่กำหนด ในที่นี้หมายถึง คลังข้อความที่ได้ มีจำนวนรูปแบบของแต่ละหน่วยเสียง เป็นไปตามอัตราส่วนเป้าหมาย และยังสามารถกำหนดเกณฑ์ของจำนวนขั้นต่ำของหน่วยเสียง การคัดเลือกประโยคจะใช้วิธีการประยุกต์ขั้นตอนเชิงละโมบ (Greedy Algorithm) คือ ต้องการประโยคที่ทำให้การกระจายตัวทางหน่วยเสียงเป็นไปตามที่ต้องการและเกิดความครอบคลุมทางหน่วยเสียงด้วย ใช้ค่าความต้องการของรูปแบบหน่วยเสียงที่เกิดขึ้นในประโยคในการให้คะแนน เพื่อให้ทำให้เกิดการกระจายตัวตามที่กำหนดได้จากรูปแบบหน่วยเสียงที่อยู่ในประโยคที่ถูกเลือก ทั้งนี้ยังสามารถกำหนดลักษณะของประโยคในคลังข้อความได้ คือ สามารถกำหนดจำนวนประโยคในคลังข้อความ กำหนดลักษณะการกระจายตัวทางหน่วยเสียง กำหนดรูปแบบหน่วยเสียง กำหนดขอบเขตความยาวของประโยค เพื่อให้ได้คลังข้อความตามที่ต้องการใช้ต้องการได้

1.2 วัตถุประสงค์ของการวิจัย

เพื่อเสนอวิธีในการสร้างคลังข้อความ จากการกระจายตัวรูปแบบหน่วยเสียงที่กำหนด เพื่อได้คลังข้อความที่มีค่าการกระจายตัวทางสถิติทางหน่วยเสียงตามที่ต้องการ สร้างประโยคออนไลน์จากบทความอินเทอร์เน็ตเพื่อความทันสมัยของประโยค ใช้วิธีการให้คะแนนและเลือกประโยคที่มีค่าความต้องการรูปแบบหน่วยเสียงมาก เพื่อให้ลักษณะการกระจายตัวทางหน่วยเสียงและความครอบคลุมหน่วยเสียงเป็นไปตามที่กำหนดได้

1.3 ขอบเขตของการวิจัย

1. ในการทดลองประโยคที่คัดเลือกเป็นมาต้องเป็นประโยคที่ประกอบด้วยตัวอักษรภาษาไทยเท่านั้น
2. ใช้รูปแบบหน่วยเสียงชนิด หน่วยเสียงคู่ ในการทดลองสร้างฐานข้อมูลข้อความ
3. ในการทดลองใช้ตัวแปลงรูปเขียนเป็นรูปอ่านไทยแบบ Probabilistic Generalized LR parser [12]
4. ในการทดลองใช้ชุดประโยคจากคลังข้อมูลข้อความ ฐานข้อมูล Large Vocabulary Continuous Speech Recognition: LVCSR [13] เลือกใช้ชุดประโยคที่ครอบคลุมภาษาไทย 5,000 คำ ในการสร้างการกระจายตัวทางหน่วยเสียงเป้าหมาย

1.4 ลำดับขั้นตอนในการเสนอผลการวิจัย

1. ขั้นตอนการศึกษาเบื้องต้น
 - 1.1. ศึกษางานวิจัยเกี่ยวกับการสร้างฐานข้อมูลเสียง
 - 1.2. ศึกษางานวิจัยเกี่ยวกับการหาค่าการกระจายตัวของหน่วยเสียง
 - 1.3. ศึกษาอัลกอริทึมที่ใช้ในการคัดเลือกประโยค
 - 1.4. ศึกษาการแปลงข้อความเป็นรูปอ่าน
 - 1.5. ศึกษากระบวนการจัดเก็บและการสืบค้นข้อมูลทางอินเทอร์เน็ต
2. ขั้นตอนการออกแบบระบบ
 - 2.1. ออกแบบการทำงานของระบบ
 - 2.2. เลือกใช้เครื่องมือในการวิจัยที่เหมาะสม
 - 2.3. ออกแบบการทดลอง
3. ขั้นตอนการทดลอง
 - 3.1. ทดสอบการทำงานของระบบ
 - 3.2. บันทึกผลการทดลอง
 - 3.3. วิเคราะห์ผลการทดลอง
4. สรุปผลและเรียบเรียงงานวิจัย

1.5 ประโยชน์ที่คาดว่าจะได้รับ

1. ได้ระบบการสร้างคลังข้อความอัตโนมัติตามรูปแบบการกระจายตัวของหน่วยเสียงที่กำหนดได้
2. ประโยชน์ที่ได้มีใจความต่อเนื่องและทันสมัยเหมาะกับการนำไปใช้
3. คาดเดาคูณภาพคลังข้อความได้
4. ช่วยลดเวลาการบันทึกเสียง เมื่อใช้ประโยคจากคลังข้อความ
5. สามารถนำระบบนี้ไปประยุกต์ใช้กับสร้างฐานข้อมูลข้อความในภาษาอื่นได้

บทที่ 2

แนวคิดทฤษฎีและงานวิจัยที่เกี่ยวข้อง

2.1 แนวคิดและทฤษฎี

ในงานวิจัยนี้ทฤษฎีที่เกี่ยวข้องแบ่งออกเป็น ระบบการเขียนภาษาไทย, ระบบการออกเสียงภาษาไทย, การแปลงข้อความภาษาไทยเป็นสัทอักษร, การค้นหาสารสนเทศบนอินเทอร์เน็ต, ขั้นตอนวิธีเชิงละโมบ: Greedy algorithm และการวัดระยะทางแบบยุคลิด: Euclidean distance

2.1.1 ระบบการเขียนภาษาไทย

ระบบการเขียนของภาษาไทยเป็นภาษาที่มีระดับของเสียง [14] พยัญชนะไทยมี 44 ตัว ได้แก่ ก ข ช ค ต พ ง จ ฉ ช ฌ ญ ฎ ฏ ฐ ท 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 มีสระพื้นฐาน 15 ตัว ได้แก่ -ะ -า -ิ -ี -ึ -ู -เ -แ -โอ -ไ -อ -ำ -อ มีเครื่องหมายบอกระดับเสียงหรือวรรณยุกต์ 4 ตัว (่, ้, ๊, ๋) ข้อความถูกเขียนจากทางด้านซ้ายไปขวาตามแนวนอน ไม่มีการเว้นวรรคแต่ละคำ สระสามารถอยู่ด้านบน, ด้านล่าง, ด้านซ้ายหรือด้านขวาของพยัญชนะ หลักไวยากรณ์ของภาษาไทยเมื่อเทียบกับภาษาของทางโลกตะวันตกมีความง่ายกว่ามาก ในแต่ละประโยคประกอบด้วย “ประธาน+กริยา+กรรม” ไม่มีคำนำหน้านาม ไม่ต้องเปลี่ยนรูปกริยาจากคำบอกเวลาหรือจำนวน ไม่มีเครื่องหมายจบประโยค ไม่มีการเว้นวรรคระหว่างคำ เช่น “น้ำหอมอบอวล” ประโยคนี้สามารถตัดคำได้อย่างหลากหลาย “น้ำ-หอ-มอ-บอ-วล” หรือ “น้ำ-หอม-อบ-อวน” ซึ่งการแก้ปัญหานี้ส่วนใหญ่จะการใช้การดูความหมายจากประโยครอบข้าง “กลิ่นน้ำหอมอบอวนเต็มห้อง” เป็นต้น [15] งานวิจัยนี้ใช้คำในภาษาไทยในการทดลองจึงเลือกใช้รูปแบบการเขียนภาษาไทยดังกล่าวเป็นมาตรฐานการเขียนประโยคภาษาไทย

2.1.2 ระบบการออกเสียงในภาษาไทย

การออกเสียงแต่ละพยางค์ในภาษาไทย [14] โดยทั่วไปแล้วมี 4 ส่วนคือ เสียงพยัญชนะต้น เสียงสระ เสียงตัวสะกด และเสียงวรรณยุกต์ พยัญชนะนำหน้าที่แทนด้วย “C” คือพยัญชนะเดี่ยว พยัญชนะนำหน้าที่แทนด้วย “CC” แทนด้วยพยัญชนะควบกล้ำ “V” แทนด้วยสระ, พยัญชนะท้ายที่แทนด้วย “Cf” คือตัวสะกดและตัวเลข (0-4) แทนด้วยเสียงวรรณยุกต์

โครงสร้างของพยางค์มี 4 รูปแบบคือ CV เช่น ปา [p-aa-0], รี [r-ii-0], CCV เช่น ปลา [pl-a-0], ครุ [khr-uu-0], CVCf เช่น ปาด [p-aa-t^h-1], กาก [k-aa-k^h-1], CCVCf เช่น ปราบ [pr-aa-p^h-1], กวาด [kw-aa-t^h-1] พยัญชนะต้นในภาษาไทยมี 44 ตัว 21 เสียง มีสามระดับเสียงคือ อักษรสูง, อักษรกลางและอักษรต่ำ พยัญชนะต้นมีทั้งพยัญชนะเดี่ยว เช่น (ก, บ, ผ) พยัญชนะควบกล้ำ (ปล-, คว-, ตร-) และบางตัวยืมมาจากต่างประเทศ (ดร-, พร-, พร-) สามารถดูการออกเสียงของพยัญชนะไทยได้ดังตารางที่ 2-1 พยัญชนะควบกล้ำ ตารางที่ 2-2 และ หน่วยเสียงยืมภาษาต่างประเทศ ดังตารางที่ 2-3

ตารางที่ 2-1 การออกเสียงของพยัญชนะภาษาไทย

ตัวอักษรไทย	หน่วยเสียงพยัญชนะต้น	หน่วยเสียงตัวสะกด
ก	k	k ^h
ข ช ค ฅ ฆ	kh	-
ง	ng	ng ^h
จ	c	-
ฉ ช ฌ	ch	-
ซ ศ ษ ส	s	-
ญ ย	j	j ^h
ฎ ด	d	t ^h
ฏ ต	t	-
ฐ ท ฒ ถ ท ธ	th	-
ณ น	n	n ^h
บ	b	p ^h
ป	p	-
พ ภ ผ	ph	-
ฟ ฝ	f	-
ม	m	m ^h
ร	r	-
ล ฬ	l	-
ว	w	w ^h
ห ฮ	h	-
อ	z	-

ตารางที่ 2-2 การออกเสียงของพยัญชนะควบกล้ำไทย

ตัวอักษรไทย	หน่วยเสียงพยัญชนะต้น	หน่วยเสียงตัวสะกด
ปร-	pr	-
ปล-	pl	-
ตร-	tr	-
กร-	kr	-
กล-	kl	-
กว-	kw	-
พร-, ผล-	phr	-
ทร	phl	-
คร-, ขร-	chr	-
คล-, ขล-	chl	-
คว	khw	-

ตารางที่ 2-3 การออกเสียงของพยัญชนะหน่วยเสียงยืมภาษาต่างประเทศ

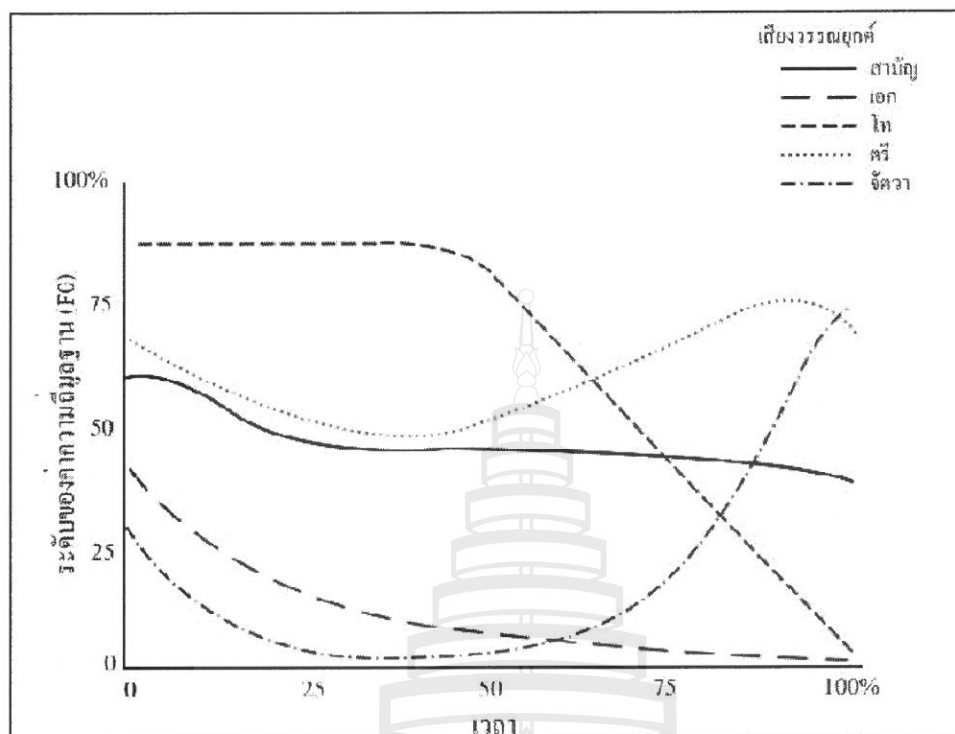
ตัวอักษรไทย	หน่วยเสียงพยัญชนะต้น	หน่วยเสียงตัวสะกด
бр-	br	-
бл-	bl	-
фр-	fr	-
פל-	fl	-
др-	dr	-
-ฟ	-	f^
-ล	-	l^
-ส	-	s^
-ช	-	ch^

สระในภาษาไทยมีสองชนิดคือ สระเดี่ยวและสระผสม สระเดี่ยวนั้นมี 9 ตัวที่เป็นเสียงยาว เช่น -า/a/; /, - ู /-ยู, /, แ- /x / และอีก 9 ตัวเป็นสระเสียงสั้นเช่น -ะ/a/, - ุ /u/, แ-ะ/x/; สระผสมมีทั้งหมด 3 ตัวที่ออกเสียงยาว คือ - ี๊ย / i; a/ , - ี๊อ / v; a/ , - ัว / u; และอีก 3 ตัวที่ออกเสียงสั้นคือ - ี๊ยะ / ia/ , - ี๊อะ / va/ , - ัว / ua/ สามารถดูการออกเสียงของสระจากตารางที่ 2-4 และสามารถเทียบวิธีการอ่านหน่วยเสียงทั้งหมดกับสัทอักษรสากล จากภาคผนวก ก

ตารางที่ 2-4 การออกเสียงสระไทย

สระเดี่ยว				สระผสม							
เสียงสั้น		เสียงยาว		เสียงสั้น		เสียงยาว					
รูปสระ	หน่วยเสียง	รูปสระ	หน่วยเสียง	รูปสระ	หน่วยเสียง	รูปสระ	หน่วยเสียง				
ะ	a	า	aa	เียะ	ia	เีย	iaa				
ิ	i	ี	ii	เือะ	va	เือ	vva				
ึ	u	ู	uu	ัวะ	ua	ัว	uuv				
ุ	u	ู	uu								
เ-ะ	e	เ-	ee								
แ-ะ	x	แ-	xx								
โ-ะ	o	โ-	oo								
เ-าะ	@	-อ	@@								
เ-อะ	q	เ-อ	qq								
สระผสม											
เสียงสั้น		เสียงยาว									
รูปสระ	หน่วยเสียง	รูปสระ	หน่วยเสียง								
ำ	am	-	-								
ไ-, ไ-	ai	-	-								
เ-า	aw	-	-								

เสียงวรรณยุกต์ในภาษาไทยแบ่งออกเป็น 5 เสียงแสดงในลำดับตัวเลข 0-4 ได้แก่ เสียงสามัญ (Mid Tone: 0), เสียงเอก (Low Tone: 1), เสียงโท (Falling tone: 2), เสียงตรี (High Tone, 3) และเสียงจัตวา (Rising Tone: 4) ซึ่งแต่ละเสียงมีลักษณะเด่นดังแสดงในภาพที่ 2-1



ภาพที่ 2-1 ลักษณะของเสียงวรรณยุกต์ในภาษาไทย [7]

ในงานวิจัยนี้เลือกใช้ระบบการออกเสียงภาษาไทยดังกล่าวเป็นมาตรฐานในการแปลงรูปเขียนเป็นรูปอ่าน ผู้ที่ต้องการนำงานวิจัยนี้ไปประยุกต์ใช้กับหน่วยเสียงภาษาอื่นก็ทำได้เช่นกัน

2.1.3 รูปแบบหน่วยเสียง

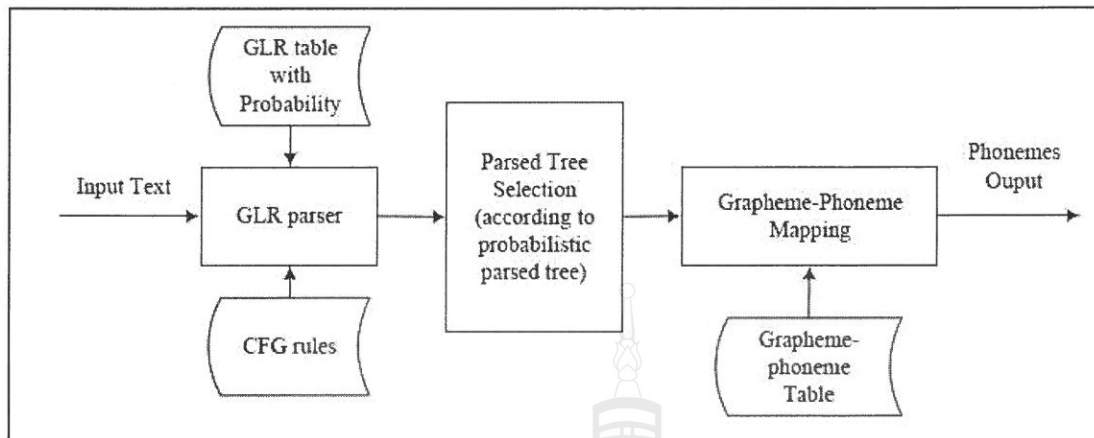
หน่วยเสียงคือหน่วยที่เล็กที่สุดของการออกเสียงเสียงในแต่ละภาษา รวมถึงหน่วยเสียงเงียบ "silence" ที่แสดงถึงการที่ไม่ได้เปล่งเสียงใด ๆ ในช่วงเวลานั้น รูปแบบหน่วยเสียงโดยทั่วไป ที่มีลักษณะต่างกันไปตามการใช้งาน ตัวอย่างรูปแบบหน่วยเสียงชนิดต่าง ๆ แสดงดังตารางที่ 2.5 หน่วยเสียงเดี่ยว ประกอบไปด้วยหน่วยที่เล็กที่สุดของหน่วยเสียง 1 หน่วย หน่วยเสียงคู่ ประกอบไปด้วยลำดับของหน่วยเสียงเดี่ยว 2 หน่วยเสียง หน่วยเสียงสาม ประกอบไปด้วยลำดับของหน่วยเสียงเดี่ยว 3 หน่วยเสียง และหน่วยเสียงที่มากกว่า 3 หน่วยเสียงขึ้นไป ซึ่งทุกรูปแบบอาจมีรูปแบบเสียงวรรณยุกต์กำกับด้วย ในการทดลองของงานวิจัยนี้ ได้ใช้ "หน่วยเสียงคู่" แบบไม่มีวรรณยุกต์ เป็นรูปแบบหน่วยเสียง

ตารางที่ 2-5 ตัวอย่างรูปแบบหน่วยเสียง

รูปแบบหน่วยเสียง	ตัวอย่างคำ	ลำดับหน่วยเสียง
หน่วยเสียงเดี่ยว	ฮัลโหล	sil, h, e, l [^] , l, o:, sil
หน่วยเสียงคู่		sil+h, h+e, e+l [^] , l [^] +l, l+o:, o:+sil
หน่วยเสียงสาม		sil-h+e, h-e+l [^] , e-l [^] +l, l [^] -l+o:, l-o:+sil...
หน่วยเสียงห้า		sil_sil-h+e=l [^] , sil_h-e+l [^] =l, h_e-l [^] =l...
หน่วยเสียงที่มี วรรณยุกต์กำกับ	สวัสดี	sil, s1, a1, w1, a1, ^t1, d0, i:0, sil

2.1.4 การแปลงข้อความภาษาไทยเป็นสัทอักษร

การแปลงรูปเขียนเป็นสัทอักษร คือ การแปลง ลำดับรูปเขียนของคำใด ๆ เป็นรูปแบบคำอ่าน เป็นหนึ่งในสิ่งสำคัญในการใช้กับการพัฒนา การสังเคราะห์ มีหลายวิธีในการแปลงแปลงรูปเขียนเป็นสัทอักษร เช่น การใช้พจนานุกรม (Dictionary-based) เป็นการแปลงรูปเขียนเป็นสัทอักษรโดยการดูคำอ่านจากพจนานุกรม วิธีนี้ต้องการ คลังคำศัพท์ที่ใหญ่และอาจเกิดปัญหาในการหาคำศัพท์ไม่เจอได้อีกด้วย การใช้กฎ (Rule-based) เป็นการแปลงรูปเขียนเป็นสัทอักษรโดยการจัดเตรียม โครงสร้างของกลุ่มพยางค์ นำไปใช้กับเครื่องสถานะจำกัด (Finite state machine) แต่วิธีนี้อาจเกิดข้อผิดพลาดได้เมื่อเจอการออกเสียงที่คลุมเครือของภาษาไทยได้ การใช้ต้นไม้ตัดสินใจ (Decision-tree) เป็นการแปลงรูปเขียนเป็นสัทอักษรโดยใช้ต้นไม้ตัดสินใจ วิธีนี้ยังคงเกิดปัญหาในการแปลงการอ่านตามขอบเขตของพยางค์อยู่ การใช้ PGLR (Probabilistic Generalized LR) [12] เป็นเทคนิคการนำค่าความน่าจะเป็นมารวมกับต้นไม้ตัดสินใจของประโยคเป็นวิธีการที่ได้ผลการแปลงแม่นยำที่สุด เมื่อเทียบกับวิธีการใช้พจนานุกรมการใช้กฎหรือการใช้ต้นไม้ตัดสินใจ หลักการทำงานแสดงดังภาพที่ 2-2



ภาพที่ 2-2 ระบบการแปลงข้อความเป็นสัทอักษรภาษาไทย [10]

ในงานวิจัยนี้จึงนำกระบวนการแปลงข้อความภาษาไทยเป็นสัทอักษรดังกล่าว มาใช้ในการแปลงรูปเขียนเป็นรูปอ่าน ทั้งในขั้นตอนการสร้างการกระจายตัวเป้าหมายและขั้นตอนการให้คะแนนประโยคออนไลน์

2.1.5 การค้นหาสารสนเทศบนอินเทอร์เน็ต

ข้อมูลสารสนเทศบนอินเทอร์เน็ตเป็นแหล่งข้อมูลทางอิเล็กทรอนิกส์ที่สำคัญและใหญ่ที่สุดที่มีการเปลี่ยนแปลงอยู่ตลอดเวลา [11] ดังนั้น ในการสืบค้นข้อมูลสารสนเทศบนอินเทอร์เน็ต ควรกำหนด วัตถุประสงค์ในการสืบค้น ให้ชัดเจนกำหนด Search engine ที่เหมาะสมกำหนด ช่วงเวลาที่ต้องการค้นหา จะทำให้การสืบค้นใช้เวลา น้อย นอกจากนั้นเราควรระบุประเภทข้อมูล สารสนเทศที่ต้องการสืบค้น เช่น ข้อมูลที่เป็นข้อความ ภาพวาด ภาพเขียนหรือภาพลายเส้น ภาพ ไดอะแกรม ภาพถ่าย เสียง เสียงสังเคราะห์ เช่น เสียงดนตรี ภาพยนตร์ ภาพเคลื่อนไหวอะนิเมชัน เครื่องมือหรือโปรแกรม สำหรับการสืบค้น (Search Engine) มีอยู่มากมายและมีให้บริการ อยู่ตามเว็บไซต์ต่างๆ ที่ให้บริการการสืบค้นข้อมูลโดยเฉพาะ การเลือกใช้นั้นขึ้นกับประเภทของข้อมูล สารสนเทศที่ต้องการสืบค้น Search Engine ต่างๆ จะให้ข้อมูลที่มีความลึกในแง่มุมหรือศาสตร์ ต่างๆ ไม่เท่ากัน

ข้อดีของการค้นหาข้อมูลจากอินเทอร์เน็ตคือ ขอบเขตของข้อมูลสารสนเทศกว้างขวาง มีความหลากหลาย ไร้พรมแดน ข้อมูลสารสนเทศที่สืบค้นได้มีความทันสมัย เนื่องจากผู้สร้างข้อมูล สามารถแก้ไข ปรับปรุงได้ง่ายและตลอดเวลา สะดวก ไม่มีข้อจำกัดในแง่ของเวลาและสถานที่ สามารถสืบค้นเวลาใดก็ได้ที่ใดก็ได้ สามารถสืบค้นได้ง่ายและรวดเร็วโดยอาศัย Search Engine การได้มาซึ่งข้อมูลผ่านอินเทอร์เน็ตใช้เวลาสั้นมาก เมื่อเทียบกับการส่งเอกสารวิธีอื่น ๆ การได้มา

ซึ่งข้อมูลนั้น ประหยัดทั้งเวลาและทรัพยากร จัดเป็นห้องสมุดที่ใหญ่ที่สุดในโลก ข้อมูลสารสนเทศที่สืบค้นมา มีประโยชน์มาก สามารถนำไปจัดหมวดหมู่ ทำฐานข้อมูล บรรณธิการ และจัดการต่อได้โดยง่าย และมีความทันสมัย งานวิจัยนี้จึงใช้ การค้นหาข้อมูลทางอินเทอร์เน็ตในการค้นหา ประโยคที่จะนำมาพิจารณาในการเลือกสุ่มคลังข้อความ

2.1.6 โครงสร้างคำสั่งในภาษา HTML

โครงสร้างคำสั่งในภาษา HTML [17] จะมีรูปแบบโครงสร้างการเขียนแบ่งออกเป็น 3 ส่วน คือ ส่วนประกาศ เป็นส่วนที่กำหนดให้เบราว์เซอร์ทราบว่า นี่คือภาษาเอชทีเอ็มแอล และจะต้องทำการแปลงอย่างไรมีคำสั่งคู่เดียวคือ `<html>` และ `</html>` ปกติที่หัวและท้ายไฟล์ ส่วนหัวเรื่อง (head) เป็นส่วนที่แสดงผลข้อความบนไตเติ้ลบาร์ของเบราว์เซอร์ และอาจมีคำสั่งสำหรับกำหนดรายละเอียดด้านเทคนิคอื่น ๆ อีก แทรกอยู่ระหว่างคำสั่ง `<head>` และ `</head>` และส่วนเนื้อหา (body) เป็นส่วนที่มีความซับซ้อนมากที่สุด และสามารถใส่เทคนิคลูกเล่นเพื่อดึงดูดความสนใจจากผู้ชมได้มาก ความแตกต่างระหว่างเว็บไซต์ต่างๆ แสดงความมีฝีมือของผู้จัดทำ ศิลปะในการออกแบบจะอยู่ในส่วนนี้ทั้งหมด ซึ่งจะแทรกอยู่ระหว่างคำสั่ง `<body>` และ `</body>`

โครงสร้าง พื้นฐานของภาษา Computer เป็นส่วนที่สำคัญที่สุด ของการเขียนภาษา Computer โดยทั่วไปแล้ว มันจะต้องถูกเขียนขึ้นทุกครั้ง ภาษา HTML ก็เหมือนกับภาษา Computer ทั่วไป ที่มี โครงสร้าง พื้นฐานเฉพาะ ของมันคำสั่งของ HTML ส่วนมากจะถูกกำหนด อยู่ในเครื่องหมาย `<` และ `>` ซึ่งถูกเรียกว่า Tag สำหรับในส่วนของคำสั่ง Tag ภายในคำสั่ง ซึ่ง Tag แบ่งออกเป็น 2 ประเภท คือ แท็กเดี่ยว และแท็กคู่ สำหรับคำสั่งหัวเรื่อง เป็น คำสั่งชื่อโฮมเพจ และข้อความอธิบายข้อมูลที่เกี่ยวข้องกับเว็บเซิร์ฟเวอร์ที่เป็นเจ้าของ โดยชื่อโปรแกรมดังกล่าวจะไปปรากฏบนเมนูของโปรแกรมเบราว์เซอร์ในขณะที่โปรแกรมได้รับการเชื่อมโยงแบบไฮเปอร์เท็กซ์ ดังนั้น หัวเรื่องจึงหมายถึงเอกลักษณ์ประจำโฮมเพจ เพราะเนื่องจากโปรแกรม HTML เป็น โปรแกรมของโฮมเพจ สามารถดูคำสั่งหัวเรื่องจากตารางที่ 2-6

ตารางที่ 2-6 คำสั่งหัวเรื่องโครงสร้าง HTML

คำสั่งหัวเรื่อง	คำอธิบาย
<code><TITLE>...</TITLE></code>	เพื่อแสดงไฟล์เอกสาร HTML หรือ ชื่อ HOMEPAGE
<code><ISINDEX></code>	เพื่อแสดงไฟล์เอกสาร HTML ชนิดที่สืบค้นได้
<code><NEXTID></code>	เพื่อแสดงเลขประจำตัวของไฟล์เอกสาร HTML
<code><LINK></code>	เพื่อกำหนดความสัมพันธ์ระหว่างไฟล์เอกสารนี้กับไฟล์เอกสารอื่นที่เกี่ยวข้อง
<code><BASE></code>	เพื่ออ้างอิงรหัสสืบค้น URL
<code><META></code>	เพื่อแสดงข่าวสารของไฟล์เอกสาร HTML

สำหรับงานวิจัยนี้ได้ใช้โครงสร้าง HTML ในการเข้าถึงเนื้อข้อมูลในรูปแบบบทความของหน้าเว็บต่าง ๆ และยังใช้ในการเชื่อมโยงไปยังหน้าถัดไปภายในเว็บไซต์นั้น ในขั้นตอนการสร้างประโยคออนไลน์

2.1.7 ขั้นตอนวิธีเชิงละโมบ: Greedy Algorithm

เป็นขั้นตอนวิธีการแก้ปัญหาที่คิดแบบง่าย ๆ และตรงไปตรงมา โดยพิจารณาว่าข้อมูลที่มีอยู่ในขณะนั้นมีทางเลือกใดที่ ให้ผลตอบแทนคัมที่สูงสุด ขั้นตอนวิธีจะหาทางเลือกที่ดีที่สุด ในขณะที่ถ้าข้อมูลนั้นพอเพียงที่จะทำให้สรุปคำตอบที่ดีที่สุด เราจะได้ขั้นตอนวิธีที่มีประสิทธิภาพ โดยทั่วไปเราจะใช้ Greedy algorithm กับปัญหาเหมาะสมที่สุด Optimization problem เพราะว่าเราต้องการการตัดสินใจว่าทางเลือกในปัจจุบันมีค่าตอบแทนมากที่สุดหรือน้อยที่สุดหรือไม่ [7]

ในงานวิจัยนี้ ขั้นตอนวิธีเชิงละโมบ (Greedy Algorithm) ถูกนำมาใช้ในขั้นตอนการให้คะแนนของแต่ละประโยคเพื่อให้ได้ประโยคที่มีคะแนนสูงและมีการกระจายตัวในแต่ละประโยคที่เหมาะสม

2.1.8 การวัดระยะทางแบบยูคลิด: Euclidean distance

ระยะทางแบบยูคลิด (Euclidean distance, Euclidean metric) [17] คือการวัดระยะทางปกติระหว่างจุดสองจุดในแนวเส้นตรง นิยามของระยะทางแบบยูคลิดคือ

ระยะทางแบบยูคลิดระหว่างจุดสองจุด p และ q คือความยาวของส่วนของเส้นตรง pq ถ้า $p = (p_1, p_2, \dots, p_n)$ และ $q = (q_1, q_2, \dots, q_n)$ ในระบบพิกัดคาร์ทีเซียน เป็นจุดสองจุดบนปริภูมิยูคลิด n มิติ ระยะทางระหว่างจุด p กับ q คำนวณได้จากสมการที่ 1

$$d(p, q) = \sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2 + \dots + (p_n - q_n)^2} = \sqrt{\sum_{i=1}^n (p_i - q_i)^2} \quad (1)$$

ค่าประจำแบบยูคลิด คือระยะทางจากจุดหนึ่งจุด p ไปยังจุดกำเนิด $(0, 0, \dots, 0)$ บนปริภูมิยูคลิด ได้ ดังสมการที่ 2

$$\|p\| = \sqrt{p_1^2 + p_2^2 + \dots + p_n^2} = \sqrt{p \cdot p} \quad (2)$$

ซึ่งสมการตัวหลังเกี่ยวข้องกับผลคูณจุด เป็นขนาดของเวกเตอร์ p จากจุดกำเนิด ระยะทางแบบยูคลิดจึงอาจนิยามได้อีกแบบหนึ่ง ดังสมการที่ 3

$$\|p - q\| = \sqrt{(p - q) \cdot (p - q)} = \sqrt{\|p\|^2 + \|q\|^2 - 2p \cdot q} \quad (3)$$

กรณีพิเศษในหนึ่งมิติ ระยะทางระหว่างจุดสองจุดบนเส้นจำนวนจริงคือค่าสัมบูรณ์ของผลต่างของสองค่านั้น ดังนั้นถ้าให้ p และ q เป็นจุดสองจุด (หรือจำนวนสองจำนวน) บนเส้นจำนวนจริงแล้ว ระยะทางระหว่าง p และ q คำนวณได้ ดังสมการที่ 4

$$d(p, q) = \sqrt{(p - q)^2} = |p - q| \quad (4)$$

ในสองมิติแบบยูคลิด ถ้า $p = (p_1, p_2)$ และ $q = (q_1, q_2)$ แล้ว ระยะทางระหว่าง p และ q สามารถคำนวณได้ดังนี้ ซึ่งมีสูตรเหมือนกับทฤษฎีบทพีทาโกรัส ดังสมการที่ 5

$$d(p, q) = \sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2} \quad (5)$$

จากนิยามแบบที่สองของระยะทางแบบยูคลิด ถ้าหาก $p = (r_1, \theta_1)$ และ $q = (r_2, \theta_2)$ ในระบบพิกัดเชิงขั้ว จะสามารถคำนวณระยะทางได้ ดังสมการที่ 6

$$\|p - q\| = \sqrt{r_1^2 + r_2^2 - 2r_1 r_2 \cos(\theta_1 - \theta_2)} \quad (6)$$

ในสามมิติแบบยูคลิด ระยะทางระหว่าง p และ q ได้ ดังสมการที่ 7

$$d(p, q) = \sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2 + (p_3 - q_3)^2} \quad (7)$$

ในงานวิจัยนี้เลือกใช้การวัดระยะทางแบบยูคลิดแบบการเปรียบเทียบค่าสัมบูรณ์ ของผลต่างของแต่ละจำนวนหน่วยเสียง จากสมการที่ 4 เพื่อใช้ในการหาค่าเบี่ยงเบนเฉลี่ยร้อยละจากเป้าหมาย (Percent average deviation of target) ในการเปรียบเทียบความแตกต่างกันระหว่างค่าการกระจายตัวทางหน่วยเสียงกำหนดกับค่าการกระจายตัวทางหน่วยเสียงปัจจุบัน

2.2 เอกสารและงานวิจัยที่เกี่ยวข้อง

Methods of Sentences Selection for Read-Speech Corpus Design (V. Radova, 1999) [18] งานวิจัยนี้เสนอวิธีการเลือกข้อความเพื่อนำไปสร้างระบบรู้จำเสียงพูดในภาษาเช็ก ใช้ประโยคจากหนังสือพิมพ์รายวันภาษาเช็ก จำนวนทั้งสิ้น 24,442 ประโยค มีการคัดเลือกประโยคเบื้องต้นที่ต้องประกอบด้วยคำมากกว่า 3 คำแต่ห้ามเกิน 15 คำ เพื่อความไม่เหลื่อมกันของ

คะแนน ไม่มีตัวเลข สัญลักษณ์ หรือภาษาอื่น ๆ ปนอยู่ ให้คะแนนแต่ละประโยคจากการกระจายตัวทางสถิติของหน่วย Tri-Phone คัดเหลือ 40 เหลือประโยคผลลัพธ์ ได้การครอบคลุมทางหน่วยเสียง Tri-Phone 73.66% จากงานวิจัยนี้จึงเกิดแนวคิดว่า การคัดเลือกประโยคที่มีการกระจายตัวทางหน่วยเสียงครอบคลุมจะทำให้ลดเวลาในการอัดเสียงได้

LOTUS: Phonetically Distributed Continuous Speech Corpus for Thai Language (C. Wutiwiwatchai, 2002) [13] มีการเลือกข้อความเพื่อสร้างฐานข้อมูลเสียงพูดต่อเนื่องภาษาไทยครอบคลุมคำศัพท์จำนวน 5,000 คำ ใช้หลักการออกแบบฐานข้อมูลเสียงเดียวกับฐานข้อมูล JNAS [22] และ WSJAM0 [24] การสร้างฐานข้อมูลเริ่มจากเลือกข้อความจากประโยคในฐานข้อมูลบทความ ORCHID (V. Sornlertlamvanich, 1998) [20] ซึ่งประกอบไปด้วย 27,634 ประโยคโดยการเลือกประโยคมีขั้นตอนการสกัดประโยคที่ประกอบด้วยตัวอักษรภาษาต่างประเทศ ออก และแปลงประโยคในรูปเขียนให้เป็นรูปอ่าน ใช้รูปแบบหน่วยเสียงคู่ จากนั้นให้คะแนนประโยคจากความครอบคลุมทางหน่วยเสียงที่เกิดขึ้น จนกระทั่งความครอบคลุมทางหน่วยเสียงผ่านเกณฑ์ที่กำหนดไว้ ผลคือได้ความครอบคลุมทางหน่วยเสียงรูปแบบหน่วยเสียงคู่เป็น 90.9% ทั้งหมด 1,628 หน่วยเสียงคู่ งานวิจัยนี้ทำให้เกิดแนวคิดเรื่อง การใช้ความครอบคลุมทางหน่วยเสียงเป็นหลักในการให้คะแนนประโยคจนกระทั่ง ผลความครอบคลุมทางหน่วยเสียงผ่านเกณฑ์ที่กำหนดไว้

TSynC-1: Thai tagged speech corpus for speech synthesis (C. Hansakunbuntheung, 2003) [19] เลือกข้อความนำไปสร้างระบบสังเคราะห์เสียงพูดภาษาไทย ใช้ประโยคจากฐานข้อมูล ORCHID (V. Sornlertlamvanich, 1998) [20] สารานุกรมสำหรับเด็ก บทความวิชาการจากงานสัมมนาด้านเทคโนโลยี 43,340 ประโยค มีการกำหนดขอบเขตจำนวนพยางค์ในแต่ละประโยค แต่ละประโยคประกอบด้วยอักษรไทยเท่านั้น ให้คะแนนแต่ละประโยคจากการกระจายตัวทางสถิติของหน่วยเสียงสาม ใช้ Greedy Algorithm คัดเลือกประโยค คัดเหลือ 5,200 ประโยค ผลลัพธ์ ได้การครอบคลุมทางหน่วยเสียงแบบ หน่วยเสียงสามเท่ากับ 39.6% งานวิจัยนี้ทำให้เกิดแนวคิดเรื่อง การนำ Greedy Algorithm มาใช้ในการคัดเลือกประโยคเป็นวิธีที่ดีในการคัดเลือก คลังข้อความที่มีขนาดจำกัดอาจเกิดปัญหาในการเพิ่มความครอบคลุมของหน่วยเสียง

Constructing a Phonetic-Rich Speech Corpus while Controlling Time-Dependent Voice Quality Variability for English Speech Synthesis (N. Jinfu, 2006) [25] การเลือก

ข้อความนำไปสร้างระบบสังเคราะห์เสียงพูดภาษาอังกฤษใช้ประโยคจากฐานข้อมูล BTEC 749,500 ประโยค และ หนังสือพิมพ์อังกฤษ 4,985,200 ประโยค ในหนึ่งประโยคมีจำนวนคำ 10-25 คำ ให้คะแนนแต่ละประโยคจากการกระจายตัวทางสถิติของหน่วยเสียงสาม คัดประโยคเหลือ 3,100 ประโยค ผลลัพธ์ได้การครอบคลุมทางหน่วยเสียงสาม 99.33% งานวิจัยนี้ทำให้เกิดแนวคิดเรื่องการเลือกข้อความจากแหล่งข้อมูลขนาดใหญ่ เพิ่มค่าความครอบคลุมทางหน่วยเสียงได้มากกว่า

TSynC-2: An intensive design of a Thai speech corpus (C. Wutiwiwatchai, 2007) [14] การเลือกข้อความนำไปสร้างระบบสังเคราะห์เสียงพูดภาษาไทย ใช้ประโยคจากฐานข้อมูล TSynC-1 (C. Hansakunbuntheung, 2003) เป็นประโยคตั้งต้นในการเลือกประโยค 5,386 ประโยค เลือกประโยคโดยเลือกคะแนนที่สูงที่สุด ให้คะแนนโดยการเพิ่มคะแนนให้หน่วยเสียงคู่ที่มีน้อยและคะแนนยังถูกหารด้วยจำนวนความยาวของหน่วยเสียงด้วย ผลลัพธ์คือได้ประโยคมาทั้งสิ้น 3,581 ประโยค 5,823 ชนิดของรูปแบบหน่วยเสียงคู่ จากนั้นนำประโยคที่ได้ไปสร้างโมเดลเสียงและสังเคราะห์ออกมา จากการประเมินได้คะแนนความพึงพอใจ อยู่ในระดับ 3.09 ซึ่งมากกว่าเสียงสังเคราะห์จาก TSynC-1 งานวิจัยนี้ทำให้เกิดแนวคิดเรื่อง การนำให้คะแนนกับหน่วยเสียงที่มีน้อยหรือหายาก จะทำให้การเลือกประโยคได้ผลลัพธ์ตามเป้าหมายได้

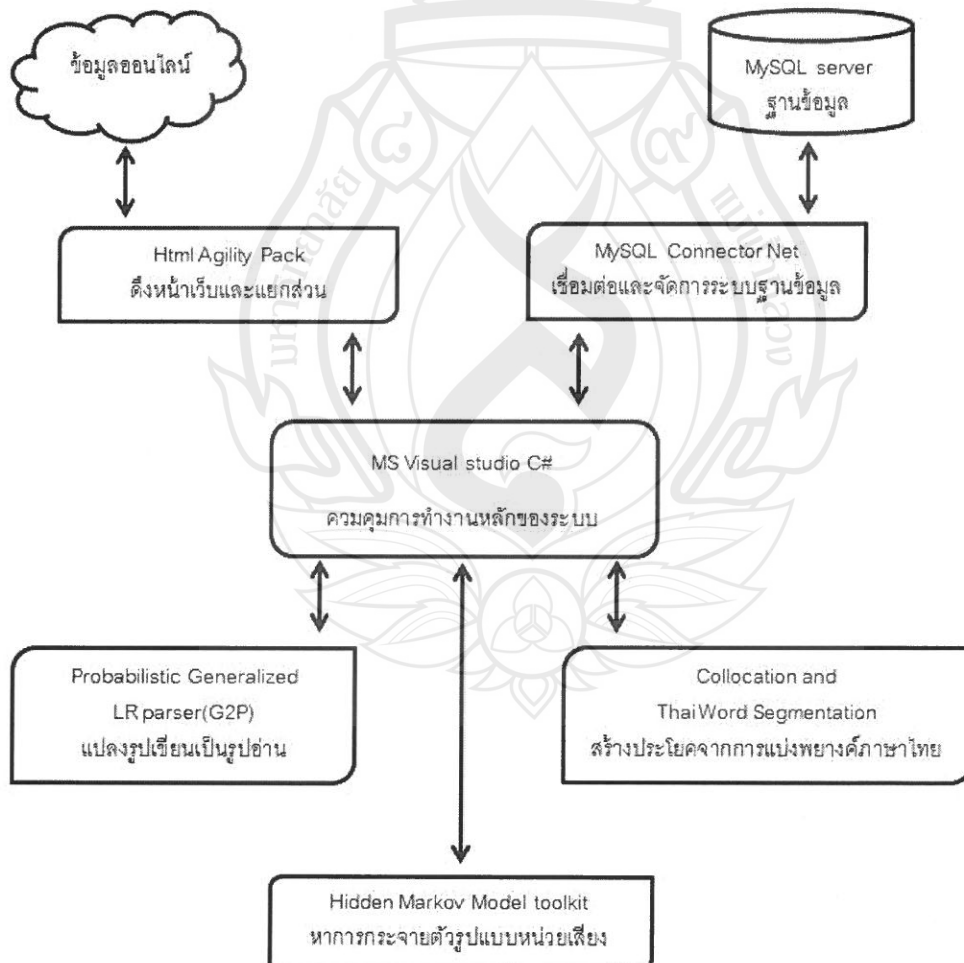
Automatic Construction for a TTS Corpus with Limited Text (W. Zhang, 2010) [11] เลือกข้อความนำไปสร้างระบบสังเคราะห์เสียงพูดภาษาอังกฤษ ใช้ประโยคจากฐานข้อมูลออนไลน์ VOA special English คัดมาแบบ Random 4,621 ประโยค (ใช้เป็น Mother Text) แปลง text เป็น Di-phone ด้วย HTK+CMU dictionary ใช้ Greedy algorithm Weight คะแนนให้ หน่วยเสียงที่หายาก (Di-phone) ให้คะแนนแต่ละ sentence ด้วย Okapi formula คัดเหลือ 1,000 ประโยค ผลลัพธ์ ได้การครอบคลุมทางหน่วยเสียงคู่ 93.52% งานวิจัยนี้ทำให้เกิดแนวคิดเรื่องการพิจารณาในเรื่องความสั้นยาวของแต่ละประโยค เนื่องจากคะแนนขึ้นอยู่กับจำนวนความยาวของ ประโยคด้วย การ Weight คะแนนให้ หน่วยเสียงหายากทำให้เข้าถึงเป้าหมายได้เร็ว เกิดแนวคิดเรื่องการดึงข้อความจากฐานข้อมูลออนไลน์

บทที่ 3

ระเบียบวิธีวิจัย

3.1 เครื่องมือที่ใช้ในการวิจัย

1. ตัวดึงหน้าเว็บและแยกส่วน HTML, Html Agility Pack (HAP) [30]
2. ตัวแปลงรูปเขียนเป็นรูปอ่าน, Probabilistic Generalized LR parser [12]
3. ตัวนับความถี่รูปแบบหน่วยเสียง, Hidden Markov Model toolkit [31]
4. ตัวแบ่งพยางค์ภาษาไทย, Collocation and Thai Word Segmentation [32]
5. ระบบเชื่อมต่อบริบบฐานข้อมูล, MySQL Connector Net 6.2.4 [33]
6. ระบบฐานข้อมูล, MySQL Server 5.1 [34]
7. เครื่องมือพัฒนาโปรแกรมภาษา C#, Microsoft Visual Studio 2010 [35]

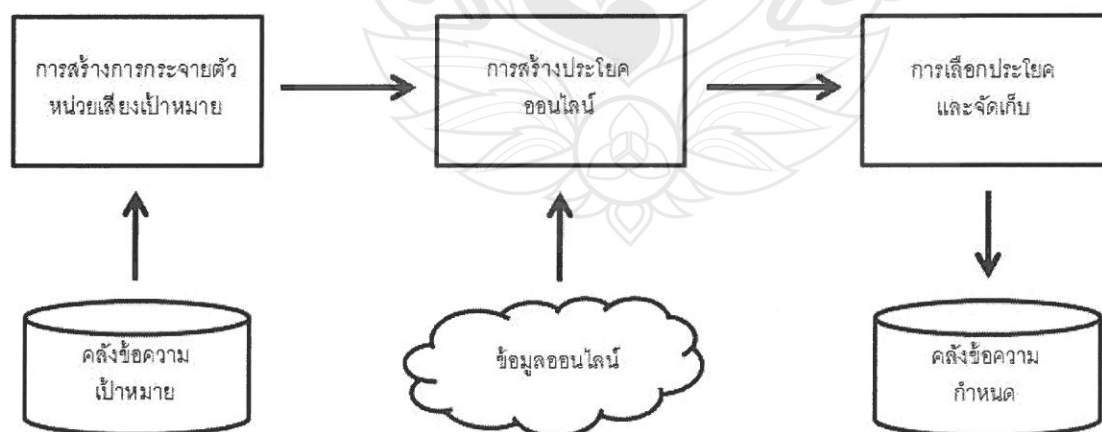


ภาพที่ 3-1 แผนภาพความสัมพันธ์ของเครื่องมือวิจัย

สำหรับเครื่องมือที่ใช้ในงานวิจัย ซึ่งการทำงานมีความสัมพันธ์กัน ดังภาพที่ 3-1 ตัวดึงหน้าเว็บและแยกส่วน มีหน้าที่ดึงหน้าเว็บในรูปแบบ HTML จากข้อมูลออนไลน์บนอินเทอร์เน็ต มาวิเคราะห์แยกส่วนที่เป็นข้อความภาษาไทยและส่วนที่เป็น URL ไปยังหน้าเว็บเพจอื่น ๆ เพื่อส่งให้ระบบหลัก ตัวแปลงรูปเขียนเป็นรูปอ่านมีหน้าที่รับประโยคในรูปเขียนมาแปลงเป็นรูปอ่านและส่งให้ระบบหลัก ตัวนับความถี่รูปแบบหน่วยเสียง มีหน้าที่นำชุดลำดับรูปอ่านทั้งหมดมาแปลงให้อยู่ในรูปแบบหน่วยเสียงที่ต้องการและนับจำนวนที่เกิดขึ้นของแต่ละรูปแบบเพื่อส่งไปให้ระบบหลักสร้างการกระจายตัวรูปแบบหน่วยเสียง ตัวแบ่งพยางค์ภาษาไทย มีหน้าที่นำชุดบทความมาแปลงเป็นลำดับของประโยคซึ่งในแต่ละประโยคมีจำนวนพยางค์ตามที่ต้องการ และส่งให้ระบบหลัก ระบบเชื่อมต่อฐานข้อมูลมีหน้าที่ ส่งชุดคำสั่ง SQL จากระบบหลักไปจัดการฐานข้อมูล สามารถส่งข้อมูลหรืออ่านข้อมูลจากฐานข้อมูลได้ ส่วนเครื่องมือพัฒนาระบบหลักใช้ชุดคำสั่งภาษา C# ในการควบคุมระบบทั้งหมด

3.2 ขั้นตอนการดำเนินงานวิจัย

ในบทนี้จะกล่าวถึงขั้นตอนการดำเนินงานวิจัยซึ่งประกอบด้วยสามส่วนหลัก ส่วนแรกคือการหาการกระจายตัวทางหน่วยเสียงเป้าหมาย โดยสามารถหาการกระจายตัวของหน่วยเสียงเป้าหมายได้จากการกำหนดจำนวนความถี่ของแต่ละหน่วยเสียงเอง หรือกำหนดจากรูปแบบการกระจายตัวทางหน่วยเสียงของคลังข้อความอื่น ๆ ส่วนที่สองคือส่วนของการสร้างประโยค ในส่วนนี้จะรวมถึงขั้นตอนการดึงข้อความจากข้อมูลทางอินเทอร์เน็ต ขั้นตอนการแบ่งประโยคจากกลุ่มของข้อความและขั้นตอนการคัดประโยคที่มีส่วนประกอบของอักขระที่ไม่ต้องการออก โดยขั้นตอนนี้จะทำงานไปเรื่อย ๆ จนกระทั่งคลังข้อความที่ต้องการ ส่วนที่สามคือส่วนของการเลือกประโยคและการจัดเก็บสู่คลังข้อความ การทำงานหลักของระบบดังกล่าว แสดงดังภาพที่ 3-2



ภาพที่ 3-2 ภาพรวมการทำงานหลักของระบบ

3.2.1 ขั้นตอนการสร้างการกระจายตัวทางหน่วยเสียงเป้าหมาย

การกระจายตัวทางหน่วยเสียงเป้าหมายคือ รูปแบบของการกระจายตัวทางหน่วยเสียงที่ต้องการให้เกิดในคลังข้อความที่จะถูกสร้างขึ้น ในงานวิจัยนี้สามารถใช้รูปแบบการกระจายตัวได้สองรูปแบบ รูปแบบแรกคือ กระบวนการกระจายตัวทางหน่วยเสียงที่สร้างจากการกำหนดความถี่จำนวนที่เกิดขึ้นของแต่ละรูปแบบหน่วยเสียงเอง อีกรูปแบบคือ ใช้การกระจายตัวทางหน่วยเสียงที่สร้างจากความถี่จำนวนที่เกิดขึ้นของแต่ละรูปแบบหน่วยเสียงจากคลังข้อความในฐานข้อมูลอื่น ๆ สำหรับงานวิจัยนี้ได้เลือกใช้ข้อความจากฐานข้อมูล Large Vocabulary Continuous Speech Recognition: LVCSR [13] หรือที่ถูกระบุง่าย ๆ ว่า LOTUS พัฒนาโดยหน่วยปฏิบัติการวิจัยวิทยาการมนุษยภาษา ศูนย์เทคโนโลยีอิเล็กทรอนิกส์และคอมพิวเตอร์แห่งชาติ (NECTEC) เลือกใช้ชุดประโยคที่ครอบคลุมภาษาไทย 5,000 คำ ซึ่งมาจากบทความทั่วไปและบทความข่าวซึ่งเป็นภาษาไทย จำนวน 3,007 ประโยค ขั้นตอนการสร้างการกระจายตัวทางหน่วยเสียงเป้าหมายสามารถแสดงภาพการทำงาน ดังภาพที่ 3-3



ภาพที่ 3-3 การสร้างการกระจายตัวทางหน่วยเสียงเป้าหมาย

ขั้นตอนแรกนำข้อความจากฐานข้อมูล LOTUS มาแปลงรูปเขียนให้เป็นรูปอ่าน โดยในงานวิจัยนี้ได้ใช้ ตัวแปลงรูปเขียนเป็นรูปอ่านของ Probabilistic Generalized LR parser [12] ผลลัพธ์จากการแปลงรูปเขียนของข้อความจากฐานข้อมูล LOTUS เป็นรูปอ่านสามารถดูได้จากตารางที่ 3-1 การแปลงรูปเขียนเป็นรูปอ่านจากประโยคทั้งหมด 3,007 ประโยคได้หน่วยเสียงทั้งหมดรวมเป็น 241,199 หน่วยเสียง หลังจากได้รูปอ่านของแต่ละประโยคมาแล้ว ขั้นตอนต่อไปคือการนับจำนวนแต่ละรูปแบบหน่วยเสียงที่เกิดขึ้นทั้งหมดจากรูปอ่านที่แปลงมา โดยในงานวิจัยนี้ได้เลือกใช้ รูปแบบหน่วยเสียงแบบหน่วยเสียงคู่ (Diphone) ในการทดลองขั้นตอนการนับจำนวนแต่ละรูปแบบหน่วยเสียงที่เกิดขึ้น จะใช้เครื่องมือ Hidden Markov Model toolkit, HTK [31] มาช่วยจัดการในขั้นตอนนี้ โดยใช้ฟังก์ชัน HLEd ในการหารูปแบบของหน่วยเสียงคู่ที่เกิดขึ้นจากรูปอ่านทั้งหมดในที่นี้มีรูปแบบหน่วยเสียงคู่ที่เกิดขึ้น 1,383 รูปแบบ จากนั้นนับการ

เกิดขึ้นของแต่ละรูปแบบหน่วยเสียงคู่ด้วยฟังก์ชัน HLStats หลังจากขั้นตอนนี้เราจะได้การกระจายตัวทางสถิติทางหน่วยเสียงของเป้าหมาย เพื่อนำไปใช้ในขั้นตอนการคัดเลือกประโยคต่อไป สามารถดูการกระจายตัวทางหน่วยเสียงเป้าหมายในรูปแบบหน่วยเสียงคู่ ได้จากตารางที่ 3-2 และภาพที่ 3-4

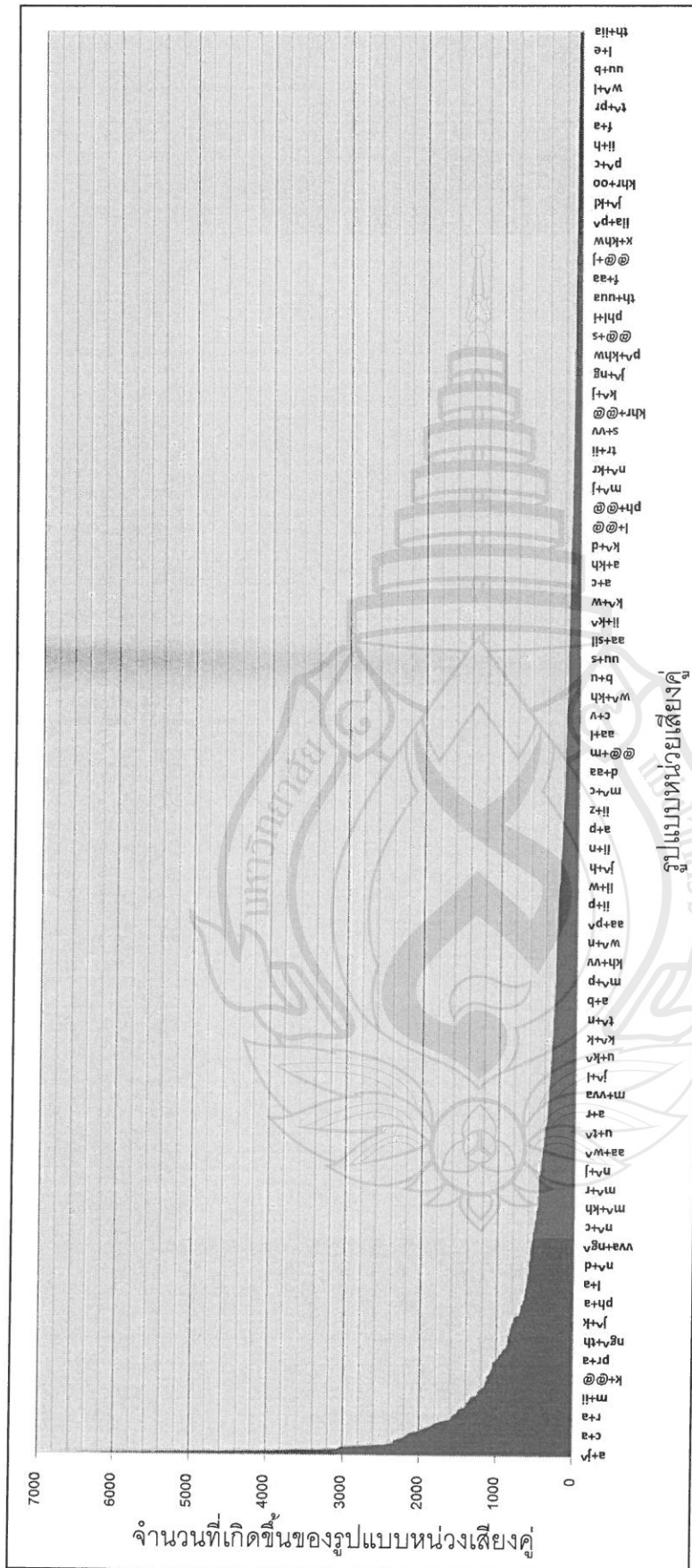
ตารางที่ 3-1 ตัวอย่างการแปลงรูปเขียนเป็นรูปอ่านจากฐานข้อมูล LOTUS

ประโยคในรูปเขียน	ประโยคในรูปอ่าน
อย่างไรก็ตามเหล่านักธุรกิจนักลงทุนที่เข้าร่วมประชุมสัมมนาครั้งนี้มีความเห็นว่าความเติบโตทางเศรษฐกิจในภูมิภาคเอเชียแปซิฟิกจะขึ้นอยู่กับทรัพยากรบุคคลที่ได้รับการฝึกฝนจนมีความชำนาญในหน้าที่การงานอย่างสูง	sil j aa ng^ r a j^ k @@ t aa m^ l a w^ n a k^ th u r a k i t^ n a k^ l o ng^ th u n^ th ii kh a w^ r uua m^ pr a ch u m^ s a m^ m a n aa kh r a ng^ n ii m ii khw aa m^ h e n^ w aa khw aa m^ t qq p^ t oo th aa ng^ s ee t^ th a k i t^ n a j^ ph uu m i ph aa k^ z ee ch iia p xx s i f i k^ c a kh v n^ j uu k a p^ s a p^ ph a j aa k @@ n^ b u k^ kh o n^ th ii d a j^ r a p^ k aa n^ f v k^ f o n^ c o n^ m ii khw aa m^ ch a m^ n i ch a m^ n aa n^ n a j^ n aa th ii k aa n^ ng aa n^ j aa ng^ s uu ng^ sil
ก็ต้องแก้ด้วยการศึกษาอบรมด้วยวิธีแปลความหมายใหม่ในความเชื่อที่เป็นพื้นเพเดิมของเขาตั้งจะกล่าวต่อไปในตอนหลัง	sil k @@ t @@ ng^ k xx d uua j^ k aa n^ s v k^ s aa z o p^ r o m^ d uua j^ w i th ii pl xx khw aa m^ m aa j^ m a j^ n a j^ khw aa m^ ch vva th ii p e n^ ph vv n^ ph ee d qq m^ kh @@ ng^ kh a w^ d a ng^ c a kl aa w^ t @@ p a j^ n a j^ t @@ n^ l a ng^ sil
ซึ่งเป็นการสร้างภาพตนเองให้เป็นบุคคลสำคัญในลักษณะที่สงสารตัวเอง	sil s v ng^ p e n^ k aa n^ s aa ng^ ph aa p^ t o n^ z ee ng^ h a j^ p e n^ b u k^ kh o n^ s a m^ kh a n^ n a j^ l a k^ s a n a th ii s o ng^ s aa n^ t uua z ee ng^ sil
แต่มันก็บอกไปอะไรไม่ได้มากกว่านั้น	sil t xx m a n^ k @@ b @@ k^ b a j^ z a r a j^ m a j^ d a j^ m aa k^ kw aa n a n^ sil
แต่ก็ตกอยู่ระหว่างเวลาของคัมภีร์ทั้งสองที่กล่าวมาแล้ว	sil t xx k @@ t o k^ j uu r a w aa ng^ w ee l aa kh @@ ng^ kh a m^ ph ii th a ng^ s @@ ng^ th ii kl aa w^ m aa l xx w^ sil
หรือชาวชนบทอพยพเข้าสู่ตัวเมืองที่ทวีจำนวนมากขึ้นในแต่ละประเทศได้ก่อให้เกิดปัญหาย่างยากมากทำให้บางประเทศเช่นสหรัฐอเมริกาต้องออกกฎหมายห้ามคนเข้าเมืองในปีคริสต์ศักราชหนึ่งพันเก้าร้อยยี่สิบ	sil r vv ch aa w^ ch o n^ b o t^ z o p^ ph a j o p^ kh a w^ s uu t uua m vva ng^ th ii th a w ii c a m^ n uua n^ m aa k^ kh v n^ n a j^ t xx l a pr a th ee t^ d a j^ k @@ h a j^ k qq t^ p a n^ h aa j u ng a j aa k^ m aa k^ th a m^ h a j^ b aa ng^ pr a th ee t^ ch ee n^ s a r a t^ z a m ee r i k aa t @@ ng^ z @@ k^ k o t^ m aa j^ h aa m^ kh o n^ kh a w^ m vva ng^ n a j^ p ii kh r i t^ s a k^ r aa t^ n v ng^ ph a n^ k a w^ r @@ j^ j ii s i p^ sil

ประโยคในรูปเขียน	ประโยคในรูปอ่าน
กับคำถามแรกที่ทำให้จารณ์บทบาท ของสภารักษาความสงบเรียบร้อย แห่งชาติและรัฐบาลในรอบสองเดือน ที่ผ่านมา	sil k a p^ kh a m^ th aa m^ r xx k^ th ii h a j^ w i c aa n^ b o t^ b aa t^ kh @@ ng^ s a ph aa r a k^ s aa khw aa m^ s a ng o p^ r iia p^ r @@ j^ h xx ng^ ch aa t^ l x r a t^ th a b aa n^ n a j^ r @@ p^ s @@ ng^ d vva n^ th ii ph aa n^ m aa sil
แต่ในการประชุมเรื่องโลกร้อนขึ้นที่ ประเทศเนเธอร์แลนด์เมื่อเดือน พฤศจิกายนปีที่แล้ว	sil t xx n a j^ k aa n^ pr a ch u m^ r vva ng^ l oo k^ r @@ n^ kh v n^ th ii pr a th ee t^ n ee th qq l xx n^ m vva d vva n^ phr v t^ s a c i k aa j o n^ p ii th ii l xx w^ sil
และมีชีวิตอยู่ในโลกด้วยการร้องเพลง ให้ความรื่นรมย์เท่านั้น	sil l x m ii ch ii w i t^ j uu n a j^ l oo k^ d uua j^ k aa n^ r @@ ng^ phl ee ng^ h a j^ khw aa m^ r vv n^ r o m^ th a w^ n a n^ sil
ไม่ใช่ความจริงที่จับต้องได้เดี๋ยวนี้	sil m a j^ ch a j^ khw aa m^ c i ng^ th ii c a p^ t @@ ng^ d a j^ d iia w^ n ii sil
คงเป็นที่ทราบกันดีว่าการจัด การศึกษาในปัจจุบันคุณภาพ การศึกษายังได้ผลเป็นที่น่าพอใจ	sil kh o ng^ p e n^ th ii s aa p^ k a n^ d ii w aa k aa n^ c a t^ k aa n^ s v k^ s aa n a j^ p a t^ c u b a n^ kh u n^ n a ph aa p^ k aa n^ s v k^ s aa j a ng^ d a j^ ph o n^ p e n^ th ii n aa ph @@ c a j^ sil
ซึ่งได้แก่บุคคลที่ไม่อาจจะเรียนรู้ ระเบียบของสังคมได้ถูกต้องหรือผู้ที่ ต่อต้านคุณค่าทางสังคมและหรือ ความเชื่อทางสังคม	sil s v ng^ d a j^ k xx b u k^ kh o n^ th ii m a j^ z aa t^ c a r iia n^ r uu r a b iia p^ kh @@ ng^ s a ng^ kh o m^ d a j^ th uu k^ t @@ ng^ r vv ph uu th ii t @@ t aa n^ kh u n^ kh aa th aa ng^ s a ng^ kh o m^ l x r vv khw aa m^ ch vva th aa ng^ s a ng^ kh o m^ sil

ตารางที่ 3-2 ตัวอย่างจำนวนของแต่ละรูปแบบหน่วยเสียงคู่ที่เกิดขึ้นจากฐานข้อมูล LOTUS

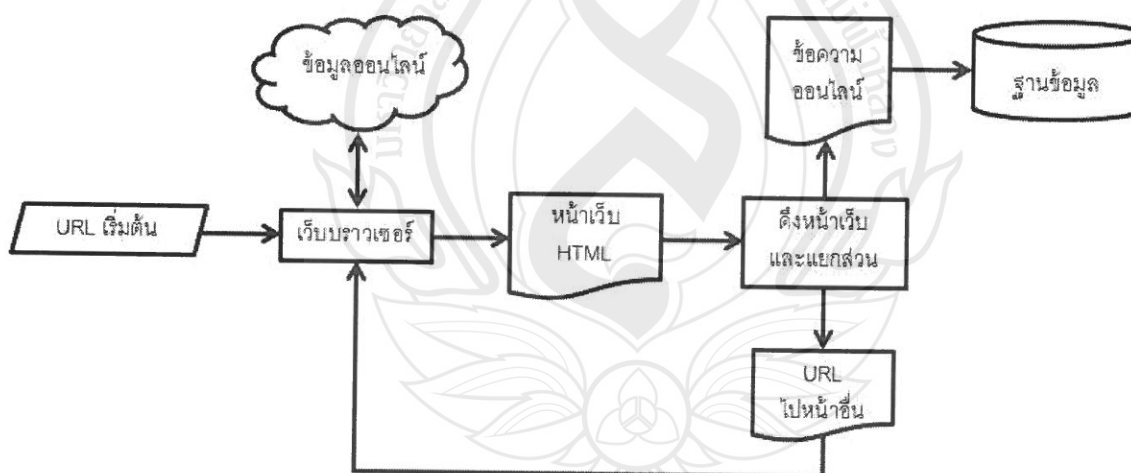
รูปแบบ	จำนวน	รูปแบบ	จำนวน	รูปแบบ	จำนวน	รูปแบบ	จำนวน
a+j^	6792	a+n	1061	l+a	617	j^+d	476
aa+n^	4136	v+ng^	1056	j^+sil	613	ng^+r	476
n+a	3076	s+aa	1055	k^+s	609	j^+kh	473
k+aa	3023	p+a	1050	ng^+m	607	ii+s	472
th+ii	2454	pr+a	1007	a+m	606	m^+r	471
a+n^	2335	j+aa	982	j^+r	596	m^+s	470
a+m^	2326	th+aa	969	n^+ph	593	t^+s	470
@@+ng^	2226	aa+k^	937	j^+th	592	z+aa	469
c+a	2165	n^+n	933	n^+d	590	ii+m	466
th+a	2131	kh+o	886	n^+r	586	ng^+ph	463
aa+ng^	1998	n^+k	870	z+a	583	a+s	462
s+a	1947	ng^+k	865	s+v	576	j+a	452
aa+m^	1844	ng^+th	863	s+i	575	n^+j	449
e+n^	1803	n+aa	851	c+aa	574	ii+k	443
o+n^	1714	n^+kh	848	o+ng^	569	aa+k	442
d+a	1595	h+a	836	j^+s	568	w+i	442
r+a	1583	n^+s	829	vva+ng^	563	a+ph	440
p+e	1548	l+x	828	n^+z	556	t+xx	440
k+a	1506	ng^+kh	817	o+m^	556	k^+kh	431
kh+@@	1504	a+k^	814	t+a	548	n^+h	429
a+ng^	1416	j^+k	791	a+k	546	aa+w^	428
m+a	1392	a+th	785	h+aa	546	j+uu	424
khw+aa	1389	t+@@	784	ng^+t	538	kh+aa	417
n^+th	1347	n+ii	715	j^+n	533	aa+th	412
m+ii	1334	ch+a	702	n^+c	529	a+d	409
kh+a	1264	t+aa	698	ph+uu	517	b+aa	405
a+t^	1242	ng^+s	695	i+ng^	516	t^+c	405
m+aa	1240	n^+m	688	xx+ng^	514	uua+n^	401
a+w^	1189	ph+a	666	n^+l	504	u+t^	397
a+p^	1152	ph+aa	661	n^+t	500	s+o	395
t^+th	1147	ee+t^	647	a+j	499	aa+s	390
i+t^	1140	n^+sil	645	n^+p	499	t^+kh	388
k+@@	1123	w+a	636	m^+kh	485	ii+c	373
w+aa	1122	ng^+n	632	ch+aa	478	uua+j^	371



ภาพที่ 3-4 แผนภาพการกระจายตัวทางสถิติทางหน่วยเสียงของฐานข้อมูล LOTUS

3.2.2 ขั้นตอนการสร้างประโยคออนไลน์

ขั้นตอนนี้มีหน้าที่สร้างประโยคเพื่อนำไปใช้ในขั้นตอนการเลือกประโยคเพื่อจัดเก็บสู่คลังข้อความกำหนด แผนภาพการทำงานโดยรวมของขั้นตอนดูได้จากภาพที่ 3-5 ในงานวิจัยนี้ได้เลือกใช้บทความจากอินเทอร์เน็ตมาสร้างประโยค เครื่องมือการดึงข้อความจากหน้าเว็บจึงถูกนำมาใช้ ในที่นี้เราใช้ตัวดึงหน้าเว็บแยกวิเคราะห์ HTML ของ Html Agility Pack (HAP) [30] เนื่องจากเป็นฟังก์ชันที่ใช้ง่ายและสามารถดึงข้อมูลจากหน้าเว็บได้เร็ว หลักการทำงานคือ URL เริ่มต้น จะถูกส่งไปให้เว็บเบราว์เซอร์สิ่งที่ได้คือ โครงสร้าง HTML ของหน้าเว็บเพจนั้นซึ่งอาจประกอบไปด้วยข้อมูลข้อความ ข้อมูลภาพ ข้อมูล URL หรือข้อมูลอื่น ๆ จากนั้นตัวดึงหน้าเว็บแยกวิเคราะห์ HTML จะทำการคัดกรองส่วนที่เราต้องการออกมา ในที่นี้ได้ดึงส่วนที่เป็น URL เก็บไว้เพื่อใช้ในการเข้าสู่หน้า HTML อื่น ๆ และดึงส่วนที่เป็นข้อความไปจัดเก็บไว้ในฐานข้อมูลเพื่อเตรียมไว้ในส่วนของการจัดการประโยคต่อไป ในงานวิจัยนี้ได้เลือกใช้บทความจากเว็บไซต์ สารานุกรมไทยสำหรับเยาวชน [36] ตัวอย่างหน้าเว็บแสดงดังภาพที่ 3-6 และจากเว็บคลังข้อมูลข้อความภาษาไทยแห่งชาติ [37] เนื่องจากมีโครงสร้าง HTML ที่ไม่ซับซ้อนและบทความมีหลากหลายลักษณะ เช่น บทความวิชาการ หนังสือแบบเรียน หนังสือพิมพ์ นิยายและอื่น ๆ สามารถดูตัวอย่างของข้อความที่เก็บไว้ในฐานข้อมูลได้ดังภาพที่ 3-7



ภาพที่ 3-5 การดึงข้อความออนไลน์ด้วยตัวดึงหน้าเว็บแยกวิเคราะห์ HTML



สารานุกรมไทยสำหรับเยาวชน

โดยพระราชประสงค์ในพระบาทสมเด็จพระเจ้าอยู่หัว

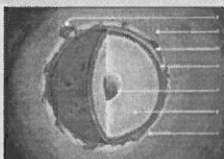
หน้าหลัก
พระราชดำริ
โครงการ
เรื่องในสารานุกรม
ฉบับสมบูรณ์
ค้นข้อมูล
ติดต่อ
ระบบงาน

สารานุกรมไทย สำหรับเยาวชน เล่มที่ ๓

- ❖ เรื่องที่ ๑ ดวงอาทิตย์
- ❖ เรื่องที่ ๒ อวกาศ
- ❖ เรื่องที่ ๓ ห้องฟิสิกส์
- ❖ เรื่องที่ ๔ นก
- ❖ เรื่องที่ ๕ ปลา
- ❖ เรื่องที่ ๖ เครื่องจักรกล
- ❖ เรื่องที่ ๗ พลังงาน
- ❖ เรื่องที่ ๘ อวกาศยาน
- ❖ เรื่องที่ ๙ คนไทย

กลับหน้าแรก

f Share สารานุกรมไทยสำหรับเยาวชนฯ / เล่มที่ ๓ / เรื่องที่ ๑ ดวงอาทิตย์ / ภาพภายในดวงอาทิตย์



ภาพภายในดวงอาทิตย์

ดังที่ได้กล่าวมาแล้วว่า ดวงอาทิตย์เป็นก้อนสสารใหญ่ก้อนหนึ่ง และรวมตัวเป็นสสารฐาน ๒ อย่างคือไฮโดรเจน และฮีเลียม โดยแรงดึงดูดระหว่างอะตอมและโมเลกุล แรงดึงดูดหรือแรงโน้มถ่วง (gravitational force) นี้ มีทิศทางเข้าหาจุดศูนย์กลาง เนื้อสสารของดวงอาทิตย์ ซึ่งอยู่ในระดับใดระดับหนึ่ง ภายในตัวดวงอาทิตย์นั้น โดยเนื้อสสารที่อยู่ชั้นใน จะมีความร้อนและความดันสูงมากจนกว่าเนื้อสสารในระดับที่อยู่ด้านนอกชั้นในนี้ กล่าวได้ว่า ความร้อนและความดันสูงของเนื้อสสารชั้นในจะดันสสารชั้นนอกออกไปทางด้านนอกชั้นในนี้

ดังนั้นถ้าเราสังเกตชั้นในของดวงอาทิตย์ เราจะพบว่าไฮโดรเจนและฮีเลียมจะรวมตัวกันเป็นสสารฐาน ๒ อย่างคือไฮโดรเจน และฮีเลียม โดยแรงดึงดูดระหว่างอะตอมและโมเลกุล แรงดึงดูดหรือแรงโน้มถ่วง (gravitational force) นี้ มีทิศทางเข้าหาจุดศูนย์กลาง เนื้อสสารของดวงอาทิตย์ ซึ่งอยู่ในระดับใดระดับหนึ่ง ภายในตัวดวงอาทิตย์นั้น โดยเนื้อสสารที่อยู่ชั้นใน จะมีความร้อนและความดันสูงมากจนกว่าเนื้อสสารในระดับที่อยู่ด้านนอกชั้นในนี้ กล่าวได้ว่า ความร้อนและความดันสูงของเนื้อสสารชั้นในจะดันสสารชั้นนอกออกไปทางด้านนอกชั้นในนี้

ดังนั้นถ้าเราสังเกตชั้นในของดวงอาทิตย์ เราจะพบว่าไฮโดรเจนและฮีเลียมจะรวมตัวกันเป็นสสารฐาน ๒ อย่างคือไฮโดรเจน และฮีเลียม โดยแรงดึงดูดระหว่างอะตอมและโมเลกุล แรงดึงดูดหรือแรงโน้มถ่วง (gravitational force) นี้ มีทิศทางเข้าหาจุดศูนย์กลาง เนื้อสสารของดวงอาทิตย์ ซึ่งอยู่ในระดับใดระดับหนึ่ง ภายในตัวดวงอาทิตย์นั้น โดยเนื้อสสารที่อยู่ชั้นใน จะมีความร้อนและความดันสูงมากจนกว่าเนื้อสสารในระดับที่อยู่ด้านนอกชั้นในนี้ กล่าวได้ว่า ความร้อนและความดันสูงของเนื้อสสารชั้นในจะดันสสารชั้นนอกออกไปทางด้านนอกชั้นในนี้

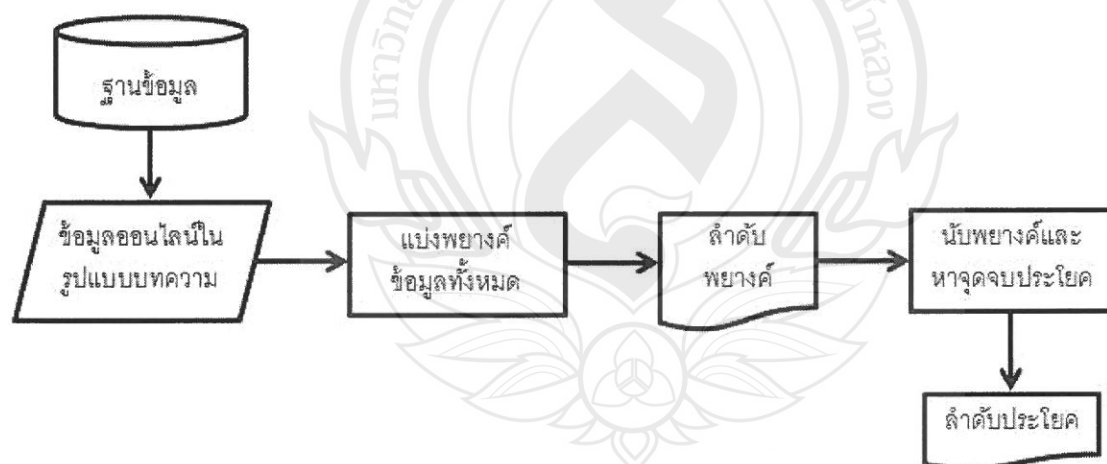
ภาพที่ 3-6 ตัวอย่างหน้าเว็บจากอินเทอร์เน็ตที่มีเนื้อความที่เกี่ยวกับการดึงข้อความ



The screenshot shows a web browser window with a list of 20 items. Each item has a 'Data_ID' and a 'Data_Text'. The text is in Thai and appears to be a list of articles or documents. The browser window shows the address bar with 'http://ling.arts.chula.ac.th/tnc' and the status bar with 'Record 1 of 853'.

ภาพที่ 3-7 ตัวอย่างข้อความออนไลน์จากอินเทอร์เน็ตที่เก็บไว้ในฐานข้อมูล

ต่อมาระบบจะนำข้อความออนไลน์ที่ถูกจัดเก็บไว้ในฐานข้อมูล มาทำการแปลงข้อความในรูปแบบบทความให้เป็นรูปแบบของประโยค โดยแต่ละประโยคต้องมีความยาวที่เหมาะสมที่จะนำไปใช้ในขั้นตอนการเลือกประโยค โดยความยาวประโยคในที่นี้สามารถกำหนดได้ โดยระบุขอบเขตค่าต่ำสุดและสูงสุดของจำนวนพยางค์ที่อยู่ในประโยค การทำงานของขั้นตอนนี้ได้จากภาพที่ 3-8 เริ่มจากระบบจะดึงบทความออนไลน์ในฐานข้อมูลมาทีละหนึ่งบทความ จากนั้นจะทำการแบ่งพยางค์ด้วยเครื่องมือ ตัวแบ่งพยางค์จากคำไทย Collocation and Thai Word Segmentation [32] เมื่อแบ่งบทความในรูปแบบของพยางค์แล้วจะได้ลำดับของพยางค์ ตัวอย่างผลลัพธ์ของการแบ่งพยางค์ดูได้จากตารางที่ 3-3 จากนั้นระบบจะทำการนำลำดับพยางค์มาจัดเรียงใหม่ตามลำดับ โดยที่จะเพิ่มพยางค์เพื่อสร้างประโยคใหม่ที่ละลำดับทำการนับพยางค์โดยที่ ถ้าผลรวมของลำดับพยางค์ มากกว่าค่าต่ำสุดที่กำหนดไว้ของจำนวนพยางค์ให้เพิ่มพยางค์ไปอีกหนึ่ง ๆ จนพบกับลำดับที่เป็นเครื่องหมายวรรคตอน จึงตัดเป็นหนึ่งประโยค ถ้าผลรวมของจำนวนพยางค์ในประโยคน้อยกว่าหรือมากกว่าขอบเขตค่าต่ำสุดและสูงสุดของพยางค์ที่กำหนดไว้ให้ตัดประโยคนั้นทิ้งไป ถือว่าใช้ไม่ได้ แต่ถ้าผลรวมของจำนวนพยางค์ในประโยคอยู่ในขอบเขตค่าต่ำสุดและสูงสุดของพยางค์ที่กำหนดไว้ ประโยคนั้นจะถือว่าใช้ได้และเข้าสู่ขั้นตอนต่อไป โดยในการทดลองของงานวิจัยนี้ได้กำหนดขอบเขตของจำนวนพยางค์ในหนึ่งประโยค คือ ต้องมากกว่า 20 พยางค์แต่ไม่เกิน 40 พยางค์ในหนึ่งประโยค เพื่อความเหมาะสมและง่ายต่อการนำไปบันทึกเสียง



ภาพที่ 3-8 การแปลงข้อความในรูปแบบบทความเป็นรูปแบบประโยค

ตารางที่ 3-3 ตัวอย่างผลลัพธ์จากการแปลงข้อความรูปแบบบทความเป็นรูปแบบลำดับพยางค์

ข้อความในรูปแบบลำดับพยางค์
จะว่าไปแล้ว ก็สุภานั่นแหละที่ทำให้ชิตจันทร์เกิดอยากจะเสี่ยงดวงแบบนี้ขึ้นมาบ้าง เพราะเมื่อวันก่อนตอนเธอก้าวเข้าประตูออฟฟิศมาก็เห็นสาว ๆ ทั้งหลายกำลังมุงอยู่รอบโต๊ะของสภา ชิตจันทร์ผู้มาทีหลังเลยอยากรู้ขึ้นมาบ้างว่าอะไรเป็นอะไร พอแทรกตัวเข้าไปได้ก็เห็นหน้าหมวย ๆ จืด ๆ ของสุภากำลังยิ้มชะจันทาหิยี้ หล่อนนั่นเองที่เป็นสาเหตุของความสับสนในตอนนี้ สุภาถูกล็อตเตอรี่ ฉันซื้อเพราะเป็นเรื่องบังเอิญใจ ที่อยู่ ๆ ก็มีรถของจันทร์มารับให้ไม่ต้องเปียกฝน มันก็ต้องเป็นรถนำโชคของฉันทันแหละ จริงไหมเล่า ชิตจันทร์ไม่ค่อยชอบรถมาทำงานหรอก เพราะรถนั้นก็ไม่ใช่ของเธอ แต่เป็นรถของพ่อ เมื่อก่อนพ่อต้องขับรถไปทำงานทุกวัน พ่อเกษียณแล้วก็เลยจอดเอาไว้ขับไปซื้อกับข้าวบ้าง ไปสังสรรค์กับเพื่อนเก่า ๆ บ้าง ชิตจันทร์ก็ได้อาศัยขับมาทำงานในวันที่ฝนตกหรือต้องออกไปพบลูกค้าข้างนอก มันไม่ใช่รถใหม่ป้ายแดงอย่างที่คนเล่นหวยชอบซื้อเลขตามทะเบียนเสียหน่อย แต่แล้วมันก็ทำให้สุภาถูกรางวัลเข้าจนได้ แถมไม่ใช่เงินน้อย ๆ เพราะเลขทะเบียนรถของพ่อชิตจันทร์นะ ผิดไปจากรางวัลที่ 1 แค่นี้เดียว สุภาจึงได้เงินร่วมครึ่งแสนมาแบบสบาย ๆ ด้วยโชคที่ควรจะเป็นของชิตจันทร์ หลังจากเข้านั้น ชิตจันทร์ผู้แค้นใจก็เลยตั้งเป้าหมายไว้ว่าจะซื้อล็อตเตอรี่กับเค้าบ้าง ทั้งที่เธอไม่เคยสนใจเรื่องการเสี่ยงดวงแบบนี้เลยแม้แต่หน่อย แต่ไฉนการที่เธอขับรถที่มีเงินครึ่งแสนติดท้ายฝาฝนมาให้คนอื่นเช็ดไปง่าย ๆ แบบนี้ก็ไม่ใช่ว่าเรื่องที่ชิตจันทร์จะยอมได้เหมือนกัน ชิตจันทร์เลยคิดว่า ถ้าทะเบียนรถพ่อออกไปแล้ว คราวนี้ก็มาถึงทะเบียนรถของแม่บ้างละ เธอเก็บเงียบไม่ยอมบอกเรื่องนี้กับใคร แล้วค่อย ๆ หาซื้อล็อตเตอรี่ที่มีเลขตรงกับทะเบียนรถของแม่เท่านั้นมาเก็บไว้ เลข 64821 นั้นหาไม่ยากนัก แต่ทุกครั้งที่เธอยืนยันจะเอาแต่เลขจำนวนนี้ ก็คงทำให้คนขายล็อตเตอรี่มองว่าเธอคงไปได้เลขเด็ดจากไหนเป็นแน่ แต่ถึงจะหาขายังไง เวลา 15 วันก่อนจะประกาศผลทำให้ชิตจันทร์มีเวลาซื้อเก็บไว้ได้พอสมควร

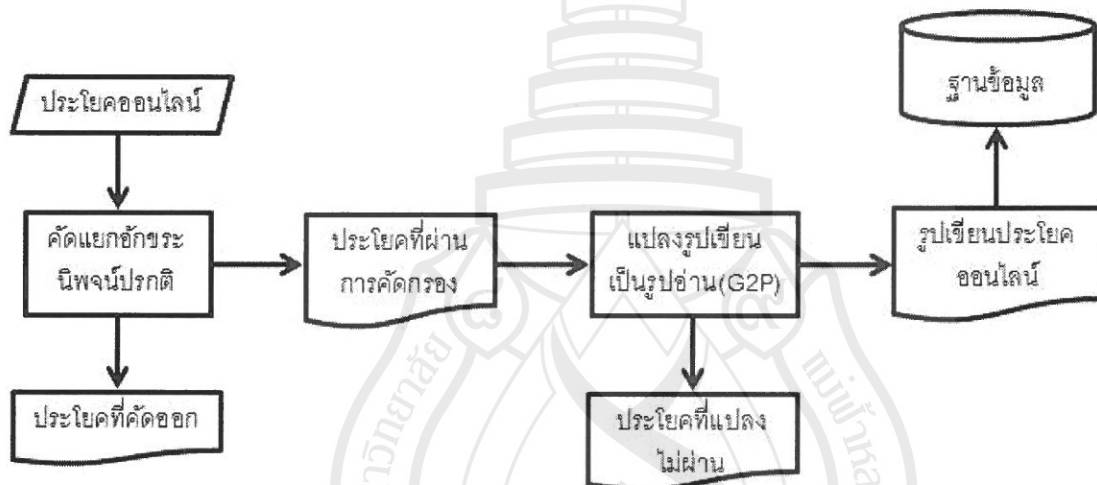
เมื่อได้สร้างประโยคแล้วประโยคเหล่านั้นจะถูกนำไปแปลงรูปเขียนเป็นรูปอ่าน สามารถดูแผนภาพขั้นตอนนี้ได้จาก ภาพที่ 3-9 โดยจะนำไปผ่านขั้นตอนการกรอกรูปแบบตัวอักษรที่ไม่ต้องการออกก่อน ซึ่งในการทดลองของงานวิจัยนี้ได้กำหนดให้ ประโยคที่ใช้ประกอบด้วยตัวอักษรไทยเท่านั้น โดยในที่นี้จะใช้หลักการระบุช่วงอักขระของนิพจน์ปกติ (Regular Expression) ในที่นี้ระบุไว้ที่ช่วง [ก-ฮ], [ะ-ู], [เ-็] และ [็-๊] เพื่อต้องการตัดอักขระ ภาษาต่างประเทศ ตัวเลข และสัญลักษณ์พิเศษออก ในการคัดกรองนี้ยังนำประโยคที่ไม่สามารถแปลงรูปเขียนเป็น

รูปอ่านสำเร็จออกไปด้วย เช่น ประโยคที่มีเครื่องหมาย [ๆ] จะทำให้การแปลงรูปเขียนเป็นรูปอ่านไม่สำเร็จ สามารถดูผลการคัดกรองประโยคได้จากตารางที่ 3-4

ตารางที่ 3-4 ตัวอย่างผลการคัดแยกประโยคตามอักขระนิพจน์ปกติที่ต้องการและไม่ต้องการ

ตัวอย่างผลการคัดแยกประโยคตามอักขระนิพจน์ปกติที่ต้องการ
- ไปขึ้นรถญี่ปุ่นของเขาที่จอดอยู่หน้าโรงสี ระหว่างทางชายหนุ่มชวน - มีของกินสารพัดอย่างกับผลไม้ตลอดปี ถ้ามีเงินซื้อกิน เขาภูมิใจที่เกิดมาเป็นคนไทย - เดินมา ฉันทชัยไปไหนไม่ได้ซาซายเป็นตะคริวเพราะย่อตัวนานเกินไป - ประกอบ น้ำเขียวอีอย่างนั้น ไม่ต้องติดย่าว่าห้ามกินห้ามอาบก็ไม่มีใครกล้าเข้าใกล้อยู่แล้ว - ยกเว้นครั้งที่ผมเห็นมันทำท่าเหมือนจะเปลี่ยนสีขน หมอดูแลรักษามผมด้วยยาสายลมอยู่หนึ่งเดือนเต็ม - สหายเพียวบอกว่าประเดี๋ยวจะพาไปตกปลา เราไม่ขัดข้อง หลังกินข้าวเช้าเสร็จเราก็เดินออกมาพร้อมกับคันเบ็ด - พ่อเผ่าอาจเอาเยื่อปลอก้านแฝดพื้นอยู่ตามร่มไม้ชายคา แม่เผ่าอาจต้มข้าวหมู - เมื่อเห็นเพื่อนสาว ไม่พอใจ ฮิลลารีจึงต้องเงียบ จัดการกับ แฮมเบอร์เกอร์ตรงหน้าจนหมดอัน
ตัวอย่างผลการคัดแยกประโยคตามอักขระนิพจน์ปกติที่ไม่ต้องการออก
- Collage ภาพปะติด งานศิลปะแบบหนึ่งที่ใช้วิธีปะแผ่นวัสดุบนพื้นในลักษณะแบนราบหรือนูนสูงจากพื้นภาพเพียงเล็กน้อย - ชังแล้วค่อยเรียบร้อยหน่อยง.มันก็มาทำให้หนูยุ่งไม่ได้ ผมเจ็บกริบ - และมิลลิสตัสจะต้องผิดหวัง เฟิร์ล ทำใจสบาย ๆ เกอะ อย่างกังวลไปเลย - จะต้องมียุคสมบัติดังนี้ ๑ เป็นหน่วยงานภาครัฐ หรือองค์กรภาคเอกชน - เลี้ยงเพื่อนเมื่อทำบุญอายุ 6 รอบ 72 ปี เมื่อเดือนตุลาคม - ชายหนุ่มจ้องหล่อนตาเขม็ง ไม่เกรงไม่กลัวใด ๆ หากเจ้าอยากแกล้งข้าเจ้า - ที่ไหนสักแห่งในโลกคู่ขนานเหมือนในนิยายวิทยาศาสตร์ โลกแห่ง <Fail>ถ้าหากว่า.</Fail> - นึกแล้วว่าเธอต้องมา เรื่องแย ๆ ที่แปลงเกษตรทำให้เธอเซ็งใช้มั้ยละ - ที่จริงก็น่าจะทำได้เพียงแต่ต้องเหนื่อยหน่อยเพราะจะต้องชุดหลุมกันคนละหลุมกว้าง ๆจนพอที่จะลงไปนอนได้ - nonlover ซึ่งจะมีโอกาสและสถิติสัมพัทธ์ที่จะเลือก เพื่อนคู่ใจ ที่มีคุณภาพสูงกว่าผู้ที่ตกอยู่ในห่วงรัก

จากนั้นประโยคที่ผ่านการคัดกรองแล้วจะถูกนำไปแปลงรูปเขียนเป็นรูปอ่าน ในการทดลองของงานวิจัยนี้ได้ใช้ ตัวแปลงรูปเขียนเป็นรูปอ่านของ Probabilistic Generalized LR parser [12] ในงานวิจัยนี้ได้ทำการประยุกต์วิธีการแปลงรูปเขียนเป็นรูปอ่านในเรื่องของการเพิ่มหน่วยเสียงเงียบ (Silence) เข้าไปในประโยคด้วยโดยการเพิ่มหน่วยเสียงเงียบไปที่ส่วนเริ่มของประโยค ส่วนที่เป็นวรรคของประโยค และส่วนท้ายของประโยค เทคนิคนี้สามารถช่วยเพิ่มรูปแบบหน่วยเสียงที่มีหน่วยเสียงเงียบเป็นส่วนประกอบได้ จากนั้นรูปเขียนของประโยคนั้น ๆ จะถูกเก็บเข้าสู่ฐานข้อมูล และในขั้นตอนนี้ประโยคออนไลน์ที่ไม่สามารถแปลงรูปเขียนเป็นรูปอ่านสำเร็จ จะถูกคัดออกอีกด้วยดังตารางที่ 3-4 ตัวอย่างผลลัพธ์ของประโยคที่แปลงรูปเขียนเป็นรูปอ่านสำเร็จในฐานข้อมูล ในภาพที่ 3-10 และจากตารางที่ 3-5



ภาพที่ 3-9 การแปลงประโยคจากรูปเขียนเป็นรูปอ่าน

Sentence_ID	Sentence_Text	Sentence_Label
1	กินฉมพรรษา หอมพรหมเทพ จุฑษรวิโรจน์ในอินทามัน มองเห็นดินแดนกระแสน้ำ	sil kin^ lo m^ cha ma wi w^ sil xx m^ ph @ ra ho m^ thee p^ sil c
2	เกาะสำหรับจับในมหาสมุทร พังงาส่วนกลางหลาย ทะเลก็สงบ	sil m @ sa m^ ra p^ ri a n^ ruu ma ha aa sa m u t^ sil ph a ng^ ng aa s a
3	ตั้งจะซิกอินตามัน มหาสมุทรอินเดีย แม่น้ำขบมิงทะเลและหมู่เกาะโดยง่ายได้เป็นสวน	sil ta ng^ pra ch it^ za n^ daa ma n^ sil ma ha aa sa m u t^ th a ra zi n
4	ส่วนแรก คือขบมิมายาไกลจากเกาะบูมึงท้ายเหมือง ส่วนที่สอง	sil s uua n^ xx k^ sil kh vv ch aa j^ fa ng^ jaa w^ kl a j^ caa k^ kh u ra t
5	คือหมู่เกาะที่หันไกล และส่วนสุดท้ายคืออ่าวพังงา หาดทราย หาดบนหมู่เกาะสิมิลัน	sil kh vv m u u k @ th ii ha ng^ kl a j^ sil l x s uua n^ s u t^ th aa j^ kh vv z
6	คือหาดทรายสวยสุดในประเทศไทย หากไม่มีความสะอาดและน้ำใสสะอาดก็เหมือน	sil kh vv h aa t^ s aa j^ s uua j^ s u t^ n a j^ pra thee t^ th a j^ sil h aa k^ i
7	ใกล้เขาหลัก คือความเปลี่ยนแปลงจากสันน้ำ ป่าชายเลน เส้นทางการจราจร	sil kl a j^ kh a w^ la k^ sil s v k^ s aa khw aa m^ pl i a n^ pl xx ng^ caa k^
8	เช่น อูระนุช ชาวพังงา จะล่องเรือพายคายักไปชายแดนในอำเภอสิมิลัน	sil ch ee n^ sil kh u ra b u r i i sil aa w^ ph a ng^ ng aa sil ca l @ ng^ r v
9	และเกาะกิ้งหรีด หมู่เกาะสุรินทร์หรือเกาะกิ้งหรีดในสิมิลันของไทย	sil n xx w^ pa ka a ra ng^ sil m u u k @ s u r i n^ kh vv n xx w^ pa ka a r
10	หมู่เกาะสิมิลันคือสุดขอบแนวปะการังน้ำลึก ยาวแห่งไปเข้าในอ่าวใหญ่	sil m u u k @ si m i l a n^ kh vv s u t^ j @ t^ n xx w^ pa ka a ra ng^ n a r
11	เช่น เกาะลันเต เกาะไข โดยนี้เรือจากกรุงเทพฯ ถึงสิมิลัน	sil ch ee n^ sil k @ d @ k^ m a j^ sil k^ kh a j^ sil d oo j a n a ng^ r v va
12	เกาะอ่าวพังงา ภูเขาพังงา เกาะตาปู เกาะพังกา เกาะรัง เกาะกูด เกาะกูดเกาะ	sil th iia w^ za a w^ ph a ng^ ng aa sil ch o m^ kh a w^ ph i ng^ ka n^ sil
13	สุดขอบเข้าใจในอ่าวตังเกี๋ยหรือจะไปทำรังในอ่าวตังเกี๋ย	sil s u t^ s a n u k^ ra w^ ca j^ n a j^ wee laa ph iia ng^ n v ng a w a n^ s
14	กระดิ่งเกาะกูดในทะเลไทย กระดิ่งเกาะกูดว่าสวยงาม ผสมผสานหมู่เกาะกัน	sil k ra b i i k @ m aa k^ s u t^ n a j^ th a lee th a j^ sil k ra b i i k @ m a
15	อันนาคายบลังก์ มีน้ำป่าชายเลนอุดมสมบูรณ์ แหล่งน้ำจืดขนาดใหญ่ในทะเล	sil za th i b aa j^ j a ph aa j^ la ng^ ca sil j a ng^ m i i pa a ch aa j^ lee n a
16	ประติมากรรมปูนปั้นในทะเลไทยงดงามยิ่งนัก ชายฝั่งทะเลอันดามัน	sil pra wa t^ s aa t^ j @ n^ kl a p^ a j^ th v ng^ kh o n^ th a j^ j a kh
17	หาดทราย หาดบนหมู่เกาะสุรินทร์หรือเกาะกูด หาดทรายที่สวยงาม	sil h aa t^ s aa j^ sil h aa t^ n a ph ra t^ th aa ra m @ s a m^ ra p^ s v k^
18	ถ้าต้องการชมทะเลในทะเล หาดสวยสุดก็อยู่บนเกาะสิมิลันไทย ไม่รวมเกาะอื่น	sil th aa t @ ng^ kaa n^ za a r o m^ r o o m xx n^ t i k^ sil h aa t^ s uua j^
19	หรือจะไปชมทะเลในทะเล หาดสวยสุดก็อยู่บนเกาะสิมิลันไทย ไม่รวมเกาะอื่น	sil ph ri m^ r o o t^ s i l v va ng^ s i i ch o m^ ph u u p x t^ sil b i t^ t @ t i
20	แม้ว่าพืชพรรณจะไม่เหมือนกับในแผ่นดิน แต่ลักษณะและกลิ่นก็เหมือน	sil m xx w aa ph vv t^ ch a ph uua k^ n i i ca phl oo naa ka n^ kh v n^ m a

ภาพที่ 3-10 ตัวอย่างประโยคออนไลน์ในรูปแบบเขียนและรูปอ่านที่เก็บไว้ในฐานข้อมูล

ตารางที่ 3-5 ตัวอย่างผลการแปลงรูปเขียนเป็นรูปอ่านของประโยคออนไลน์

ประโยคในรูปเขียน	ประโยคในรูปอ่าน
เธอขอให้เพลงจบ จะเล่าเรื่องเด็กหนุ่มนักเปียโนให้สามีพัง แต่ประพจน์เหมือนไม่ได้ยินอะไรทั้งสิ้น	sil th qq r @ @ h a j^ phl ee ng a c o p^ sil c a l a w^ r vva ng^ de k^ n u m^ n a k^ p iia n oo h a j^ s aa m i i f a ng^ sil t xx pra ph o t^ m vva n^ m a j^ d a j^ j i n^ z a r a j^ th a ng^ s i n^ sil
เจอแล้ว ผมร้องออกมาแล้วก็เปิดประตู ล้วงเข้าไป สัมผัสกระทั้งว่าจะต้องเคาะประตูก่อน	sil c qq l xx w^ sil ph o m^ r @ @ ng^ z @ @ k^ m aa l xx w^ k @ @ p qq t^ pra t u u phl ua kh a w^ p a j^ sil l vv m^ m xx kr a th a ng^ w aa c a t @ @ ng^ kh @ pra t u u k @ @ n^ sil
ทั้งที่จัดของเกือบเสร็จเรียบร้อย ของฝากญาติโกโหติกา ผองเพื่อนก็เตรียมไว้เสร็จสรรพ	sil th a ng^ th ii c a t^ kh @ @ ng^ k vva p^ s e t^ r iia p^ r @ @ j^ sil kh @ @ ng^ f aa k^ j aa t^ k oo h oo t i k aa sil ph @ @ ng^ ph vva n^ k @ @ tr iia m^ w a j^ s e t^ s a p^ sil
อ้าว คุณปู่ตอบเกมหัวเราะ นี่แหละบทกวีของกวีเอกชาว อิตาลี	sil z aa w^ sil kh u n^ p u u t @ @ p^ k xx m^ h uua r @ sil n i i l x b o t^ kw ii kh @ @ ng^ kw ii z ee k^ ch aa w^ sil z i t aa l i i sil
กบบัวสะบัดปาก ปล่อยเหยื่อปลา นกกวัก กู้ฟาลันบึงน้ำ กระสาขาส่ายคล้ายยีนสัน	sil k aa p^ b uua s a b a t^ p aa k^ sil pl @ @ j^ j vva pl aa sil n o k^ k a w a k^ k u u f aa l a n^ b v ng^ n a m^ sil kr a s aa kh aa s aa j^ khl aa j^ j a j vv n^ s a n^ sil
ถ้าฮิตเลอร์ไม่หันปลายปืนไปทางมอสโก เยอร์มันก็คงไม่แพ้สงคราม	sil th aa h i t^ l qq m a j^ h a n^ pl aa j^ p vv n^ p a j^ th aa ng^ m @ @ t^ k oo sil j qq r a m a n^ k @ @ kh o ng^ m a j^ ph xx s o ng^ khr aa m^ sil
ล้างจาน ปิดไฟนอน คืนนี้ทั้งคู่วมรักกัน อย่างดูเดือด เป่ญคว่หน้า	sil l aa ng^ c aa n^ sil p i t^ f a j^ n @ @ n^ sil kh vv n^ n a n^ th a ng^ kh uu r uua m^ r a k^ k a n^ j aa ng^ d u d

ประโยคในรูปเขียน	ประโยคในรูปอ่าน
	vva t ^h sil p a j ee khw a m ^h n aa sil
พี่จิวทะเลาะกับพี่พาย พี่สะใภ้มีเวรดูแลคุณพ่อ แต่เผลอหลับจนน้ำเกลือหมดขวด	sil ph ii c i w ^h th a l @ k a p ^h ph ii ph aa j ^h sil ph ii s a ph a j ^h m ii w ee n ^h d uu l xx kh u n ^h ph @@ sil t xx phl qq l a p ^h c o n ^h n a m ^h kl vva m o t ^h kh uua t ^h sil
ปรอดปราดหลีกเสียปีกบัต สาลิกากูร้องก่อนล่องลัด เขาชาวคล้ายหัวหวาดขันรื่อง	sil pr @@ t ^h pr aa t ^h l ii k ^h s iia p ii k ^h b a t ^h sil s aa l i k aa k uu r @@ ng ^h k @@ n ^h l @@ ng ^h l a t ^h sil kh a w ^h ch a w aa khl aa j ^h w a t ^h w aa t ^h kh a n ^h r @@ ng ^h sil
กลับกลายเป็นตรงกันข้าม ต่อมาทั้งสองได้รับความช่วยเหลือจากเบดูอินกลุ่มหนึ่ง	sil kl a p ^h kl aa j ^h p e n ^h tr o ng ^h k a n ^h kh aa m ^h sil t @@ m aa th a ng ^h s @@ ng ^h d a j ^h r a p ^h khw aa m ^h ch uua j ^h l vva c aa k ^h b ee d uu z i n ^h kl u m ^h n v ng ^h sil
แบบว่าขัดขี้โคลในเวลาเดียวกัน แต่ก็หายเร็วกว่าความนุ่มนวลโดยการทาสีแบบของยาย	sil b xx b a w aa kh a t ^h kh ii khl a j ^h n a j ^h w ee l aa d iia w ^h k a n ^h sil t xx k @@ h aa j ^h r e w ^h kw aa khw aa m ^h n u m ^h n uua n ^h d oo j ^h k aa n ^h th aa kh a m i n ^h b xx p ^h kh @@ ng ^h j aa j ^h sil
เหตุการณ์ได้เข้มข้นยิ่งขึ้นเมื่อทั้งสองถูกพวกโจรเบดูอินปล้นและทำร้ายบาดเจ็บ	sil h ee t ^h t u k aa n ^h d a j ^h kh ee m ^h kh o n ^h n a j i ng ^h kh v n ^h m vva th a ng ^h s @@ ng ^h th uu k ^h ph uua k ^h c oo n ^h b ee d uu z i n ^h pl o n ^h l x th a m ^h r aa j ^h b aa t ^h c e p ^h sil
ผมจะทำไม้ตีป่า บอกหน่อย จะบ้าตายอยู่แล้ว จะทำอะไรน้องป้อปก็ตามประจบจนผมทำอะไรไม่ถูกแล้ว	sil ph o m ^h c a th a m ^h ng a j ^h d ii p aa sil b @@ k ^h n @@ j ^h sil c a b aa t aa j ^h j uu l xx w ^h sil c a th a m ^h z a r a j ^h n @@ ng ^h p @@ p a k @@ t aa m ^h pr a k o p ^h c o n ^h ph o m ^h th a m ^h z a r a j ^h m a j ^h th uu k ^h l xx w ^h sil

Sentence Segment

Data_ID: 10

Source: <http://ling.arts.chula.ac.th/>

Word Boundaries Ok...!

Word: 493/493 Units.

Sentences: 19/19 Utterances.

Sentences Boundaries Ok...!

Filter: 3 Utterances.

Sentences: 16/19 Utterances.

Sentences Filter Ok...!

G2P Convert Ok...!

Input Text

การดื่มเหล้ามีหลายชนิดทั้งที่ปรุงและกลั่นแล้ว แต่วิธีกลั่นของโปรตุเกสจะดื่มไป แทนที่จะดื่มในกระป๋องเหมือนคนอื่นเขา เพราะคนโปรตุเกสจะดื่มเหล้าในกระป๋องแล้วจะดื่มในกระป๋องอีกทีหนึ่ง... (The text continues with a description of drinking habits and the G2P process.)

Selected Sentences

การดื่มเหล้ามีหลายชนิดทั้งที่ปรุงและกลั่นแล้ว แต่วิธีกลั่นของโปรตุเกสจะดื่มไป แทนที่จะดื่มในกระป๋องเหมือนคนอื่นเขา เพราะคนโปรตุเกสจะดื่มเหล้าในกระป๋องแล้วจะดื่มในกระป๋องอีกทีหนึ่ง... (The text continues with a description of drinking habits.)

Filter Out

การดื่มเหล้ามีหลายชนิดทั้งที่ปรุงและกลั่นแล้ว แต่วิธีกลั่นของโปรตุเกสจะดื่มไป แทนที่จะดื่มในกระป๋องเหมือนคนอื่นเขา เพราะคนโปรตุเกสจะดื่มเหล้าในกระป๋องแล้วจะดื่มในกระป๋องอีกทีหนึ่ง... (The text continues with a description of drinking habits.)

G2P

sil k s s n^h d v v m^h n o m^h t h i m m i c u l^h m a a j^h p h q q m^h p r o o t i i n^h l x k h x n^h s i a m^h k h a w^h n a j^h t u a s i l t x m i s a s^h s u a n^h k h @ @ n^h s i l t x x n^h t h i c a d a j^h p h o n^h k a j^h k o o n^h p @ @ n g^h k a n^h k r a d u u k^h p h u n^h s i l c v n g^h z a a t^h k q q i^h p h o n^h n o j^h t h a a n g^h k l a p^h k a s i s^h @ n g^h w a n^h p h o n^h l a n g^h b o i^h k h w s o m^h k i e w^h k a p^h k h @ @ m u n^h m o j^h k h @ @ n g^h n o m^h s i l p h v a k a a n^h p h i c a r a s i l p e n^h s i l p e n^h c a a k^h p h u u z a a n^h t h i m m i z a a w s o o s i l t h a k^h t h u o n k g^h m a a k h w a a n^h w a a t h a a n^h h a j^h j u u l x w^h s i l n s i l t x s a m^h r a p^h r v v a n g^h n o m^h s i l n a m a j a n g^h p e n^h z a a h a a n^h z u i^h t a s k h u n^h k h a a s i l k o n^h j i p u n^h t u o u s u n k g^h j i p u n^h k h v n s i r a m^h r a p^h k h o n^h t h a j^h j a n g a n g^h m i z i k^h m a a k^h t h i k h a a l^h z a a h a a n^h s i l c v n g^h m a j^h k h u a n^h l o p^h l a n g^h k h w a a n^h d i k h s i l r a l^h t h a b a a n^h c h u a n^h k @ @ t h u m^h t h o e n g o p^h p r a m a a n^h p a j^h j x p h v a s o n g^h s o d m^h h a j^h k h o n^h t h a j^h d v v m^h n o m^h s i l s i l k h a z a n^h @ m^h r a p^h n a j^h k h @ @ k h i t^h c a a k^h p h u u z a a n^h s i l n o m^h p e n^h z a a h a a n^h m i p r a j o o^h r a w^h t u u k a n^h m a a g a a s i l p h o m^h z e e n g^h k @ @ j a n g h a j^h l u u k^h d v v m^h n o m^h s o t^h n a j^h w a j^h t h i k a m^h l a n g^h c a r q n^h l q q p^h i o o s i l t x n a n^h l x k h r a s i l d a a n^h t h i p e n^h k h @ @ d i s i l k a p^h d a a n^h t h i p e n^h k h @ @ s i a s i l m x t x r v v a n g^h n o m^h s i l n o m^h s i l s a m^h k h a n^h j u u t h i w a a s i l k l a a w^h k a n^h w a a n a j^h s a m a j^h t h i p r a c h a a k @ @ r a s u a n^h j a n a j^h s o n g^h k h a m a j a n g^h d @ @ i s v n g^h p h i l a n g^h w i k h r f

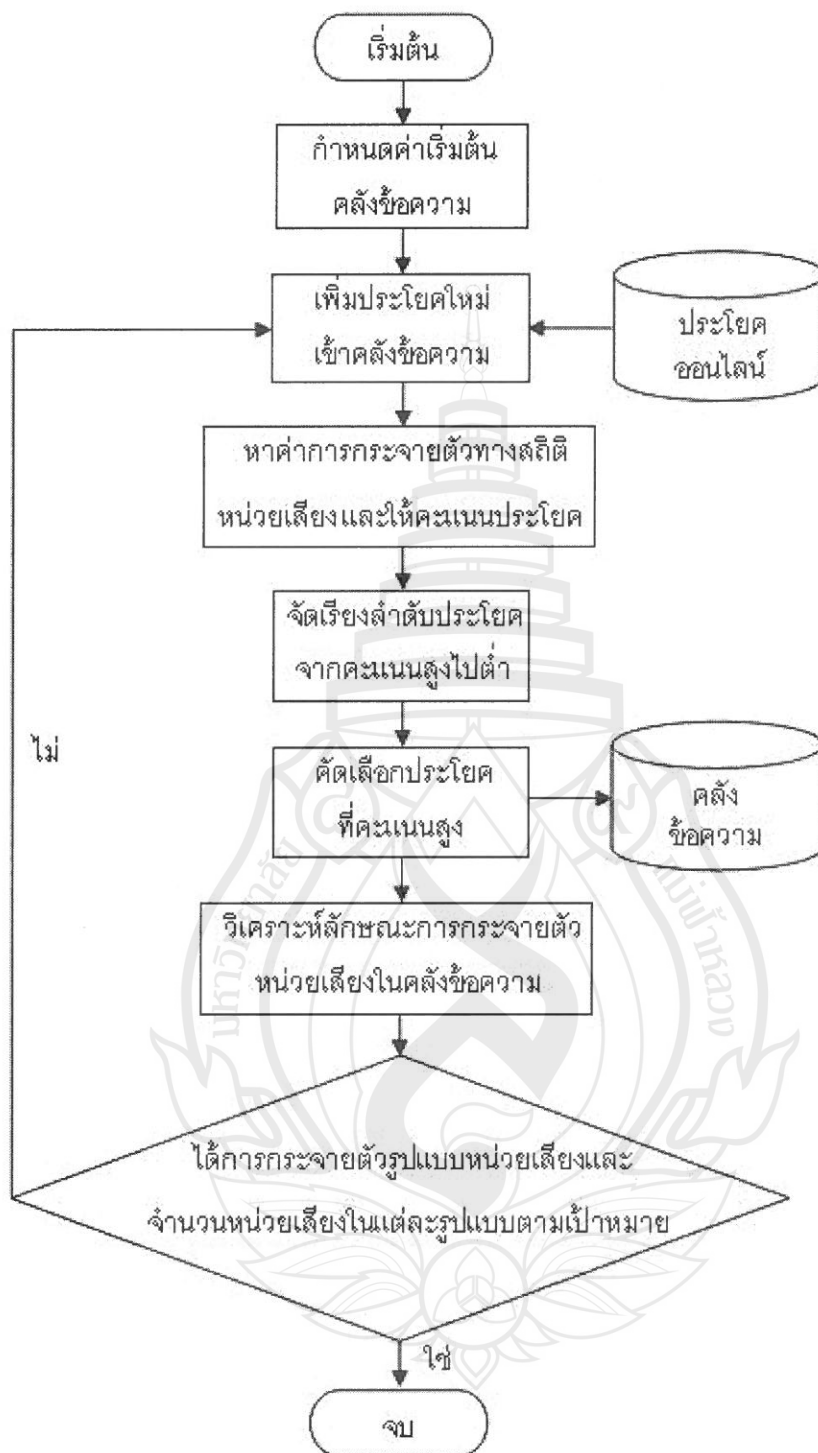
Sentence Segment by Surapol V.

ภาพที่ 3-11 หน้าต่างโปรแกรมขั้นตอนการสร้างประโยคออนไลน์

หลังจากที่บทความจากฐานข้อมูลข้อความออนไลน์ได้ถูกจัดเก็บลงสู่ฐานข้อมูลรูปแบบประโยคและแปลงรูปเขียนของแต่ละประโยคเป็นรูปอ่านแล้ว รูปอ่านเหล่านี้จะถูกนำไปวิเคราะห์และให้คะแนนเพื่อคัดเลือกประโยคดังกล่าวเข้าสู่คลังข้อความที่ต้องการสร้าง ในขั้นตอนการเลือกประโยคและจัดเก็บข้อมูลต่อไป สามารถดูหน้าต่างโปรแกรมนี้จากภาพที่ 3-11

3.2.3 ขั้นตอนการเลือกประโยคและจัดเก็บคลังข้อความ

ขั้นตอนการเลือกประโยคและจัดเก็บคลังข้อความนี้ เป็นขั้นตอนที่สำคัญที่สุดในการทำงานวิจัยนี้ ขั้นตอนนี้มีหน้าที่สร้างฐานข้อมูลข้อความตามที่ต้องการ โดยการนำเอารูปเขียนของแต่ละประโยคจากขั้นตอนการสร้างประโยคออนไลน์ มาให้คะแนนแต่ละประโยค ขั้นตอนวิธีเชิงละโมภ (Greedy Algorithm) ถูกนำมาใช้ในการให้คะแนนดังกล่าว เพื่อให้คลังข้อความที่ถูกสร้างมีจำนวนรูปแบบหน่วยเสียงที่ต้องการและให้การกระจายตัวทางหน่วยเสียงของคลังข้อความที่ถูกสร้างขึ้นคล้ายกับการกระจายตัวทางหน่วยเสียงของต้นแบบมากที่สุด ประโยคที่ได้คะแนนสูงจะถูกดึงเข้าสู่ฐานข้อมูล และประโยคที่ได้คะแนนต่ำจะถูกดึงออกไป ขั้นตอนนี้จะถูกทำแบบวนซ้ำไปเรื่อย ๆ จนกระทั่งคลังข้อความที่ต้องการสร้าง มีคุณสมบัติตรงกับสิ่งที่กำหนด แผนภาพการทำงานของขั้นตอนการเลือกประโยคและการจัดเก็บลงฐานข้อมูลนี้ สามารถดูได้จากภาพที่ 3-12



ภาพที่ 3-12 แผนภาพขั้นตอนการเลือกประโยคและจัดเก็บคลังข้อความ

ขั้นตอนการเลือกประโยคและจัดเก็บข้อมูล เริ่มจากการกำหนดค่าเริ่มต้นต่าง ๆ เพื่อใช้ในการสร้างคลังข้อความที่ต้องการ สิ่งที่ต้องกำหนดในขั้นตอนนี้ใช้ในการ กำหนดรูปแบบของคลังข้อความที่ต้องการสร้าง ซึ่งได้แก่

1. **รูปแบบของหน่วยเสียง** คือ ชนิดของรูปแบบหน่วยเสียงที่จะใช้ในการคำนวณทางสถิติในการเลือกประโยค เช่น หน่วยเสียงเดี่ยว (Monophone) หน่วยเสียงคู่ (Diphone) หน่วยเสียงสาม (Triphone) หรือหน่วยเสียงรูปแบบอื่น ๆ สำหรับงานวิจัยนี้ได้ใช้ รูปแบบหน่วยเสียงชนิดหน่วยเสียงคู่ ในการทดลอง
2. **จำนวนขั้นต่ำของรูปแบบหน่วยเสียง** คือ จำนวนของแต่ละรูปแบบหน่วยเสียงขั้นต่ำที่ต้องการให้เกิดขึ้นในคลังข้อความที่ต้องการสร้าง สำหรับในการทดลองของงานวิจัยนี้ได้กำหนดให้รูปแบบของหน่วยเสียงเกิดอย่างน้อยหนึ่งครั้ง
3. **ขอบเขตของจำนวนพยางค์ในแต่ละประโยค** คือ การกำหนดความยาวของประโยคออนไลน์ด้วยจำนวนพยางค์ ซึ่งในการทดลองของงานวิจัยนี้ ในแต่ละประโยคกำหนดให้ประกอบด้วยจำนวนพยางค์ที่มากกว่า 20 พยางค์แต่ไม่เกิน 40 พยางค์
4. **ขนาดของจำนวนประโยคในคลังข้อความ** คือ จำนวนของประโยคที่จะเกิดขึ้นในคลังข้อความที่ต้องการ จำนวนของประโยคในคลังข้อความนี้มีผลต่อจำนวนความครอบคลุมของรูปแบบหน่วยเสียง โดยถ้าจำนวนประโยคน้อยเกินไปอาจทำให้ความครอบคลุมของแต่ละรูปแบบหน่วยเสียงและจำนวนขั้นต่ำของรูปแบบหน่วยเสียง ไม่ได้ตามเป้าหมายที่ต้องการได้ สำหรับการทดลองในงานวิจัยนี้ได้กำหนดจำนวนประโยคในคลังข้อความไว้ที่ 1,000 ประโยค
5. **ค่าเบี่ยงเบนเฉลี่ยร้อยละจากเป้าหมาย** คือ ค่าผลต่างที่ยอมรับได้ของการกระจายตัวรูปแบบหน่วยเสียงคลังข้อความที่กำลังสร้างเทียบกับการกระจายตัวรูปแบบหน่วยเสียงของเป้าหมายในรูปแบบของร้อยละ ในที่นี้ถ้าค่าเบี่ยงเบนเฉลี่ยจากเป้าหมายมีค่าน้อยจะทำให้คลังเสียงที่ต้องการสร้างมีการกระจายตัวทางหน่วยเสียงที่คล้ายกับเป้าหมายมากขึ้น แต่อย่างไรก็ตาม ถ้ากำหนดค่าเบี่ยงเบนเฉลี่ยจากเป้าหมายน้อยเกินไปอาจทำให้การสร้างคลังข้อความเป็นไปได้ยากหรือใช้ระยะเวลามากไปได้ ในการทดลองของงานวิจัยนี้ค่าเบี่ยงเบนเฉลี่ยจากเป้าหมายที่ ร้อยละ 3.0

เมื่อกำหนดรูปแบบของคลังข้อความที่ต้องการแล้ว ขั้นตอนต่อไประบบจะทำการดึงประโยคจากฐานข้อมูลประโยคออนไลน์เพิ่มเข้าไปในฐานข้อมูลวิเคราะห์ประโยค ฐานข้อมูลนี้เป็นฐานข้อมูลชั่วคราว ใช้ในการคำนวณค่าทางสถิติทางการกระจายตัวหน่วยเสียง เพื่อหาความต้องการในแต่ละรูปแบบของหน่วยเสียงเพื่อให้คะแนนประโยค และใช้ในการหาค่าเบี่ยงเบนเฉลี่ยจากเป้าหมาย วิธีเพิ่มประโยคออนไลน์เข้าไปในฐานข้อมูลวิเคราะห์ประโยคนี้ ในรอบแรกของขั้นตอนนี้เราจะเพิ่มไปเท่ากับ 110% ของขนาดคลังข้อความ เช่นถ้าคลังข้อความที่ต้องการสร้างกำหนดขนาดไว้ที่ 1,000 ประโยคในการเพิ่มประโยคครั้งแรก 1,100 ประโยค และในรอบต่อไป

ของขั้นตอนนี้จะเราเพิ่มประโยคเข้าไปอีกทีละ 10% ของจำนวนคลังข้อความที่ต้องการสร้างเช่น ถ้าคลังข้อความมีขนาด 1,000 จะเพิ่มประโยคเข้าไปอีก 100 เป็นต้น ทั้งนี้อีก 10% ของประโยคจะถูกคัดเลือกทั้งไปในขั้นตอนการคัดเลือกประโยค เมื่อเพิ่มประโยคเข้าไปแล้วระบบจำทำการคำนวณค่าทางสถิติของการกระจายตัวทางหน่วยเสียง จากฐานข้อมูลวิเคราะห์ประโยค ตัวอย่างข้อมูลทางสถิติของการกระจายตัวทางหน่วยเสียงดังกล่าว ดูได้จากตารางที่ 3-6

ข้อมูลทางสถิติของการกระจายตัวทางหน่วยเสียงนี้ ใช้เพื่อการพิจารณาในการหาค่าความต้องการในแต่ละรูปแบบของหน่วยเสียงและใช้ในการหาค่าเบี่ยงเบนเฉลี่ยจากเป้าหมาย ซึ่งได้มาจากการการนำรูปอ่านของประโยคออนไลน์ในคลังเสียงที่กำลังสร้างมาแจกแจงและนับจำนวนเทียบกับรูปแบบหน่วยเสียงของเป้าหมาย ข้อมูลทางสถิติทั้งหมดนี้ ประกอบด้วย

1. **Patterns** คือ รูปแบบของหน่วยเสียงที่ปรากฏ สร้างมาจากขั้นตอนการสร้างการกระจายตัวทางหน่วยเสียงเป้าหมายได้มาจากการใช้ฟังก์ชัน HLEd จากเครื่องมือ Hidden Markov Model toolkit, HTK [31] ในการหารูปแบบของหน่วยเสียงคู่ ที่เกิดขึ้นจากรูปอ่านทั้งหมดจากคลังข้อความเป้าหมาย ในที่นี้ปรากฏรูปแบบหน่วยเสียงคู่ที่เกิดขึ้น 1,383 หน่วยเสียง
2. **Target_Freq** คือ จำนวนความถี่ที่เกิดขึ้นของหน่วยเสียงในเป้าหมาย สร้างมาจากขั้นตอนการสร้างการกระจายตัวทางหน่วยเสียงเป้าหมายเช่นกัน โดยใช้ฟังก์ชัน HLStats จากเครื่องมือ Hidden Markov Model toolkit, HTK [31] ในการนับจำนวนที่เกิดขึ้นในแต่ละรูปแบบของหน่วยเสียงในคลังข้อความเป้าหมาย
3. **Target_Prob** คือ โอกาสที่รูปแบบของแต่ละหน่วยเสียงจะเกิดขึ้นในคลังข้อความเป้าหมาย เพื่อให้อัตราการเกิดรูปแบบหน่วยเสียงของเป้าหมายและอัตราการเกิดรูปแบบหน่วยเสียงในคลังข้อความที่ต้องการสร้างอยู่ในบรรทัดฐานเดียวกัน ใช้ในการเปรียบเทียบการกระจายตัวรูปแบบหน่วยเสียงของทั้งสอง คำนวณได้จากสมการที่ (1)

ตารางที่ 3-6 ข้อมูลทางสถิติการกระจายตัวทางหน่วยเสียงที่ใช้ในการคำนวณจากฐานข้อมูล

Patterns	Target_Freq	Target_Prob	Custom_Freq	Custom_Prob	Weight	Delta_Prob	Deviation
ng ⁺ z	269	-2.89743	73	-2.98	0.0825672	0.0825672	0.0284967
ph+uu	517	-2.61407	162	-2.63543	0.0213612	0.0213612	0.00817161
b+u	157	-3.1307	17	-3.60325	0.472546	0.472546	0.15094
n+v	282	-2.87697	92	-2.88014	0.00317279	0.00317279	0.00110282
vv+s	64	-3.51843	26	-3.42304	-0.0953913	0.0953913	0.0271119
t ⁺ +ng	11	-4.26729	13	-3.71595	-0.551339	0.551339	0.129201
ii+c	373	-2.75569	61	-3.05741	0.301719	0.301719	0.10949
m ⁺ +pr	70	-3.4798	16	-3.6288	0.149001	0.149001	0.0428187
@+s	16	-4.1105	8	-3.91687	-0.193634	0.193634	0.0471073
k ⁺ +k	311	-2.83453	91	-2.88486	0.0503334	0.0503334	0.0177572
i+d	35	-3.77776	15	-3.65595	-0.121807	0.121807	0.0322432
ee+l	84	-3.40113	29	-3.37646	-0.0246675	0.0246675	0.00725273
i+kh	16	-4.1105	6	-4.03337	-0.0771288	0.0771288	0.0187639
m ⁺ +m	290	-2.86484	133	-2.7208	-0.144037	0.144037	0.0502774
uu+l	61	-3.53911	22	-3.4941	-0.045008	0.045008	0.0127173

$$P(Target_i) = \log_{10} \frac{F_i + 0.5}{n + 0.5} \quad (1)$$

โดยที่ F_i คือ จำนวนความถี่ที่รูปแบบหน่วยเสียงที่ i เกิดขึ้นในคลังข้อความเป้าหมาย และ n คือผลรวมทั้งหมดของหน่วยเสียงในคลังข้อความเป้าหมาย และด้วยลักษณะการกระจายตัวของหน่วยเสียงในภาษาธรรมชาติมีความสัมพันธ์กันแบบลอการิทึม จึงใช้ฟังก์ชันลอการิทึมฐานสิบ ส่วนค่าคงที่ 0.5 ใช้ในการบังคับค่าจากการคำนวณไม่ให้เกิดความผิดพลาดในกรณีที่ $F_i = 0$

4. **Custom_Freq** คือ จำนวนความถี่ที่เกิดขึ้นของหน่วยเสียงในคลังเสียงที่กำลังสร้าง สร้างจากการนับจำนวนที่เกิดขึ้นในแต่ละรูปแบบของหน่วยเสียงในคลังข้อความที่กำลังสร้าง โดยใช้ ฟังก์ชัน HLStats จากเครื่องมือ Hidden Markov Model toolkit, HTK [31]
5. **Custom_Prob** คือ โอกาสที่หน่วยเสียงนี้เกิดขึ้นในคลังข้อความที่กำลังสร้าง คำนวณได้จากสมการที่ (2)

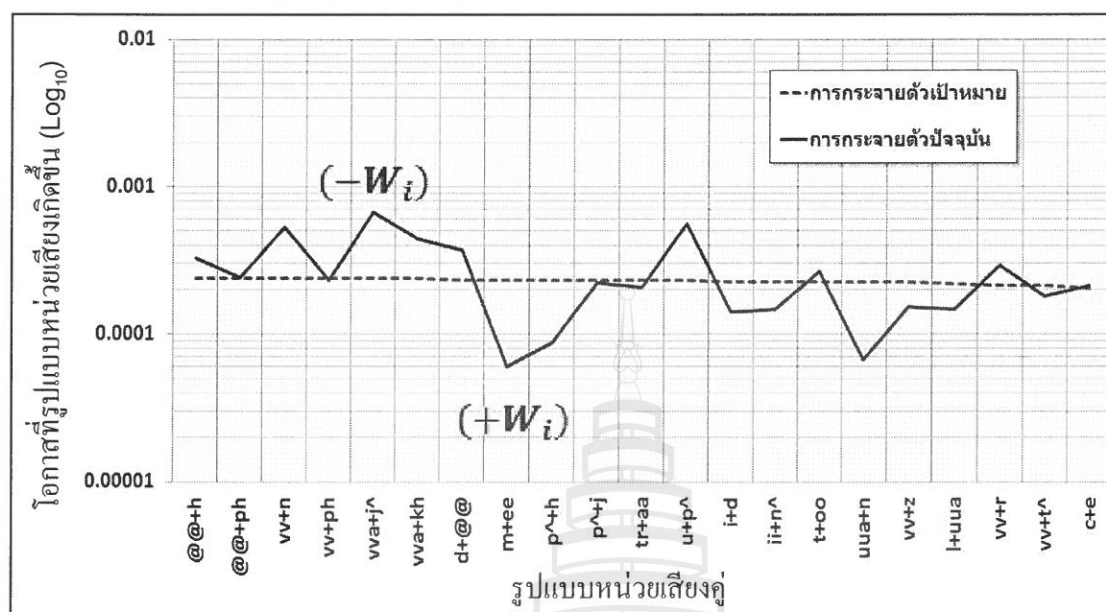
$$P(Custom_i) = \log_{10} \frac{F_i + 0.5}{n + 0.5} \quad (2)$$

โดยที่ F_i คือ จำนวนความถี่ที่รูปแบบหน่วยเสียงที่ i เกิดขึ้นในคลังข้อความที่กำลังสร้าง และ n คือผลรวมทั้งหมดของหน่วยเสียงในคลังข้อความที่กำลังสร้าง

6. **Weight** คือ ค่าของความต้องการในของรูปแบบหน่วยเสียง เกิดจากผลต่างของโอกาสการเกิดของรูปแบบหน่วยเสียงนั้น การคำนวณค่าของความต้องการรูปแบบหน่วยเสียงสำหรับงานวิจัยนี้มีสองกรณี กรณีแรกคือ จำนวนของรูปแบบหน่วยเสียงผ่านเกณฑ์ขั้นต่ำที่กำหนดแล้ว จะคำนวณค่าของความต้องการรูปแบบหน่วยเสียงได้จากสมการที่ (3)

$$W_i = P(Target_i) - P(Custom_i) \quad (3)$$

โดยที่ $P(Target_i)$ คือ โอกาสที่รูปแบบของหน่วยเสียงที่ i เกิดขึ้นในคลังข้อความเป้าหมาย และ $P(Custom_i)$ คือ โอกาสที่รูปแบบของหน่วยเสียงที่ i เกิดขึ้นในคลังข้อความที่กำลังสร้าง ถ้าผลของ W_i เป็นบวก (+) แสดงว่าหน่วยเสียงนั้นในคลังข้อความที่ต้องการสร้างยังขาดอยู่ ดังนั้นระบบจึงต้องการรูปแบบหน่วยเสียงนี้เพิ่มอีก แต่ถ้าค่าของ W_i ออกมาเป็นค่าลบ (-) แสดงว่าหน่วยเสียงนั้นในคลังข้อความที่ต้องการสร้าง มีมากเกินไป ระบบจึงต้องการลดรูปแบบหน่วยเสียงนี้ออกเพื่อความเข้าใจมากขึ้นสามารถดูภาพที่ 3-13 ประกอบได้



ภาพที่ 3-13 ใช้ค่าความต้องการรูปแบบหน่วยเสียงในการให้คะแนนประโยค

ส่วนการคำนวณค่าความต้องการรูปแบบหน่วยเสียง ในกรณีที่สอง คือ จำนวนรูปแบบหน่วยเสียงนั้นยังต่ำกว่าเกณฑ์ขั้นต่ำที่กำหนด เนื่องจากเราต้องการรักษาหน่วยเสียงที่หายากไว้เพื่อให้ได้ความครอบคลุมหน่วยเสียงตามเป้าหมาย การคำนวณค่าของความต้องการรูปแบบหน่วยเสียง จากสมการที่ (4) ในการคำนวณ

$$W_i = |P(\text{Custom}_i)| \times \alpha \quad (4)$$

หลังจากทำการหาค่าสมบูรณ์ของโอกาสที่หน่วยเสียงนี้เกิดขึ้นในคลังข้อความที่กำลังสร้างแล้ว เราจะทำการการเพิ่มระดับความสำคัญให้รูปแบบหน่วยเสียงหายาก ด้วยการคูณด้วยค่าคงที่ α ถ้ากำหนดให้ค่า α มีค่ามากจะทำให้คลังข้อความที่ถูกสร้างเข้าถึงการครอบคลุมทางหน่วยเสียงเป้าหมายได้เร็วขึ้น แต่อย่างไรก็ตามถ้ากำหนดค่า α มากเกินไปจะทำให้ การกระจายตัวของคลังข้อความที่ต้องการสร้างเป็นไปตามเป้าหมายยากมากขึ้น ในงานวิจัยนี้ได้ใช้ ค่า α ที่ 1.5

7. Delta_Prob คือ ผลต่างระหว่างโอกาสที่เกิดขึ้นของรูปแบบหน่วยเสียงเป้าหมายและปัจจุบัน ใช้เพื่อการคำนวณค่าเบี่ยงเบนจากเป้าหมายของแต่ละหน่วยเสียง คำนวณจากสมการที่ (5)

$$\Delta P_i = |P(\text{Target}_i) - P(\text{Custom}_i)| \quad (5)$$

โดยที่ $P(Target_i)$ คือ โอกาสที่รูปแบบของหน่วยเสียงที่ i เกิดขึ้นในคลังข้อความเป้าหมาย และ $P(Custom_i)$ คือ โอกาสที่รูปแบบของหน่วยเสียงที่ i เกิดขึ้นในคลังข้อความที่กำลังสร้าง

8. **Deviation** คือ ผลต่างระหว่างโอกาสที่เกิดขึ้นของรูปแบบหน่วยเสียงเป้าหมายและปัจจุบัน เทียบกับโอกาสที่เกิดขึ้นของเป้าหมาย ใช้คำนวณ ค่าเบี่ยงเบนเฉลี่ยร้อยละจากเป้าหมาย (Percent average deviation of target) ดูได้จากสมการที่ (6)

$$\sigma_i = \frac{\Delta P_i}{P(Target_i)} \quad (6)$$

โดยที่ σ_i คือ ผลต่างระหว่างโอกาสที่เกิดขึ้นของรูปแบบหน่วยเสียงเป้าหมายและปัจจุบัน และ $P(Target_i)$ คือ โอกาสที่รูปแบบของหน่วยเสียงที่ i เกิดขึ้นในคลังข้อความเป้าหมาย ถ้าค่าของ σ_i มีค่ามากรูปแบบหน่วยเสียงนั้นห่างจากเป้าหมายอยู่มากและถ้าค่าของ σ_i มีค่าน้อย รูปแบบหน่วยเสียงนั้นเข้าใกล้เป้าหมายมากขึ้น

ขั้นตอนต่อไปคือการให้คะแนนประโยชน์จากข้อมูลการกระจายตัวทางสถิติ เพื่อใช้พิจารณาในการเลือกประโยชน์ที่จะเก็บไว้ในคลังข้อความที่ต้องการสร้าง การให้คะแนนประโยชน์ในงานวิจัยนี้ ได้ใช้สมการที่ (7) ในการให้คะแนน

$$Score_c = \sum_{i=1}^n \frac{F_i \times W_i}{L_c} \quad (7)$$

ในที่นี้ F_i คือ จำนวนรูปแบบของหน่วยเสียง i ที่เกิดในคลังข้อความที่กำลังสร้างปัจจุบัน ทั้งนี้ในประโยชน์ที่กำลังพิจารณาอาจไม่ปรากฏรูปแบบของหน่วยเสียงที่ i ก็ได้ ส่วน W_i คือ ค่าของความต้องการในของรูปแบบหน่วยเสียงนั้นจากสมการที่ (3) และ L_c คือ ความยาวของประโยชน์นับจากผลรวมของจำนวนหน่วยเสียงในประโยชน์ c ค่า L_c ใช้ในเพื่อทำให้การให้คะแนนไม่ขึ้นอยู่กับความยาวของประโยชน์ n คือจำนวนรูปแบบของหน่วยเสียงทั้งหมดที่ปรากฏในคลังข้อความเป้าหมาย สำหรับคะแนนประโยชน์นี้ ถ้าค่า $Score_c$ ในประโยชน์มีค่ามาก ประโยชน์นั้นจะประกอบด้วยรูปแบบหน่วยเสียงที่ต้องการหรือหน่วยเสียงที่ขาดอยู่มาก แต่ถ้าประโยชน์มีค่าน้อย ประโยชน์นั้นอาจประกอบด้วยรูปแบบหน่วยเสียงไม่จำเป็นหรือมีมากเกินไป สำหรับการสร้างคลังข้อความที่ต้องการ สามารถดูตัวอย่างผลการให้คะแนนประโยชน์ได้จากตารางที่ 3-7

ตารางที่ 3-7 ผลคะแนนประโยชน์จากขั้นตอนการให้คะแนน

ประโยชน์ในฐานข้อมูล	คะแนน
ถึงจะเหนื่อย ฉันก็ไม่มีความเหนื่อย ฉันจะบุกไปจนกว่าจะถึงบึงหญ้าใหญ่	0.1113

ประโยคในฐานข้อมูล	คะแนน
ในแง่สุขภาพทางจิต ปรากฏว่า คนชราที่รู้สึกว่ตนถูกทอดทิ้ง	-0.0107
หากเมื่อเจ้าตัวไม่ยอมเอ่ยปากพูดอะไรสักคำ ชายหนุ่มก็สนใจ ได้แต่หวังว่าสาวไทยจะสัมผัสได้ถึงความห่วงใยเหลือล้นของเขา	0.0257
มือหนึ่งถือมีดเหลาด้ามยาว อีกมือถือซีกไข่แห้งผิวเหลืองมัน หลวงลุงใช้ให้ถากเหลาเกลารูปซีกไข่เป็นไม้เหน็บประกับธรรม	-0.0151
น่าเสียดายนะที่แกได้เรียนน้อยไปหน่อย แต่ฉันดูแกฉลาดเฉลียวกว่าคนอื่น	0.1585
ให้คำปรึกษาแนะนำเกี่ยวกับการเสริมสร้างวินัยและพัฒนาให้ข้าราชการมีวินัยแก่	-0.0138
ผู้ส่งสาร กลายเป็นสิ่งที่มีตัวตน มีมิติด้านเศรษฐกิจการเมือง	0.0495
เดินออกมาจากห้อง ผมใช้ทิศทางเดิมที่ผมสำรวจไว้และทดลองมาแล้วเมื่อคืนก่อน	-0.0338
การผดุงครรภ์ หรือการพยาบาลและการผดุงครรภ์ ตามระเบียบที่รัฐมนตรีกำหนดโดยประกาศในราชกิจจานุเบกษา	0.0864
ความมีชีวิตชีวา สดใส แม้จะทำให้เขาปวดหัวในช่วงแรกที่ได้พบ	-0.0184
ชอบนึกว่าตัวเองดีจนคนอื่นต้องคอยอิจฉา ฉันพยักหน้า แอมมิเลียจะไม่เชื่อก็ช่างหล่อน	-0.0067
ทำใจดีมาเถยส คุณเป็นซุบให้ฉันไม่ได้เธอ นี่ขบ่นเสียงแหลม เมื่อเห็นประกาศจากทางคณะตั้งแต่ต้นเทอมสอง	-0.0329
เข้ามาแต่งเติมหัวใจที่เว้าแหว่งเพราะพิษความทุกข์ได้อย่างคาดไม่ถึง จากที่เคยคิดว่า เป็นเพียงเด็กสาวสอดรู้สอดเห็น	0.0533
ในใจได้บ้าง หากภาพของผู้คนแม้จะบางตา แต่ก็มากขึ้นเป็นครอบครัวส่วนใหญ่	-0.0032
ข้ากำเนิดดินน้ำลมและไฟหนึ่งเกิดอาจดับไหม้ในพริบตา ข้าท้องฟ้าเมฆหมอกนอกหนนหาหนึ่งดวงดาวพรราวแสงแห่งเวหา	0.2096
ในการอำนวยความสะดวก คุ่มครองสิทธิเสรีภาพประชาชน ป้องกันและควบคุมอาชญากรรม	0.0151
เพื่อนคนหนึ่งเลยชักชวนไปขายผลไม้รถเข็นที่ภูเก็ต ฉันไม่กล้าไปทำงานแบบนั้นหรอก	0.0773
เป้าหมายอย่างมีประสิทธิภาพและประสิทธิผล ปฏิบัติงานร่วมกับหรือสนับสนุนการปฏิบัติงานของหน่วยงานอื่นที่เกี่ยวข้องหรือที่ได้รับมอบหมาย	0.0161
หรือหน่วยกรรมรองก็ได้ วิจินตนเรียกคำเชื่อมที่เชื่อมอนุพากย์คุณศัพท์กับคำนามที่มันขยายว่า	-0.0175
ผลดีอีกอย่างคือได้มีเวลาดูแลนักเรียนใกล้ชิดยิ่งขึ้น ครูวิชัยพยายามรวบรวมนักเรียนให้มากขึ้นช่วยด้วยกันที่ระเบียบ	-0.0179
ไม่กี่นาทีก็รู้กันทั้งโรงเรียนแล้ว ว่าแต่เธอเห็นนั่นมั้ย รินชี้ไปที่พอลกับแซมซึ่งยังคง	-0.0111

ประโยคในฐานข้อมูล	คะแนน
มองไปที่เวที	
อยากเป็นหอเป็นเรือนเยี่ยะไรไถนา มีลูกสักสองสามคน หน้าแดงขึ้นวูบหนึ่ง	0.1036
เกี่ยวกับขั้นตอนในการดำเนินงานของจังหวัด นายอารีย์ วงศ์อารยะ	-0.0160
หญิงสาวไม่เข้าใจว่ากำลังเกิดอะไรขึ้นกับตัวเอง พอรู้ตัวตั้งสติได้	-0.0270
หาทราย หาดนพรัตน์ธาราเหมาะสำหรับศึกษาธรรมชาติ หากอยากได้แสงสีต้องไปหาดอ่าวนาง	-0.0095
ท่าทางจะอายุจริง อายุ คือตอนนี้ก็กำลังติดหนังสืออย่างรุนแรงนะค่ะพี่แหวน	0.1767
ขัดกับภาพสาวมาดแครงที่มาพร้อมกับแว่นตากรอบดำทรงภูมิ ก็ทำเอาคนโดนถามอดอมยิ้มกับงานสปาเกิตตี้ตรงหน้าไม่ได้	0.0238
ริมถนน รถเข็นขายขนมสี่สันลุดฉาดจอดนิ่งอยู่ สีสนของขนมบนรถดูน่ากลัวมากกว่าน่ากิน	0.1445
เพื่อการพิจารณาก่อนดื่ม ที่ผมเขียนในคอลัมน์นี้ออกวางตลาด เป็นเสียงจากผู้อ่านที่มีอาวุโส	0.0650
ลามไหม้ทั้งด้านเหนือและใต้มุ่งไปทางทิศตะวันตกเข้าหาลำน้ำแม่ยมซึ่งมีวัดราชธานีที่เต็มไปด้วยโบราณวัตถุขวางอยู่	0.1996
ยังผู้หญิงทั้งคู่นี้อีกเล่า อาจเสแสร้งแกล้งเข้ามาตีสนิท เพื่อล้วงเอาความสงสัยออกมาให้ล้น	0.1254
โครงการศึกษาภาษาและวัฒนธรรมท้องถิ่นของข้าราชการจังหวัดชายแดนภาคใต้	0.1184
ทิพภฤตาเห็นฝ่ายนั้นเงยหน้าขึ้นมองท้องฟ้าสีหม่น มือใหญ่ยกขึ้นปาดหยดน้ำซึ่งเริ่มจับตัวบนศีรษะได้รูปนั้น	0.1077

เมื่อระบบคำนวณคะแนนของแต่ละประโยคจนครบทุกประโยคแล้ว ประโยคเหล่านั้นจะถูกทำการจัดอันดับประโยคตามคะแนนของประโยคจากมากไปหาน้อย เพื่อความรวดเร็วในการจัดอันดับประโยค ได้ใช้ฟังก์ชันการจัดเรียงตามลำดับประโยคอัตโนมัติด้วยคำสั่งจากฐานข้อมูล, MySQL Server 5.1 [30]

เมื่อจัดเรียงประโยคจากมากไปน้อยแล้วระบบจะทำการคัดประโยคที่คะแนนตกอันดับออก ในที่นี้ได้ตัดประโยคออก 10% จากคลังข้อความปัจจุบันออก เช่น ถ้าปัจจุบันในคลังข้อความที่กำลังสร้างมีขนาด 1,100 ประโยค ประโยคที่ได้คะแนน 100 อันดับสุดท้ายจะถูกตัดออก เมื่อคัดประโยคออกแล้ว ระบบจะคัดลอกประโยคที่เหลือในฐานข้อมูลคลังข้อความวิเคราะห์นำไปใส่ในฐานข้อมูลคลังข้อความที่ต้องการสร้าง

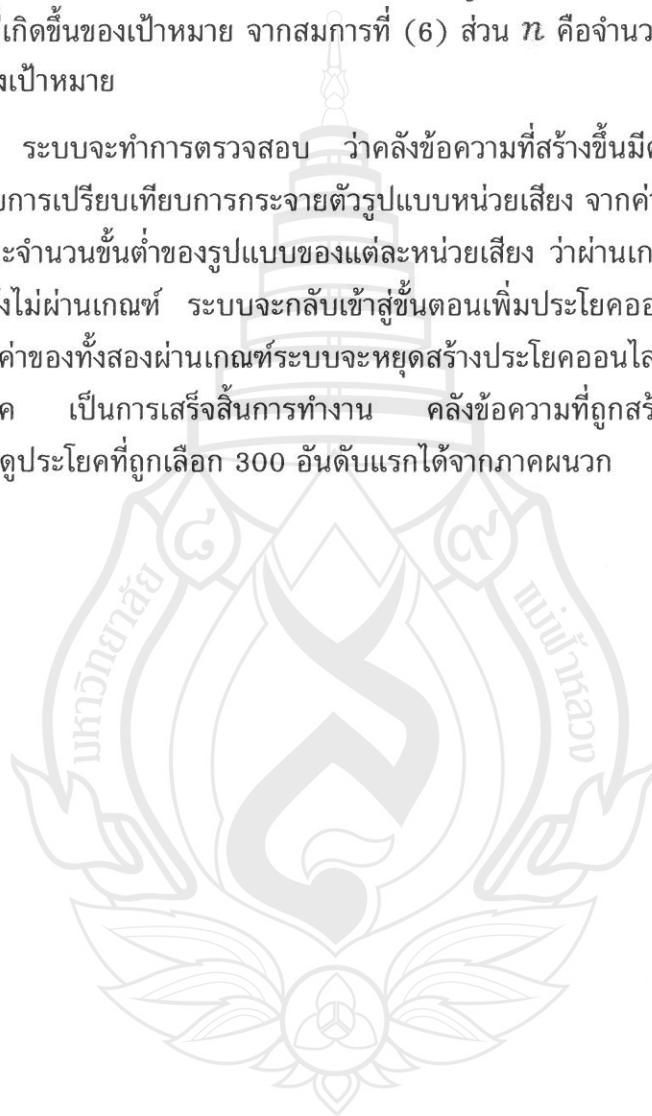
จากนั้นระบบจะทำการวิเคราะห์ลักษณะการกระจายตัวหน่วยเสียงในคลังข้อความที่สร้าง เพื่อตรวจสอบการกระจายตัวรูปแบบหน่วยเสียงของเป้าหมายและการกระจายตัวที่เกิดขึ้นในคลัง

ข้อความที่สร้าง ดังนั้นขั้นตอนนี้จึงทำการหาค่าเบี่ยงเบนเฉลี่ยร้อยละจากเป้าหมาย (Percent average deviation of target) จากสมการที่ (8)

$$Avg\ Deviation(\%) = \frac{\sum_{i=1}^n \sigma_i \times 100}{n} \quad (8)$$

โดยที่ σ_i คือ ผลต่างระหว่างโอกาสที่เกิดขึ้นของรูปแบบหน่วยเสียงเป้าหมายและปัจจุบันเทียบกับโอกาสที่เกิดขึ้นของเป้าหมาย จากสมการที่ (6) ส่วน n คือจำนวนรูปแบบหน่วยเสียงที่เกิดขึ้นทั้งหมดของเป้าหมาย

ขั้นตอนต่อไป ระบบจะทำการตรวจสอบ ว่าคลังข้อความที่สร้างขึ้นมีคุณสมบัติตามที่ต้องการสร้างหรือไม่ โดยการเปรียบเทียบการกระจายตัวรูปแบบหน่วยเสียง จากค่าเบี่ยงเบนเฉลี่ยร้อยละจากเป้าหมาย และจำนวนขั้นต่ำของรูปแบบของแต่ละหน่วยเสียง ว่าผ่านเกณฑ์ที่กำหนดไว้หรือไม่ ถ้าค่าดังกล่าวยังไม่ผ่านเกณฑ์ ระบบจะกลับเข้าสู่ขั้นตอนเพิ่มประโยคออนไลน์เข้าสู่คลังข้อความใหม่ แต่ถ้าค่าของทั้งสองผ่านเกณฑ์ระบบจะหยุดสร้างประโยคออนไลน์เพิ่มและหยุดกระบวนการเลือกประโยค เป็นการเสร็จสิ้นการทำงาน คลังข้อความที่ถูกสร้างขึ้นนี้สามารถนำไปใช้ได้ทันที สามารถดูประโยคที่ถูกเลือก 300 อันดับแรกได้จากภาคผนวก



บทที่ 4

ผลการวิจัย

4.1 การทดลอง

ในการทดลองระบบการเลือกข้อความอัตโนมัติเพื่อสร้างคลังข้อความตามการกระจายตัวหน่วยเสียงที่กำหนดได้นั้น ได้แบ่งการทดลองเป็นสี่การทดลองคือ การทดลองแรกเป็นการทดสอบระบบกับรูปแบบการกระจายตัวทางหน่วยเสียงของเป้าหมายจากคลังข้อความต้นแบบ การทดลองที่สองเป็นการทดสอบระบบกับรูปแบบการกระจายตัวทางหน่วยเสียงแบบกำหนดเอง โดยไม่ได้นำรูปแบบการกระจายตัวมาจากรูปแบบการกระจายตัวของคลังข้อความอื่น ๆ ซึ่งในที่นี้กำหนดให้ทุกรูปแบบหน่วยเสียงมีโอกาสเกิดเท่ากันทุกหน่วย การทดลองที่สามเป็นการทดลองเพื่อพิจารณาผลของการปรับพารามิเตอร์ค่าคงที่ความสำคัญรูปแบบหน่วยเสียงหายาก (α) ซึ่งใช้ในการคำนวณค่าความต้องการรูปแบบหน่วยเสียง การทดลองที่สี่เพื่อทดสอบประสิทธิภาพของงานวิจัยนี้วัดผลเทียบกับการเลือกประโยคแบบเก่าจากงานวิจัย W. Zhang, 2010: Automatic Construction for a TTS Corpus with Limited Text [11] ซึ่งเป็นงานวิจัยที่ใช้หลักการเลือกประโยคคล้ายกันแต่ต่างกันที่วิธีการคิดคะแนน

4.1.1 การสร้างคลังข้อความจากรูปแบบการกระจายตัวทางหน่วยเสียงจากการกระจายตัวเป้าหมาย

ในการทดลองนี้ ต้องการสร้างฐานข้อมูลข้อความภาษาไทยขนาด 1,000 ประโยค แต่ละประโยคประกอบด้วยจำนวนพยางค์มากกว่า 20 พยางค์แต่ไม่เกิน 40 พยางค์ ชนิดของหน่วยเสียงใช้รูปแบบ หน่วยเสียงคู่ ความครอบคลุมทางหน่วยเสียง 100% เทียบจากจำนวนรูปแบบหน่วยเสียงที่เกิดขึ้นจากรูปแบบหน่วยเสียงของเป้าหมาย แต่ละหน่วยเสียง ต้องเกิดมากกว่า 1 ครั้ง ค่าคงที่ความสำคัญรูปแบบหน่วยเสียงหายาก (α) ไว้ที่ 1.5 ค่าเบี่ยงเบนเฉลี่ยร้อยละจากเป้าหมายต้องน้อยกว่า 3.0 % ของรูปแบบของการกระจายตัวเป้าหมาย การทดลองมีขั้นตอนดังนี้

1. การสร้างการกระจายตัวทางหน่วยเสียงเป้าหมาย

ในการทดลองได้เลือกใช้การกระจายตัวของหน่วยเสียงเป้าหมาย จากคลังเสียงพูด Large Vocabulary Continuous Speech Recognition (LVCSR) corpus for Thai language[18] เลือกใช้ชุดครอบคลุมคำศัพท์ 5,000 คำ ฐานข้อมูลข้อความนี้คัดเลือกประโยคมาจากรานข้อมูลบทความภาษาไทย ORCHID (Open Linguistic Resource Channelled toward InterDisciplinary research) [20] และบทความจากฐานข้อมูลอื่น ๆ ประกอบด้วย 3,007 ประโยค ระบบจะ

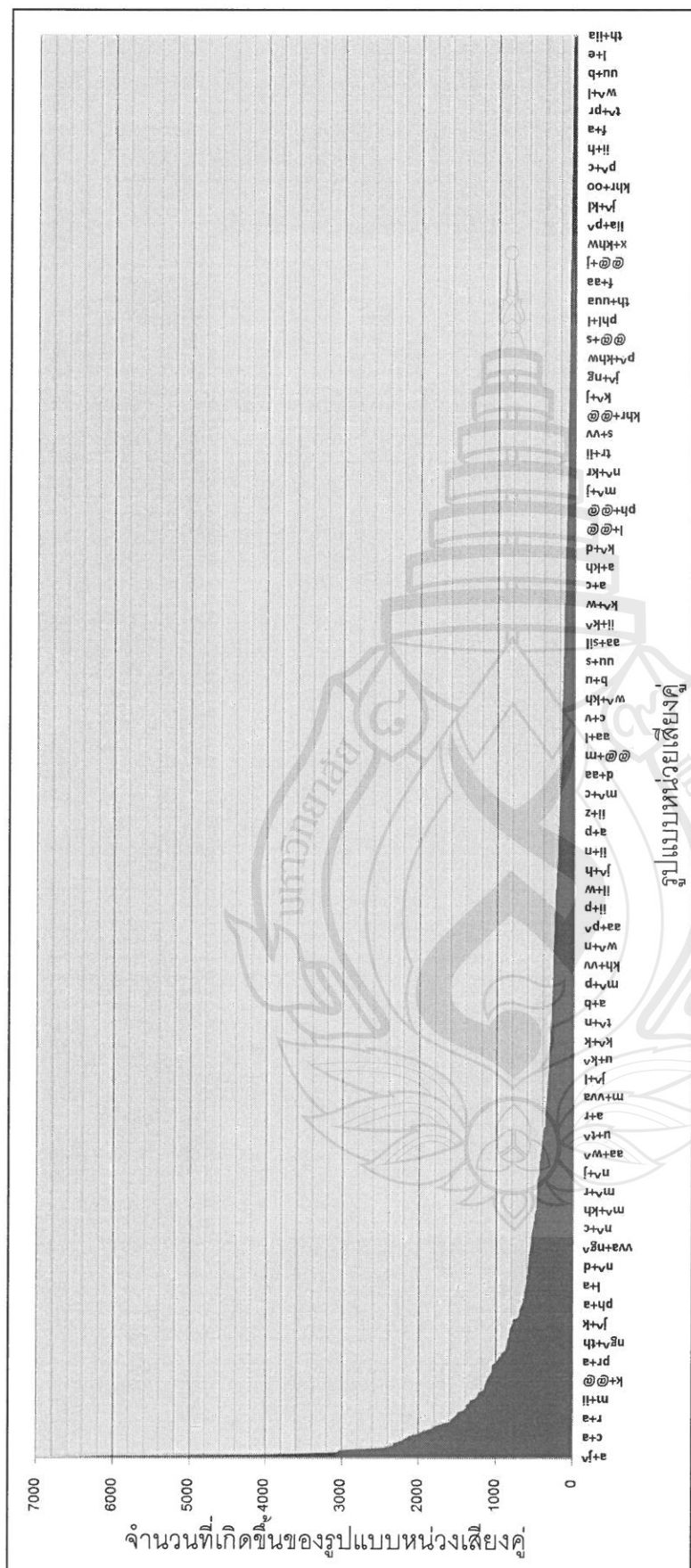
คำนวณค่าการกระจายตัวของหน่วยเสียง โดยการเริ่มจากนำข้อความทั้งหมดเหล่านั้น ไปแปลงรูปเขียนให้เป็นคำอ่านไทยโดยใช้ PGLR (Probabilistic generalized LR parser) [12] สิ่งที่ได้คือหน่วยเสียงทั้งหมดจำนวน 246,199 หน่วยเสียง จากนั้นจะทำการหาจำนวนการเกิดของหน่วยเสียงคู่ทั้งหมด โดยใช้เครื่องมือ Hidden Markov toolkit (HTK) [31] ได้หน่วยเสียงคู่ที่เกิดขึ้นทั้งหมดคือ 1,383 รูปแบบ ใช้เวลาคำนวณเพียง 1.4 วินาทีความถี่ของจำนวนการเกิดขึ้นบางส่วนของแต่ละหน่วยเสียง ดูได้จากภาพที่ 4-1 การกระจายตัวนี้สามารถนำไปใช้เป็นรูปแบบการกระจายตัวของหน่วยเสียงเป้าหมายได้ทันที

2. การสร้างประโยคออนไลน์

ขั้นตอนนี้จะเริ่มทำการดึงข้อมูลข้อความจากหน้าเว็บบนอินเทอร์เน็ต โดยกลุ่มข้อมูลข้อความที่ได้สืบค้นมาจาก เว็บไซต์สารานุกรมไทยออนไลน์ [36] และจากคลังข้อมูลภาษาไทยออนไลน์ [37] ซึ่งประกอบไปด้วยข้อมูลจากบทความวิชาการ หนังสือแบบเรียน หนังสือพิมพ์ และนิตยสาร หาดูขอบเขตของประโยคภาษาไทยโดยนับจากจำนวนพยางค์ด้วยเครื่องมือ Collocation and Thai Word Segmentation [32] โดยที่ในแต่ละประโยคต้องประกอบด้วยจำนวนพยางค์ที่มากกว่า 20 พยางค์ แต่น้อยกว่า 40 พยางค์ จากนั้นประโยคจะถูกนำไปคัดกรอง โดย ในที่นี้ตัวอักษรที่ยอมให้อยู่ในประโยคคือ ตัวอักษรภาษาไทยเท่านั้น จากนั้นประโยคผ่านการคัดกรองจะถูกนำไปแปลงแปลงรูปเขียนให้เป็นคำอ่านไทยโดยใช้ PGLR (Probabilistic generalized LR parser) [12] จากนั้นระบบจะทำการดึงข้อมูลจากอินเทอร์เน็ตไปเรื่อย ๆ อย่างต่อเนื่องรอการสั่งหยุดในกรณีที่ระบบทำการจัดทำคลังข้อความที่เสร็จสมบูรณ์แล้ว ประโยคดังกล่าวจะถูกนำไปเก็บไว้ใน ฐานข้อมูลเพื่อเตรียมนำไปคัดเลือกประโยคต่อไป

3. การเลือกประโยคและจัดเก็บคลังข้อความ

ขนาดของฐานข้อมูลคลังข้อความที่กำหนดไว้คือคือ 1,000 ประโยค ข้อความในแต่ละประโยคต้องไม่ซ้ำกัน จากนั้นนำประโยคจากฐานข้อมูลที่เก็บมาจากอินเทอร์เน็ตถูกนำเข้ามายังคลังข้อความปัจจุบัน อีกครั้งละ 100 ประโยค จากนั้นจะทำการให้คะแนนของประโยค จากการทดลองนี้การให้คะแนนแต่ละประโยคใช้เวลาเฉลี่ย 0.38 วินาทีต่อประโยค (CPU: Intel Core i3 processor 2.13 GHz, Memory: RAM DDR3 4 GB) ขั้นตอนนี้ใช้เวลาไปทั้งสิ้น 11 ชั่วโมง 37 นาที



ภาพที่ 4-1 แผนภาพการกระจายตัวทางสถิติทางหน่วยเสียงของฐานข้อมูล LOTUS [18]

ผลของการเปรียบเทียบคลังข้อความที่ถูกสร้างขึ้นกับคลังข้อความเป้าหมาย จากการเลือกดึงข้อมูลจากอินเทอร์เน็ตไปทั้งสิ้น 100 รอบได้มา 10,000 ประโยค แสดงในตารางที่ 4-1 ความยาวของประโยคในคลังข้อความที่ถูกสร้างขึ้นมาได้ความยาวที่เหมาะสม ผลของการครอบคลุมทางหน่วยเสียงคู่อยู่ที่ 99.13 % จากการดึงประโยคจากอินเทอร์เน็ต พบว่ารูปแบบหน่วยเสียงที่หาไม่ได้ คิดเป็น 0.87 % ของรูปแบบหน่วยเสียงทั้งหมด ส่วนมากเกิดจากการแปลงรูปเขียนเป็นคำอ่านที่ผิดพลาด ดังตารางที่ 4-2 และปัญหาจากข้อเฉพาะที่ไม่พบ ตารางที่ 4-3 ปัญหาด้านการแปลงรูปเขียนเป็นคำอ่านที่ผิดพลาดสามารถแก้ไขได้โดยหาใช้เครื่องมือแปลงรูปเขียนเป็นคำอ่านที่มีประสิทธิภาพมากขึ้น ส่วนข้อเฉพาะที่หาไม่พบ คาดว่าถ้าเลือกประโยคจากอินเทอร์เน็ตเข้ามาเพิ่มอีก อาจทำให้เจอคำศัพท์นั้นจากแหล่งข้อมูลทางอินเทอร์เน็ตเพิ่มขึ้นได้

ตารางที่ 4-1 คุณลักษณะของคลังข้อความที่สร้างขึ้น

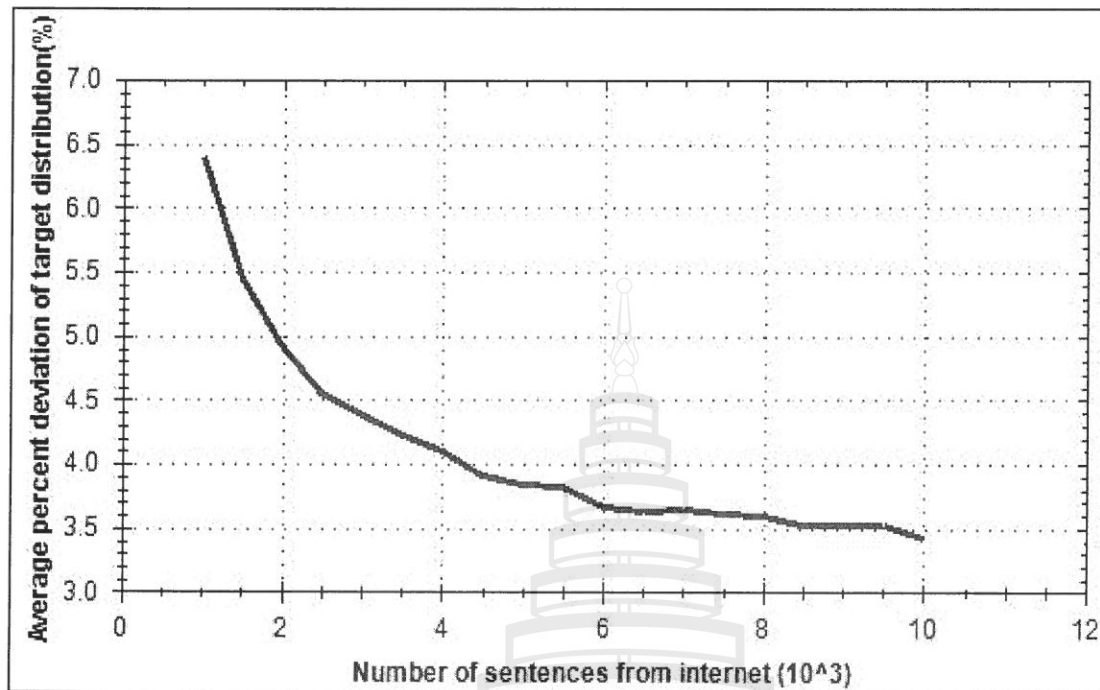
คุณลักษณะ	คลังข้อความที่สร้างขึ้น
จำนวนประโยค	1,000
จำนวนพยางค์ที่มากที่สุดในประโยค	40
จำนวนพยางค์ที่น้อยที่สุดในประโยค	20
ค่าเฉลี่ยจำนวนพยางค์ต่อประโยค	25.1
จำนวนหน่วยเสียงคู่ทั้งหมด	72,458
จำนวนรูปแบบหน่วยเสียงคู่	1,371
ความครอบคลุมทางหน่วยเสียง(%)	99.13

ตารางที่ 4-2 รูปแบบหน่วยเสียงที่ไม่ปรากฏจากการแปลงรูปเขียนเป็นรูปอ่านที่ผิด

รูปแบบที่ไม่พบ	ตัวอย่าง	แปลงรูปเสียงเป็นรูปอ่าน	
		คลังข้อความเป้าหมาย	คลังข้อความกำหนด
i+khw	ตีความ	t i khw a m [^]	t ii khw a m [^]
uu+br	ผู้บริโภค	ph uu br i ph oo k [^]	ph uu b @@ r i ph oo k [^]
@+br	เพราะบริษัท	phr @ br i s a t [^]	phr @ b @@ r i s a t [^]
ii+br	ที่บริเวณ	th ii br i w ee n [^]	th ii b @@ r i w ee n [^]

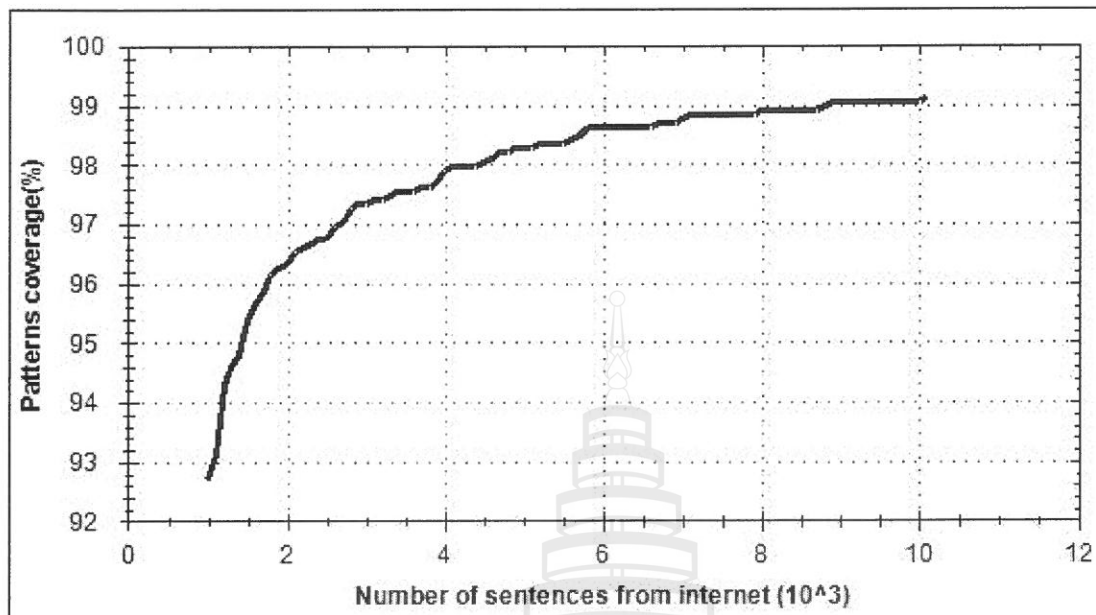
ตารางที่ 4-3 รูปแบบหน่วยเสียงที่ไม่ปรากฏจากชื่อเฉพาะ

รูปแบบ ที่ไม่พบ	ตัวอย่าง	แปลงรูปเสียงเป็นรูปอ่าน	
		คลังข้อความเป้าหมาย	คลังข้อความกำหนด
t+@	เตาะเตะ	t @ t x	ไม่พบ
q+b	เดอะบอสส์	d q b a z o t ^	ไม่พบ
dr+ia	อเล็กซานเดรีย	z a l e k ^ s a a n ^ d r i a	ไม่พบ
q+kl	เดอะแกลเลอรี	d q kl x x l q q r ii	ไม่พบ
vva+kw	เมื่อกว่านซื้อ	m vva kw aa n ^ s vv	ไม่พบ
ia+pr	เสียเปรียบ	s ia pr ia p ^	ไม่พบ
ng+@	เจ้าเงาะ	c a w ^ ng @	ไม่พบ
kw+ee	วัดตันเกวัน	w a t ^ t o n ^ kw ee n ^	ไม่พบ
kl+oo	กระดุกรูปโกลน	kr a d uu k ^ r uu p ^ kl oo n ^	ไม่พบ
bl+@	หนึ่งบล็อด	n v ng ^ bl @ k ^	ไม่พบ
ua+kr	ตัวกระตุ้น	t ua kr a t u n ^	ไม่พบ
ua+kw	น่ากลัวกว่า	n aa kl ua kw aa	ไม่พบ
pr+ua	แปรปรวน	pr xx pr ua n ^	ไม่พบ
f+ia	มาเฟีย	m aa f ia	ไม่พบ
kw+xx	แกว่ง	kw xx ng ^	ไม่พบ
h+@@	เหาะ	h @@	ไม่พบ
ia+khw	เสียความ	s ia khw aa m ^	ไม่พบ



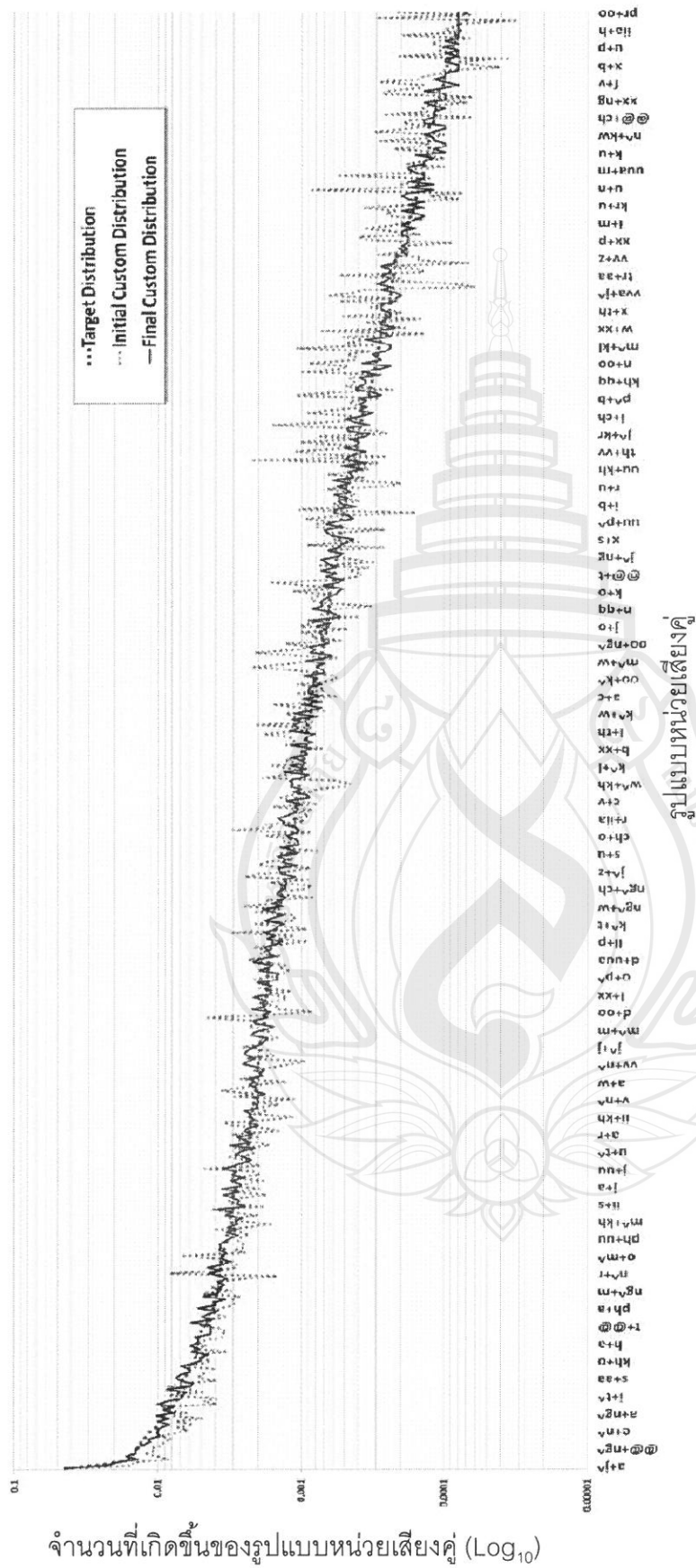
ภาพที่ 4-2 ผลของค่าเฉลี่ยผลต่างของกระจายตัวทางหน่วยเสียงเป้าหมายเทียบกับจำนวนประโยคออนไลน์ที่เข้ามาใหม่

ผลของค่าเฉลี่ยผลต่างของกระจายตัวทางหน่วยเสียงเป้าหมายจากประโยคที่ดึงมาจาก internet ดังภาพที่ 4-2 จากการทดลองผลของค่าเฉลี่ยความต่างจากเป้าหมาย จะลดลงมากในช่วงแรกและค่อย ๆ ลดลงอย่างช้า ๆ ในช่วงหลัง ส่วนการครอบคลุมทางหน่วยเสียง แสดงออกมาดังภาพที่ 4-3 จะเห็นได้ว่าการครอบคลุมทางหน่วยเสียงเพิ่มขึ้นอย่างมากในช่วงแรกและค่อย ๆ เพิ่มขึ้นช้า ๆ ในช่วงหลัง เนื่องจากข้อมูลข้อความจากอินเทอร์เน็ตที่ได้มาในแต่ละช่วงการรับเข้าสู่ระบบมีผลต่อค่าเฉลี่ยผลต่างของกระจายตัวทางหน่วยเสียงเป้าหมายและค่าการครอบคลุมทางหน่วยเสียง ถ้าข้อมูลช่วงนั้นประกอบด้วยประโยคที่มีหน่วยเสียงที่ระบบต้องการมากจะทำให้ค่าของทั้งสองดีขึ้นอย่างมาก แต่ถ้าข้อมูลที่รับเข้ามาประกอบด้วยประโยคที่มีหน่วยเสียงที่ต้องการน้อยระบบจะรับประโยคใหม่เข้าแทนที่ได้น้อย จึงทำให้ค่าเฉลี่ยผลต่างของกระจายตัวทางหน่วยเสียงเป้าหมายค่อย ๆ ลดลง และค่าของการครอบคลุมทางหน่วยเสียงค่อย ๆ เพิ่มขึ้น



ภาพที่ 4-3 ความครอบคลุมทางหน่วยเสียงเทียบกับจำนวนประโยคออนไลน์ที่เข้ามาใหม่

เมื่อพิจารณาการกระจายตัวของหน่วยหลังจากการคัดเลือกประโยคแล้ว ดังภาพที่ 4-4 จะเห็นได้ว่าการกระจายตัวของหน่วยเสียงได้ถูกปรับ ให้ได้เป็นไปตามรูปแบบของการกระจายตัวของหน่วยเป้าหมายตามอัตราส่วนของจำนวนขนาดหน่วยเสียงรวมระบบ ได้ดีขึ้น เมื่อเทียบค่าจากการกระจายตัวตั้งต้น และระบบยังสามารถรักษาจำนวนของหน่วยเสียงขั้นต่ำให้อยู่จำนวนที่ต้องการได้อีกด้วย จึงเป็นข้อพิสูจน์ได้ว่าการเลือกประโยคจากฐานข้อมูลออนไลน์ ซึ่งเป็นคลังข้อความที่ไม่จำกัด สามารถสร้างคลังข้อความที่มีการกระจายตัวของหน่วยเสียงตามที่ต้องการได้อย่างมีประสิทธิภาพ

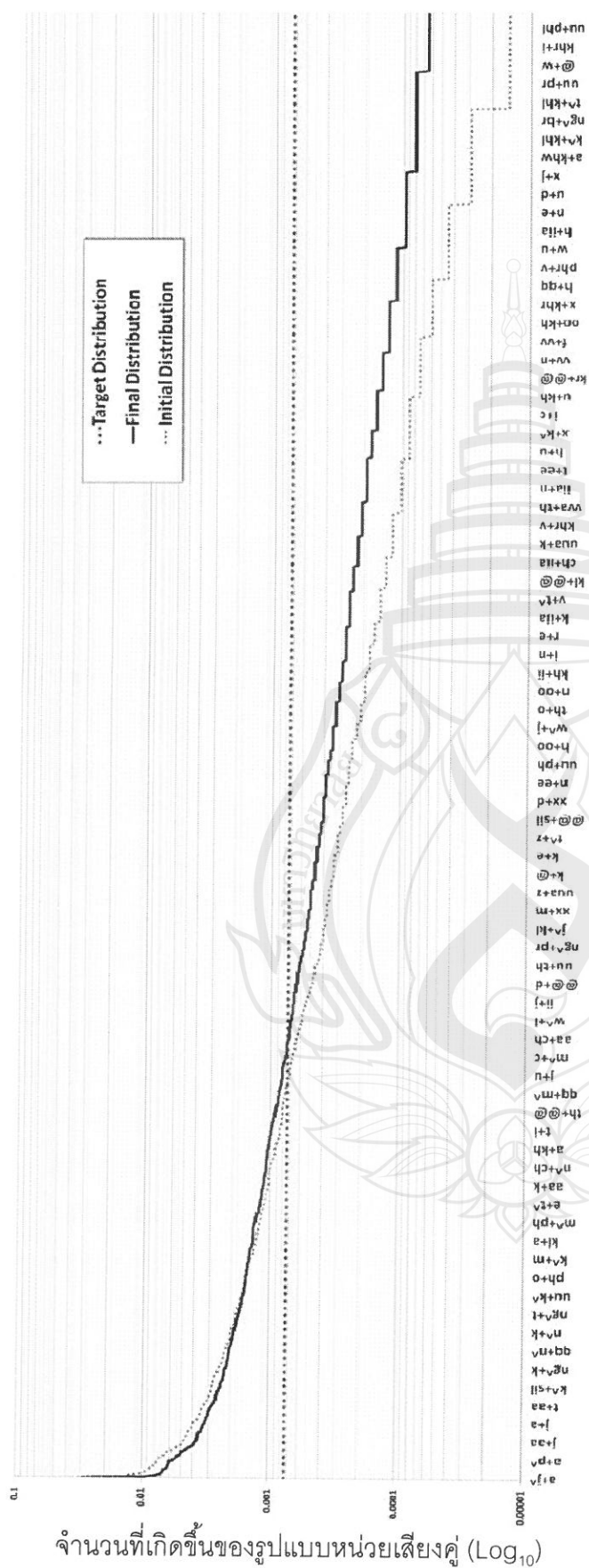


ภาพที่ 4-4 แผนภาพการกระจายตัวของหน่วยหลังจากการคัดเลือกประโยค

4.1.2 การสร้างคลังข้อความโดยใช้การกำหนดรูปแบบการกระจายตัวทางหน่วยเสียงแบบกำหนดเอง

ในการทดลองนี้ต้องการที่จะเปลี่ยนมาใช้ในการกระจายตัวทางหน่วยเสียงแบบกำหนดเอง โดยที่ใช้รูปแบบหน่วยเสียงคู่ 1,383 รูปแบบที่ได้มาจากการสร้างคลังข้อความจากรูปแบบการกระจายตัวทางหน่วยเสียงจากการกระจายตัวเป้าหมายในการทดลองแรก มากำหนดการกระจายตัวเป้าหมายโดยให้โอกาสที่ทุกรูปแบบของหน่วยเสียงจะเกิดขึ้นได้เท่ากันเพื่อที่จะได้การกระจายตัวเป้าหมายที่ต้องการ สามารถดูรูปแบบการกระจายตัวเป้าหมายได้จากภาพที่ 4-5 การกำหนดขนาดคลังข้อความไว้ที่ 1,000 ประโยค แต่ละประโยคประกอบด้วยจำนวนพยางค์มากกว่า 20 พยางค์แต่ไม่เกิน 40 พยางค์ ชนิดของหน่วยเสียงใช้รูปแบบ หน่วยเสียงคู่ ความครอบคลุมทางหน่วยเสียง 100% เทียบจากจำนวนรูปแบบหน่วยเสียงที่เกิดขึ้นจากรูปแบบหน่วยเสียงของเป้าหมาย แต่ละหน่วยเสียง ต้องเกิดมากกว่า 1 ครั้ง ค่าคงที่ความสำคัญรูปแบบหน่วยเสียงหายาก (α) ไว้ที่ 1.5 ค่าเบี่ยงเบนเฉลี่ยร้อยละจากเป้าหมายต้องน้อยกว่า 3.0 % ของรูปแบบของการกระจายตัวเป้าหมาย การทดลอง ใช้ลำดับขั้นตอนการทดลองเหมือนการทดลองที่แรกทั้งหมด

จากการทดลองกำหนดการกระจายตัวเป้าหมายโดยให้โอกาสที่ทุกรูปแบบของหน่วยเสียงจะเกิดขึ้นได้เท่ากันเพื่อที่จะได้การกระจายตัวเป้าหมายที่ต้องการนั้น พบว่าเมื่อพิจารณาการกระจายตัวทางหน่วยเสียงจากประโยคออนไลน์ในคลังข้อความที่ถูกสร้างเริ่มต้น การกระจายตัวนั้นเป็นตามลักษณะรูปแบบคล้ายกราฟของฟังก์ชันลอการิทึม ดังภาพที่ 4-5 และเมื่อทำการเลือกประโยคจากอินเทอร์เน็ตไปเรื่อย ๆ พบว่าการกระจายตัวของรูปแบบหน่วยเสียงนั้น ลู่เข้าสู่การกระจายตัวทางหน่วยเสียงเป้าหมายได้ไม่มากเท่าที่ควรทั้งนี้เพราะว่าลักษณะธรรมชาติของการกระจายตัวทางหน่วยเสียงภาษาไทยนั้น ไม่ได้เป็นแบบการกระจายตัวที่เรากำหนดไว้ เพราะฉะนั้นในการทดลองนี้อาจต้องใช้เวลามาก กว่าจะได้รูปแบบการกระจายตัวในคลังข้อความตามที่ต้องการ เพราะฉะนั้นการกำหนดลักษณะการกระจายตัวจึงมีผลต่อระยะเวลาในการสร้างคลังข้อความ ถ้าลักษณะการกระจายตัวขัดแย้งกับธรรมชาติของการกระจายตัวของภาษามากเกินไป จะทำให้การสร้างคลังข้อความเป็นไปได้ยากมากขึ้น

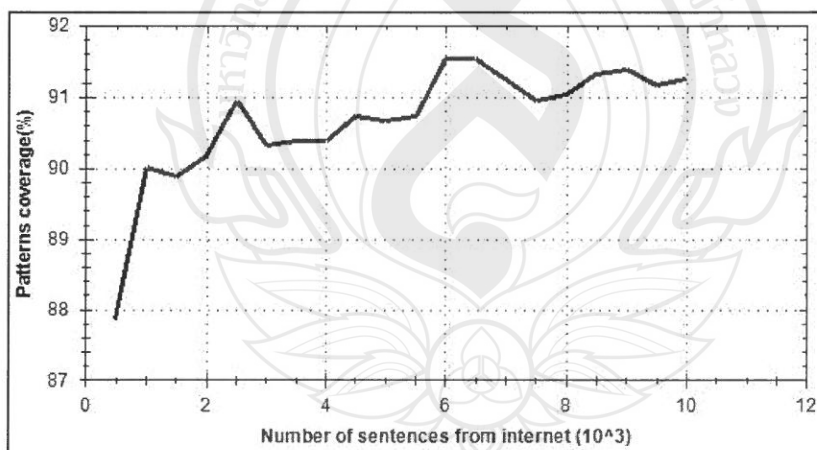


ภาพที่ 4-5 แผนภาพการกระจายตัวของหน่วยหลังจากการคัดเลือกประโยชน์ใช้สอยจากการกระจายตัวหน่วยเสี่ยงเป้าหมายกำหนดเอง

4.1.3 ทดสอบผลของการสร้างคลังข้อความจากการปรับพารามิเตอร์ค่าคงที่ ความสำคัญรูปแบบหน่วยเสียงหายาก

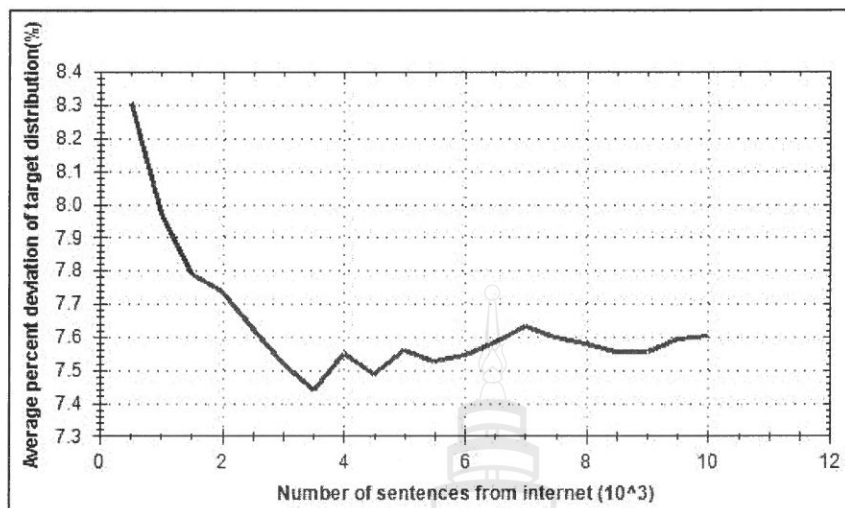
ในการทดลองนี้ต้องการวัดผลกระทบของการเปลี่ยนแปลงในการปรับพารามิเตอร์ค่าคงที่ความสำคัญรูปแบบหน่วยเสียงหายาก ทดลองโดยการสร้างคลังข้อความ 1,000 ประโยค แต่ละประโยคประกอบด้วยจำนวนพยางค์มากกว่า 20 พยางค์แต่ไม่เกิน 40 พยางค์ ชนิดของหน่วยเสียงใช้รูปแบบ หน่วยเสียงคู่ ความครอบคลุมทางหน่วยเสียง 100% เทียบจากจำนวนรูปแบบหน่วยเสียงที่เกิดขึ้นจากรูปแบบหน่วยเสียงของเป้าหมาย แต่ละหน่วยเสียง ต้องเกิดมากกว่า 1 ครั้ง ค่าเบี่ยงเบนเฉลี่ยร้อยละจากเป้าหมายต้องน้อยกว่า 3.0 % ของรูปแบบของการกระจายตัวเป้าหมาย การทดลอง ใช้ลำดับขั้นตอนการทดลองเหมือนการทดลองที่แรกทั้งหมด แต่เปลี่ยนแปลงค่าคงที่ความสำคัญรูปแบบหน่วยเสียงหายาก (α) โดยการทดลองครั้งแรกกำหนดไว้ที่ 0.0 การทดลองครั้งที่สองกำหนดไว้ที่ 1.5 และการทดลองครั้งที่สามกำหนดไว้ที่ 6.0

ผลการทดลองครั้งที่หนึ่งใช้ค่าคงที่ความสำคัญรูปแบบหน่วยเสียงหายาก (α) ที่ 0.0 เป็นการลดความสำคัญของหน่วยเสียงหาพบว่า ความครอบคลุมทางหน่วยเสียงเข้าสู่เป้าหมายได้ช้าและเป็นไปได้อย่างที่ผลลัพธ์ของความครอบคลุมทางหน่วยเสียงจะเข้าสู่เป้าหมายได้ในระยะเวลานั้น ๆ สามารถดูการเปลี่ยนแปลงความครอบคลุมทางหน่วยเสียงได้จากภาพที่ 4-6



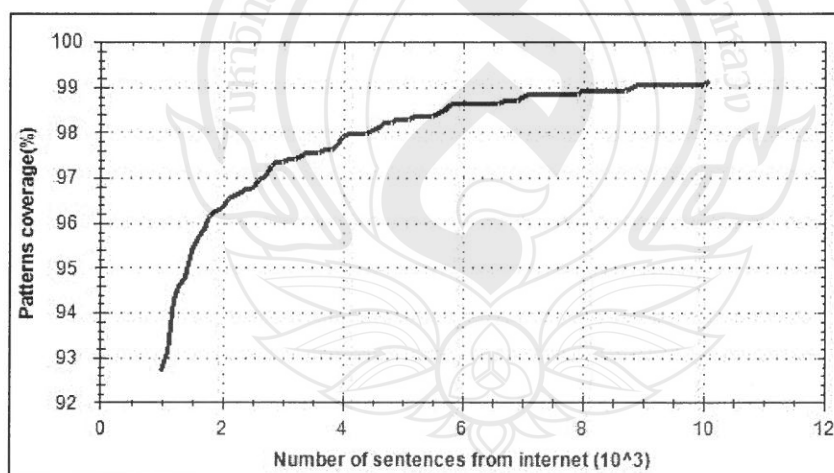
ภาพที่ 4-6 ความครอบคลุมทางหน่วยเสียงเมื่อกำหนดให้ค่า α เท่ากับ 0.0

เมื่อพิจารณาค่าเบี่ยงเบนเฉลี่ยร้อยละจากการกระจายตัวทางหน่วยเสียงเป้าหมายพบว่า สามารถเข้าสู่ค่าการกระจายตัวทางหน่วยเสียงที่ต้องการได้ดีในระยะเวลานั้น ๆ สามารถดูการเปลี่ยนแปลงได้ ดังภาพที่ 4-7



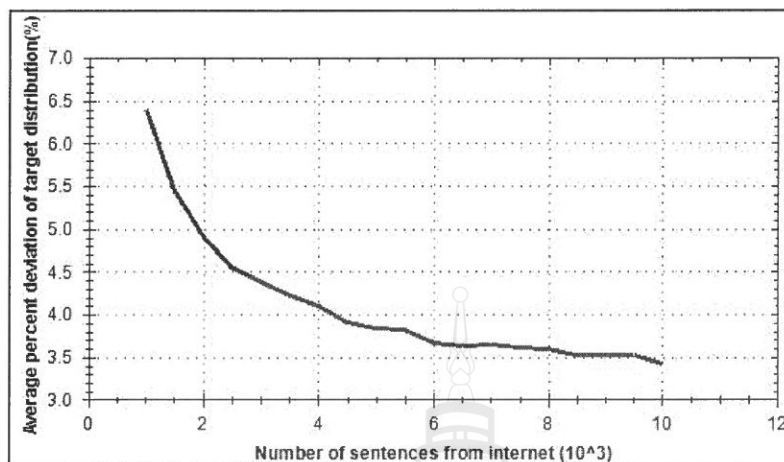
ภาพที่ 4-7 ผลของค่าเฉลี่ยผลต่างของการกระจายตัวทางหน่วยเสียงเป้าหมายเมื่อกำหนดให้ค่า α เท่ากับ 0.0

ผลการทดลองครั้งที่สองใช้ค่าคงที่ความสำคัญรูปแบบหน่วยเสียงหายาก (α) ที่ 1.5 เป็นค่าที่ใช้ในการทดลองหลักของงานวิจัยนี้ พบว่าความครอบคลุมทางหน่วยเสียงเข้าสู่เป้าหมายได้เร็วกว่าการกำหนดค่าคงที่ความสำคัญรูปแบบหน่วยเสียงหายาก (α) ที่ 0.0 และสามารถเข้าถึงความครอบคลุมทางหน่วยเสียงเป้าหมายได้ แสดงดังภาพที่ 4-8



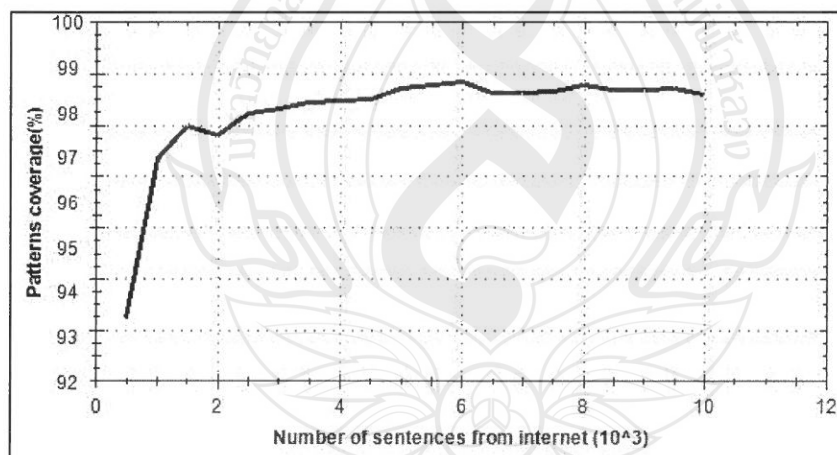
ภาพที่ 4-8 ความครอบคลุมทางหน่วยเสียงเมื่อกำหนดให้ค่า α เท่ากับ 1.5

ผลของความเปลี่ยนแปลงค่าเบี่ยงเบนเฉลี่ยร้อยละจากการกระจายตัวทางหน่วยเสียงเป้าหมาย แสดงดังภาพที่ 4-9 พบว่าสามารถเข้าสู่เข้าสู่การกระจายตัวทางหน่วยเสียงที่ต้องการได้ และได้รูปแบบการกระจายตัวทางหน่วยเสียงที่ไวกว่าการใช้ค่า α เท่ากับ 0.0



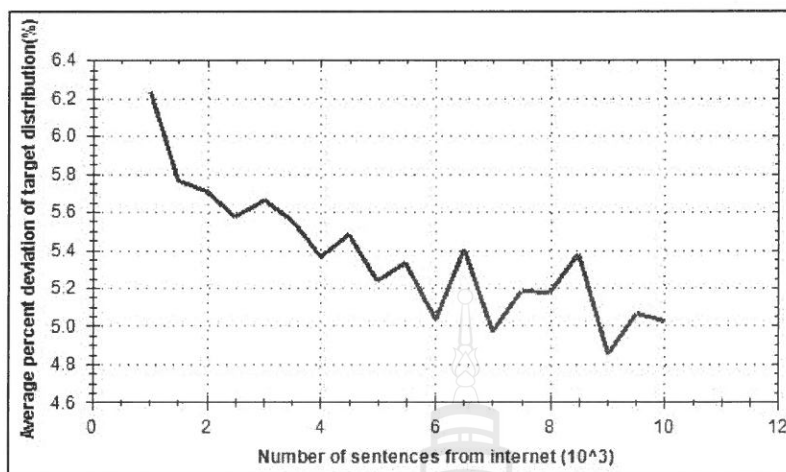
ภาพที่ 4-9 ผลของค่าเฉลี่ยผลต่างของการกระจายตัวทางหน่วยเสียงเป้าหมายเมื่อกำหนดให้ค่า α เท่ากับ 1.5

ผลการทดลองครั้งที่สามกำหนดค่าคงที่ความสำคัญรูปแบบหน่วยเสียงหายาก (α) ที่ 3.0 พบว่าความครอบคลุมทางหน่วยเสียงเข้าสู่เป้าหมายได้เร็วมากที่สุด เทียบกับกว่าการกำหนดค่าคงที่ความสำคัญรูปแบบหน่วยเสียงหายาก (α) ที่ 0.0 และ 1.5 และสามารถเข้าถึงความครอบคลุมทางหน่วยเสียงเป้าหมายได้ แสดงดังภาพที่ 4-10



ภาพที่ 4-10 ความครอบคลุมทางหน่วยเสียงเมื่อกำหนดให้ค่า α เท่ากับ 3.0

แต่เมื่อพิจารณาผลของความเปลี่ยนแปลงค่าเบี่ยงเบนเฉลี่ยร้อยละจากการกระจายตัวทางหน่วยเสียงเป้าหมาย แสดงดังภาพที่ 4-11 พบว่าความสามารถในการเข้าสู่การกระจายตัวทางหน่วยเสียงที่ต้องการขาดความต่อเนื่องและขึ้น ๆ ลง ๆ อย่างคาดการไม่ได้ ทั้งนี้เนื่องจากการกำหนดค่าคงที่ความสำคัญรูปแบบหน่วยเสียงหายาก (α) ที่มากเกินไปทำให้การคัดเลือกประโยชน์มุ่งเน้นการให้คะแนนไปที่ประโยชน์ที่มีหน่วยเสียงหายากมากเกินไป

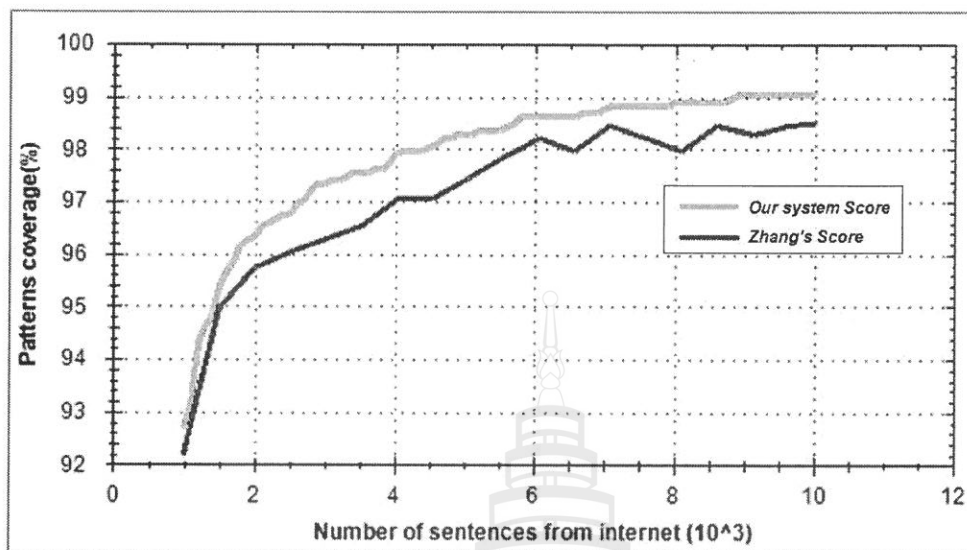


ภาพที่ 4-11 ผลของค่าเฉลี่ยผลต่างของกระจายตัวทางหน่วยเสียงเป้าหมายเมื่อกำหนดให้ค่า α เท่ากับ 3.0

ดังนั้นในการสร้างคลังข้อความแต่ละครั้งควรกำหนดค่าคงที่ความสำคัญรูปแบบหน่วยเสียงหายาก (α) ให้พอดี ถ้าเรากำหนดน้อยเกินไปจะทำให้ความครอบคลุมทางหน่วยเสียงเข้าสู่เป้าหมายได้ช้า แต่ถ้าเรากำหนดค่าคงที่ความสำคัญรูปแบบหน่วยเสียงหายาก (α) มากเกินไปจะทำให้กระจายตัวทางหน่วยเสียงในคลังข้อความที่สร้างเข้าสู่การกระจายตัวทางหน่วยเสียงเป้าหมายได้ช้าเกินไปอีกด้วย

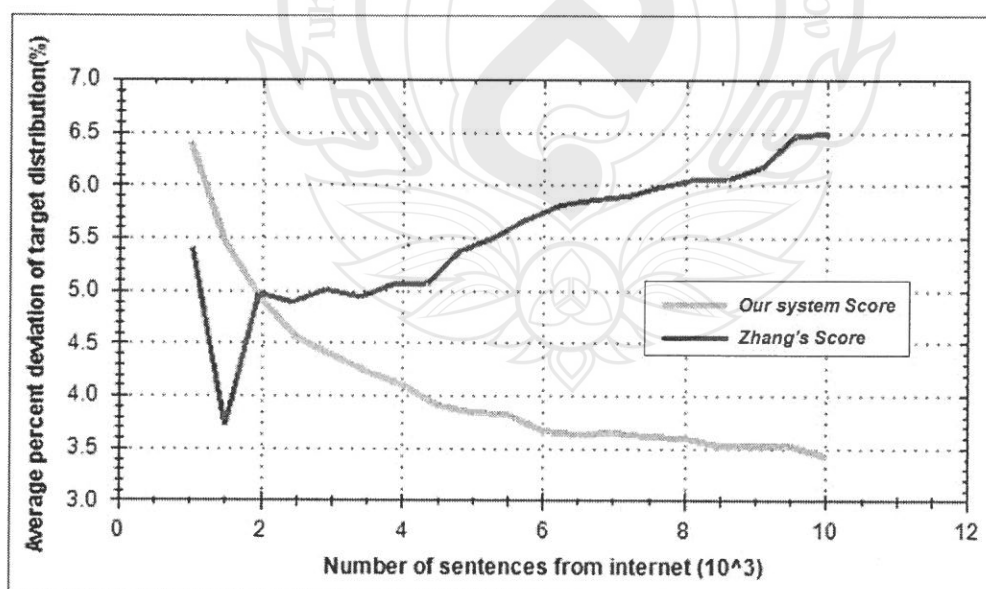
4.1.4 เปรียบเทียบผลจากวิธีการให้คะแนนในงานวิจัยกับการให้คะแนนของงานวิจัย Automatic Construction for a TTS Corpus with Limited Text (W. Zhang, 2010)

ในการทดลองนี้เป็นการเปรียบเทียบผลลัพธ์ของการเลือกประโยคโดยเทียบผลกับงานวิจัย Automatic Construction for a TTS Corpus with Limited Text (W. Zhang, 2010) [11] ซึ่งเป็นงานวิจัยที่ใช้วิธีการเลือกประโยคคล้ายกลับวิธีการเลือกประโยคของงานวิจัยนี้ต่างกันที่การให้คะแนนประโยคโดยวิธีของ Zhang [11] ให้คะแนนประโยคโดยการมุ่งเน้นไปที่หน่วยเสียงหายากเพียงอย่างเดียว การทดลองนี้ทำโดยการกำหนดการสร้าง คลังข้อความขนาด 1,000 ประโยค แต่ละประโยคประกอบด้วยจำนวนพยางค์มากกว่า 20 พยางค์แต่ไม่เกิน 40 พยางค์ ชนิดของหน่วยเสียงใช้รูปแบบ หน่วยเสียงคู่ ความครอบคลุมทางหน่วยเสียง 100% เทียบจากจำนวนรูปแบบหน่วยเสียงที่เกิดขึ้นจากรูปแบบหน่วยเสียงของเป้าหมาย แต่ละหน่วยเสียง ต้องเกิดมากกว่า 1 ครั้ง ค่าคงที่ความสำคัญรูปแบบหน่วยเสียงหายาก (α) ไว้ที่ 1.5 ค่าเบี่ยงเบนเฉลี่ยร้อยละจากเป้าหมายต้องน้อยกว่า 3.0 % ของรูปแบบของการกระจายตัวเป้าหมาย ผลการทดลองออกมาดังภาพที่ 4-12 และ 4-13



ภาพที่ 4-12 ความครอบคลุมทางหน่วยเสียงเมื่อเปรียบเทียบวิธีการเลือกประโยคของงานวิจัยกับวิธีการเลือกประโยคแบบเก่า

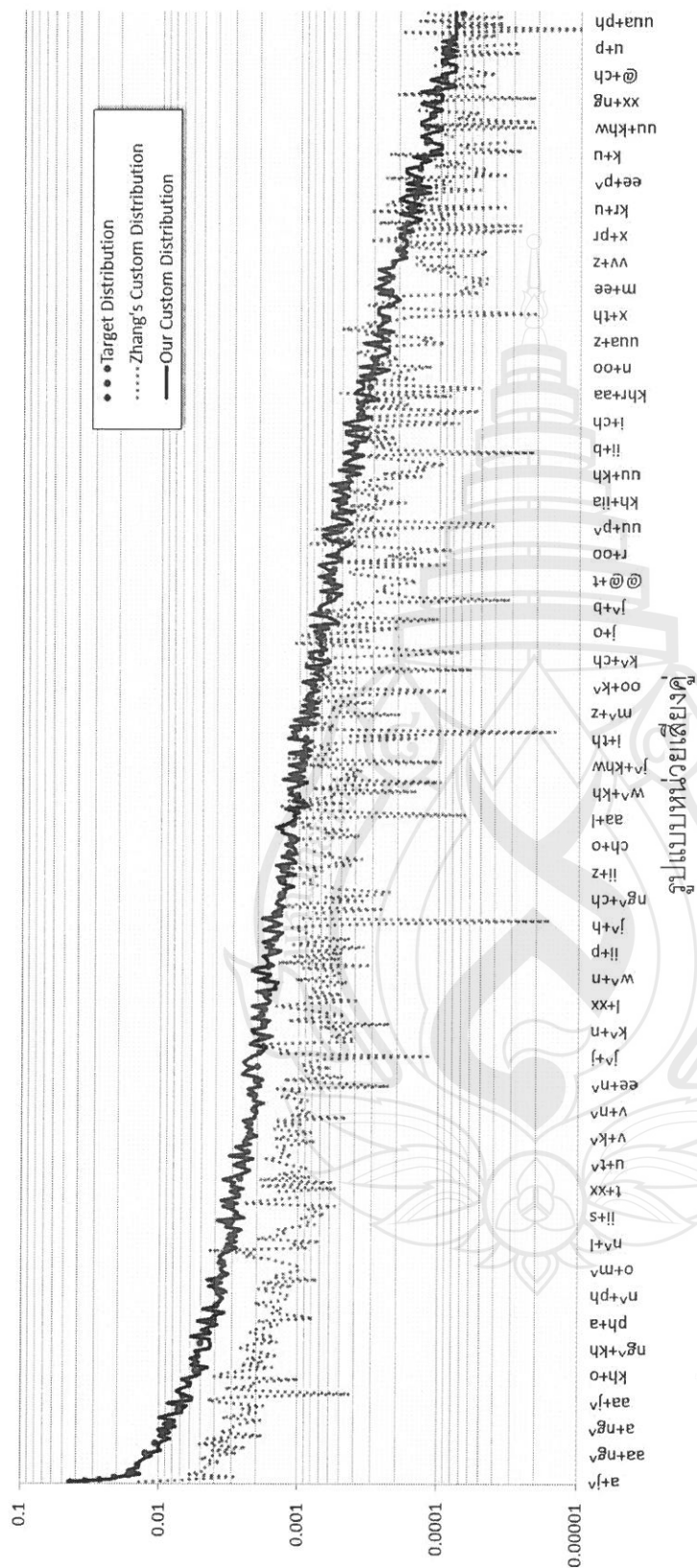
ผลความครอบคลุมทางหน่วยเสียงเมื่อเปรียบเทียบวิธีการเลือกประโยคของงานวิจัยกับวิธีการเลือกประโยคแบบเก่าแสดงดังภาพที่ 4-12 จะเห็นได้ว่าวิธีการให้คะแนนประโยคของ Zhang [11] สามารถทำให้ความครอบคลุมทางหน่วยเสียงเป็นไปตามที่ต้องการได้เนื่องจากใช้ระบบการให้คะแนนไปที่หน่วยเสียงหายากจึงทำให้ความครอบคลุมทางหน่วยเสียงได้ออกมาตามเป้าหมายได้



ภาพที่ 4-13 ผลของค่าเฉลี่ยผลต่างของกระจายตัวทางหน่วยเสียงเป้าหมายเมื่อเปรียบเทียบวิธีการเลือกประโยคของงานวิจัยกับวิธีการเลือกประโยคแบบเก่า

แต่เมื่อพิจารณาผลของความเปลี่ยนแปลงค่าเบี่ยงเบนเฉลี่ยร้อยละจากการกระจายตัวทางหน่วยเสียงเป้าหมาย แสดงดังภาพที่ 4-13 พบว่าวิธีของ Zhang [11] ไม่สามารถเข้าถึงการกระจายตัวทางหน่วยเสียงตามที่ต้องการได้ จากการเปรียบเทียบผลของความเปลี่ยนแปลงค่าเบี่ยงเบนเฉลี่ยร้อยละจากการกระจายตัวทางหน่วยเสียงเป้าหมายที่มีค่ามากขึ้นเรื่อย ๆ เมื่อมีประโยคใหม่เข้ามาทั้งนี้เนื่องจากการให้คะแนนประโยคของ Zhang [11] มุ่งสู่ความครอบคลุมทางหน่วยเสียงเท่านั้น จึงไม่สามารถเข้าถึงการกระจายตัวตามหน่วยเสียงเป้าหมายได้ ส่วนวิธีการให้คะแนนของงานวิจัยนี้สามารถเข้าถึงการกระจายตัวทางหน่วยเสียงตามเป้าหมายได้ เนื่องจากวิธีการให้คะแนนประโยค ที่นอกจากจะให้คะแนนประโยคที่ประกอบด้วยหน่วยเสียงหายากแล้วยังให้คะแนนประโยคโดยการพิจารณาว่าประโยคนั้นจะทำให้การกระจายตัวทางสถิติเข้าสู่เป้าหมายได้หรือไม่อีกด้วย สามารถดูลักษณะการกระจายตัวทางหน่วยเสียงของวิธีการให้คะแนนประโยคของงานวิจัยนี้เปรียบเทียบกับวิธีการให้คะแนนประโยค Zhang [11] ได้ แสดงดังภาพที่ 4-14





ภาพที่ 4-14 แผนภาพการกระจายตัวของหน่วยเสียงจากการคัดเลือกเปรียบเทียบวิธีของ Zhang

จำนวนที่เกิดขึ้นของรูปแบบหน่วยเสียงคู่ (Log_{10})

บทที่ 5

สรุปผลการวิจัยและข้อเสนอแนะ

5.1 การสรุปผลการวิจัย

บทความนี้ได้เสนอวิธีการการสร้างคลังข้อความอัตโนมัติ จากการกระจายตัวของหน่วยที่กำหนด โดยใช้การเลือกประโยคจากอินเทอร์เน็ตเพื่อให้ได้คลังข้อความตามที่ต้องการ จากการทดลองจึงสรุปได้ว่าการเลือกประโยคจากอินเทอร์เน็ตมาสร้างคลังข้อความสามารถสร้างคลังข้อความตามการกระจายตัวของหน่วยที่กำหนดได้อย่างมีประสิทธิภาพ และได้ประโยชน์ที่ทันสมัยในการทดสอบระบบในเรื่องการกำหนดค่าพารามิเตอร์ต่าง ๆ ซึ่งผู้ใช้สามารถกำหนดได้เอง สามารถสรุปผลการปรับค่าพารามิเตอร์ได้ดังตารางที่ 5-1

ตารางที่ 5-1 ผลของการเปลี่ยนแปลงค่าพารามิเตอร์

พารามิเตอร์	หน่วย	ผลของการปรับค่าพารามิเตอร์	
		เพิ่มค่าพารามิเตอร์	ลดค่าพารามิเตอร์
รูปแบบของหน่วยเสียง	หน่วยเสียง/ รูปแบบ	- พบรูปแบบหน่วยเสียงได้ยาก - ใช้เวลาในการสร้างคลังข้อความมากขึ้น - ใช้กับงานสังเคราะห์เสียงได้ดีขึ้น เสียงเป็นธรรมชาติมากขึ้น	- พบรูปแบบหน่วยเสียงได้ง่ายกว่า - ใช้เวลาในการสร้างคลังข้อความน้อยลง - น้อยเกินไปจนทำให้ไม่เหมาะกับงานสังเคราะห์เสียง เสียงขาดความธรรมชาติ
จำนวนขั้นต่ำในแต่ละรูปแบบหน่วยเสียง	หน่วยเสียง	- กำหนดมากเกินไป สะสมรูปแบบหน่วยเสียงที่หายากได้ไม่พอกับความครอบคลุมทางหน่วยเสียงที่ต้องการ - ใช้เวลาในการสร้างคลังข้อความมากขึ้น	- ใช้เวลาหารูปแบบหน่วยเสียงที่หายากกว่าเข้าถึงความครอบคลุมทางหน่วยเสียงได้ดีกว่า - ใช้เวลาในการสร้างคลังข้อความน้อยลง
ขนาดของจำนวนประโยคในคลังข้อความ	ประโยค	- ความครอบคลุมทางหน่วยเสียงเพิ่มขึ้นได้เร็วขึ้น	- ความครอบคลุมทางหน่วยเสียงเข้าถึงเป้าหมายได้ช้าลง

พารามิเตอร์	หน่วย	ผลของการปรับค่าพารามิเตอร์	
		เพิ่มค่าพารามิเตอร์	ลดค่าพารามิเตอร์
		- ใช้เวลาในการสร้างคลังข้อความมากขึ้น - การกระจายตัวทางหน่วยเสียงเป็นไปตามเป้าหมายได้ดีกว่า	- ใช้เวลาในการสร้างคลังข้อความน้อยลง - การกระจายตัวทางหน่วยเสียงเป็นไปตามเป้าหมายได้น้อยลง
ขอบเขตขั้นต่ำจำนวนพยางค์ในแต่ละประโยค	พยางค์/ประโยค	- จำนวนพยางค์ที่มากเกินไปอาจทำให้ การให้คะแนนประโยคผิดพลาด	- จำนวนพยางค์ที่น้อยเกินไปอาจทำให้ประโยคไม่ได้ใจความ
ค่าคงที่ความสำคัญรูปแบบหน่วยเสียงหายาก(α)	-	- เข้าสู่เป้าหมายของความครอบคลุมทางหน่วยเสียงได้เร็ว - รักษาประโยคที่มีหน่วยเสียงที่หายากไว้ได้ - การกระจายตัวทางหน่วยเสียงเข้าสู่เป้าหมายได้ยากขึ้น	- เข้าสู่เป้าหมายของความครอบคลุมทางหน่วยเสียงได้ช้ากว่า - รักษาประโยคที่มีหน่วยเสียงที่หายากไว้ไม่ได้ ต้องไปหาใหม่ - การกระจายตัวทางหน่วยเสียงเข้าสู่เป้าหมายได้ดีขึ้น
ความครอบคลุมทางหน่วยเสียง	เปอร์เซ็นต์	- คลังข้อความมีความครอบคลุมทางหน่วยเสียงที่ดี เหมาะสำหรับการใช้งานในทุก ๆ ด้าน - ใช้เวลาในการสร้างคลังข้อความมากขึ้น	- คลังข้อความมีความครอบคลุมทางหน่วยเสียงที่ดี เหมาะสำหรับการใช้งานในทุก ๆ ด้าน - ใช้เวลาในการสร้างคลังข้อความน้อยลง
ค่าเบี่ยงเบนเฉลี่ยร้อยละจากเป้าหมาย	เปอร์เซ็นต์	- เข้าสู่การกระจายตัวทางหน่วยเสียงเป้าหมายได้น้อยลง - ใช้เวลาในการสร้างคลังข้อความน้อยลง	- เข้าสู่การกระจายตัวทางหน่วยเสียงเป้าหมายได้มากขึ้น - ใช้เวลาในการสร้างคลังข้อความมากขึ้น

การกำหนดพารามิเตอร์ต่าง ๆ ดังที่กล่าวมาเป็นสิ่งสำคัญที่จะทำให้คลังข้อความออกมาตามเป้าหมายที่เราต้องการ ในบางงานอาจต้องการสร้างคลังข้อความที่มีการกระจายตัวทางสถิติ

และความครอบคลุมทางหน่วยเสียงในระดับดีมาก แต่ในบางงานอาจต้องการสร้างคลังข้อความ ในระยะเวลาที่กำหนดเนื่องจากเวลาน้อย ในที่นี้ผู้ใช้จึงต้องกำหนดพารามิเตอร์ให้ดี

คลังข้อความที่ได้นี้จึงเป็นคลังข้อความที่ดีสามารถคาดการณ์ประสิทธิภาพได้ หลักการเลือกข้อความของงานวิจัยนี้ ยังสามารถนำไปประยุกต์ใช้การสร้างคลังข้อความกับภาษาอื่น ๆ ได้อีกด้วย โดยการกำหนดแหล่งข้อมูลออนไลน์ให้ตรงกับภาษาที่ต้องการ กำหนดตัวกรองประโยคใหม่ และเลือกใช้เครื่องมือตัวแปลงรูปเขียนเป็นรูปอ่านของภาษาให้ตรงกับภาษาที่กำหนด

5.2 ข้อเสนอแนะ

ในขั้นตอนการคัดเลือกประโยคของงานวิจัยนี้ เมื่อพิจารณาประโยคที่ถูกคัดออกในการคัดเลือกส่วนใหญ่แล้วเป็นประโยคที่น่าจะนำไปใช้ได้ แต่ประโยคเหล่านั้นประกอบไปด้วยตัวอักษรหรือสัญลักษณ์พิเศษต่าง ๆ ที่ไม่สามารถแปลงรูปเขียนเป็นรูปอ่านได้ ทำให้ระบบต้องทิ้งประโยคที่ดีเหล่านั้นออกไปเป็นจำนวนมาก จึงอยากเสนอแนะให้ปรับปรุงตัวแปลงรูปเขียนเป็นรูปอ่านที่ครอบคลุมการแปลงของตัวอักษรที่มีสัญลักษณ์พิเศษที่พบอยู่ในแต่ละภาษาด้วย โดยอาจใช้เทคนิคการแยกสัญลักษณ์พิเศษประเภทที่จะทำให้การแปลงรูปเขียนเป็นรูปอ่านไม่สำเร็จออกก่อน เมื่อทำการแปลงรูปเขียนเป็นรูปอ่านส่วนประโยคหลักผ่านแล้ว ค่อยนำสัญลักษณ์เหล่านั้นมาใส่ในประโยคหลักทีหลัง

จากปัญหาการค้นหารูปแบบหน่วยเสียงบางหน่วยที่หายาก อาจเกิดมาจากข้อเฉพาะที่เกิดขึ้นในคลังข้อความเป้าหมาย ทำให้การเข้าถึงความครอบคลุมทางหน่วยเสียงเป็นไปได้ช้า จึงอยากเสนอแนะให้ใช้วิธีการนำคำศัพท์ที่เกิดจากรูปแบบหน่วยเสียงที่มาจากข้อเฉพาะเหล่านั้น ใช้เป็นคำสำคัญในการค้นหาประโยคจากอินเทอร์เน็ต จะทำให้การได้มาของหน่วยเสียงที่หายากหรือหน่วยเสียงที่มาจากข้อเฉพาะเป็นไปอย่างรวดเร็ว ทำให้สามารถเข้าถึงความครอบคลุมทางหน่วยเสียงเร็วขึ้น

ปัญหาที่คาดว่าจะเกิดขึ้นจากความต้องการสร้างคลังข้อความในการใช้งานจริงคือ การสร้างคลังข้อความที่ต้องการนั้น อาจจะไม่สามารถหาการกระจายตัวทางหน่วยเสียงของเป้าหมายที่เหมาะสมได้ งานวิจัยในอนาคตของเรามุ่งไปที่การสร้างการกระจายตัวเป้าหมายที่เหมาะสมอัตโนมัติ โดยพิจารณารูปแบบและจำนวนหน่วยเสียงตามโอกาสที่เกิดขึ้นในแต่ละภาษา

มีแนวคิดเรื่องการทดสอบคุณภาพของคลังข้อความที่ถูกสร้างขึ้นเพิ่มเติม โดยนำข้อความจากคลังข้อความที่สร้างขึ้นจากกระบวนการของงานวิจัยนี้ นำไปอ่านบันทึกเสียงโดยอาสาสมัครเพื่อสร้างคลังเสียงจริง และนำไปประยุกต์ใช้ในการสร้างระบบรู้จำเสียงพูดอัตโนมัติหรือระบบสังเคราะห์เสียงพูดภาษาอื่น ๆ เพื่อทดสอบคุณภาพ และประเมินผลของคลังข้อความ

บรรณานุกรม

- [1] Wijaya, T., and Wijaya, D. Limited Speech Recognition for Controlling Movement of Mobile Robot Implemented on ATmega162 Microcontroller. International Conference on Computer and Automation Engineering (2009) : 47–350.
- [2] NV Access Inc. NonVisual Desktop Access (NVDA) [Online]. 2011. Available from: <http://www.nvda-project.org/> [2011,July 6]
- [3] NECTEC. TVIS Traffic Talk [Online]. 2011. Available from: <http://www.tvis.nectec.or.th/> [2011, August 8]
- [4] Stan, A., Yamagishi, J., King, S., and Aylett, M. The Romanian Speech Synthesis (RSS) corpus: building a high quality HMM-based speech synthesis system using a high sampling rate. Speech Communication 53,3 (2011): 442-450.
- [5] Wei, Z., Ranran, D., Minhui, P., and QiuHong, W. Automatic Speech Corpus Construction from Broadcasting Speech Databases. Computational Intelligence and Security (CIS), 2010 International Conference (2010) : 639–643.
- [6] Hansakunbuntheung, C., Rugchatjaroen, A., and Wutiwiwatchai, C. An intensive design of a Thai speech corpus. International Symposium on Natural Language Processing (SNLP) (2007) : 27-132.
- [7] Cormen, Leiserson, and Rivest. Introduction to Algorithms. Chapter 16 Greedy Algorithms (1990) : 329-360.
- [8] Mandal, S., Das, B., Mitra, P., and Basu, A. Developing Bengali Speech Corpus for Phone Recognizer Using Optimum Text Selection Technique. Asian Language Processing (IALP), 2011 International Conference (2011) : 268 –271.
- [9] Chulalongkorn University. ChulaDaisy: Daisy full-text audio book creator [Online].2010.Available from:<http://www.chuladaisy.eng.chula.ac.th> [2011,July 1]
- [10]Kominek, J., and Black, A.W. CMU ARCTIC database for speech synthesis, Technical Report CMU LTI-03-177 (2003).
- [11]Zhang W., Liu Y., Deng Y., Pang M., Automatic Construction for a TTS Corpus with Limited Text. International Conference on Measuring Technology and Mechatronics Automation (2010) : 707-710.
- [12]Tarsaku, P., Sornlertlamvanich, V., and Thongpresirt, R., Thai Grapheme-to-Phoneme Using Probabilistic GLR Parser. Eurospeech 2001 2 (2001) : 1057-1060.

- [13] Kasuriya, S., Sornlertlamvanich, V., Cotsomrong, P., Kanokphara, S., Thatphithakkul, N. Thai Speech Corpus for Speech Recognition. Proc. Of Oriental-COCOSDA 2003 (2003) : 54-61.
- [14] Wutiwiwatchai, C., Saychum, S., Rugchatijaroen, A. An Intensive Design of a Thai Speech Synthesis Corpus. Proceeding of the SNLP 2007 (2007).
- [15] Mittapiyanurak, P., Hansakunbuntheung, C. and Teprasit, V. and Sornlertlamvanich, V. Issues in Thai Text-to-Speech Synthesis: The NECTEC Approach. Proceedings of NECTEC Annual Conference (2000) : 483-495.
- [16] Wiboon, T. The Engineering Training Report at NECTEC. Engineering Faculty, Khasetart University (1990).
- [17] Deza, E., Deza, M. Dictionary of Distances, Elsevier, ISBN 0444520872 (2006)
- [18] Vlasta V., Petr V. Methods of Sentences Selection for Read-Speech Corpus Design TSD'99, LNAI 1692 (1999) : 165-170.
- [19] Hansakunbuntheung, C., Tesprasit V., and Sornlertlamvanich, V. Thai Tagged Speech Corpus for Speech Synthesis. the Oriental COCOSDA 2003 (2003) : 97-104.
- [20] Sornlertlamvanich, V., Charoenporn, T., and Isahara, H., ORCHID: Thai Part-of Speech Tagged Corpus. National Electronics and Computer Technology Center Technical Report (1997) : 5-19.
- [21] Sornlertlamvanich, V. Probabilistic Language Modeling for Generalized LR Parsing. Technical Report, Department of Computer Science, Tokyo Institute of Technology (September 1998).
- [22] Itou, K., Yamamoto, M., Takeda, K., Takezawa, T., Matsuka, T., Kobayashi, T., Shikano, K., and Itahashi, S. JNAS: Japanese Speech Corpus for Large Vocabulary Continuous Speech Recognition Research. In Journal of Acoustic Society of Japan 20,3 (1999).
- [23] Fransen, J., Pye, D., Robinson, T., Woodland, P., and Young, S. WSJCAM0 Corpus and Recording Description, Cambridge University (1994).
- [24] Luksaneeyanawin, S. A Thai Text to Speech System. Proceeding of the Conference of the Regional Workshops on Computer Processing of Asian Languages (1990) : 305-315.
- [25] Ni, J.F., Hirai, T., and Kawai, H. Constructing a Phonetic-Rich Speech Corpus while Controlling Time-Dependent Voice Quality Variability for English Speech Synthesis. Proc. ICASSPU (2006) : 881-884.
- [26] Wutiwiwatchai, C., and Furui, S. Thai speech processing technology: a review. Speech Communication 49,1 (2007) : 8-27.
- [27] Mittrapiynuruk, P., Hansakunbuntheung, C., Tesprasit, V., and Sornlertlamvanich, V. Issues in Thai-to-Speech Synthesis The NECTECT Approach. NECTEC Technical Journal 2,7 (2000) : 36-47.

- [28] Chitturi, R., Sebsibe H.M, Kumar, R., Elissa, K. Rapid methods for optimal text selection, [Online]. 2011. Available from: <http://www.rohitkumar.net/publications/RANLP2005.pdf> [2011,July 6]
- [29] Ling, S., Yu, H., Ren-hua, W. Corpus design for the Chinese speech synthesis system, In ICSLP-2000 2 (2000) : 391-394.
- [30] Microsoft Public License (Ms-PL). HTML parser Html Agility Pack, [Online]. 2011. Available from: <http://htmlagilitypack.codeplex.com/> [2011, July 6]
- [31] Young, S., Evermann, G., Galse, M., Kershaw, D., Moore, G. Hidden Markov model toolkit – speech recognition toolkit. [Online]. 2011. Available from: <http://htk.eng.cam.ac.uk/> [2011,July 12]
- [32] Aroonmanakun, W. Collocation and Thai Word Segmentation, Proc. SNLP and Oriental COCOSDA Workshop (2002) : 68-75.
- [33] Oracle Corporation, MySQL Connector Net 6.2.4, [Online]. 2011. Available from: <http://dev.mysql.com/downloads/connector/net/> [2011,July 12]
- [34] Oracle Corporation, MySQL Community Server 5.1, [Online]. 2011. Available from: <http://dev.mysql.com/downloads/mysql/5.1.html/> [2011,Jan 12]
- [35] Microsoft Corporation, Microsoft Visual Studio 2010, [Online]. 2011. Available from:<http://www.microsoft.com/visualstudio/th-th/download> [2011,Jan 12]
- [36] Royal Command of H.M. the King, Thai junior encyclopedia project, [Online]. 2011. Available from:<http://kanchanapisek.or.th/kp6/New/index.php> [2011,Feb 13]
- [37] Department of Linguistics Chulalongkorn University. TNC: Thai national corpus. [Online]. 2011. Available from: <http://ling.arts.chula.ac.th/tnc2> [2011,Jan 17]
- [38] Mohammad, A., Abushariah, M., Raja, N.A., Zainuddin, R., Moustafa, E. Phonetically rich and balance speech corpus for Arabic speaker-independent continuous automatic speech recognition system. 10th International Conference on Information Science, Signal Processing and their Applications (2010) : 65-68.



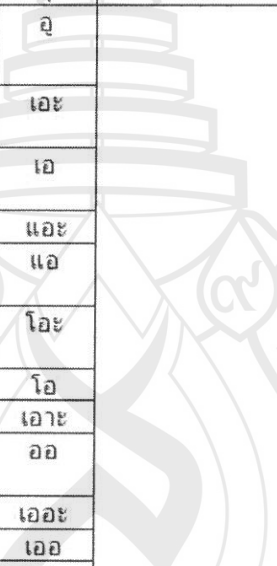
ภาคผนวก

ภาคผนวก ก
ตารางเทียบสัทอักษรสากลกับหน่วยเสียงไทยที่ใช้



Thai Phoneme

พยัญชนะต้น			
เดี่ยว	ตัวอย่าง	ผสม	ตัวอย่าง
p	ปาก	pr	ประสาน
t	เด่น, กฏี	ph r	พราณ
c	จะ	tr	เตรียม
k	ก่อน	kr	กราบ
z	อ่าน	kh r	ครา
ph	พบ, ภัย, ผ่าน	pl	ปลา
th	ทั้ง, ธง, เตา,ฐาน, มกโช	phl	พลาต
ch	ชอบ, เจอ	thr	จันทร์
kh	คน, เขิน , ชา	kl	เกลอ
b	บอก	khl	เคลือบ
d	ด้าน, ชญา	kw	กว้าง
m	ไม	kh w	ขวา
n	นาน, เกร	เสียงท้ายศัพท์	
ng	เงิน	br	บรรณ
l	เลน, กีฬา	bl	บล
r	รอ, ฤทัย	fr	ฟราย
f	ฟัน, ฟัน	fl	เฟลม
s	สาย, สีลา, รักษา, ชวน	dr	ดราคอน
h	โหน, เฮฮา	17 หน่วย	
w	ว่า		
j	ย่อน, หญิง		
21 หน่วย			

สระ			
เดี่ยว	ตัวอย่าง	ผสม	ตัวอย่าง
a	อะ	ia	เอียะ
aa	อา	ila	เอีย
i	อี	va	เอือะ
ii	อี	vva	เอือ
v	อึ	ua	อัวะ
vv	อึ	uua	อัว
u	อุ	6 หน่วย	
uu	อุ		
e	เอะ		
ee	เอ		
x	แอะ		
xx	แอ		
o	โอะ		
oo	โอ		
@	เอาะ		
@@	ออ		
q	เออะ		
qq	เออ		
18 หน่วย			

ตัวสะกด	
เดี่ยว	ตัวอย่าง
p^	พบ
t^	เทร็ด
k^	ปาก
g^	หาร
m^	ลม
ng^	ฟาง
j^	ยาย
w^	กา
เสียงท้ายศัพท์	
f^	กราฟ
l^	แอล
s^	เอส
ch^	คลัช
12 หน่วย	

THE INTERNATIONAL PHONETIC ALPHABET (2005)

CONSONANTS (PULMONIC)

	LABIAL		CORONAL				DORSAL			RADICAL		LARYNGEAL
	Bilabial	Labio-dental	Dental	Alveolar	Palato-alveolar	Retroflex	Palatal	Velar	Uvular	Pharyngeal	Epi-glottal	Glottal
Nasal	m	ɱ	n			ɳ	ɲ	ŋ	ɴ			
Plosive	p b	ɸ β	t d			ʈ ɖ	c ɟ	k ɡ	q ɢ	ʔ ʕ		
Fricative	ɸ β	f v	θ ð	s z	ʃ ʒ	ʂ ʐ	ç ʝ	x ɣ	χ ʁ	ħ ʕ	ħ ʕ	h ɦ
Approximant		ʋ	ɹ			ɻ	j	ɰ				
Trill	ʙ		r						ʀ			
Tap, Flap		ɹ̥	ɾ			ɽ						
Lateral fricative			ɬ ɮ			ɭ						
Lateral approximant			l			ɭ	ʎ	ʟ				
Lateral flap			ɭ									

Where symbols appear in pairs, the one to the right represents a modally voiced consonant, except for murmured *h*. Shaded areas denote articulations judged to be impossible. *ɸ* and *β* are unofficial extensions of the IPA.

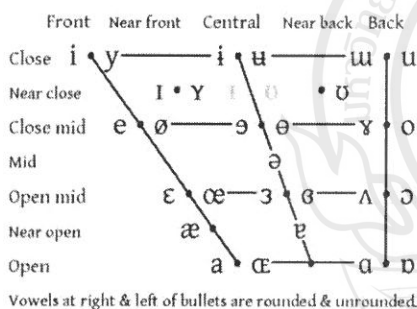
CONSONANTS (NON-PULMONIC)

Anterior click releases (require posterior stops)	Voiced implosives	Ejectives
◌ Bilabial fricated	ɓ Bilabial	ʼ Examples:
◌ Laminar alveolar fricated ("dental")	ɗ Dental or alveolar	◌ Bilabial
◌ Apical (post)alveolar abrupt ("retroflex")	ɟ Palatal	◌ Dental or alveolar
◌ Laminar postalveolar abrupt ("palatal")	ɠ Velar	◌ Velar
◌ Lateral alveolar fricated ("lateral")	ɣ Uvular	◌ Alveolar fricative

CONSONANTS (CO-ARTICULATED)

- ɱ Voiceless labialized velar approximant
- ʋ Voiced labialized velar approximant
- ɰ Voiced labialized palatal approximant
- ɕ Voiceless palatalized postalveolar (alveolo-palatal) fricative
- ʑ Voiced palatalized postalveolar (alveolo-palatal) fricative
- ɧ Simultaneous x and ʃ (disputed)
- kp ts Affricates and double articulations may be joined by a tie bar

VOWELS



Vowels at right & left of bullets are rounded & unrounded.

SUPRASEGMENTALS

- Primary stress ˈ
- Extra stress ˈˈ
- Level tones: ˥ Top, ˦ High, ˧ Mid, ˨ Low, ˩ Bottom
- Contour-tone examples: ˥˥ Rising, ˥˦ Falling, ˥˧ High rising, ˥˨ Low rising, ˥˩ High falling, ˥˪ Low falling, ˥˫ Peaking, ˥ˬ Dipping
- Secondary stress ˌ
- Long eː, Half-long eˑ, Extra-short e̘
- Short e, Linking (no break) ̤
- Syllable break ˑ
- Minor (foot) break ˑ
- Major (intonation) break ˑ
- Global rise ↗, Global fall ↘, Downstep ˥˥˥

DIACRITICS

Diacritics may be placed above a symbol with a descender, as ɲ̥. Other IPA symbols may appear as diacritics to represent phonetic detail: ʰ (fricative release), ʙ̥ (breathily voiced), ʔ̥ (glottal onset), ʔ̥ (epenthetic schwa), ɔ̥ (diphthongization).

SYLLABILITY & RELEASES	PHONATION	PRIMARY ARTICULATION	SECONDARY ARTICULATION
ɲ̥	Syllabic	ɲ̥	ɲ̥
ɲ̥	Non-syllabic	ɲ̥	ɲ̥
ɲ̥	(Pre)aspirated	ɲ̥	ɲ̥
ɲ̥	Nasal release	ɲ̥	ɲ̥
ɲ̥	Lateral release	ɲ̥	ɲ̥
ɲ̥	No audible release	ɲ̥	ɲ̥
ɲ̥	Lowered (β is a bilabial approximant)	ɲ̥	ɲ̥

ภาคผนวก ข
ตัวอย่างประโยคจากคลังข้อความที่สร้างขึ้น



ลำดับ	ประโยคในฐานข้อมูล
1	เธอรอให้เพลงจบ จะเล่าเรื่องเด็กหนุ่มนักเปียโนให้สามีฟัง แต่ประพจน์เหมือนไม่ได้ยินอะไรทั้งสิ้น
2	เจอแล้ว ผมร้องออกมาแล้วก็เปิดประตูล้วยเข้าไป สัมผัสกระทั่งว่าจะต้องเคาะประตูก่อน
3	ทั้งที่จิตของเกือบเสร็จเรียบร้อย ของผาณิตโกโหติกา ผองเพื่อนก็เตรียมไว้เสร็จสรรพ
4	อ้าว คุณปุดอบแกมหัวเราะ นี่แหละบทกวีของกวีเอกชาว อิตาลี
5	กาบบัวสะบัดปาก ปล่อยเหยื่อปลา นกกวักกู่ฟาลันบึงน้ำ กระสาขาส่ายคล้ายย่นสั้น
6	ถ้าอิตเลอร์ไม่หันปลายปืนไปทางมอสโก เยอรมันก็คงไม่แพ้สงคราม
7	ล้างจาน ปิดไฟนอน ค็นนั้นทั้งคู่ร่วมรักกันอย่างดุเดือด เปญคว่าหน้า
8	พี่จิวทะเลาะกับพี่พาย พี่สะใภ้มีเวรดูแลคุณพ่อ แต่เปลือหลังจนน้ำเกลือหมดขวด
9	ปรอดปราดหลีกเสียปีกัด สาลิกากู่ร้องก่อนล่องลัด เขาขวากคล้ายหวัดหวาดขันร้อง
10	กลับกลายเป็นตรงกันข้าม ต่อมาทั้งสองได้รับความช่วยเหลือจากเบตุนกลุ่มหนึ่ง
11	แบบว่าขัดซี่ไคลในเวลาเดียวกัน แต่ก็หายเร็วกว่าความนุ่มนวลโดยการทามันแบบของยาย
12	เหตุการณ์ได้เข้มข้นยิ่งขึ้นเมื่อทั้งสองถูกพวกโจรเบตุนปล้นและทำร้ายบาดเจ็บ
13	ทั้งนี้ ในความควบคุมของเจ้าหน้าที่ผู้ฝึกหัดหรือผู้ให้การฝึกอบรมซึ่งเป็นผู้ประกอบวิชาชีพการพยาบาล
14	รวมทั้งน้ำเสียงอบอุ่นที่ปลุกปลอบเด็กหญิงพิการนำเสนอคนนั้นอยู่ในหัวใจเธอแล้ว
15	นางถือกระเช้าหยาบบรรจุอาหารเข้าประเภทแซนวิช และขนมปังแครกเกอร์
16	เคยเห็นเต่าทะเลใช้ไหมครับ ถ้าเทียบกับเต่าบก เต่าแม่น้ำเต่าทะเลจะต้องจะกลมกว่า
17	สงสารมึงนัก กูตัดใจ กุยกมือนี่ให้มึง นี่คือแขนพร้อมมือที่มีนิ้วครบ
18	ความล้มเหลวทางการเมืองที่ประเทศไม่สามารถผลักดันประวัติศาสตร์ทางการเมืองไปสู่ยุคปฏิรูปได้
19	หากผู้ผลิตรายการเป็นหน่วยงานรัฐ ก็น่าจะผลิตรายการที่ให้บริการแก่คนทุกชั้น
20	เตือนเธอให้ระลึกถึงวันที่เขาเอ่ยปากขอเธอแต่งงาน แต่เพราะวันทวยวิษัณห์แต่คิดแผนหลอกเธออยู่นั่น
21	เพราะยังไม่กลับตัวกัน แค่ความซื่อสัตย์สุจริตก็เป็นคุณภาพที่ขาดตกบกพร่องกันเสียแล้ว
22	หนึ่งชั่วโมงวันจันทร์คือไฟ ข่าเปลวไฟในตะวันกัณฑ์จักรหนึ่งแกนหลักจักรวาลขับเคลื่อน
23	แกแหละจำเขาไม่ได้ ผมโกรธฟุ่ขึ้นมาทันทีที่ได้ยินประโยคนี้
24	เขาบอกว่าดีใจที่ได้เจอฉันซะที แม่เล่าเรื่องของฉันให้ฟังเยอะมาก

ลำดับ	ประโยคในฐานข้อมูล
25	ถึงดวงหทัยที่สดใสและอยู่ใกล้ฉัน สังคมมีทั้งความกลมกลืนและความแตกต่าง
26	ทีเซอร์รั้งฉันไว้นั่นไม่ยอมปล่อย เธอมากับฉันก็ต้องกลับกับฉันสิ
27	ซึ่งเพลโตได้นำชื่อพีธีรสมาเป็นชื่อเรื่องหนังสือของเขา กล่าวถึงผู้ไม่ตกอยู่ในห้วงรัก
28	ว่าเกิดอะไรขึ้นกับหญิงสาวที่เป็นความหวังของมิสซิสคลีฟ คุณเห็นวิกตอเรียใช่ไหม
29	ผมเข้าห้องน้ำซึ่งตอนนั้นว่างคน คราวนี้ผมต้องเจียหูฟังเป็นพิเศษเพราะกลัวจะไม่ได้ยินเสียงวิทยุจากห้องข้าง
30	แต่เมืองไทยเป็นอย่างนี้ ช่างอุดมสมบูรณ์ด้วยความเป็นธรรมชาติที่น่ากลางแฉงและชวนให้คิด
31	เชิญฟังเพลงบรรเลงหวานประสานเสียง เสนาะสำเนียงใต้ต้นไทรพุ่มไผหนา เมื่อแสงเรือเจือชมพูลู่ท้องฟ้า
32	ดวงตาคมกริบเปลือยทอดนึ่งมายังภาพสาวตรงหน้า ก่อให้ความเจ็บโรยตัวโอบล้อมทั้งห้อง
33	พ่อ ฉันมีลาภ เขาชูขึ้นอวด อ้วนปี่เลย นี่ไง มันเชื่อ
34	เสียง ฉันล่องท่องไพรกว้างฟังเสียงค่างบางชะนี เสียงนกพลอดจูจืดจงดนตรีในป่าดอย
35	โลกนับสิบจะออกมาคล้ายคลึงกันเป็นโลกแฝดซึ่งสุดท้ายจะถูกกลืนรวมเป็นหนึ่งเดียว
36	พญาเมงฮายสร้างเสียงใหม่ขึ้นมาเพื่อเป็นศูนย์กลางล่านาเนื้อลูกเนื้อ ก็ดังที่พ่อได้บอกเจ้าแล้วว่าพญาเมงฮายสร้างเสียงใหม่เมื่อ
37	น้องหมวยได้ยินเข้าก็หันไปทางอื่นทันที เพราะเพราะเรื่องนี้เป็นแผลในใจมาโดยตลอด
38	กระบี่เกาะมากสุดในทะเลไทย กระบี่มีเกาะมากกว่าสุราษฎร์ ผสมผสานทั้งหมู่เกาะหินปูนและหมู่เกาะหินแกรนิต
39	สัตยาปรากฏแม่เกลี้ยกลบ สัตว์อย่ามา หมายถึงองค์พยพ สัตว์อย่าหมาย
40	บนฝั่งเกาะเกาลูน เลี้ยงลูกสาวคนเดียวจนเติบโตเป็นสาว เงินหยวนทุกเหรียญ
41	เพลโตใช้วิธีการนี้เพราะคิดว่าเป็นวิธีเดียวที่จะทำให้ค้นพบความจริง
42	หกโมงเย็นตรงเป๊ะ คุณสามาเคาะประตูเรียกให้ไปกินข้าว พอโผล่หน้าเข้าไปคุณยายนั่งที่โต๊ะกินข้าวเรียบร้อยแล้ว
43	หากเมื่อบทสนทนาดำเนินไปได้ไม่นาน มันก็เวียนวนกลับไปอยู่ในรูปเดิม
44	นานและนานหลายนานผมเป็นผื่นคันเนื่องจากเหงื่อและกระเือก ความหมักหมม แม่กลับจากสามชุก
45	รอยยิ้มกว้างปรากฏบนใบหน้าก่อนระโหยจากการกรำศึกสอบ เมื่อนึกถึง
46	พริ้มโรสสีเหลืองสีชมพูแปด บิสตอตำที่ดูเหมือนไมโครโฟนของมิสฟิงค์ก็
47	ข้ามวิถีข้าเถียงภูผาเกินพลัดเซ กว้างเทือกเกินถ้ำเทพเกินเหลิงเล่ห์หลงลมลง
48	และไม่รู้ตัวเลยว่าอีกไม่กี่วินาทีคนส่วนใหญ่จะสิ้นชีวิตเพราะตึกระเบิด

ลำดับ	ประโยคในฐานข้อมูล
49	พื้ดินฉันไม่ชอบให้มีคนนอกมาวุ่นวายที่นี่ คราวก่อนองศาทำเรื่องไว้ก็โกรธแทบแย่
50	เลยไปถึงบริเวณบาร์ข้างสระที่มีเก้าอี้หมู่ให้นั่ง ต่างก็เหม็นมองความงามตระหง่านของ
51	อีกฝ่ายก็สูญเสียเองและชวนเพนกวิ้นคุยอย่างสนิทสนมเสียด้วย มันทำให้เขาเสียดาย ช่วงเวลานี้มาก
52	ยุคนั้นไม่มีตราเรือใบและตรามะลิ หมาสาวหมาหนุ่มหรือหมาแก่ไม่เคยถูกจองไปอาบนํ้า
53	ในทันทีหลังจากนั้นถ้าได้กลิ่นของกินมันไปคลุกและค้นกองขยะเน่า เปื้อนขนตัวเหม็น
54	ฝรั่งหัวเราะแล้วไม่พูดพาลำทำเพลงอีกต่อไป วางสบู่อะผ้าเช็ดตัวพาดเอาไว้
55	ฉันละก็ไม่ได้ ถึงไหนถึงกันเขียว เหม็นกลิ่นกระถางไฟคะ ยุพราตอเรียบ
56	จิต ทรมานแต่คอยรั้งความ นํ้า ฝนหล่นฟ้ามาฝาก มิตร ภาพพลันหลากเลี้ยวย่าน
57	เสียงเครื่องจักรที่ลั่นสมัยและฝุ่นข้างเปลือกที่กระจายคลุ้งอยู่ทั่วไปในโรงสีมิได้ทำให้ เข้าเกิดความขย้อท้อแต่อย่างใด
58	อู้งมาเดินข้างฉัน เธอบอกฉันว่า ข้างมันเถอะ แม่นี่ก็เรื่องมากอย่างนี้แหละ
59	หากว่าสองวันก่อนเธอจะไม่ล้มเลิกแผนการกลับเมืองไทยอย่างกะทันหันเสียก่อน
60	แต่งตัว กินอาหารเข้าเมื่อได้เวลาเช้างาน เขาจะเดินออกไปตรวจดูกระบวนการสีข้าวทุก ขั้นตอนอย่างละเอียด
61	มินะมะตะ ปลาเล็กกินแพลงตัน ปลาขนาดกลางกินปลาเล็ก ปลาใหญ่กินปลาขนาด กลาง
62	คะ คือแบบว่าโอย อายจัง กิรดาหลบยิ้มกับท่าทางของหนูน้อยตรงหน้า
63	มุกขัยจะก้าวพรวดไปข้างหน้าเพื่อดูให้แน่ใจอีกครั้ง หากเซเสียหลักเสียก่อน
64	ส่งโน่นส่งนี่มาให้บอกให้ทันยกครัวชนิดนี้ตอบสนอง กะจิงหะไม่ให้ลูกอย่างเหี้ยมโหด
65	มองเฟิร์นเพลินพิศใบดูลงดูไวกวามน่าดู เก้งปามาเป็นคู่ช่างแสนรู้ระวังภัย
66	แต่วิธีของยายนั้นละลายก้ามะถันกับน้ำหน่อไม้ต้องเปรี้ยวขโหลม เผลอไม่นานหันมองก็ พบว่าโรคนั้นล่อนออกเกลี้ยงเกลา
67	สงครามของหมาวัดค่อนข้างโหดร้ายและโหด หมาในสภาพต่างเอาตัวรอดสั่งสมความ เลวมากกว่าความดี
68	ทันทีที่มีแสงสว่างก็เห็นนาฬิกาทรงโบราณพร้อมทั้งเสียงที่ดังขึ้น
69	และได้สูญเสียมิตรภาพไป เขากลายเป็นคนเงิบขริ้ม ซึมเศร้านานเท่าไรหนอ
70	ตั้งชื่อลูกสาวว่า จาง หรือ ลินลี่ หลังจากคลอดลูกได้ไม่นานเสียงก็เสียชีวิตไป
71	การเลี้ยงสัตว์ก็ตกอยู่ในชะตากรรมเดียวกันกับการเพาะปลูก หลังสงคราม
72	และแนวทางของคณะกรรมการพัฒนาระบบราชการ ซึ่งครอบคลุมถึงการปรับบทบาท
73	อาจอ้หมินป่วยเป็นโรคมะเร็งหมอวินิจัยว่าอาการขั้นสุดท้ายจะมีชีวิตอยู่ไม่เกินหก เดือน

ลำดับ	ประโยคในฐานข้อมูล
74	มีเมียน้อยอีกคนละซี มีเมียน้อยอีกสองคนกระมัง เขาแก้งกระซิบถามอีก
75	ด้านหลังมหาวิหาร ผ่านพิพิธภัณฑสถาน ดุโอโมโดยไม่แวะชม เพราะเธอบอกว่าได้เข้าไปดูภายในวิหารแล้วเมื่อวาน
76	การประสานงานกับหน่วยรับตรวจและหน่วยตรวจสอบภายนอก และงานประชาสัมพันธ์งานคณะกรรมการตรวจสอบ
77	และซาริฟได้กลับไปช่วยกษัตริย์ของประเทศเขาปราบกบฏสำเร็จ ทั้งสองได้เข้าพิธีแต่งงาน
78	เพราะต่อมวัดระดับความหล่อของฉันทน์มันร้องเตือนว่าความหล่อของเขาถึงขีดมาตรฐาน
79	แทบไม่รู้ตัวว่ามาถึงบ้านได้อย่างไร ก่อนจะก้าวขึ้นเรือน พ่อชะงักเหมือนถูกตรึงกับแผ่นดินเมื่อได้ยินบุญเพ็งเรียกเสียงดัง
80	เป็นการใส่ร้ายทั้งนั้น สาบานได้ ให้ฟ้าผ่าลิเอ้า ถ้าผมเลวจริง
81	จะเห็นปรากฏการณ์นี้ได้ชัดเจน ทางด้านเหนือฝั่งฝรั่งเศสจะมีป่ามากกว่า
82	เรานอนตากลมเล่นที่ได้ต้นไม้ชายทุ่งอีกสักพักก็กลับ เพื่อให้ทันอาหารกลางวัน
83	หลายคนหิมนแคลนเสียแค่นับ แต่มียอมมอบรับสิ่งอัปรีดิ์ ฟ้า เมฆาณ
84	ค่านิยมของนักการเมือง ซึ่งเป็นตัวแปรสำคัญในเรื่องนี้ บางคนยอมรับว่า
85	ก็จบการศึกษาระดับปริญญาตรีทางด้านธุรกิจและการบัญชีจากมหาวิทยาลัยมีชื่อที่สุดแห่งหนึ่งของเกาะฮองกงด้วยวัย
86	มันเตะแล้วกระต๊อบฉันทน์จมน้ำตายเกือบตาย สู้มันบนหาดทรายไม่ได้ พอมันเพลดตอนอาบน้ำ
87	อะริโยะมิได้พูดถึงสารเคมีที่ใช้ในการ เพาะปลูกว่ามีมากมายหลายชนิด
88	ซึ่งมีใบอนุญาตเป็นผู้ประกอบวิชาชีพการพยาบาล การผดุงครรภ์
89	เพื่อให้การปฏิบัติงานเกี่ยวกับการพัฒนาระบบราชการบรรลุตามวัตถุประสงค์
90	ผมพลิกตัวกลับมาทันที ตามองไปที่กรงนกที่แขวนลอยอยู่ เจ้าจิวโผล่หน้าทำตาบ้องแบ้วอยู่
91	คือตัวกลางที่จะต้องเอานโยบายมาแปลงเป็นภาคปฏิบัติให้ได้ตามที่รัฐบาลเขาต้องการ
92	ผมชันสายลูงขึ้นนั่งมองผ่านหน้าต่างไปทางก้อนน้ำ จับพลันที่เห็นภาพ
93	ลูกหามไม่ได้ให้คำตอบและเรียกเจ้าหญิงว่า เจ้าหญิง ชื่อเจ้าหญิงซึ่งเป็นชื่อหนังสือรวมเรื่องสั้นเล่มนี้จึงมาจากเรื่องสั้นเรื่องนี้นั่นเอง
94	รัฐบาลญี่ปุ่นเร่งที่จะผลิตเนื้อสัตว์เพื่อเป็นอาหารให้ได้จำนวนมาก ทำให้เกิดปัญหาต่าง
95	และจะพยายามกีดกันคนรักของตนออกจากปรัชญาที่เทพเจ้าประทาน วาตะซันที่
96	รอยยิ้มแฉ่งยิ้มที่ไม่เชิงว่ามีไมตรีหรือรังเกียจก้ำกึ่งกันอยู่บนริมฝีปากแดงเข้ม
97	เจ้าหกไปเป็นตราให้กระทรวงเกษตร ส่วนเจ้าเจ็ดเป็นนายแบบแฟชั่นถ่ายรูปลงในนิตยสาร

ลำดับ	ประโยคในฐานข้อมูล
	ชิดไฟ
98	โดยเฉพาะอย่างยิ่งสามารถให้บริการกับนักเรียน นักศึกษา เด็ก เยาวชน
99	หลักสูตรการศึกษาในโรงเรียนทุกระดับยังไม่สอดคล้องกับความเป็นอยู่ของคนในชนบท
100	กับข้าวเยอะแยะ ถ้าไม่ถือว่าเป็นปิ่นโต ก็มากินด้วยกัน เย็นวันนั้นคุณยายเจริญอาหารเป็นพิเศษ
101	ข้าก่าเน็ดดินน้ำลมและไฟหนึ่งเกิดอาจดับไหม้ในพริบตา ข้าท้องฟ้าเมฆหมอกนอกหนหนาวหนึ่งดวงดาวพราวแสงแห่งเวหา
102	ข้าวอดโอโอหังพ่ายพังเพหนึ่งเวลากัปกลัป้อาสัญโลก ข้าเป็นหนึ่งในมวลล้วนชาคศพหนึ่งจักรภพประสพสุขทุกขโคก
103	เกี่ยวข้องกับชี้แจงรายละเอียดหรือแสดงหลักฐานเพิ่มเติม หรือจัดส่งเจ้าหน้าที่ไปตรวจสอบข้อเท็จจริงยังสำนักงาน
104	ประกอบกับลักษณะทางภูมิศาสตร์ที่เต็มไปด้วยเขาสูงและหุบลึก มีสังคมพืชเขตกึ่งร้อนเกือบทุกระดับ
105	ป่า ยิ่งยิ่งใหญ่ใกล้ศึก คือ ความเคร่งเครียดโครมครึก เมือง หลับหลงดึกยาวนาน
106	ต้องไล่ตะครุบ ส่วนสหายเพียวนั้นจะดำน้ำจะว่ายอย่างไรผ้าก็ไม่หลุด
107	เมื่อวิเคราะห์กันต่อไปถึงทางออก กลับเป็นเรื่องค่อนข้างแปลก ที่ผู้ซึ่งมองเห็นว่า
108	และกลัวว่าพ่อแม่จะทิ้งตนไป จึงไม่ยอมห่างพ่อแม่ ในเรื่องของคนไข้เด็กนี้
109	เพราะเป็นกระบวนการสื่อสารแบบทางเดียว นอกจากนั้นยังมีข้อเท็จจริงเพิ่มเติมที่ว่า
110	เนื่องด้วยมีดอกกันอยู่ปลายเท้า เป็นดอกอาบน้้างกล่าวนั้น แลแคร่ด้านซ้ายก็กันด้วยแผงไม้ไผ่
111	จนถึงหุบเขาลึกแคบรูปตัววีที่ถูกร่องน้ำเซาะ เพมาโคจึงเป็นแหล่งรวมพันธุ์ไม้ที่หลากหลายน่าอัศจรรย์
112	เกี่ยวกับการโอนนักโทษ ให้เป็นไปตามที่กำหนดในกฎกระทรวง มาตรา
113	กับข้าวของคนจีนจะต้องมีแกงจืดขามใหญ่ขามหนึ่งไว้ชดน้ำด้วยเสมอและไม่มีช้อนกลาง
114	เหตุเพราะความรู้เท่าไม่ถึงการณ์ สรรพสิ่งมีสองข้าง แม้แต่เรื่องนม นม
115	เพื่อกระชับการบรรยายให้กะทัดรัด เข้าใจง่าย และเห็นภาพกว่าการบรรยายธรรมดา
116	กล่าวกันว่าในสมัยที่ประชากรส่วนใหญ่ในสังคมยังด้อยซึ่งพลังวิเคราะห์
117	อาการยังทรงอยู่ แต่หมอบอกว่าให้เตรียมใจไว้อาจไม่ถึงสัปดาห์หน้า
118	ลามไหม้ทั้งด้านเหนือและใต้มุ่งไปทางทิศตะวันตกเข้าหาลำน้ำแม่ยมซึ่งมีวัดราชธานีที่เต็มไปด้วยโบราณวัตถุขวางอยู่

ลำดับ	ประโยคในฐานข้อมูล
119	ใช่ นั่นเสียงซึ่งอ้ายอินถา ข้าพเจ้าเจียหู่ฟัง เสียงซึ่งดังจ๋ง จ๋งโกล่เข้ามา
120	จำชื่อนี้ได้ แก่นแก้วนั้ ก็เพราะความรู้ภาษาอังกฤษที่เคยร่ำเรียนมากับอาจารย์
121	หรือองค์กรของผู้สูงอายุ ที่ดำเนินการตามวัตถุประสงค์ และกิจกรรมเกี่ยวกับการ คุ้มครอง
122	แม่ท้องฟ้าเบื้องนอกจะยังคลุ้มคลั่งด้วยม่านกำมะหยี่สีด้าอยู่ หากเขาก็ไม่อาจชมตาหลับ ได้อีก
123	บำบัดแก้ไขฟื้นฟูและสงเคราะห์ผู้กระทำผิด สนับสนุนการแก้ไขปัญหายาเสพติดและ สังคมของประเทศ
124	จึงได้ทราบว่เกษตรกรยังต้องออกไปทำงานทำในเมืองใหญ่เพื่อนำเงินมาซื้อเครื่องจักร และปุ๋ยสำหรับใช้ในการเพาะปลูก
125	แล้วควายมันก็ลงน้ำนอน โผล่แต่หัวที่มีเขาเกะกะขึ้นมาสะบัด
126	โดนดอกเบ๊ยไล่จี้ทุกวัน บางคนต้องเปลี่ยนอาชีพมาทำงานกลางคืนเพื่อความอยู่รอด
127	โดยมิเชลล์ยังไหวหวั่น กลัวชาร์ฟจะมีภรรยาอีกตามที่ศาสนานุญาต
128	ความไม่สบายใจจากสิ่งที่เพิ่งได้ยินเข้าเกาะกุมจิตใจของเขาจนหนักอึ้ง
129	คุณปู่จึงต้องแก้แค้นโดยการจับหัวประสิทธิ์ศักดิ์อีกแล้ว น้องหมวยเห็นภาพนี้เป็นครั้งที่ สอง
130	เคย ต่อ บรรยากาศที่นี้คล้ายตลาดนัดในเมืองไทย แต่สินค้าหลากหลายกว่า
131	ผมแทบช็อค เพราะฝรั่งพเนจรคนนั้นกำลังเปลือยกายล่อนจ้อนอาบน้ำอยู่คนเดียว
132	นิทรรศการ ศูนย์รับบริจาคหนังสือ เป็นหน่วยงานที่ทำหน้าที่บริการรับบริจาคหนังสือ
133	ธนะรัชต์ ไม่พอใจแค่การผสมผสานผู้นำหรือลดความรู้สึกต่อต้านในหมู่ผู้นำต่ออำนาจ รัฐ
134	ชายหนุ่มยังไม่สามารถทำอะไรลงแปลงได้ คงต้องอาศัยเวลาอีกนานทีเดียว
135	ผมมากินก๋วยเตี๋ยว ไม่ได้มาดูโชว์ กลางคืนของสีลม ตามฟุตปาธทางเดินตั้งแต่หน้า กรุงเทพคริสเตียนถึงหน้าห้างโรบินสัน
136	และกระทรวงมหาดไทย โดยมีเนื้อหารายละเอียดของการดำเนินการที่ดูจะแตกต่างกัน ออกไป
137	ทรงเชี่ยวชาญเรื่องดนตรีไทยและเพลงไทย จอห์นกระซิบ อัสธามองตามสายตาของเขา
138	โดยบรรยายประกอบรูปไว้ว่า พญายักษ์ทั้งเจ็ดต่างสำนึกตัวกลับเป็นงูดี
139	เรนอดรู้สึกไม่ได้ว่ามันทั้งสกปรก ทั้งขุ่นข้น ดูน่ากลัวพิลึก
140	แล้วก็ได้เห็นไฟประลัยกัลป์ฉายแวว เข้มโหดเท่าที่คนสุโขทัยเคยประสบกลางตลาด ใหญ่เป็นห้องแถวเรือนไม้เก่าสร้างต่อเนื่องกัน
141	คิดแต่ว่าชีวิตของตัวเองจบสิ้นตามการจากไปของคนที่ยรักยิ่ง หากมาวันนี้เขาสามารถ

ลำดับ	ประโยคในฐานข้อมูล
	ปลดปล่อยความทุกข์โศกนั้นได้
142	นกร้องถี่ แล้วเราก็อโผล่ถึงที่โล่ง หย่าต่ำ มันเขี้ยวอะไรอย่างนั้น
143	ศึกษา ค้นคว้า วิเคราะห์ วิจัย รวบรวมสถิติข้อมูลเพื่อปรับปรุงกฎหมายและระเบียบ
144	ยังชีพด้วยการเป็นพ่อค้าขายสมุนไพรเงินจากเงินแผ่นดินใหญ่แถวย่านชิมชาซุ่ย
145	ยังไม่กำหนดไว้เด็ดขาดตายตัว คงจะต้องเตรียมอะไรบางอย่างให้พร้อมเสียก่อน
146	และงอกขนใหม่สวย การอาบและฟอกสบู่ยากกับไอ้มิดอีหรือวันนั้นจำเป็นต้องเรียกมันขึ้นบ้าน
147	ท่าทางจะอายุจริง อายุ คือตอนนี้ก็กำลังติดหนังสืออย่างรุนแรงนะค่ะพี่แหวน
148	ออกจากซาเคร์ท สปริง ก็เคลื่อนเดินไปดูบริเวณที่เรียกว่าเท็มเพิล คอร์ทยาร์ด
149	พวกเครื่องบวงสรวงทั้งหลายก็เอาไปตั้งตรงบริเวณที่เป็นแท่นบูชา ที่นี้มีจอโทรทัศน์ให้ดูภาพและเรื่องราวประกอบด้วย
150	เสียงระเบิดกึ่งก้อง แผ่นสังกะสีคมกริบนับไม่ถ้วนปลิวว่อนลอยจวดเฉวียนขึ้นไปกับกลุ่มควันนำหวาดเสียว
151	ตุลล่ายหมอนปักเข็ม การกอดติดกันช่วยให้มันเก็บความอบอุ่นไว้ข้างในพุ่ม
152	อัมชันเงินเปล่า เงยหน้าตะเบ็งเสียงตอบ หญิงวัยกลางคนท่าทางเรียบร้อย
153	กระดาด ผ้า เศษวัสดุ พื้นระนาบที่ใช้แผ่นวัสดุปะติด อาจเป็นแผ่นไม้
154	ตัวอย่างเช่น ระบบมัธยมศึกษาเป็นหลักสูตรที่ถูกบีบลงมาจากระบบอุดมศึกษา
155	ใจอยากปฏิเสธ หากความมืดและเปลี่ยวเล็กน้อยบริเวณนั้นทำให้จำเป็นต้องพยักหน้ารับความช่วยเหลือที่
156	ผมนึกถึงต็อกเตอร์ปัทมา เธอเกาะแขนเหมือนต้องการเครื่องยึดเหนี่ยวชีวิตของคนเรา มีวิธีดำเนินที่พิสดารอย่างไม่น่าจะเป็นไปได้
157	พลอยเดินเข้าไปใกล้ ได้ยินเสียงอีกฝ่ายบรรยาย แต่ก่อนชาวเลเช่นหล่อโบงด้วยเต้าทะเล
158	สาวหน้าหวานที่ไม่เคยอับอายกับการอ่านนิยายโรมานซ์ซึ่งเธอถือว่ามันเป็นนิยายแปลประเภทหนึ่งทำหน้าที่แปลกใจ
159	ซึ่งล้วนเป็นอันตรายต่อมนุษย์ทั้งสิ้น และได้พูดถึงสภาพความเป็นอยู่ของเกษตรกรว่ายังลำบากอยู่
160	ซึ่งเป็นย่านการค้าของหมู่บ้านประมาณครึ่งกิโล ตรงนั้นมีร้านขายของและมีผู้ขายอาหารประเภทส้มตำ
161	เพราะการเดินทางรอบโลกแบบอนาถาอย่างเขาไม่เตรียมของพวกนี้มาด้วยก็แย่เต็มที โดยเฉพาะมีโครงการผ่านเมืองยูงอย่างเมืองไทยด้วยแล้ว
162	หน้าเธอก็ไวที่เดิมแหละ ไม่รู้จะย้ายทำไม ทีเซอร์กวนประสาทฉันอีก

ลำดับ	ประโยคในฐานข้อมูล
163	ชายวัยหกสิบร่างผอมผิวดำ ผมเป็นสีดอกเลาเกือบทั้งศีรษะ ความเก่าแก่ของโรงสีที่
164	แต่ความเห็นส่วนใหญ่ยังคงเน้นอยู่ที่ การทุจริตในกาหาเสียงเลือกตั้งของนักการเมือง กับประสิทธิภาพของ
165	เครือข่ายการสื่อสาร ชนิดหนึ่งเท่านั้น เนื่องจากการก่อกำเนิดความเจริญ เด็บโตของอิน เทอร์เน็ต
166	และจะค้นพบได้ก็โดยการใช่เหตุผลเท่านั้นเอง นอกจากวิธีภาษวิธี
167	บางทีก็มีล่ำมมาด้วยเป็นคนไทยที่เป็นแฟนของลูกชายเฟิน รู้จักกันที่ภูเก็ตได้อาทิตย์ กว่า
168	แต่นาฬิกาบนเวทีเดินเร็วกว่าเวลาจริงเพราะต้องให้เนื้อเรื่องจบภายในเวลาที่กำหนด
169	แต่ผมคัดค้านบอกว่า อย่าเลยเพราะอาบน้ำเสร็จกว่าจะเดินถึงบ้านได้ก็ต้องร้อนอีก
170	หนึ่งศูนย์แปด แปดศูนย์หนึ่ง แล้วทำไมจึงกลายเป็นเจ้าจิวไปได้ แต่ผมก็ดีใจ เหลือหลายแล้ว
171	เสียงใสข้างตัวบอก เมื่อทั้งสองเดินมาถึงอ่างน้ำใหญ่มหิมานอกอาคารที่เรียกว่าเกรท บาร์
172	ทันใดนั้น โรงหนังทั้งหลังก็ลอยขึ้น ปีกหลังคาส่ายคลายนกยักษ์โคลงเคลง
173	แต่ก็กล่าวต่อ รู้สึกว่าคุณไม่ค่อยเต็มใจมาด้วยซ้ำ สีน่านางสาวทเองก็พลอยเจื่อน
174	และ همینอย่างว่าถึงโดนเจ้าของตวาดขับไล่ไปไกลจนกรอว่าหาย همینแล้วจึงให้เข้าบ้าน
175	ได้อารมณ์ ได้สาระ และได้กลิ่นอายประวัติศาสตร์ เช่น ให้ตัวละครชาย
176	ประเดี๋ยวกี่ขึ้นซี่แล้วทำกิจกรรมสืบเชื้อสืบทอด ประเดี๋ยวกี่ลง แล้วก็แห่เข้าหา
177	ทำไมเครื่องบินโคลงเคลง ทำไมฉันดูหนังไม่รู้เรื่อง ทำไมพวกเขาบอกว่าให้รัดเข็มขัดอีก ทั้งที่ฉันรู้สึกอึดอัด
178	น่าเสียดายนะที่แกได้เรียนน้อยไปหน่อย แต่ฉันดูแกฉลาดเฉลียวกว่าคนอื่น
179	เล็บดินเล่าหลุดเรียงมน้อยอยู่ที่ใด ยังดีแขนข้างที่เหลือยังมีมือ
180	หรือผู้รับการฝึกอบรม ในความควบคุมของสถาบันการศึกษาวิชาการพยาบาล
181	จะได้ฟังคำสอน หรือไม่กับทกวิไพเราะของกรมหมื่นพิทยาลงกรณ์
182	ผลที่เกิดขึ้นทันที คือ คนที่โดนสารปรอทจะมีอาการเปลือกตาบวม
183	เมื่อไม่ติดยึดอยู่กับข้อสมมุติเหล่านี้ หมา หมี หมู หนู ก็สามารถใส่สติปัญญา
184	ด้วยการให้ประชาชนมีส่วนร่วมในการปกครองและตรวจสอบการใช้อำนาจรัฐ
185	ซ้อนใครซ้อนมันจ้วงตักน้ำแกงจืดใส่ปากได้เลย คนจีนกินข้าวไม่ค่อยได้ใช้ซ้อน
186	สำนักทฤษฎีวิพากษ์ได้วิจารณ์ทฤษฎีสื่อมวลชนกระแสหลัก ที่กล่าวถึงการทำหน้าที่ของ สื่อมวลชนว่า
187	ไม่มีบ้านพัก ต้องอาศัยวัดอยู่ ที่วัดคนก็มาก ไหนจะต้องเสียค่าใช้จ่ายแพงอีก

ลำดับ	ประโยคในฐานข้อมูล
188	ประโยคสุดท้ายกิริดาแน่นเสียงหนัก มือบางบีบกระชับมันกับมือของคนที่นั่งข้างตัว
189	ความนึกคิดคำนึงล่องลอยมือเลยซาลง เอ็งอย่าใจลอย ทำอะไรอย่าใจลอย
190	ศิระชนุนผ้าโพกผมต่างหมอน ผ้าห่มหนาคลุมจากเจ้าขึ้นมาถึงคอ แต่ก็แสนขัดข้องอยู่ในใจ
191	แม้ว่าพืชพวกนี้จะโผล่หน้ากันขึ้นมาในหน้าร้อน แต่ลักษณะใบและต้นของมันก็ออกแบบมาให้กันน้ำค้างแข็งแม่อะนึ่งและลมหนาว
192	มาสบมากลบทับ ข้าพร้อมยอมสิ้นถูกหมิ่นค่า ด้อยต่ำธรรมดาไร้ระดับ
193	คล้ายติดคราบตมขุนต่างวุ่นวาย ห่วงชีวิตติดวนไปจนวาย กว่าชีพตรมล้มตายมิได้พบ
194	เที่ยงค้อยยังชั่วขึ้นใหม่ลูก นางสาวทถามเมื่อตามกันมาจนทะลุถึงด้านหน้าโรงแรม
195	เอกสาร และสิ่งพิมพ์ที่มีคุณค่าจากประชาชนที่มีจิตศรัทธา เพื่อจัด
196	อย่างไรก็ตาม นับว่ารัฐบาลสมัยจอมพลสฤษดิ์ ธนะรัชต์ ได้แสดงให้เห็นประจักษ์ว่ามีความสนใจต่อปัญหาจังหวัดชายแดนภาคใต้
197	การปฏิรูปการเมือง โดยรวมจะปรากฏเกินขีดว่าการมองโลกในแง่ดีว่าความสำเร็จของการปฏิรูปการเมืองอยู่ใกล้แค่เอื้อม
198	สังคมและวัฒนธรรม ไม่ว่าจะเป็นผู้ส่งสารในระดับปัจเจกบุคคล
199	ไปในคืนนี้ ยิ่งเมื่อตวงหน้าที่คุ่นเคยนิวหน้าครุ่นคิดหนักอย่างไม่รู้ตัวขณะที่เก็บกวาดจานบนโต๊ะ
200	หญิงสาวรีบบอก เมื่อน้ำเสียงของเจ้านายราวกับจะตำหนิว่าเป็นความผิดของเธออย่างนั้นล่ะ
201	และองค์การเอกชนที่เป็นไปในรูปแบบของโรงงานหรือห้างร้านบริษัทใหญ่
202	เรื่องอะไรจะหนักกลับไปคนเดียวได้ไง แล้วใครจะช่วยฉันขึ้นไอ้แบทแมนล่ะ
203	หรือการให้ยาอันตราย ยาควบคุมพิเศษ วัตถุออกฤทธิ์ต่อจิตและประสาท
204	ซูกระดาดร่อน อาโด้งหยิบกระดาดของประสิทธิ์ศักดิ์มาดู มองเส้นยุกยิกกลางกระดาดอยู่พักหนึ่ง
205	ขอเครดิตสได้พูดถึงผู้ที่ตกอยู่ในห่วงรัก จากญาติที่พี่ได้สั่งให้
206	ด้านใต้ทางเพมาโคหันรับมรสุมเต็มที่ เพมาโคจึงเป็นพื้นที่ที่มีฝนตกตลอดปี
207	สักสามสี่วันเสียก่อนจึงค่อยเริ่มลุยทำวิทยานิพนธ์ คิดว่ามาเยี่ยสอจารย์ที่ปรึกษาคงเข้าใจ
208	โดยไม่ได้พูดปดเลยสักคำเดียว และผู้กองก็ไม่ได้เฉลียวใจสักนิดว่ามีอะไรเคลือบคลุมวจาเหล่านั้นอยู่บ้าง
209	ข้าเกาะแก่งแอ่งผาพนาโตรกหนึ่งหุบเหวต่ำโศกซ้ำชะตากรรม ข้าภูผามหาสมุทรสุดคะเนหนึ่งว่าเหวระอุปะทุระล่า

ลำดับ	ประโยคในฐานข้อมูล
210	ริมถนน รถเข็นขายขนมสี่สันลุดจาดจอตึงอยู่ สีสันของขนมบนรถดูน่ากล้วมากกว่าน่ากิน
211	มองผ่านหน้าต่างก็เห็นบีเอ็มดับเบิลยูสีน้ำเงินซึ่งคันตาเหลือเกิน แม่ฟุ้งเหมือนติดจรวดไปที่ประตูพร้อมรอยยิ้มกว้าง
212	เมื่อเปิดเปลือกตาขึ้นมา รับรู้ว่ามีคนรอบข้างกำลังเคลื่อนไหว ชีวิตดำเนินไปตามวิถีของตนเอง
213	ใบหน้าหวานเซตหน้าราวกับหมดเรื่องพูดแล้ว และแน่นอนเธอต้องเพิกเฉยต่อเสียงนกเสียงกาที่ร้องวิ๊ดวิ๊ดกับท่าทางของเธอและตาเกือบอู
214	พูดพลางเอื้อมมือไขประตู่ กลับไปอยู่ป่า พบคู่ ช่วยเขากกไข
215	ตู้พระธรรม ถ้วยชามสังข์โลก ลวดลายปูนปั้น งานแกะสลักไม้ พระพุทธรูปใหญ่น้อย
216	จอกแห่นลอยฟ่องกำลังกระเพื่อมคลื่น เราตะลึงยั้งนิ่ง แต่ว่าตรงหน้านั้นยังมีป่าอ้อกว้างหลายไร่บังเงื่อนน้ำไว้
217	คืนนั้นผมยังโดนคนเชียร์แซกตามตื้อให้ดูโชว์หลายครั้งต้องบอกว่าไม่ ไม่
218	ก็จะให้ทำยังไงล่ะ เรื่องมันเกิดไปแล้ว เธอไม่ได้ใช้งานคอมพิวเตอร์อยู่นี่นะ
219	ตอบข้อหารือและให้คำปรึกษาแก่ส่วนราชการที่เกี่ยวข้อง จัดทำนิติกรรมสัญญาแก่กระทรวงและหน่วยงานในสังกัด
220	เพมาโค ตำนานแห่งพันธุ์พืช หุบเขาหลังดาซิงลาเป็นเทือกหิมาลัยที่หันหน้าเข้าทิศใต้
221	ความสำคัญในการดำรงชีวิตของมนุษย์ คือการต่อสู้เพื่อความอยู่รอด
222	แต่ไม่มีเงินซื้อตัวเครื่องบินกลับบ้าน นี่ก็กำลังส่งข่าวคราวไปบอกลูก
223	แล้วชำระล้างฝุ่นโคล พรงนี้กองทัพอจะหยุดพักหนึ่งวัน แรมคืนอีกหนึ่งคืน
224	ดูแล้วไม่น่าเชื่อว่าเขาจะเดินทางมาได้ถึงค่อนโลกเช่นนี้ รุ่งเช้าผมตื่นนอนเช้กว่าปกติ
225	นอกจากจะเพิ่มความน่าสนใจให้ผู้อ่านคิดไตร่ตรองตามแล้ว ยังเป็นหนทางไปสู่ความเข้าใจแนวคิดของเรื่องได้อย่างลึกซึ้งขึ้นด้วย
226	ฝนเริ่มโปรยเม็ดหนาขึ้นจนร่มไม้ใหญ่ก็ดูจะปกป้องเจ้าของเสื้อเชิ้ตสีฟ้าอมเทาไว้ไม่ได้
227	พาไปวัดไทยให้รู้จักเพื่อนคนไทยที่มาจากภาคอีสาน หลังแต่งงานแล้วเขาส่งฉันไปเรียนภาษาเยอรมันเกือบ
228	การบริหารภายในสำนักงานปลัดกระทรวงและกระทรวง ให้สอดคล้องกับนโยบายของรัฐบาล
229	นี่ก็อยากจะปรับปรุงกิจการโรงสีให้ดีขึ้นและทันสมัยกว่าที่เป็นอยู่ในปัจจุบัน
230	ถ้าต้องการอารมณ์โรแมนติก หาดสวยสุดที่อยู่บนแผ่นดินไทย ไม่รวมเกาะชื่อเสียงไพเราะว่า
231	เป็นบ้า ฉันเอาไม้หวดเอา เด็กพูดพลางเดินรีเขามาทางตัวโซคลงในมือผู้เป็นพ่อ

ลำดับ	ประโยคในฐานข้อมูล
232	มีกองเพลิงอยู่ใกล้กายพร้อมคบไฟโซนแสง ปักเรียงกันตลอดแนว แล้วจะเลยดำหัว
233	เขาก็เพียงปรายตามองเธออย่างรำคาญใจก่อนจะส่งสายตาหึงไยไปดูว่าป้ายนั้นเสียหายหรือเปล่า
234	น้ำมะเฟืองทำเสียงมีลบลมคมนัย เสียงกรรณตะโกนเสริมขึ้น น้ำมะเฟืองเรื่องมาก
235	จริงอย่างที่รินว่า วันนี้แซมใส่เสื้อยืดแขนยาวสีดำกับกางเกงผ้าสีเดียวกับเสื้อ
236	เขาเปิดกระโปรงรถไฟร์วีล ตรวจเช็คความเรียบร้อยอยู่ขณะที่ ในชุดเสื้อกางเกงยีนทั้งชุดหิ้วกระเป๋าเดินทางใบย่อมตามไปที่รถ
237	โดยคำนึงถึงประโยชน์สูงสุดที่ประชาชนจะได้รับเป็นสำคัญและ มาตรา
238	แล้วอีกอย่าง เราก็มิใช่แฟนกันสักหน่อย เพราะฉะนั้นมันไม่ใช่เรื่องธรรมดาแน่นอน
239	สายน้ำหั้นเข้าหาฟ้าสีหมากสุกด้านตะวันตก กอไผ่บนฝั่งที่สูงเด่นดูโปร่งเบาเหนือทิวไม้อื่น
240	โดยปรับปรุงให้สมบูรณ์ชัดเจน และสอดคล้องกับสถานการณ์ในปัจจุบันยิ่งขึ้น
241	ผมว่าคนผู้นั้นอันตรายมาก อันตรายต่อประเทศชาติ ต่อความมั่นคง
242	ความเข้าใจในสิ่ง ที่เปลี่ยนแปลงทางด้านวิทยาศาสตร์และเทคโนโลยี จึงได้จัดแหล่งความรู้ที่สำคัญยิ่งในการจัดการศึกษาทางด้าน
243	หรือว่าฝูงวัวควายลงมากินน้ำเพ็งจะขึ้น ข้างบนมีบ้านคน เมื่อไรปุ้จะซื้อขนมมากินบ้างล่ะ
244	กลุ่มควันเต็มท้องฟ้า พระดอนเก็บกลดมุงหน้าข้ามทุ่งไปทางทิศนั้นด้วยความอยากรู้
245	ได้แปรสภาพฐานะของ ผู้ส่งสาร ที่เป็นรูปธรรมที่มีตัวตนจับต้องได้
246	พ่อยืนซึมมองร่างที่กำลังเดินโผล่โผล่เข้ามา แม่ก้าวลงจากบันไดและเริ่มร้องไห้
247	แต่ถ้าสาวคนนั้นบุกมาอีกรอบ ก็ตัวใครตัวมัน ธรรมชาติยกหน้าอย่างโล่งอกที่หญิงสาวเข้าใจแต่โดยดี
248	ก่อนจะปล่อยตัวให้เข้าสู่การพักผ่อนอย่างแท้จริง เขาสะดุ้งตื่นตอนที่ได้ยินเสียงแม่เรียกหา
249	ยังผู้หญิงทั้งคู่นี้อีกเล่า อาจเสแสร้งแกล้งเข้ามาตีสนิท เพื่อล้วงเอาความสงสัยออกมาให้สิ้น
250	เขารู้วันรุ่งขึ้น ยังแต่งชุดนอนเสื้อแขนสั้นกางเกงขาวาว ผมดำยุ่งเพราะเจ้าตัวเสียมืออย่างลวก
251	วิทยาศาสตร์และเทคโนโลยีให้กับประชาชนทั้งในระบบโรงเรียนและนอกระบบโรงเรียน
252	ไฟที่เจเสียงทางเดินในตึกเปิดสว่างตามปกติ แต่ไม่มีผู้คน ผมถอนใจอย่างโล่งอก
253	ผมเห็นท่าทางเขาอ่อนเพลียมาก จึงบอกให้เขานอนก่อน เขากลางคำสวัสดีกับผมแล้วคลานเข้ามาง

ลำดับ	ประโยคในฐานข้อมูล
254	ไม่อาจจำนำของได้ หรือได้ก็คงต้องมีวิธีการยุ่งยากพิลึก
255	ประชาชนมีหน้าที่อย่างเดี๋ยวนั้นคือรับเฉพาะข้อมูลของคนที่ฉลาดกว่าป้อนให้เชื่อและปฏิบัติตาม
256	นี่เธอคนนั้นนะ ครูภารดีที่มายุ่งกับเราเพียงสามเดือนก็ได้ย้ายเข้าอำเภอก็เพราะเป็นหลานป่าไม้จังหวัดเรียกคำผาง
257	ชนิดที่ไม่มีใครคาดฝันรวมทั้งตัวเธอเองด้วย ต้นเหตุของโศกนาฏกรรมเกิดขึ้นในวันก่อนออกเดินทาง
258	อีแป้ว ปากแป้ว อีพลอย ตาแดง อีนิล ต่า ไอ้ตูบ ไอ้หมอก
259	วิทยาศาสตร์เพื่อการศึกษา เป็นการจัดการศึกษาตามอัธยาศัยในรูปแบบของวิทยาศาสตร์เพื่อเป็นการพัฒนาคนให้มีความรู้
260	ไปอยู่เฉพาะที่ทำให้การเลือกตั้งสะอาดบริสุทธิ์เพียงอย่างเดียว เป็นผลให้ภารกิจของ
261	เช่นในตอนหนึ่งที่ทวนเข้ามาในความทรงจำของฉินตะโรก็คือตอนที่สงคราม
262	เหลียงซึ่งมีอายุน้อยกว่าจางอิมันนับสิบปีจึงให้กำเนิดบุตรสาวคนหนึ่ง
263	หัวโตและหุตก ไอ้เทา อีสาลี ขาวปุย อีเตี้ย ขาหน้าหรือขาหลังกะเผลก
264	ทักท้วงมาความว่าท่านเห็นด้วยอยู่แล้ว ในวิธีที่จะปรับชีวิตเปลี่ยนอาหารแก่ผู้คนด้วยวิธีการอย่างเฉียบ
265	แต่บัดนี้มันก็ได้มาทอดท่อนนอนสนิทเป็นระเบียบอยู่เฉพาะหน้าแล้ว เปลี่ยนสภาพจากต้นสักบนดอยสูงมาเป็นแพชงที่มีค่ามหาศาล
266	ฉินตะโรคิดว่าตัวเองไม่ชอบพ่อเพราะได้รับอิทธิพลจากแม่ การที่ตัวเองไม่ชอบพ่อก็จะต้องได้รับอิทธิพลจากแม่อย่างแน่นอน
267	รวมตลอดถึงเป็นปัญหาสืบเนื่องมาจากนโยบายของรัฐบาลที่ขาดประสิทธิภาพ
268	และไม่มืองค์กรควบคุมการเลือกตั้งอย่างในปัจจุบันเสียอีก จึงเป็นเรื่องปกติที่ความสิ้นหวังต่อ
269	การปฏิรูปการเมืองจะต้องไปเริ่มที่เด็กเกิดใหม่ ด้วยการให้การศึกษาสอนเรื่องประชาธิปไตยบริสุทธิ์
270	โครงการศึกษาภาษาและวัฒนธรรมท้องถิ่นของข้าราชการจังหวัดชายแดนภาคใต้
271	กำหนดดีชั่วของสรรพสิ่ง และนำพาสังคมหรือรัฐนาวาไปด้วยปรัชญาของชนชั้นผู้นำ
272	นอกจากเต้าหู้ อาหารอย่างอื่นที่ทำขึ้นเพื่อเป็นการถนอมอาหารให้อยู่ได้นาน
273	คือครูใหม่ซึ่งผมขอบอกตรงนี้เลยว่าชื่อครูวิชัย ครูวิชัยพักที่บ้ายครูใหญ่ได้ชวนครูใหญ่ห่อข้าวไปกินที่โรงเรียน
274	เจ้าหญิงที่ไม่ติดขัดกับความเป็นเจ้าหญิงก็สามารถเป็นครูสอนหนังสือเด็กยากจนได้

ลำดับ	ประโยคในฐานข้อมูล
275	กลุ่มผู้ส่งสาร นั้น เป็นเพียงกลุ่มคนกลุ่มน้อยที่ได้รับการฝึกฝนเรื่องการส่งสาร
276	เกือบไม่ทัน ทำอย่างนั้นยิ่งอยากมอง ไม่รู้หรือหรือ เขาแก๊วแล้วเล่นและจ้องมองหน้า ขาว
277	อันใดก็อกใจ กลัวไปหมด แต่ก็อยากมีชีวิตที่ดีกว่าเดิม
278	ของจังหวัดที่เราควรมีอุดมการณ์หรือแผนการทำงานอย่างไร เพื่อให้แผนสำเร็จผล หรือประสบความสำเร็จ
279	สุขาภิบาล องค์การบริหารส่วนท้องถิ่นตามที่รัฐมนตรีประกาศในราชกิจจานุเบกษา
280	มองเห็นประชาชนเป็นเพียงนั่งร้านที่จะใช้ปีนขึ้นสู่จุดหมาย เมื่อประสบผลตามความ ต้องการแล้วก็สลัดนั่งร้านทิ้ง
281	ลมเย็นยามบ่ายทำให้หัวใจที่อัดแน่นหนักอึ้งของเขาผ่อนคลายลง มีเรื่องอะไรเกี่ยวกับ พ่อหรือสะ
282	วินาทีนี้เธอผู้นั้นรวมทั้งตัวเขาด้วยได้เป็นอิสระแล้ว อิสระจากความเศร้าหมองที่ พันนาการใจเขาไว้นับจากวันที่เธอจากไป
283	หากเป็นเอกชนก็น่าจะผลิตรายการที่เป็นสินค้า เพื่อขายให้แก่ผู้ที่มีอำนาจซื้อมากกว่า
284	อยู่เมืองจีนในปัจจุบันนี้มีของกินอุดมสมบูรณ์ ขอให้เงินซื้อ
285	ใกล้เขาหลัก ศึกษาความเปลี่ยนแปลงจากสินามิ ป่าชายเลน เส้นทางศึกษาธรรมชาติ
286	ถึงจะเหนื่อย ฉันก็ไม่มีความย่อ ฉันจะบุกไปจนกว่าจะถึงบึงหญ้าใหญ่
287	งานวิจัยของอารยาได้พบว่ากระบวนการควบคุมให้การสื่อความหมายเป็นไปตามความ ตั้งใจของผู้ส่งสารนั้น
288	คู่สนทนารู้สึกตัวเบาเกิดความเข้าใจแจ้ง ว่าเราทุกคนล้วนเกิดมาร่วมทุกข์
289	ทิพกฤตาเห็นฝ่ายนั้นเงยหน้าขึ้นมองท้องฟ้าสีหม่น มือใหญ่ยกขึ้นปาดหยดน้ำซึ่งเริ่มจับ ตัวบนศีรษะได้รูปนั้น
290	เพื่อมิให้อาหารเสียเร็ว นั่นคือ เน้นความสะดวกที่จะเก็บอาหารไว้นาน
291	แม่เป็นบุตรสาวของพนักงานธนาคาร เกิดในโตเกียว และมาเติบโตที่โอซากา
292	แต่อาชีพของเธอยังดูดีกว่าพวกมีการศึกษาสูงใส่สูทผูกไท เข้าไปนั่งในสภาและรวมหัว ทุจริตโกงกินบ้านเมืองจนฉิบหาย
293	น้อย มาจากการที่เกษตรกรขายผลผลิตในราคาต่ำ ไม่คุ้มต่อการซื้อหาปัจจัยการผลิตมา ใช้
294	แม้มีความลึกลับที่จะพูดความจริงและไม่ยอมให้ผืนตะรุตตะอะไรด้วย
295	ใจของเขาได้รับการปลดปล่อยและหญิงสาวในกรอปรูปก็เช่นกัน รอยยิ้มที่สวยงามจึง ปรากฏให้เห็นเป็นครั้งสุดท้าย
296	และช่วยกันทำนุบำรุงรักษาธรรมชาติเหล่านี้ให้ดูสวยงามตามธรรมชาติ

ลำดับ	ประโยคในฐานข้อมูล
297	อะริโยะมิได้พูดถึงสารปรอทในยาฆ่าแมลงของญี่ปุ่นว่ามีมากกว่าของอเมริกาถึง
298	จนคนที่นั่งอ่านหนังสือรออยู่ที่โต๊ะทำงานเพียงคนเดียวในห้องสี่เหลี่ยมกว้างใหญ่นั้นอดหัวเราะออกมาไม่ได้
299	เช่น ตลาดที่วุ่นวายจอแจ การจราจรที่ติดขัด สภาพบ้านเมืองที่เก่าแก่
300	หรือที่ได้รับอนุญาตจากทางราชการให้จัดตั้ง หรือสถาบันการศึกษาที่คณะกรรมการรับรอง



ประวัติคณะผู้วิจัย

หัวหน้าโครงการวิจัย

1. ชื่อ - สกุล (ภาษาไทย/ ภาษาอังกฤษ)

สุรพล วรภัทรাত্র/Surapol Vorapatratorn

2. รหัสประจำตัวนักวิจัยแห่งชาติ (ถ้ามี) หรือ รหัสบัตรประจำตัวประชาชน

1409900349240

3. ตำแหน่งปัจจุบัน หน่วยงานที่อยู่ติดต่อได้สะดวก หมายเลขโทรศัพท์ โทรสาร และ E-mail address

สำนักวิชาเทคโนโลยีสารสนเทศ มหาวิทยาลัยแม่ฟ้าหลวง 333 หมู่ 1 ต.ท่าสูด อ.เมือง จ. เชียงราย 57100

โทรศัพท์ 053 916762 โทรสาร 053 916743

Email: Surapol.god@gmail.com

4. ประวัติการศึกษา

B.Eng. Computer Engineering, Mae Fah Luang University, 2009

M.Eng. Computer Engineering, Chulalongkorn University, 2011

5. ประวัติการทำงาน

2010 - 2011, Research Assistance, Spoken Language System and Human Assistive Technology Laboratory, Faculty of Engineering, Chulalongkorn University

2012 - Present, Lecturer, School of Information Technology, Mae Fah Luang University

6. สาขาวิชาการที่มีความชำนาญพิเศษ (แตกต่างจากวุฒิการศึกษา) ระบุสาขาวิชาการ

- Speech Processing Technology
- Embedded System
- Human Assistive Technology

7. ประวัติการนำเสนอผลงานวิจัยทั้งภายในและภายนอกประเทศ โปรตรับซื้อเรื่องของผลงาน
ซื้อการประชุม สถานที่ วัน เวลา ตามระบบสากล

- S. Vorapatratorn., K. Nambunmee, “An Obstacle Warning Device for the Blind,” The Journal of Assistive, Rehabilitative & Therapeutic Technologies 2013–The Abstracts for International Convention for Rehabilitation Engineering and Assistive Technology, 1:22501, eISSN 2001-6824, Vol 1, Nov2013.
- S. Vorapatratorn, K. Nambunmee, “iSonar: An Obstacle Warning Device for the Blind.” In the proceedings of the seventh International Convention for Rehabilitation Engineering and Assistive Technology (i-CREATe 2013), Korea, Aug 29–31, 2013.
- S. Vorapatratorn , N. Kertkeidkachorn, S. Chanjaradwichai, T. Suri, K. Likitsupin, P. Hirankan, W. Limpanadusadee, S. Chuetanapinyo, K. Pitakpawatkul, N. Puangsri, N. Tangsirirat, K. Trakulsuk, P. Punyabukkana, and A. Suchato, “The CU-MFEC Corpus For Thai And English Spelling Speech Recognition,” In the Proceedings of the Fifteenth Oriental COCOSDA (International Committee for the Co-ordination and Standardization of Speech Databases and Assessment Techniques) Conference (OCOCOSDA 2012), Macau, China, December 9–12, 2012.
- S. Vorapatratorn, N. Kertkeidkachorn, A. Suchato, and P. Punyabukkana, “Contribution of Spectral Shapes to Tone Perception,” In the Proceedings of the Thirteenth Annual Conference of the International Speech Communication Association (Interspeech 2012), Portland, Oregon, September 9–13, 2012.
- S. Vorapatratorn, P. Punyabukkana, N. Lertwongkhanakool, N. Kertkeidkachorn, P. Hirankan, and A. Suchato, “Chula DAISY: The Implementation and A Case Study.” In the Proceedings of the Sixth International Convention for Rehabilitation Engineering and Assistive Technology (i-CREATe 2012), Singapore, July 24–26, 2012.
- S. Vorapatratorn, A. Suchato, and P. Punyabukkana, “Automatic Online Text Selection for Constructing Text Corpus with Custom Phonetic Distribution,” In the Proceedings of the Ninth International Joint Conference on Computer Science and Software Engineering (JCSSE 2012), Bangkok, Thailand, May 30 – June 1, 2012.
- S. Vorapatratorn, A. Suchato, S. Chanjaradwichai, N. Kertkeidkachorn, P. Hirankan, T. Suri, K. Likitsupin, S. Chuetanapinyo, and P. Punyabukkana, “Effects of Acoustic Mismatches on Speech Recognition Accuracies due to Playback-recorded Speech Corpus,” In the Proceedings of the Ninth International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology (ECTI-CON 2012), Hua Hin, Thailand, May 16–18, 2012.

8. ประวัติการเผยแพร่ผลงานวิจัย ทั้งภายในและภายนอกประเทศ โปรตรับซื้อเรื่องของผลงาน
วารสาร ตามระบบสากล

- ไม่มี

9. ผลงานวิจัยที่ได้รับการจดทะเบียนทรัพย์สินทางปัญญา หรือผลงานวิจัยที่อยู่ระหว่างการยื่นขอจดทะเบียนทรัพย์สินทางปัญญา

- เครื่องบอกเตือนสิ่งกีดขวางสำหรับผู้พิการทางสายตา

10. ประสบการณ์ที่เกี่ยวข้องกับการบริหารงานวิจัยทั้งภายในและต่างประเทศ

10.1 งานวิจัยที่ทำเสร็จแล้ว : ชื่อข้อเสนอโครงการวิจัย สัดส่วนที่ทำงานวิจัย (%)
คณะผู้วิจัยและสถาบันร่วมวิจัย แหล่งทุน ปีที่ได้รับทุน การเผยแพร่ผลงานวิจัย

- ไม่มี

10.2 งานวิจัยที่กำลังทำ : ชื่อข้อเสนอโครงการวิจัย สัดส่วนที่ทำงานวิจัย (%) คณะผู้วิจัย
และสถาบันร่วมวิจัย แหล่งทุน ปีที่ได้รับทุน และสถานภาพของงานวิจัย

- ไม่มี

ผู้ร่วมวิจัยคนที่ 1

1. ชื่อ - สกุล (ภาษาไทย/ ภาษาอังกฤษ)

ศิริกานต์ ชูเชิด /Sirikan Chucherd

2. รหัสประจำตัวนักวิจัยแห่งชาติ (ถ้ามี) หรือ รหัสบัตรประจำตัวประชาชน

3640700566332

3. ตำแหน่งปัจจุบัน หน่วยงานที่อยู่ติดต่อได้สะดวก หมายเลขโทรศัพท์ โทรสาร และ E-mail address

สำนักวิชาเทคโนโลยีสารสนเทศ มหาวิทยาลัยแม่ฟ้าหลวง

333 หมู่ 1 ต.ท่าสูด อ.เมือง จ.เชียงราย 57100

โทรศัพท์ 053 916763 โทรสาร 053 916743

Email: sirikan@mfu.ac.th, s.chucherd@gmail.com

4. ประวัติการศึกษา

B.Eng. Telecommunication, King Mongkut's Institute of Technology
Ladkrabang, 1998

M.Eng. Electrical Engineering, King Mongkut's Institute of Technology
Ladkrabang, 2001

Ph.D. Computer Science, Sirindhorn International Institute of Technology,
2011

5. ประวัติการทำงาน

2002 – Present, Lecturer, School of Information Technology, Mae Fah Luang University

6. สาขาวิชาการที่มีความชำนาญพิเศษ (แตกต่างจากวุฒิการศึกษา) ระบุสาขาวิชาการ

–ไม่มี–

7. ประวัติการนำเสนอผลงานวิจัยทั้งภายในและภายนอกประเทศ โปรดระบุชื่อเรื่องของผลงาน
ชื่อการประชุม สถานที่ วัน เวลา ตามระบบสากล

- S. Chucherd, and S.S. Makhanov, “Multiresolution phase portrait analysis for segmentation of ultrasound images for detection of breast cancer”, In International MultiConference of Engineers and Computer Scientists (IMECS 2011), Hong Kong. 16-18 March 2011, pp. 460-465.

- S. Chucherd, and S. S. Makhanov, “Detection of Breast Tumor in Ultrasound Images by Phase Portrait Analysis”, In proceedings of the Third International Conference of Information and Communication Technology for Embedded Systems (ICICTES2012) [CD-ROM], Bangkok. 22-24 March 2012, 5 p.

8. ประวัติการเผยแพร่ผลงานวิจัย ทั้งภายในและภายนอกประเทศ โปรดระบุชื่อเรื่องผลงาน ชื่อวารสาร ตามระบบสากล

- S. Chucherd, and S.S. Makhanov, “Sparse phase portrait analysis for preprocessing and segmentation of ultrasound images of breast cancer”, IAENG International Journal of Computer Science, Vol.38, No.2, May 2011, pp. 146-159.

- S. Chucherd, A. Rodtook, S. S. Makhanov, “Phase Portrait Analysis for Multiresolution Generalized Gradient Vector Flow”, IEICE Transactions on Information and Systems, 93-D(10), 2012, pp. 2822-2835.

9. ผลงานวิจัยที่ได้รับการจดทะเบียนทรัพย์สินทางปัญญา หรือผลงานวิจัยที่อยู่ระหว่างการยื่นขอจดทะเบียนทรัพย์สินทางปัญญา

–ไม่มี–

10. ประสบการณ์ที่เกี่ยวข้องกับการบริหารงานวิจัย

10.1 งานวิจัยที่ทำเสร็จแล้ว :

–ไม่มี–

10.2 งานวิจัยที่กำลังทำ :

–มี–เมธีวิจัยองค์ความรู้ใหม่ (2 ปี)

Research ID: BRG5380016 (Thailand Research Fund)

ชื่อเรื่อง: Continuous Force Field Analysis and Generalized Vector Flow Fields for Ultrasound and CT-scan Image Segmentations

–มี–

Research ID: NRU (National Research University Project of Thailand Office of Higher Education Commission)

โครงการวิจัยเรื่อง ศูนย์วิจัยและพัฒนาสารสนเทศอัจฉริยะ (Center for Intelligent Informatics)

โครงการย่อยเรื่อง โครงการวิจัยเทคโนโลยีด้านวิศวกรรมชีวการแพทย์

สถาบันเทคโนโลยีนานาชาติสิรินธร มหาวิทยาลัยธรรมศาสตร์

ผู้ร่วมวิจัยคนที่ 2

1. ชื่อ - สกุล (ภาษาไทย/ ภาษาอังกฤษ)

ลักษณะ เจริญวัฒนา /Luck Charoenwatana

2. รหัสประจำตัวนักวิจัยแห่งชาติ (ถ้ามี) หรือ รหัสบัตรประจำตัวประชาชน

3100700058102

3. ตำแหน่งปัจจุบัน หน่วยงานที่อยู่ติดต่อได้สะดวก หมายเลขโทรศัพท์ โทรสาร และ E-mail address

อาจารย์ สำนักวิชาเทคโนโลยีสารสนเทศ มหาวิทยาลัยแม่ฟ้าหลวง

333 หมู่ 1 ต.ท่าสุต อ.เมือง จ.เชียงราย 57100

โทรศัพท์ 053 916763 โทรสาร 053 916743

Email: lcharoen@gmail.com

4. ประวัติการศึกษา

B.Sc. Computer Science, Thammasat University, 1996

M.Sc. Computer and Information Networks, University of Essex, UK, 2001

Ph.D. Computer Networks, University of Essex, UK, 2006

5. ประวัติการทำงาน

1996-2000, Research Assistance, Network Technology Laboratory, NECTEC

1996-2000, Lecturer, Faculty of Science, Ubon Ratchathani University

2011 – Present, Lecturer, School of Information Technology, Mae Fah Luang University

6. สาขาวิชาการที่มีความชำนาญพิเศษ (แตกต่างจากวุฒิการศึกษา) ระบุสาขาวิชาการ

–ไม่มี–

7. ประวัติการนำเสนอผลงานวิจัยทั้งภายในและภายนอกประเทศ โปรดระบุชื่อเรื่องของผลงาน ชื่อการประชุม สถานที่ วัน เวลา ตามระบบสากล

- Bandwidth Contention Behaviors of UDP TCP SCTP and DCCP Transport Layer Protocols, 2009 Sixth International Joint Conference on Computer Science and Software Engineering (JCSSE 2009), May 2009, Phuket, Thailand
- Coexistence of SCTP and TCP Variants under Self-Similar Network, 2011 Eighth International Joint Conference on Computer Science and Software Engineering (JCSSE 2011), May 2011, Nakhorn Prathom, Thailand
- Measuring Self-Healing Delay of Java-Implemented Multi-Homed SCTP Association, 2012 Ninth International Joint Conference on Computer Science and Software Engineering (JCSSE 2012), May 2012, Bangkok, Thailand

8. ประวัติการเผยแพร่ผลงานวิจัย ทั้งภายในและภายนอกประเทศ โปรดระบุชื่อเรื่องผลงาน ชื่อวารสาร ตามระบบสากล

- L. Charoenwatana and P. Chaibuppa, “Bandwidth Contention Behaviors of UDP TCP SCTP and DCCP Transport Layer Protocols”, Sixth International Joint Conference on Computer Science and Software Engineering (JCSSE 2009), May 2009, Phuket, Thailand
- W. Padtum and L. Charoenwatana, “Performance of SCTP Multi-Steamed Data File Transfer using Java API”, Seventh International Joint Conference on Computer Science and Software Engineering (JCSSE 2010), May 2010, Bangkok, Thailand
- L. Charoenwatana and S. Rattanabumrung, “Coexistence of SCTP and TCP Variants under Self-Similar Network”, Eighth International Joint Conference on Computer Science and Software Engineering (JCSSE 2011), May 2011, Nakhorn Prathom, Thailand

- L. Charoenwatana and K. Phithakwongyotin, “Measuring Self-Healing Delay of Java-Implemented Multi-Homed SCTP Association”, Ninth International Joint Conference on Computer Science and Software Engineering (JCSSE 2012), May 2012, Bangkok, Thailand

- W. WongYai and L. Charoenwatana, “Examining the Network Traffic of Facebook Homepage Retrieval: An End User Perspective”, Ninth International Joint Conference on Computer Science and Software Engineering (JCSSE 2012), May 2012, Bangkok, Thailand

9. ผลงานวิจัยที่ได้รับการจดทะเบียนทรัพย์สินทางปัญญา หรือผลงานวิจัยที่อยู่ระหว่างการยื่นขอจดทะเบียนทรัพย์สินทางปัญญา

–ไม่มี–

10. ประสบการณ์ที่เกี่ยวข้องกับการบริหารงานวิจัยทั้งภายในและต่างประเทศ

10.1 งานวิจัยที่ทำเสร็จแล้ว : ชื่อข้อเสนอโครงการวิจัย สัดส่วนที่ทำงานวิจัย (%) คณะผู้วิจัยและสถาบันร่วมวิจัย แหล่งทุน ปีที่ได้รับทุน การเผยแพร่ผลงานวิจัย

–ไม่มี–

10.2 งานวิจัยที่กำลังทำ : ชื่อข้อเสนอโครงการวิจัย สัดส่วนที่ทำงานวิจัย (%) คณะผู้วิจัยและสถาบันร่วมวิจัย แหล่งทุน ปีที่ได้รับทุน และสถานภาพของงานวิจัย

–ไม่มี–