การแบ่งประโยคภาษาไทยโดยแคททิกอเรียลแกรมม่าและหลักเกณฑ์ไวยากรณ์

นายณัฐชา ตังศิริรัตน์

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิศวกรรมศาสตรมหาบัณฑิต
สาขาวิชาวิศวกรรมคอมพิวเตอร์ ภาควิชาวิศวกรรมคอมพิวเตอร์
คณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย
ปีการศึกษา 2555

Thai Sentence Segmentation using Categorial Grammar and Grammar Rules

Mr. Nathacha Tangsirirat

A Thesis Submitted in Partial Fulfillment of the Requirements

for the Degree of Master of Engineering Program in Computer Engineering

Department of Computer Engineering

Faculty of Engineering

Chulalongkorn University

Academic Year 2012

หัวข้อวิทยานิพนธ์	การแบ่งประโยคภาษาไทยโดยแคททิกอเรียลแกรมม่าและหลักเกณฑ์ไวยากรณ์
โดย	นายณัฐชา ตังศิริรัตน์
สาขาวิชา	วิศวกรรมคอมพิวเตอร์
อาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก	รองศาสตราจารย์ ดร.อติวงศ์ สุชาโต
อาจารย์ที่ปรึกษาวิทยานิพนธ์ร่วม	ผู้ช่วยศาสตราจารย์ ดร.โปรดปราน บุณยพุกกณะ
อาจารย์ที่ปรึกษาวิทยานิพนธ์ร่วม	ดร.ชัย วุฒิวิวัฒน์ชัย

คณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย อนุมัติให้นับวิทยานิพนธ์ฉบับนี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญามหาบัณฑิต

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

คณะกรรมการสอบวิทยานิพนธ์

……ประธานกรรมการ
(ผู้ช่วยศาสตราจารย์ ดร.เศรษฐา ปานงาม)

……... อาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก
(รองศาสตราจารย์ ดร.อติวงศ์ สุชาโต)

……อาจารย์ที่ปรึกษาวิทยานิพนธ์ร่วม
(ผู้ช่วยศาสตราจารย์ ดร.โปรดปราน บุณยพุกกณะ)

……อาจารย์ที่ปรึกษาวิทยานิพนธ์ร่วม
(ดร.ชัย วุฒิวิวัฒน์ชัย)

……กรรมการภายนอกมหาวิทยาลัย
(ดร.ชัชวาลย์ หาญสกุลบรรเทิง)

ณัฐชา ตังศิริรัตน์: การแบ่งประโยคภาษาไทยโดยแคททิกอเรียลแกรมม่าและหลักเกณฑ์ไวยากรณ์. (Thai Sentence Segmentation using Categorial Grammar and Grammar Rules) อ. ที่ปรึกษาวิทยานิพนธ์หลัก : รศ.ดร.อติวงศ์ สุชาโต, อ. ที่ปรึกษาวิทยานิพนธ์ร่วม : ผศ.ดร.โปรดปราน บุณยพุกกณะ, อ. ที่ปรึกษาวิทยานิพนธ์ร่วม : ดร.ชัย วุฒิวิวัฒน์ชัย, 58 หน้า.

ประโยคจัดได้ว่าเป็นองค์ประกอบพื้นฐานที่สำคัญมากในงานด้านการประมวลผลข้อความ เช่น การแปลภาษาอัตโนมัติ (Machine translation) การค้นคืนสารสนเทศ (Information retrieval) และการสรุปข้อความ (Text summarization) ประสิทธิภาพของการประมวลผลดังกล่าวขึ้นอยู่กับความถูกต้องของประโยคที่ใช้เป็นสิ่งเข้า (Input) โดยเฉพาะอย่างยิ่งในภาษาไทยซึ่งไม่มีการแสดงการสิ้นสุดประโยคอย่างชัดเจน ดังนั้นวิทยานิพนธ์นี้จึงเสนอการใช้แคททิกอเรียลแกรมม่า จำนวนคำระหว่างการเว้นวรรคที่พิจารณากับการเว้นวรรคใกล้เคียง และจำนวนคำระหว่างการเว้นวรรคที่กำลังพิจารณากับจุดสิ้นสุดของข้อความ เป็นลักษณะสำคัญในระเบียบวิธีทางสถิติและเสนอการประยุกต์ใช้กฎบางส่วนจากหลักเกณฑ์การใช้เครื่องหมายวรรคตอนและหลักเกณฑ์การเว้นวรรคที่กำหนดโดยราชบัณฑิตยสถาน เพื่อเพิ่มความถูกต้องให้กับผลลัพท์ที่ได้จากระเบียบวิธีเรียนรู้ทางสถิติ เพื่อแก้ปัญหาการแบ่งประโยคภาษาไทย โดยการทดลองได้ใช้ข้อความและการกำกับข้อความจากฐานข้อมูล Thai speech corpus for speech synthesis (TsynC) และได้ผลการทดลองดังนี้ ความถูกต้องของการแบ่งประโยค (sentence-break-recall) เท่ากับ 84.11%, ความถูกต้องโดยรวม (space-correct) เท่ากับ 93.54% และ ความผิดพลาดของการแบ่งประโยค (false-break) เท่ากับ 2.99%

ภาควิชา วิศวกรรมคอมพิวเตอร์ ลายมือชื่อนิสิต.....

สาขาวิชา วิศวกรรมคอมพิวเตอร์ ลายมือชื่อ อ.ที่ปรึกษาวิทยานิพนธ์หลัก.....

ปีการศึกษา 2555 ลายมือชื่อ อ.ที่ปรึกษาวิทยานิพนธ์ร่วม.....

ลายมือชื่อ อ.ที่ปรึกษาวิทยานิพนธ์ร่วม.....

557 01870 21 : MAJOR COMPUTER ENGINEERING

KEYWORDS : Sentence segmentation / Sentence boundary / Categorial grammar / Statistical approach / Rule-based method

NATHACHA TANGSIRIRAT : THAI SENTENCE SEGMENTATION USING CATEGORIAL GRAMMAR AND GRAMMAR RULES : ASSOC.PROF. ATIWONG SUCHATO, Ph.D., CO-ADVISOR: ASST.PROF. PROADPRAN PUNYABUKKANA, Ph.D., CO-ADVISOR: CHAI WUTIWIWATCHAI, Ph.D., 58 PP.

A sentence is regarded as a key fundamental element in many text processing tasks such as Machine translation, Information retrieval, and text summarization. So, performance of many text processing tasks relies on correct sentences used as input especially in Thai which has no explicit sentence boundary. This thesis proposes to use the integration of statistical method using Categorial grammar, number of words between the considering space and the preceding and succeeding space, and number of words between the considering space and the previous sentence-break as features and rule-based method derived from “Rules for punctuation, space, and abbreviation” composed by The royal institute to improve accuracy of Thai sentence-breaking. Rule-based method is applied to statistical method’s results in order to minimize false-break and increase total accuracy. This research uses Thai speech corpus for speech synthesis (TsynC) as training and testing data. The sentence-break-recall, space-correct and false-break scores are 84.11%, 93.54% and 2.99% respectively.

Department : Computer Engineering	Student’s Signature .....
Field of Study : Computer Engineering	Advisor’s Signature .....
Academic Year : 2012	Co-advisor’s Signature .....
	Co-advisor’s Signature .....

กิตติกรรมประกาศ

การศึกษานี้สำเร็จลุล่วงด้วยดีด้วยความช่วยเหลือจากผู้ที่มีส่วนเกี่ยวข้อง ดังนั้น ข้าพเจ้าขอขอบคุณ อาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก รองศาสตราจารย์ ดร.อติวงศ์ สุชาโต และอาจารย์ที่ปรึกษาวิทยานิพนธ์ร่วม ผู้ช่วยศาสตราจารย์ ดร.โปรดปราน บุณยพุกกณะ และ ดร.ชัย วุฒิวิวัฒน์ชัย สำหรับความรู้และคำแนะนำ รวมทั้งความช่วยเหลือในทุกๆด้าน ตลอดระยะเวลาของการศึกษาวิจัยนี้ ข้าพเจ้าขอขอบคุณ รองศาสตราจารย์ ดร.วิโรจน์ อรุณมานะกุล, ดร.ชัชวาลย์ หาญสกุลบรรเทิง, ผู้ช่วยศาสตราจารย์ ดร.เศรษฐา ปานงาม และศูนย์เทคโนโลยีอิเล็กทรอนิกส์และคอมพิวเตอร์แห่งชาติ (NECTEC) ที่ให้คำแนะนำและแนวคิดในการทำวิทยานิพนธ์นี้ อีกทั้งหน่วยปฏิบัติการวิจัยวิทยาการมนุษยภาษา และภาควิชาวิศวกรรมคอมพิวเตอร์ คณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย ที่สนับสนุนทุนการศึกษาตลอดระยะเวลาที่ศึกษา

สารบัญ

หน้า

สารบัญตาราง

สารบัญภาพ

บทที่ 1

บทนำ

1. ความเป็นมาและความสำคัญของปัญหา

ในปัจจุบันมีการใช้งานคอมพิวเตอร์ในด้านต่างๆเป็นไปอย่างกว้างขวางและคอมพิวเตอร์นั้นถือได้ว่ามีบทบาทและเข้าไปเกี่ยวข้องอย่างมากในแทบทุกสาขาอาชีพหรืออาจกล่าวโดยสรุปได้ว่า คอมพิวเตอร์เป็นสิ่งที่อยู่ควบคู่ไปกับการพัฒนาของโลกสมัยใหม่ มีการนำคอมพิวเตอร์มาช่วยงานด้านต่างๆ อาทิเช่น ด้านการท่องเที่ยว กฎหมาย การแพทย์ การทหาร การขนส่ง วิทยาศาสตร์ วิศวกรรม การศึกษาวิจัย และงานด้านอื่นๆอีกมากมาย สำหรับสังคมสมัยใหม่ที่มีการติดต่อสื่อสารกันอย่างครอบคลุมและกว้างไกลเพื่อก่อให้เกิดการกระจายตัวของธุรกิจทั่วโลกแล้ว งานทางด้านภาษาศาสตร์ถือเป็นงานที่มีความสำคัญ ดังนั้นงานด้านภาษาศาสตร์คอมพิวเตอร์ (Computational linguistics) ซึ่งเป็นการนำคอมพิวเตอร์มาช่วยในงานด้านภาษาศาสตร์ จึงจัดได้ว่ามีความสำคัญเป็นอย่างมาก งานทางด้านภาษาศาสตร์คอมพิวเตอร์นั้นมีการศึกษาและวิจัยกันในแทบทุกภาษาทั่วโลก สำหรับภาษาที่ใช้กันอยู่นั้น ทั้งภาษาเขียน ภาษาพูด หรือแม้กระทั่งภาษากายที่สื่อสารด้วยมือหรือท่าทางในลักษณะต่างๆ มีหลายภาษาที่เมื่อนำคอมพิวเตอร์เข้ามาประยุกต์ใช้ด้วยแล้วประสบปัญหาไม่มากก็น้อย ด้วยเหตุนี้จึงมีการพัฒนาด้านภาษาศาสตร์คอมพิวเตอร์กันอย่างต่อเนื่องและในหลากหลายแง่มุม

สำหรับการประมวลผลภาษาธรรมชาติ (Natural language processing) ประโยคจัดได้ว่าเป็นองค์ประกอบขั้นพื้นฐานที่มีความสำคัญต่อการนำไปประมวลผลในขั้นตอนต่อไป สำหรับภาษาที่มีเครื่องหมายวรรคตอนที่แสดงถึงขอบเขตของประโยคที่ชัดเจนเช่น ภาษาอังกฤษ เครื่องหมายวรรคตอนที่แสดงถึงการจบประโยคซึ่งสำหรับภาษาอังกฤษคือเครื่องหมาย “.” อาจก่อให้เกิดความกำกวมเมื่อใช้เครื่องคอมพิวเตอร์ในการประมวลผล เนื่องจากเครื่องหมาย “.” นอกจากจะแสดงถึงการจบประโยคแล้วยังสามารถใช้งานในลักษณะอื่นๆได้อีกด้วย เช่น ใช้กับตัวย่อ ใช้แสดงจุดทศนิยม ดังนั้นจึงมีการเสนอหลากหลายวิธีการในการแบ่งประโยคภาษาอังกฤษ [1, 2] ซึ่งผลลัพธ์ที่ได้นั้นถือว่ามีความถูกต้องแม่นยำที่ค่อนข้างสูงเนื่องจากความหลากหลายของการใช้งานเครื่องหมาย “.” ที่ใช้ในภาษาอังกฤษนั้นมีไม่มาก สำหรับภาษาไทย ปัญหาการตัดประโยคถือได้ว่ามีความซับซ้อนมาก เนื่องจากภาษาไทยไม่มีเครื่องหมายวรรคตอนที่ชัดเจนสำหรับแสดงขอบเขตของประโยคหรือแสดงการสิ้นสุดประโยค ปัญหาดังกล่าวนี้ก่อให้เกิดความยากลำบากในการประมวลผลข้อความภาษาไทยสำหรับงานภาษาศาสตร์คอมพิวเตอร์ที่ใช้ประโยคเป็นองค์ประกอบเริ่มต้นสำหรับการประมวลผล เช่น การค้นคืนสารสนเทศ (Information retrieval) การสรุปข้อความ

(Text summarization) การแปลอัตโนมัติ (Machine translation) และงานทางด้านอื่นๆ แต่หากสังเกตการใช้ภาษาไทยในงานเขียนและศึกษาไวยากรณ์การเขียนภาษาไทย [3, 4] แล้วจะพบว่า การเว้นวรรค (White space) มักถูกใช้เพื่อแสดงถึงการจบประโยค แต่อย่างไรก็ตาม การเว้นวรรคไม่ได้ถูกใช้สำหรับการแสดงการจบประโยค (Sentence) เพียงอย่างเดียว การเว้นวรรคยังถูกใช้ในเพื่อแสดงหน้าที่อื่นๆอีก เช่น การเว้นวรรคก่อนและหลังตัวเลข การเว้นวรรคเพื่อแสดงการจบความของอนุประโยค (Clause) หรือการจบความของวลี (Phrase) ดังนั้นการแบ่งประโยคภาษาไทย (Thai sentence segmentation) จึงสามารถจัดได้ว่าเป็นงานของการแยกแยะการเว้นวรรคว่าการเว้นวรรคนั้นเป็นการเว้นวรรคเพื่อแสดงถึงการจบประโยคหรือเพื่อแสดงการใช้ในลักษณะอื่นๆ

สำหรับการศึกษาวิจัยเกี่ยวกับการวิเคราะห์ข้อความ (Text analysis) ในภาษาไทย มีการเสนอหลากหลายระเบียบวิธีสำหรับการตัดแบ่งคำในข้อความภาษาไทย (Thai word segmentation) [5 - 7] และผลลัพธ์ที่ได้มีความถูกต้องแม่นยำที่ค่อนข้างสูง แต่สำหรับการแบ่งประโยคภาษาไทยยังมีงานศึกษาวิจัย [8 - 10] อยู่ไม่มาก และส่วนใหญ่แล้วให้ความถูกต้องแม่นยำที่ไม่สูงมาก ดังนั้นการศึกษาวิจัยหาระเบียบวิธีที่เหมาะสมในการตัดแบ่งประโยคภาษาไทยเพื่อให้ได้ผลลัพธ์ที่มีความถูกต้องแม่นยำสูงจึงเป็นงานที่จำเป็นและน่าสนใจมาก

การศึกษาจำนวนมากด้านการวิเคราะห์ข้อความทั้งในภาษาไทยและภาษาอื่นๆ มีระเบียบวิธีที่นิยมอยู่สองวิธีคือ ระเบียบวิธีทางสถิติ (Statistical approach) และการใช้กฎ (Rule-based method) ซึ่งทั้งสองรูปแบบมีข้อดีและข้อเสียที่แตกต่างกัน การเลือกใช้งานขึ้นอยู่กับปัญหาที่ต้องการแก้และวัตถุประสงค์ของการแก้ปัญหานั้นรวมถึงข้อจำกัดต่างๆ สำหรับปัญหาการแบ่งประโยคภาษาไทยนั้นมีทั้งงานวิจัยที่ใช้กฎในการแก้ปัญหา [8] และการใช้ระเบียบวิธีทางสถิติ [9 - 11] ซึ่งการใช้ระเบียบวิธีทางสถิติสามารถพัฒนาให้มีความถูกต้องแม่นยำที่สูงขึ้นได้จากการเลือกใช้ลักษณะสำคัญ (Feature) ที่เหมาะสม ยิ่งไปกว่านั้นการประยุกต์ใช้หลักเกณฑ์การใช้งานภาษาไทยที่มีมีการบัญญัติไว้อย่างเหมาะสมมาช่วยในการตัดสินใจของระบบจะยิ่งสามารถทำให้ได้ความถูกต้องแม่นยำมากยิ่งขึ้น ดังนั้น งานวิจัยนี้จึงเสนอการแก้ปัญหาการแบ่งประโยคภาษาไทยด้วยกฎและระเบียบวิธีทางสถิติ โดยการเลือกใช้กฎที่เหมาะสมและการคัดกรองลักษณะสำคัญที่สามารถสะท้อนให้เห็นถึงรูปแบบการใช้งานภาษาที่เกิดขึ้นจริงรวมทั้งการวิธีการสำหรับบูรณาการกฎและระเบียบทางวิธีทางสถิติเข้าด้วยกัน เพื่อให้ได้ผลของการแบ่งประโยคโดยคอมพิวเตอร์ที่มีความใกล้เคียงกับการแบ่งประโยคจากการตัดสินใจของเจ้าของภาษา

2. วัตถุประสงค์ของการวิจัย

นำเสนอระเบียบวิธีในการแบ่งประโยคภาษาไทยโดยใช้หลักไวยากรณ์ร่วมกับระเบียบวิธีทางสถิติ โดยนำเสนอทั้งกฎและรูปแบบของลักษณะสำคัญที่เหมาะสมเพื่อให้ได้ผลของการแบ่งประโยคที่มีความใกล้เคียงกับการตัดสินใจของเจ้าของภาษา

3. ขอบเขตของการวิจัย

1. สิ่งเข้า (Input) ของระบบเป็นข้อมูลที่มีการตรวจสอบความถูกต้องและระบุข้อมูลสำคัญต่างๆแล้ว
2. การตัดแบ่งประโยคพิจารณาเฉพาะส่วนของการเว้นวรรคเท่านั้น ไม่ครอบคลุมถึงข้อความที่มีความกำกวมและไม่แสดงการเว้นวรรคเพื่อบ่งชี้ข้อสังเกตสำหรับการแบ่งประโยค
3. ประสิทธิภาพของระบบไม่รวมถึงประสิทธิภาพในเชิงเวลาของการประมวลผล
4. ระบบนี้รองรับการตัดแบ่งข้อความภาษาไทยเท่านั้น

4. ขั้นตอนในการเสนอผลงานวิจัย

1. ศึกษาการแบ่งประโยคของภาษาไทยและภาษาใกล้เคียงที่มีอยู่ในปัจจุบัน
2. ศึกษาไวยากรณ์และระเบียบวิธีการใช้ภาษาในภาษาไทย
3. ออกแบบลักษณะสำคัญที่ใช้สำหรับการแบ่งประโยค
4. ออกแบบกฎที่ใช้สำหรับการแบ่งประโยค
5. สร้างระบบสำหรับทำการทดลอง
6. ทดสอบและวัดผล
7. สรุปและวิเคราะห์ผลลัพธ์ที่ได้
8. จัดทำเอกสารวิทยานิพนธ์

5. ประโยชน์ที่คาดว่าจะได้รับ

ระเบียบวิธีที่มีประสิทธิภาพสำหรับการแบ่งประโยคภาษาไทยเพื่อที่จะสามารถนำไปใช้ร่วมกับงาน อื่นๆด้านการประมวลผลภาษาธรรมชาติ เช่น การแปลภาษาอัตโนมัติ การสรุปความอัตโนมัติ เป็นต้น

ภาพที่ 1.1 แผนการดำเนินงานวิจัย

6. ผลงานตีพิมพ์จากวิทยานิพนธ์

ส่วนหนึ่งของวิทยานิพนธ์นี้ตีพิมพ์เป็นบทความทางวิชาการในหัวข้อเรื่อง “Thai Sentence-Breaking using Categorial Grammar” จัดทำโดย “Nathacha Tangsirirat, Atiwong Suchato, Proadpran Punyabukkana” ถูกนำเสนอในงานประชุมวิชาการ “The Sixteenth International Computer Science and Engineering Conference: ICSEC'2012” ณ โรงแรม Garden Cliff Resort & Spa ระหว่างวันที่ 17 ตุลาคม 2555 ถึงวันที่ 19 ตุลาคม 2555 และ “Contextual Behaviour Features and Grammar Rules for Thai Sentence-Breaking” จัดทำโดย “Nathacha Tangsirirat, Atiwong Suchato, Proadpran Punyabukkana, Chai Wutiwiwatchai” ถูกนำเสนอในงานประชุมวิชาการ ECTI-CON’2013 ณ โรงแรม Maritime Park & Spa Resort ระหว่างวันที่ 15 พฤษภาคม 2556 ถึง 17 พฤษภาคม 2556

7. ลำดับการจัดเรียงเนื้อหาในวิทยานิพนธ์

วิทยานิพนธ์นี้มีการจัดแบ่งเนื้อหาเป็น 5 บท คือ

1. บทที่ 1 บทนำ กล่าวถึง ความเป็นมาและความสำคัญของปัญหา, วัตถุประสงค์ของการวิจัย, ขอบเขตของการวิจัย, ขั้นตอนในการเสนอผลการวิจัย และผลงานตีพิมพ์จากวิทยานิพนธ์
2. บทที่ 2 ทฤษฎีและงานวิจัยที่เกี่ยวข้อง ประกอบด้วย ทฤษฎีที่เกี่ยวข้อง คือ Classification and regression tree (CART), แคททิกอเรียลแกรมม่า (Categorial Grammar) และไวยากรณ์สำหรับการแบ่งประโยคภาษาไทย และงานวิจัยที่เกี่ยวข้อง คือ การศึกษาการแบ่งประโยคภาษาไทยเชิงคุณภาพ, งานวิจัยที่เกี่ยวข้องโดยตรง และงานวิจัยที่ใกล้เคียง
3. บทที่ 3 วิธีการวิจัย กล่าวถึง แนวทางการวิจัย, ข้อมูลที่ใช้ในการวิจัย, เครื่องมือที่ใช้ในการวิจัย และการวัดผล
4. บทที่ 4 การทดลองและอภิปรายผล กล่าวถึง การทดลองเพื่อศึกษาลักษณะสำคัญ, การทดลองเพื่อศึกษาผลของการใช้กฎ
5. บทที่ 5 สรุปผลการวิจัยและข้อเสนอแนะ กล่าวถึง สรุปผลการวิจัย ข้อเสนอแนะ

บทที่ 2

ทฤษฎีและงานวิจัยที่เกี่ยวข้อง

1. ทฤษฎีที่เกี่ยวข้อง

1.1 Classification and regression trees (CART) [8]

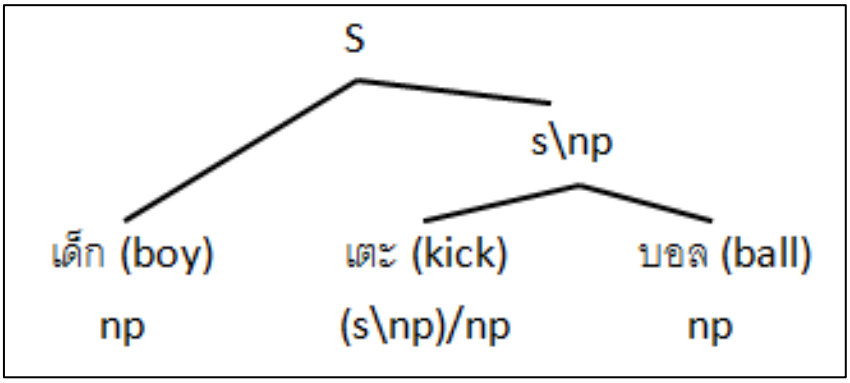
Classification and regression trees คิดค้นโดย Breiman ในปีค.ศ.1984 CART คือต้นไม้ตัดสินใจ (Decision tree) ที่สร้างจากอัลกอริทึม CART เป็นต้นไม้ทวิภาค (binary tree) ซึ่งประกอบด้วย กิ่งหรือแขนง 2 กิ่ง สำหรับแต่ละโหนด กระบวนการเทคนิคแบบ CART ดำเนินการการแบ่ง เรคคอร์ด (Record) ในชุดของข้อมูลสำหรับฝึก (Training Data Set) ออกเป็น เรคคอร์ดย่อยที่ให้ค่าเป้าหมาย (target) ที่เหมือนกัน CART มีข้อดีคือสามารถจัดการได้ทั้งลักษณะสำคัญของค่าจริง (Real value feature) และ สัญลักษณ์ (Symbol) นอกจากนี้ CART ยังสามารถ นำมาฝึกฝนได้ดีแม้ว่าข้อมูลที่นำมาเป็นข้อมูลสำหรับฝึกนั้นมีความหลากหลายกระจัดกระจายไม่มีแนวโน้มไปในทางเดียวกัน CART ถือได้ว่าเป็นเครื่องมือทางสถิติที่มีประสิทธิภาพในการเรียนรู้จากข้อมูลจำนวนไม่มากและทำนายความเป็นไปได้ที่จะเกิดขึ้นกับข้อมูลที่ใช้ในการทำนาย

1.2 แคททิกอเรียลแกรมม่า (Categorial grammar)

Categorial Grammar (CG หรือ Classical categorial grammar) [9, 10] คือแบบแผนหรือรูปแบบอย่างหนึ่งในไวยากรณ์ของภาษาธรรมชาติ (Natural language syntax) ซึ่งเป็นการพัฒนามาจากหลักการของการรวมและจัดระบบของส่วนประกอบทางไวยากรณ์ (Syntactic element) ในภาษา โดยส่วนประกอบทางไวยากรณ์ถูกนำมาจัดแบ่งประเภทโดยดูจากความสามารถหรือความเป็นไปได้ที่จะรวมเข้ากับส่วนประกอบทางไวยากรณ์อื่นๆ เพื่อที่จะสร้างไวยากรณ์หรือรูปแบบทางภาษาที่ใหญ่ขึ้น โดยดูจากความสัมพันธ์ของ ฟังก์ชัน (Function) และ อาร์กิวเมนต์ (Argument) ที่เกี่ยวข้อง ซึ่งทุกประเภทของวากยสัมพันธ์ (Syntactic category) ของแคททิกอเรียลแกรมม่านั้นสามารถแบ่งได้เป็นสองประเภทหลัก คือ

- อาร์กิวเมนต์ คือ ประเภท (Category) พื้นฐานที่สุด ตัวอย่างเช่น ประโยค (**s**) และ นามวลี (**np**)
- ฟังก์ชัน (Functor หรือ Function category) คือ ชนิดประเภท (Category type) ที่ประกอบขึ้นจาก อาร์กิวเมนต์และตัวดำเนินการ ซึ่งมีอยู่สองแบบคือ '/' และ '\' โดย ฟังก์ชันจะถูกกำหนดให้กับคำที่มีความซับซ้อนเพื่อที่จะช่วยสร้างความเชื่อมโยงและเกี่ยวข้องกับอาร์กิวเมนต์ในการที่จะสร้างประโยค ยกตัวอย่างเช่น **s\np** (กริยาที่ต้องการกรรม) ต้องการนามวลีจากทางด้านซ้ายเพื่อที่จะทำให้เป็นประโยคที่สมบูรณ์ สัญลักษณ์ **α/β** คือการรวมฟังก์เตอร์จากทางด้านขวาบนโดเมน **α** ถึงพิสัย (Range) **β** สัญลักษณ์ **α\β** คือการรวมฟังก์เตอร์จากทางด้านซ้ายบนโดเมน **β**

ถึงพิสัย **α** โดยที่ทั้ง **α** และ **β** คืออาร์กิวเมนต์ของประเภทของวากยสัมพันธ์ หลักการโดยทั่วไปของ CG คือ การหาแกนหลักของการรวมและแทนที่ด้วยตัวดัดแปลงทางไวยากรณ์ (Grammatical modifier) ซึ่งอยู่ในเซตของประเภทของวากยสัมพันธ์ที่กำหนดไว้ซึ่งมีหลักการคล้ายกับหลักการของเศษส่วนในทางคณิตศาสตร์ ยกตัวอย่างเช่น กริยาที่ไม่ ต้องการกรรมจำเป็นที่จะต้องรวมกับประธานเพื่อที่จะทำให้ประโยคสมบูรณ์ ดังนั้นกริยาที่ไม่ต้องการกรรมจึงเขียนได้เป็น **s\np** ซึ่งหมายความว่าต้องการนามวลีจากทางด้านซ้ายเพื่อที่จะทำให้ประโยคสมบูรณ์ ถ้ามีนามวลีจากทางด้านซ้าย กฎการหักล้างกันของเศษส่วน ก็จะนำมาใช้ดังนี้ **np*s\np = s** สำหรับ CG ทุกคำ (lexicon) สามารถถูกอธิบายหรือแทนได้ด้วยประเภทของวากยสัมพันธ์แต่อย่างไรก็ตาม คำหนึ่งๆสามารถมีได้มากกว่าหนึ่งประเภทของวากยสัมพันธ์ถ้าคำเหล่านั้นสามารถนำไปใช้ได้ในหลายๆลักษณะของการใช้ในภาษาและ CG สามารถนำมาเขียนในรูปแบบของต้นไม้ได้ดังแสดงต่อไปนี้



ภาพที่ 2.1 ตัวอย่างต้นไม้ไวยากรณ์ของแคททิกอเรียลแกรมม่า

จากรูปด้านบน คำทั้งสามคำสามารถกำหนดประเภทของวากยสัมพันธ์ได้ดังต่อไปนี้ “เด็ก” และ “บอล” สามารถกำหนดให้เป็นประเภทพื้นฐาน คือ **np** ส่วนคำว่า “เตะ” สามารถกำหนดให้เป็น **(s\np)/np** ซึ่งคือฟังก์ชันที่ต้องการสองอาร์กิวเมนต์ซึ่งอาร์กิวเมนต์ทั้งสองคือ **np** โดยที่ตัวหนึ่งมาจากทางด้านขวาและอีกตัวหนึ่งมาจากทางด้านซ้าย สำหรับการที่จะได้มาซึ่งประโยคนั้นเริ่มจากการรวมไปทางด้านหน้าระหว่าง **(s\np)/np** กับ **np** ซึ่งจะได้ **s\np** ต่อจากนั้นทำการรวมไปทางด้านหลังระหว่าง **s\np** กับ **np** จะได้ประเภทของวากยสัมพันธ์ “**s**”

1.2.1 แคททิกอเรียลแกรมม่าสำหรับภาษาไทย

แม้ว่าแนวคิดแคททิกอเรียลแกรมม่าจะสามารถใช้ได้กับทุกภาษาแต่อากิวเมนต์และฟังก์ชันของแคททิกอเรียลแกรมม่านั้นขึ้นอยู่กับภาษาที่นำไปใช้ ดังนั้นการกำหนดอากิวเมนต์และฟังก์ชันจึงเป็นขั้นตอนแรกเริ่มของการนำแคททิกอเรียลแกรมม่ามาใช้กับภาษาใด ๆ สำหรับภาษาไทยได้มีการกำหนดอากิวเมนต์ไว้ 7 ประเภท [13] ดังแสดงใน ตารางที่ 2.1 และตัวดำเนินการ “/” หรือ “\” ถูกใช้เพื่อแสดงความต้องการส่วนเติมเต็มจากทางซ้ายหรือขวาตามลำดับ ในปัจจุบันมีการกำหนดประเภทของวากยสัมพันธ์สำหรับภาษาไทยทั้งสิ้น 120 ประเภท

ตารางที่ 2.1 แสดงอากิวเมนต์ของแคททิกอเรียลแกรมม่าสำหรับภาษาไทย

วากยสัมพันธ์พื้นฐาน	**นิยาม**	**ตัวอย่าง**
np	คำนามหรือนามวลี	เด็ก (boy) ผม (I, me)
num	ตัวเลข	สอง (two), 3 (three)
spnum	คำแสดงจำนวนซึ่งใช้ตามหลังส่วนขยายหรือลักษณะนาม	เดียว (one)
pp	บุพบทหรือบุพบทวลี	ข้างกล่อง (beside box)
S	ประโยค	เด็กเตะบอล (Boy kicks ball)
ws	อนุประโยคที่ขึ้นต้นด้วยคำว่า “ว่า”	เธอคิด**ว่า**เขาจะไม่มา (She think **that** he will not come)
PUNC	เครื่องหมายวรรคตอน	-

อากิวเมนต์ถือได้ว่าเป็นพื้นฐานของการสร้างรูปแบบการกำหนดแคททิกอเรียลแกรมม่าที่ซับซ้อนมากยิ่งขึ้นโดยใช้ “/” และ “\” เพื่อแสดงความต้องการขององค์ประกอบโดยรอบหรือส่วนเติมเต็มที่จะทำให้ฟังก์ชันใดๆมีความสมบูรณ์ ซึ่งในที่นี้ขอยกตัวอย่างของฟังก์ชัน 5 ฟังก์ชันที่พบมากในภาษาไทย ดังแสดงในตารางที่ 2.2

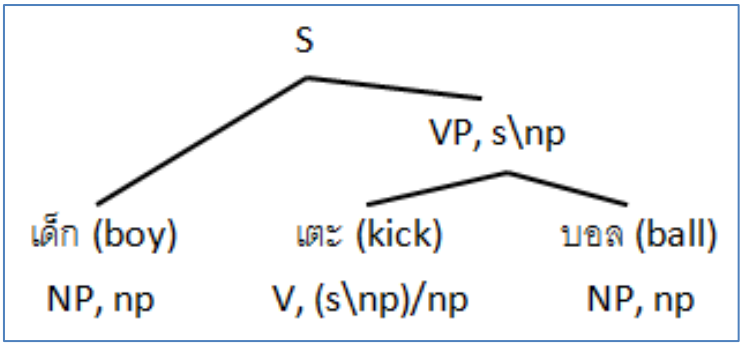
ตารางที่ 2.2 แสดงตัวอย่างของประเภทของวากยสัมพันธ์สำหรับภาษาไทย

ประเภทของไวยากรณ์	**ประเภทของวากยสัมพันธ์**	**ตัวอย่าง**
คำกริยาที่ไม่ต้องการกรรม (Intransitive verb)	s\np	เธอ**เดิน** (She **walks**.)
คำกริยาที่ต้องการกรรม (Transitive verb)	(s\np)/np	เธอ**หยิบ**เงิน (She **picks** money.)
คำกริยานุเคราะห์ (Auxiliary verb)	(s\np)\(s\np) (s\np)/(s\np)	เธอ**วิ่ง**แล้ว (She **ran** already.) เธอจะ**วิ่ง** (She will **run**.)
คำคุณศัพท์ (Adjective)	np\np	สาว**สวย**กำลังเดิน (**beautiful** woman is walking)
คำบุพบท (Preposition)	(s\np)\(s\np)/np (np\np)/np	ปากกา**ของ**ฉัน (book **of** mine) เธอยืน**บน**กล่อง (She **stands** on a box.)

1.2.2 แคททิกอเรียลแกรมม่ากับการแบ่งประโยคภาษาไทย

ข้อดีของการใช้แคททิกอเรียลแกรมม่ากับการแบ่งประโยคมีอยู่ด้วยกันสองประการ ประการแรกคือการใช้แคททิกอเรียลแกรมม่าสามารถแสดงความสัมพันธ์ของแต่ละคำในบริบทได้ดีกว่าการใช้ชนิดของคำ (Part-of-speech, POS) ยกตัวอย่างเช่น ในรูปที่ 2.2 แสดงต้นไม้ไวยากรณ์ของแคททิกอเรียลแกรมม่าและชนิดของคำ (CG parsed tree and Part-of-speech parsed tree) ของประโยค “เด็กเตะบอล” ซึ่งคำว่า “เตะ” และ “บอล” ถูกกำหนดชนิดของคำเป็นคำกริยา (**V**) และนามวลี (**NP**) ตามลำดับ แต่อย่างไรก็ตามชนิดของคำของทั้งสองคำนี้ไม่ได้แสดงให้เห็นถึงรายละเอียดที่เกี่ยวกับการรวมคำแต่ละคำให้กลายเป็นองค์ประกอบที่ใหญ่ขึ้น ในทางตรงกันข้ามแคททิกอเรียลแกรมม่าสามารถแสดงให้เห็นถึงการที่จะรวมองค์ประกอบย่อยๆให้กลายเป็นองค์ประกอบที่ใหญ่ขึ้น ดังตัวอย่างในรูปที่ 2.2 ซึ่งแสดงต้นไม้ไวยากรณ์ของแคททิกอเรียลแกรมม่า คำว่า “เตะ” ถูกกำหนดให้มีประเภทของวากยสัมพันธ์เป็น **(s\np)/np** ซึ่งหมายความว่า ต้องการนามวลีจากทางด้านขวาและซ้ายเพื่อที่จะทำให้เป็นประโยคที่สมบูรณ์ ประการที่สองคือ แคททิกอเรียลแกรมม่ามีความหลากหลายของการเป็นตัวแทนของคำต่างๆในภาษาไทยมากกว่าชนิดของคำ

ซึ่งในปัจจุบันมีการกำหนดประเภทของวากยสัมพันธ์ของแคททิกอเรียลแกรมม่าทั้งสิ้น 120 ประเภท ในขณะที่ชนิดของคำมีเพียง 47 ประเภท ดังนั้นแคททิกอเรียลแกรมม่าจึงมีความสามารถในการแสดงรายละเอียดต่างๆในภาษาได้ดี



ภาพที่ 2.2 แสดงต้นไม้ไวยากรณ์ของแคททิกอเรียลแกรมม่าและชนิดของคำ

1.3. ไวยากรณ์สำหรับการแบ่งประโยคภาษาไทย

ภาษาไทยเป็นภาษาที่ไม่มีการกำหนดชัดเจนในเรื่องของการแบ่งประโยค หรืออาจกล่าวได้ว่าการแบ่งประโยคของภาษาไทยมีความกำกวมอยู่ แต่หากพิจารณาให้ละเอียดถี่ถ้วนแล้วจะพบว่าภาษาไทยใช้การเว้นวรรคเพื่อแสดงถึงการจบประโยคแต่การเว้นวรรคยังสามารถแสดงถึงการใช้งานในลักษณะอื่นๆได้อีกด้วย ซึ่งมีกำหนดไว้ในหลักเกณฑ์การเว้นวรรคในภาษาไทย [14] ซึ่งมีรายละเอียดดังนี้

การเว้นวรรคแบ่งออกเป็น

- การเว้นวรรคเล็ก มีระยะห่างระหว่างวรรคประมาณเท่ากับความกว้างของพยัญชนะ ก
- การเว้นวรรคใหญ่ มีระยะห่างระหว่างวรรคประมาณ ๒ เท่าของการเว้นวรรคเล็ก

โดยสามารถแบ่งออกเป็น 2 กรณีคือ กรณีที่ต้องเว้นวรรคและกรณีที่ไม่ต้องเว้นวรรค

1. กรณีที่ต้องเว้นวรรค

1.1 เว้นวรรคใหญ่เพื่อแสดงการจบใจความของประโยค ตัวอย่างเช่น

- เดินให้เรียบร้อย อย่าเตะเท้า
- แม่ออกไปเล่นกีฬาที่สโมสร เสร็จแล้วก็กลับมาอาบน้ำ

1.2 เว้นวรรคเล็ก สำหรับกรณีต่อไปนี้

1.2.1 ประโยคความรวมให้มีการเว้นวรรคระหว่างประโยคย่อยที่มีใจความสมบูรณ์ และเชื่อมกับประโยคอื่นๆ ที่ขึ้นต้นด้วยคำสันธาน “แต่” “และ” “หรือ” ตัวอย่างเช่น

- ฉันชอบเล่นกีฬาในร่ม แต่เพื่อนๆของฉันชอบเล่นกีฬากลางแจ้ง
- การไม่อ่านหนังสือ หรือการไม่ตั้งใจเรียนล้วนแต่เป็นต้นเหตุของการสอบตก

แต่สำหรับกรณีที่เป็นประโยคสั้นหรือมีการรวมประธาน กริยาหรือกรรมเข้าไว้ด้วยกันให้เขียนติดกัน ตัวอย่างเช่น

- เก่งและก้องไปพักผ่อนที่หัวหิน
- ฝนตกแต่แดดออก

1.2.2 เว้นวรรคเล็กระหว่างชื่อและนามสกุล ตัวอย่างเช่น

- นายณัฐชา ตังศิริรัตน์

1.2.3 หลังคำนำหน้าพระบรมวงศานุวงศ์ให้มีการเว้นวรรคเล็ก ตัวอย่างเช่น

- สมเด็จพระเจ้าบรมวงศ์เธอ กรมพระยาดำรงราชานุภาพ

1.2.4 ระหว่างชื่อของธนาคาร บริษัท กับคำว่า จำกัด ให้มีการเว้นวรรคเล็ก ตัวอย่างเช่น

- ธนาคารไทยพาณิชย์ จำกัด
- บริษัทหลักทรัพย์สินทรัพย์ จำกัด

1.2.5 เว้นวรรคเล็กระหว่างคำว่า “ห้างหุ้นส่วนจำกัด” และ “ห้างหุ้นส่วนสามัญนิติบุคคล” กับชื่อที่ตามมา ตัวอย่างเช่น

- ห้างหุ้นส่วนจำกัด ฮั้วเซ่งเฮง
- ห้างหุ้นส่วนสามัญนิติบุคคลจำกัด ถนอมมิตร

1.2.6 ระหว่างชื่อสถานที่ต่างๆให้มีการเว้นวรรคเล็ก ตัวอย่างเช่น

- หมู่บ้านเศรษฐสิริ ถนนประชาชื่น อำเภอเมืองนนทบุรี

1.2.7 ระหว่างคำนำหน้านามกับนามให้มีการเว้นวรรคเล็ก ตัวอย่างเช่น

- ศาสตราจารย์ นายแพทย์ หม่อมราชวงศ์วีกิจ สนิทมุตา

1.2.8 เว้นวรรคระหว่างยศกับชื่อ ตัวอย่างเช่น

- จอมพล ป. พิบูลสงคราม
- ร้อยเอกหญิง แสนสวย รวยยิ่ง

1.2.9 ระหว่างกลุ่มอักษรย่อต้องมีการเว้นวรรคเล็ก ตัวอย่างเช่น

- นายเสริม วินิจฉัยกุล ป.จ. ม.ป.ช. ม.ว.ม.

1.2.10 เว้นวรรคเล็กระหว่างการเขียนตัวหนังสือกับตัวเลข ตัวอย่างเช่น

- บ้านของเขามีสุนัข 4 ตัว

1.2.11 เว้นวรรคเล็กระหว่างวันกับเวลา ตัวอย่างเช่น

- นักศึกษาต้องเข้าประชุมทุกวันอังคาร เวลา 13.00 น

1.2.1.12 เว้นวรรคเล็กหลังข้อความที่เป็นหน่วยมาตราต่าง ตัวอย่างเช่น

- กล่องมีขนาดกว้าง 0.90 เมตร ยาว 1.50 เมตร สูง 0.80 เมตร

1.2.1.13 ระหว่างตัวหนังสือภาษาอื่นกับตัวหนังสือภาษาไทยให้มีการเว้นวรรคเล็ก ตัวอย่างเช่น

- มีการใช้ Finasteride กันอย่างแพร่หลายในหมู่คนที่มีอาการผมร่วง

1.2.1.14 เว้นวรรคเล็กระหว่างรายการต่างๆที่มีการกล่าวต่อกัน หรือลำดับของเลข เพื่อแยกรายการออกเป็นรายการย่อยๆ ตัวอย่างเช่น

- ทุกข์ สมุทัย นิโรธ มรรค ทั้งหมดนี้รวมกันแล้วคือ อริยสัจ 4
- 2 3 5 7 11 ล้วนแล้วแต่เป็นจำนวนเฉพาะ

1.2.1.15 ระหว่างเครื่องหมายทางภาษาต่างๆให้มีการเว้นวรรคเล็ก

1.2.1.15.1 เว้นวรรคเล็กหน้าและหลังเครื่องหมายไปยาลใหญ่ ไม้ยมก เสมอภาค ทวิภาค วิภัชภาค ตัวอย่างเช่น

- ในทางศาสนามี อริยสัจ อิทธิบาท ขันธ์ ฯลฯ
- เด็ก ๆ มาเรียนกันสายมาก
- อเปหิ = อป + เอหิ
- กฤษณา : กฤษณาสอนน้อง แบบเรียนกวีนิพนธ์

1.2.1.15.1 หน้าเครื่องหมายอัญประกาศเปิดและวงเล็บเปิดให้เว้นวรรคเล็ก ตัวอย่างเช่น

- นาวสาวบี (นามสมมติ) ได้ให้การว่า “มีคนร้ายเข้ามาในบ้าน”

1.2.1.15.2 หลังเครื่องหมาย จุลภาค ไปยาลน้อย อัฒภาค วงเล็บปิด และอัญประกาศปิดให้เว้นวรรคเล็ก ตัวอย่างเช่น

- ศีล, สมาธิ, ปัญญา เป็นหลักเบื้องต้นของผู้ที่ศึกษาธรรมะ
- ชีวิตของ ตนเป็นที่รักยิ่งฉันใด ชีวิตของผู้อื่นก็ปานนั้น; สัตบุรุษเอาตนเข้าไปเทียบดังนี้ จึงกระทำความเมตตากรุณาในสัตว์มีชีวิตทั่วไป
- โอ้ว ! มันยอดมาก
- สมเด็จพระนางเจ้าฯ พระบรมราชินีนาถ
- สตรีในอินเดียถูกมองว่าเป็น “เอาวัลย์” หรือ “ไม้เลื้อย” ซึ่งจำเป็นต้องอาศัยที่พึ่งไม่สามารถช่วยตัวเองได้
- คุณสามี (กำมะลอ) ที่รัก เป็นละครที่ผู้หญิงหลายคนชอบดู

1.2.1.16 หลังข้อความที่เป็นหัวข้อให้เว้นวรรคเล็ก ตัวอย่างเช่น

- **การแบ่งประโยคในภาษาไทย** การแบ่งประโยคในภาษาไทยถูกมองว่ามีลักษณะที่กำกวมและยากต่อการแบ่งในหลายๆกรณี

1.2.1.17 หน้าและหลังคำว่า ณ และ ธ ให้เว้นวรรคเล็ก ตัวอย่างเช่น

- ภาพถ่ายช่วยให้ระลึกความหลัง ณ สถานที่นั้นๆ
- ผลพระคุณ **ธ** รักษา ปวงประชาเป็นศุขสานต์

1.2.1.18 หน้าและหลังคำว่า “ได้แก่” ที่ตามมาด้วยรายการที่มากกว่าหนึ่งรายการ ให้เว้นวรรคเล็ก ตัวอย่างเช่น

- มีอาหารสามอย่าง ได้แก่ ไก่ทอด แกง ผัดผัก

1.2.1.19 เว้นวรรคเล็กหน้าและหลังคำว่า “เช่น” ในความหมายของการยกตัวอย่าง ตัวอย่างเช่น

- มีงานหลายด้านที่ต้องการผลลัพธ์ที่ดีจากการแบ่งประโยคภาษาไทย เช่น การแปลภาษาอัตโนมัติ การค้นคืนสารสนเทศ

1.2.1.20 หน้าคำว่า “และ”, “หรือ” ในกรณีที่ใช้อยู่ในรายการให้เว้นวรรคเล็ก ตัวอย่างเช่น

- วันนี้คุณครูทำอาหาร 3 อย่างประกอบด้วย ข้าวต้ม ขนมครก และขนมเทียน
- ในการขึ้นตึกผู้มาติดต่อต้องใช้หลักฐานแสดงตัว เช่น บัตรประชาชน ใบขับขี่ หรือ บัตรสำหรับผู้มาติดต่อ

สำหรับรายการที่มีน้อยกว่า 3 รายการ ไม่ต้องเว้นวรรคเล็กหน้าคำว่า “และ”, หรือ

1.2.1.21 เว้นวรรคเล็กหน้าคำว่า “เป็นต้น” ที่อยู่หลังรายการ ตัวอย่างเช่น

- ภาษาที่ไม่มีการแบ่งคำในตัวเอง เช่น ภาษาไทย ญี่ปุ่น เกาหลี เป็นต้น จำเป็นต้องอาศัยขบวนการของ parsing

1.2.1.22 หลังคำว่า “ว่า” ถ้าส่วนที่ตามต่อมาคือประโยค ให้เว้นวรรคเล็กหลังคำว่า “ว่า”

- สังเกตได้ว่า คนที่มีสุขภาพดีอะไรๆก็จะดีตามมาด้วย

2. กรณีที่ไม่ต้องเว้นวรรค

2.1 ระหว่างคำนำหน้าชื่อและชื่อไม่ต้องมีการเว้นวรรค ตัวอย่างเช่น

- นายสมชาย ตั้งใจเรียน
- คุณหญิงสมหญิง สมจริง

2.2 ระหว่างบรรดาศักดิ์ สมณศักดิ์ ฐานันดรศักดิ์ กับนาม หรือราชทินนาม ไม่ต้องเว้นวรรค ตัวอย่างเช่น

- หลวงวิศาลศิลปกรรม
- สมเด็จพระพุทธโฆษาจารย์

2.3 ระหว่างคำนำหน้าชื่อที่เป็นตำแหน่งหรืออาชีพกับชื่อไม่ต้องเว้นวรรค ตัวอย่างเช่น

- ศาสตราจารย์เฉิน แซ่เจียง
- นายแพทย์สมชาย เจิงตระกูล

2.4 ระหว่างคำหน้าชื่อที่แสดงฐานะของนิติบุคคล หน่วยงาน หรือกลุ่มบุคคลกับชื่อไม่ต้องเว้นวรรค ตัวอย่างเช่น

- สมาคมผู้ไม่ประสงค์ออกนาม
- กระทรวงศึกษาธิการ
- โรงเรียนหอวัง

2.5 หลังเครื่องหมายไปยาลน้อยในกรณีที่มีเครื่องหมายอื่นตามมา ไม่ต้องเว้นวรรค ตัวอย่างเช่น

- เที่ยวบินจากกรุงเทพฯ-กระบี่

2.6 หน้าและหลังเครื่องหมายยัติภังค์ ยัติภาค ไม่ต้องเว้นวรรค ตัวอย่างเช่น

- -สะทก ใช้คู่กับคำสะทก เป็น สะทกสะท้าน
- เขามีเชื้อสายอังกฤษ-เยอรมัน

2. งานวิจัยที่เกี่ยวข้อง

งานวิจัยที่เกี่ยวข้องสามารถแบ่งออกเป็น 3 ส่วนหลักคือ งานวิจัยที่ศึกษาการแบ่งประโยคภาษาไทยเชิงคุณภาพ งานวิจัยที่เกี่ยวข้องกับการแบ่งประโยคภาษาไทยโดยตรง และงานวิจัยที่สามารถนำมาประยุกต์ใช้กับการแบ่งประโยคภาษาไทยได้หรือมีความเกี่ยวข้องกันในลักษณะที่ใกล้เคียง

2.1 การศึกษาการแบ่งประโยคภาษาไทยเชิงคุณภาพ [15] ได้ข้อสรุปว่า

1. เมื่อมีการเปลี่ยนหัวข้อของเนื้อหาในบริบทส่วนใหญ่แล้วผู้ทำการทดสอบจะมีความเห็นว่าเป็นการเริ่มต้นประโยคใหม่
2. ถึงแม้ว่าหัวข้อหรือว่าเนื้อหาในบริบทจะยังคงเหมือนเดิมไม่มีความแตกต่าง แต่หากมีการใช้คำนามหรือคำสรรพนามที่กล่าวมาแล้วอีกครั้งอย่างชัดเจน จุดนี้จะถือได้ว่าเป็นการเริ่มต้นของประโยคใหม่
3. ส่วนใหญ่แล้วอนุประโยคไม่มีประธานมักจะถูกรวมเข้าเป็นส่วนหนึ่งของประโยคหลัก
4. คำบางประเภทเช่น “และต่อมา”, “ตลอดระยะเวลาดังกล่าวนี้”, “ในสมัยนี้” โดยส่วนใหญ่แล้วมักจะใช้เพื่อขึ้นต้นประโยค ดังนั้นคำกลุ่มนี้จึงมีความสามารถในการบอกถึงการสิ้นสุดเนื้อความเดิมและขึ้นประโยคใหม่

5. คำสันธานเช่น “เพราะ”, “แต่”, “จึง” โดยส่วนใหญ่แล้วมักจะไม่ถือเป็นการขึ้นประโยคใหม่ แต่มีในบางกรณีที่คำเหล่านี้นับเป็นการขึ้นประโยคใหม่ จากการสังเกตพบว่าหากข้อความที่ตามหลังคำเหล่านี้มีความยาวมากๆและมีเนื้อความที่ชัดเจนสมบูรณ์ในตัวเองแล้วข้อความที่ตามมาจึงถือได้ว่าเป็นประโยค
6. ประพันธสรรพนาม (Embedded clause marker) เช่น “ที่”, “ซึ่ง”, “อัน” ถือได้ว่าเป็นส่วนหนึ่งของประโยคหลัก ไม่นับเป็นอีกหนึ่งประโยคถึงแม้ข้อความที่ตามหลังคำเหล่านี้จะมีเนื้อความสมบูรณ์ก็ตาม
7. จากกลุ่มตัวอย่างในการศึกษาพบว่า อนุประโยคที่มีการเว้นวรรคก่อนเริ่มอนุประโยค นับได้ว่าเป็นอีกประโยคหนึ่ง หรืออาจกล่าวได้ว่า การเว้นวรรคมีผลต่อการพิจารณาความเป็นประโยคที่สมบูรณ์ของอนุประโยคที่ตามหลังการเว้นวรรค

2.2 งานวิจัยที่เกี่ยวข้องโดยตรง

ตั้งแต่อดีตได้มีการศึกษาการแบ่งประโยคในภาษาโดยจัดเป็นสองระเบียบวิธี คือ การใช้กฎ (Rule-based approach) และการใช้ระเบียบวิธีทางสถิติ (Statistical-based approach) ทั้งสองวิธีมีรายละเอียดและการศึกษาต่างๆที่แตกต่างกันออกไป ดังจะได้กล่าวต่อไป แต่ทั้งสองรูปแบบใช้การพิจารณาการเว้นวรรคเพื่อตัดสินใจว่าการเว้นวรรคที่พิจารณาอยู่นั้นเป็นการเว้นวรรคสำหรับจบประโยคหรือไม่

2.2.1 การแบ่งประโยคภาษาไทยโดยใช้กฎ

การใช้กฎหลักเกณฑ์ในการแบ่งประโยค [8] มีการนำเสนอขึ้นเพื่อแก้ปัญหาของสิ่งเข้าที่ยาวเกินไปสำหรับการแปลภาษาอัตโนมัติ โดยระเบียบวิธีที่เสนอสามารถสรุปเป็นขั้นตอนย่อยๆคือ

1. รับสิ่งเข้าในรูปแบบของย่อหน้า (Paragraph)
2. ตรวจสอบจากกฎที่กำหนดขึ้นเพื่อหากริยาหลักทั้งหมดที่มีในข้อความ
3. ตรวจสอบจากกฎที่กำหนดขึ้นเพื่อหาคำสันธานและคำเชื่อมประโยคและอนุประโยคต่างๆที่ปรากฏในข้อความ
4. ตรวจหาจำนวนการเว้นวรรคที่ปรากฏในข้อความ
5. นำข้อมูลที่ได้จาก 2-4 ไปประมวลผลจากกฎที่ได้กำหนดขึ้นและนำเสนอสิ่งออก (Output) ในลักษณะของประโยคหลายๆประโยค

ระเบียบวิธีที่กล่าวไปให้ความถูกต้องประมาณ 72% สาเหตุที่ความถูกต้องแม่นยำไม่สูงมากนั้นอาจเป็นผลเนื่องจาก กฎและหลักเกณฑ์ที่ใช้ยังไม่ครอบคลุมการเกิดขึ้นจริงของภาษา วิธีการกำหนดกฎในการตัดสินใจสำหรับคอมพิวเตอร์มีข้อดีคือ หากกฎที่ใช้นั้นมีความตายตัวและมีการเกิดขึ้นในลักษณะที่แน่นอน ความถูกต้องของกฎที่นำมาใช้จะสูง แต่ในทางกลับกัน กฎไม่สามารถอธิบายการ

เกิดขึ้นจริงในทุกกรณีของภาษาได้ ดังนั้น ความครอบคลุมของกฎที่นำมาใช้จึงเป็นข้อด้อยของกฎ นำกฎทางภาษาศาสตร์เข้ามาประยุกต์ใช้ในงานภาษาศาสตร์คอมพิวเตอร์

2.2.2 การแบ่งประโยคโดยใช้ระเบียบวิธีทางสถิติ

การใช้ระเบียบวิธีทางสถิติในการแก้ปัญหาการแบ่งประโยคภาษาไทยได้มีการนำเสนอไว้ 3 งานวิจัย ซึ่งทั้ง 3 งานวิจัยมีการใช้ระเบียบวิธีเรียนรู้ (Learning algorithm) และลักษณะสำคัญที่แตกต่างกัน ผลที่ได้ก็มีข้อเด่นที่แตกต่างกัน ซึ่งจะกล่าวโดยเรียงจากวิธีการที่ให้ประสิทธิภาพจากน้อยไปมากตามลำดับ

- **งานวิจัยที่เสนอการใช้ Tri-gram model ร่วมกับชนิดของคำ เพื่อแบ่งประโยคภาษาไทย [9]**

  งานวิจัยนี้สามารถสรุปสาระสำคัญได้คือ

  1. ใช้ระเบียบวิธีการเรียนรู้ที่มีชื่อว่า Tri-gram ซึ่งเป็นระเบียบวิธีการเรียนรู้ที่นิยมใช้กันเป็นวงกว้างในหลากหลายงานทางด้านการศึกษาวิจัย โดยหลักการของระเบียบวิธีการเรียนรู้นี้คือ ดูว่าค่าสถิติของค่าสามค่าติดต่อกันมีเกิดขึ้นมากน้อยเพียงใด และเก็บไว้ในรูปแบบของค่าความน่าจะเป็นของการเกิดขึ้นของเหตุการณ์
  2. ใช้ชนิดของคำเป็นลักษณะสำคัญของระเบียบวิธีทางสถิติ
  3. ทำการทดลองบน ORCHID Thai part-of-speech tagged corpus [16] โดยทำ 10-fold cross-validation
  4. ผลที่ได้มีค่าความถูกต้องแม่นยำ 85.26% และค่าความผิดพลาดที่ 8.75% ซึ่งนับได้ว่ามีความผิดพลาดที่สูงมาก

งานวิจัยนี้สังเกตได้ว่ามีการใช้ข้อมูลในการเรียนรู้เพียงแค่ข้อมูลก่อนการเว้นวรรคที่พิจารณาเท่านั้น ไม่ได้มีการพิจารณาข้อมูลในบริบทอย่างรอบด้าน ทำให้ค่าความถูกต้องนั้นไม่สูงมากนัก

- **การศึกษาที่เสนอการใช้วินนาว (Winnow) เป็นระเบียบวิธีการเรียนรู้ เพื่อแบ่งประโยคภาษาไทย [10]**

  งานวิจัยนี้สามารถสรุปสาระสำคัญได้คือ

  1. ใช้ระเบียบวิธีการเรียนรู้ที่มีชื่อว่า วินนาว
  2. ใช้ชนิดของคำ คำโดยรอบ และจำนวนคำโดยรอบการเว้นวรรคที่พิจารณา เป็นลักษณะสำคัญของระเบียบวิธีทางสถิติ โดย จากจุดที่พิจารณาจะสนใจบริบทรอบข้างไปทางด้านซ้าย 2 คำ และไปทางด้านขวา 2 คำ
  3. ทำการทดลองบน ORCHID Thai part-of-speech tagged corpus [16] โดยทำ 10-fold cross-validation และเปรียบเทียบผลกับงานวิจัยก่อนหน้า [9]

4. ผลที่ได้มีค่าความถูกต้องแม่นยำ 89.13% และค่าความผิดพลาดที่ 1.74% ซึ่งเป็นค่าที่ดีขึ้นมากเมื่อเปรียบเทียบกับงานวิจัยก่อนหน้านี้

สังเกตได้ว่า งานวิจัยนี้มีการเลือกใช้ลักษณะสำคัญที่เป็นไปได้ทั้งหมดในการทดลองเดียว ซึ่งหากจะพิจารณาโดยละเอียดแล้ว อาจไม่สามารถสรุปจากการทดลองนี้ได้อย่างชัดเจนว่า การใช้วินนาวร่วมกับชนิดของคำเป็นลักษณะสำคัญหลักแล้วทำให้ผลการแบ่งประโยคมีประสิทธิภาพที่ดี เนื่องจาก ไม่มีการทดลองโดยแยกการใช้ลักษณะสำคัญออกจากกันและดูผลของแต่ละลักษณะสำคัญเปรียบเทียบกัน

- **งานวิจัยที่เสนอการใช้ แมกซิมัม เอนโทรปี (Maximum Entropy) เป็นระเบียบวิธีการเรียนรู้ร่วมกับการใช้คำโดยรอบเป็นลักษณะสำคัญ เพื่อแบ่งประโยคภาษาไทย [11]**

งานวิจัยนี้ทำการแบ่งประโยคภาษาไทยโดยมีจุดประสงค์หลักคือเพื่อเพิ่มความถูกต้องแม่นยำของระบบแปลภาษาโดยวิธีการทางสถิติ ปัญหาของระบบแปลภาษาแทบทุกแบบนั้นคือ การที่สิ่งเข้าของระบบมีความยาวมากเกินไป หรือกล่าวได้ว่า สิ่งเข้าของระบบไม่อยู่ในรูปแบบที่เหมาะสมกับการนำไปประมวลผล สำหรับรูปแบบที่เหมาะสมนั้นขึ้นอยู่กับระบบที่ออกแบบ โดยส่วนใหญ่แล้วจะอยู่ในรูปแบบของ ประโยค วลี หรือ อนุประโยค นอกจากแบ่งประโยคแล้ว งานวิจัยนี้ยังทำกระบวนการอื่นๆอันจะนำไปสู่การได้มาของสิ่งเข้าที่สมบูรณ์และเหมาะสม เช่น การระบุชื่อเฉพาะของคำนาม (Name entity recognition) การแก้ไขข้อผิดพลาดของการสะกดคำในบางรูปแบบที่ไม่ยากเกินไป การแบ่งคำ เป็นต้น ซึ่งจะสรุปสาระสำคัญเฉพาะในส่วนของการแบ่งประโยคภาษาไทยได้ดังนี้

1. ใช้ระเบียบวิธีการเรียนรู้ที่มีชื่อว่า แมกซิมัม เอนโทรปี
2. ใช้คำโดยรอบและจำนวนคำโดยรอบการเว้นวรรคเป็นลักษณะสำคัญของระเบียบวิธีการเรียนรู้
3. ใช้ประโยคทั้งสิ้น 361,802 ประโยคเป็นข้อมูลสำหรับการเรียนรู้ และทดสอบกับ ORCHID Thai part-of-speech tagged corpus [16]
4. ผลที่ได้มีค่าความถูกต้องแม่นยำ 91.19% และค่าความผิดพลาดที่ 3.91% ซึ่งเมื่อเปรียบเทียบกับการทดลองที่ใช้ระเบียบวิธีเรียนรู้แบบวินนาว [10] พบว่า ความถูกต้องแม่นยำเพิ่มขึ้น แต่ความผิดพลาดก็เพิ่มขึ้นด้วยเช่นกัน

การทดลองนี้แตกต่างจากการทดลองอื่นๆที่กล่าวมาก่อนหน้านี้คือ ข้อมูลสำหรับการเรียนรู้นั้นใช้ข้อมูลที่ไม่ใช้ ORCHID Thai part-of-speech tagged corpus [16] และเป็นข้อมูลที่มีจำนวนมากกว่าหลายเท่าหากเทียบกับ ORCHID corpus แต่ผลการทดลองมีค่าความถูกต้องแม่นยำไม่ได้สูงขึ้นอย่างชัดเจนมาก อาจเป็นผลเนื่องจากลักษณะสำคัญที่ใช้นั้นเป็นลักษณะสำคัญที่เป็นพื้นฐานกว่างานวิจัย

อื่น แต่เนื่องด้วยข้อมูลจำนวนมากในการเรียนรู้ทำให้ชดเชยจุดอ่อนตรงนี้ได้ สาเหตุที่การวิจัยนี้เลือกใช้ลักษณะสำคัญเช่นนี้เนื่องจาก ต้องการการประมวลผลที่รวดเร็วเพื่อที่จะเป็นสิ่งเข้าของระบบแปลภาษา จึงจำเป็นต้องลดขั้นตอนกระบวนการต่างๆลงให้มากเท่าที่จะเป็นไปได้

2.3 งานวิจัยที่ใกล้เคียง

งานทางด้านการแบ่งประโยคถือเป็นรูปแบบหนึ่งของการวิเคราะห์ข้อความ (Text analysis) ดังนั้นงานวิจัยต่างๆในหัวข้อของการวิเคราะห์ข้อความจึงมีความเกี่ยวโยงอย่างยิ่งกับการแบ่งประโยค ทั้งในด้านของรูปแบบของระเบียบวิธีที่ใช้ ลักษณะสำคัญที่ใช้ กฎต่างๆที่เกี่ยวข้อง มีงานที่ใกล้เคียงกับการแบ่งประโยคอาทิเช่น การแบ่งคำ (Word segmentation), การแบ่งวลี (Phrase chunking), การหาอนุประโยค (Clause recognition) เป็นต้น ซึ่งหัวข้องานวิจัยที่กล่าวมานี้มีการศึกษาที่กว้างขวางทั้งในภาษาไทยเองและภาษาที่มีลักษณะของภาษาและปัญหาที่ใกล้เคียงกับภาษาไทย ซึ่งจะขอกล่าวแบ่งตามลักษณะของปัญหา ดังต่อไปนี้

- **งานวิจัยทางด้านการแบ่งคำ**

การศึกษาทางด้านนี้สำหรับภาษาไทยและภาษาอื่นๆ สามารถจัดกลุ่มการศึกษาออกเป็นสองประเภทหลักคือ การใช้พจนานุกรม (Dictionary-based, DCB) โดยการจับคู่ยาวที่สุด (Longest matching) และการใช้การเรียนรู้ทางสถิติ (Machine learning-based, MLB) ซึ่งมีการศึกษาค้นคว้ากันอย่างแพร่หลายทั้งสองระเบียบวิธี ในภาษาไทยมีงานวิจัยที่ศึกษาเปรียบเทียบระหว่างการใช้พจนานุกรมกับการใช้การเรียนรู้ทางสถิติ [17] โดยใช้ระเบียบวิธีเรียนรู้ที่แตกต่างกันเพื่อเปรียบเทียบด้วย การใช้ระเบียบวิธีทางพจนานุกรมโดยการจับคู่ยาวที่สุดให้ผลที่ดีกว่าการใช้ระเบียบวิธีเรียนรู้ประเภท เนอิฟเบย์ (Naïve bayes) และ ซับพอร์ตเวกเตอร์แมคชีน (Support vector machine) แต่การใช้ระเบียบวิธีเรียนรู้ประเภท คอนดิชันนอล แรนดอม ฟิวต์ (Conditional random fields) ให้ผลลัพธ์ที่ดีที่สุดจากการทดลอง จากการศึกษานี้สามารถสรุปได้ว่า ทั้งระเบียบวิธีที่ใช้กฎและระเบียบวิธีเรียนรู้ทางสถิติล้วนมีข้อดีข้อด้อยแตกต่างกันในแต่ละแง่มุม มีการศึกษาที่นำเอาทั้งระเบียบวิธีทางพจนานุกรมและระเบียบวิธีเรียนรู้ทางสถิติมาใช้ร่วมกันเพื่อให้ได้ผลลัพธ์ที่ดี [18] ซึ่งการศึกษานี้แสดงให้เห็นว่าการใช้ทั้งสองระเบียบวิธีร่วมกันอย่างเหมาะสมสามารถให้ผลลัพธ์ที่ดีกว่าการใช้ระเบียบวิธีการใดวิธีการหนึ่ง

- **งานวิจัยด้านการแบ่งวลี (Phrase break)**

นักวิจัยจำนวนมากทำการศึกษาเกี่ยวกับการแบ่งวลี เนื่องจาก วลีถือเป็นองค์ประกอบทางภาษาที่ใหญ่กว่าคำและมีความหมายที่ชัดเจนและสื่อความได้ดีมากกว่าคำ สำหรับภาษาไทยมีการศึกษาจำนวนมากและใช้วิธีการแตกต่างกันเพื่อที่จะทำให้ได้ผลลัพธ์ที่ดีที่สุดซึ่งส่วนใหญ่แล้วเป็นการใช้ระเบียบวิธีเรียนรู้ทางสถิติร่วมกับลักษณะสำคัญที่เหมาะสม โดยมีการศึกษาในแง่มุม

ต่างๆทั้งการหาระเบียบวิธีการเรียนรู้ที่เหมาะสม [19, 20] การหาลักษณะสำคัญที่เหมาะสม [21, 22] ซึ่งในขณะนี้สำหรับภาษาไทยอาจจะสามารถสรุปจากหลายๆงานวิจัยได้ว่าการใช้ลักษณะสำคัญคือ แคททิกอเรียลแกรมม่า ให้ผลลัพธ์ที่น่าพึงพอใจที่สุดสำหรับการแบ่งวลีในภาษาไทย

- **งานวิจัยด้านการรู้จำชื่อและเอนทิตี (Name entity recognition)**

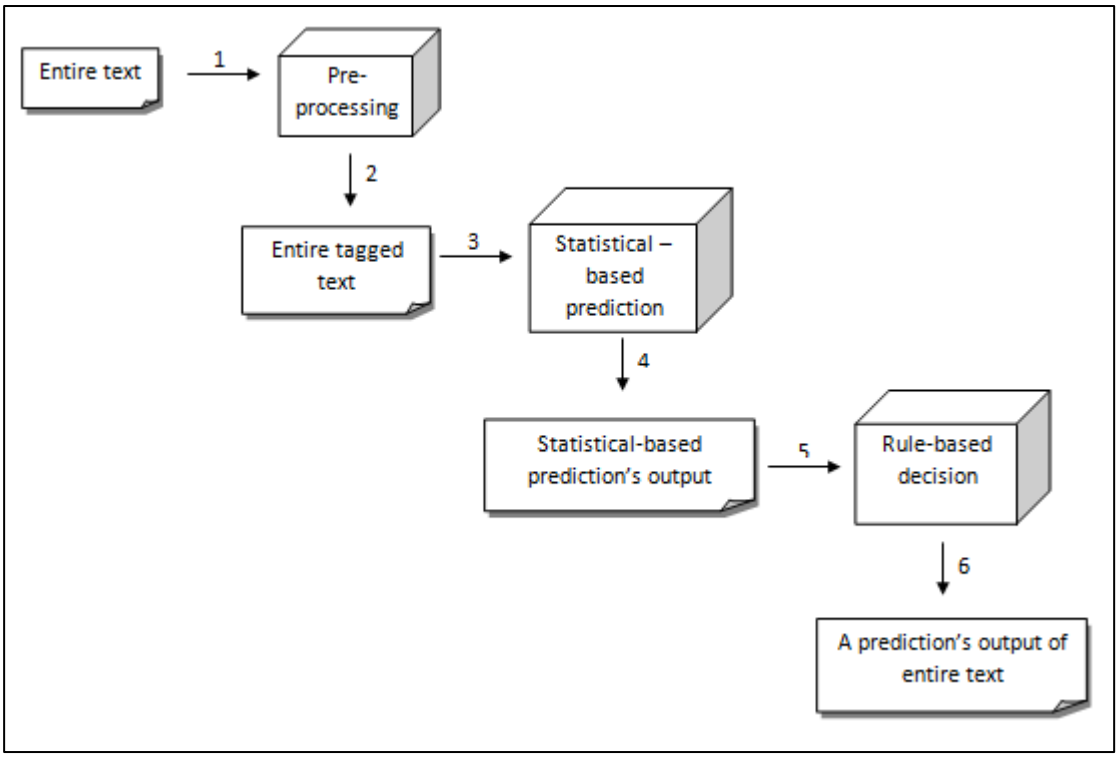
การศึกษางานด้านนี้สำหรับภาษาไทยมีการศึกษาทั้งในรูปแบบของการใช้ระเบียบวิธีเรียนรู้ทางสถิติและการใช้กฎเพื่อวิเคราะห์คำในประโยคว่าคำใดบ้างเป็นชื่อเฉพาะ การรู้จำชื่อเฉพาะนั้นมีความจำเป็นมากต่อการประมวลผลในงานด้านต่างๆ อาทิเช่น การแปลภาษา การแบ่งคำ เป็นต้น มีการเสนอการใช้กฎเพื่อการรู้จำชื่อและเอนทิตีในประโยค [23] ซึ่งกฎที่ใช้นั้นได้มาจากการสังเกตการใช้งานจริงของภาษาไทย โดยอาศัยการบริบทโดยรอบเพื่อพิจารณาว่าคำที่กำลังพิจารณาอยู่นั้นเป็นชื่อเฉพาะหรือไม่ ยกตัวอย่างเช่น หากพบบริบทที่ขึ้นต้นด้วยคำว่า “นาย” คำต่อไปมีความเป็นไปได้สูงที่จะเป็นชื่อของบุคคล เป็นต้น ซึ่งผลลัพธ์ที่ได้จากการทดลองมีค่า F-measure อยู่ที่ 91.442% ซึ่งจุดที่ผิดพลาดนั้นส่วนใหญ่แล้วเป็นจุดที่กฎที่นิยามไว้ไม่ครอบคลุมหรือตรงกับกฎที่นิยามไว้แต่เป็นข้อปลีกย่อยที่เกิดขึ้นในภาษา นอกจากนี้แล้วยังมีการใช้ระเบียบวิธีทางสถิติมาใช้เพื่อการรู้จำชื่อเฉพาะในข้อความภาษาไทยเพื่อเพิ่มประสิทธิภาพของการแบ่งคำ [5] โดยใช้ระเบียบวิธีเรียนรู้ที่มีชื่อว่า คอนดิชันนอล แรนดอม ฟิวต์ ร่วมกับลักษณะสำคัญคือ คำบริบทรอบๆที่ปรากฏในข้อความ ผลที่ได้มีค่าความถูกต้องที่ 93.96%

นอกจากกลุ่มงานวิจัยที่กล่าวมาข้างต้นแล้วยังมีการศึกษาอีกมากมายที่เกี่ยวข้องใกล้เคียงและสามารถนำแนวคิด มาใช้ร่วมกับการแบ่งประโยคได้ เช่น งานวิจัยเรื่อง การรู้จำประโยคเชิงซ้อนในภาษาไทยที่ประกอบด้วยประพันธสรรพนาม [24] ซึ่งใช้กฎหรือหลักทางไวยากรณ์ภาษาไทยในการรู้จำประโยคเชิงซ้อนในประโยค

บทที่ 3
วิธีการวิจัย

1. แนวทางการวิจัย

การแบ่งประโยคภาษาไทยที่นำเสนอประกอบด้วย 3 ส่วนประกอบหลักคือ การเตรียมข้อมูลเบื้องต้น (Pre-processing) การตัดสินโดยใช้ระเบียบวิธีทางสถิติ (Statistical-based decision) และการตัดสินโดยใช้หลักทางไวยากรณ์ทางภาษาศาสตร์ (Rule-based decision) ดังภาพที่ 3.1



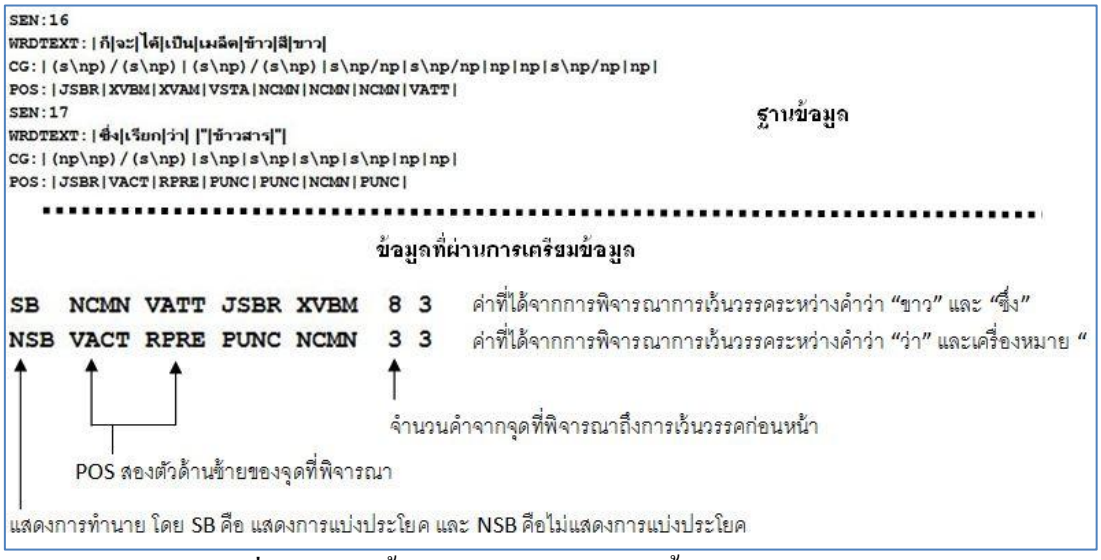
ภาพที่ 3.1 แสดงขั้นตอนโดยสังเขปของการแบ่งประโยคภาษาไทย

จากรูปที่ 3.1 แสดงให้เห็นว่ากระบวนการแบ่งประโยคภาษาไทยที่นำเสนอประกอบด้วย 3 ส่วนประกอบหลักคือ การเตรียมข้อมูลเบื้องต้น (Pre-processing) การตัดสินโดยใช้ระเบียบวิธีทางสถิติ (Statistical-based decision) และการตัดสินโดยใช้หลักทางไวยากรณ์ทางภาษาศาสตร์ (Rule-based decision)

1.1 การเตรียมข้อมูลเบื้องต้น

การเตรียมข้อมูลเบื้องต้นคือ การระบุค่าต่างๆที่จำเป็นให้กับข้อมูลสำหรับใช้ในกระบวนการต่อไป เช่น ชนิดของคำ ประเภทของวากยสัมพันธ์ จำนวนคำในแต่ละช่วง เป็นต้น

รวมถึงการจัดรูปแบบของข้อมูลให้สอดคล้องกับการประมวลผลในขั้นต่อไปด้วย ซึ่งในการศึกษานี้ใช้ข้อมูลที่มีการระบุชนิดของคำและประเภทของวากยสัมพันธ์ไว้อย่างถูกต้องแล้วซึ่งจะกล่าวต่อไปในส่วนของข้อมูลที่ใช้ในการทดลอง ดังนั้นการประมวลผลในสองส่วนนี้จึงไม่จำเป็นต้องดำเนินการ ดังนั้นในขั้นตอนการเตรียมข้อมูลเบื้องต้นนั้นจึงเหลือเพียงส่วนของ การเตรียมลักษณะสำคัญให้ได้รูปแบบตามที่ต้องการเพื่อใช้ในขั้นตอนของการตัดสินใจโดยใช้ระเบียบวิธีทางสถิติ ซึ่งแต่ละลักษณะสำคัญที่ใช้มีขั้นตอนการประมวลผลที่แตกต่างกัน ดังนั้นขึ้นตอนของการเตรียมข้อมูลจึงต้องทำใหม่ทุกครั้งเมื่อมีการเปลี่ยนแปลงการใช้งานลักษณะสำคัญ ตัวอย่างของขั้นตอนการเตรียมข้อมูลเบื้องต้น แสดงดังภาพที่ 3.3



ภาพที่ 3.2 แสดงขั้นตอนการเตรียมข้อมูลเบื้องต้นจากฐานข้อมูล

1.2 การตัดสินโดยใช้ระเบียบวิธีทางสถิติ

การใช้ระเบียบวิธีทางสถิติมีสองส่วนหลักที่ต้องคำนึงถึงคือ ระเบียบวิธีเรียนรู้ (Learning algorithm) และลักษณะสำคัญ (Feature)

1.2.1 การเลือกระเบียบวิธีเรียนรู้ที่ใช้กับงานวิจัยนี้

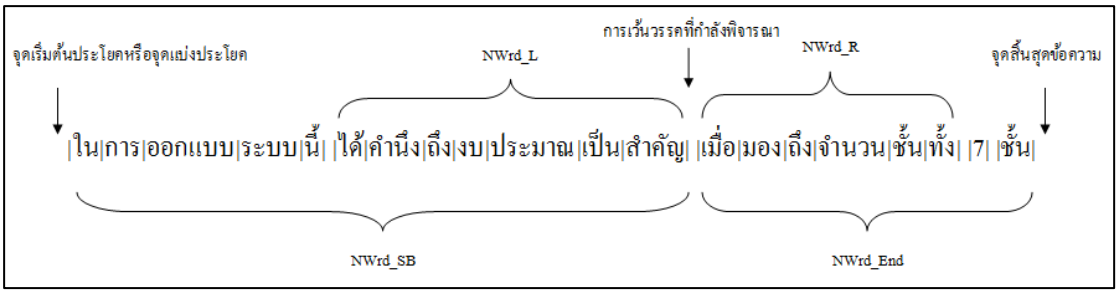
จากการศึกษางานวิจัยที่ผ่านมา [21, 25] CART เป็นระเบียบวิธีที่ให้ผลลัพธ์ที่ดีเมื่อศึกษาเปรียบเทียบกับระเบียบวิธีอื่นๆ สำหรับงานทางด้านการแบ่งวลีภาษาไทยซึ่งมีความใกล้เคียงกับการแบ่งประโยคภาษาไทย และ CART ยังสามารถให้ผลการตัดสินที่ค่อนข้างแม่นยำแม้ข้อมูลที่ใช้เรียนรู้มีความกระจายตัวมากก็ตาม ยิ่งไปกว่านั้น CART ยังถูกนำไปใช้ในหลากหลายงานที่เกี่ยวข้องในด้านของการประมวลผลข้อความและให้ผลลัพธ์ที่ดีมาก [26, 27] ซึ่งเป็นการยืนยันได้ว่า CART เป็นระเบียบวิธีเรียนรู้ที่มีประสิทธิภาพ ดังนั้นในงานวิจัยนี้จึงเลือกใช้ CART เป็นระเบียบวิธีเรียนรู้

1.2.2 การเลือกลักษณะสำคัญที่ใช้ในการแบ่งประโยค

การทดลองในงานวิจัยนี้จะมีการเลือกลักษณะสำคัญแต่ละชนิดและดูผลลัพธ์ของการใช้ลักษณะสำคัญแต่ละชนิดรวมทั้งการนำลักษณะสำคัญชนิดต่างๆมารวมกันเพื่อดูผลลัพธ์ที่ได้ โดยลักษณะสำคัญที่ใช้ประกอบด้วย

1. ชนิดของคำ (ถูกใช้ในงานวิจัยการแบ่งประโยคภาษาไทย [9, 10])
2. แคททิกอเรียลแกรมม่า (เสนอใหม่เพื่อใช้สำหรับการแบ่งประโยคภาษาไทย)
3. คำที่อยู่โดยรอบ (ถูกใช้ในงานวิจัยการแบ่งประโยคภาษาไทย [11])
4. จำนวนคำระหว่างการเว้นวรรคที่พิจารณากับการเว้นวรรคใกล้เคียง (ถูกใช้ในงานวิจัยการแบ่งประโยคภาษาไทย [11])
5. จำนวนคำระหว่างการเว้นวรรคที่กำลังพิจารณากับการเว้นวรรคก่อนหน้าที่เป็นการแบ่งประโยค (เสนอใหม่เพื่อใช้สำหรับการแบ่งประโยคภาษาไทย)
6. จำนวนคำระหว่างการเว้นวรรคที่กำลังพิจารณากับจุดสิ้นสุดของข้อความ (ในการทดลองคือจุดสิ้นสุดของย่อหน้า) (เสนอใหม่เพื่อใช้สำหรับการแบ่งประโยคภาษาไทย)
7. การปรากฏของคำกริยาระหว่างการเว้นวรรคที่พิจารณากับการเว้นวรรคก่อนหน้าที่เป็นการแบ่งประโยค (เสนอใหม่เพื่อใช้สำหรับการแบ่งประโยคภาษาไทย)

ทั้งนี้รายละเอียดที่มาและแนวคิดในการออกแบบลักษณะสำคัญทั้งหมดจะอยู่ในเนื้อหาส่วนถัดไป สำหรับชนิดของคำ แคททิกอเรียลแกรมม่า และคำที่อยู่โดยรอบ การวิจัยในครั้งนี้เลือกพิจารณาสองคำทางด้านซ้ายและสองคำทางด้านขวาของการเว้นวรรคที่พิจารณา การที่เลือกพิจารณาสองคำทางด้านซ้ายและสองคำทางด้านขวาเนื่องจาก การศึกษาหลายงานที่ลักษณะงานมีความใกล้เคียงกับการแบ่งประโยคภาษาไทย [9 - 11, 21] ได้แสดงให้เห็นว่าขนาดของหน้าต่างเท่ากับ 2 คือขนาดที่เหมาะสม ด้วยเหตุผลสองประการคือ บริบทที่มีนัยสำคัญต่อการตัดสินใจส่วนใหญ่แล้วมักอยู่บริเวณโดยรอบที่ไม่ห่างมากนักและการประมวลผลที่ไม่ก่อให้เกิดการ กระจัดกระจายของข้อมูลที่มากเกินไปเมื่อทำการตัดสินใจด้วยระเบียบวิธีทางสถิติ ลักษณะสำคัญนอกเหนือจากชนิดของคำและแคททิกอเรียลแกรมม่าได้แสดงตัวอย่างดังภาพที่ 3.2 เพื่อให้เข้าใจได้ชัดเจนขึ้น



ภาพที่ 3.3 แสดงตัวอย่างของการใช้งานลักษณะสำคัญ

จากภาพที่ 3.2 แสดงตัวอย่างของการใช้งานลักษณะสำคัญที่เกี่ยวเนื่องกับการนับจำนวนคำโดยเริ่มต้นจากจุดของการเว้นวรรคที่กำลังพิจารณา NWrd_L (จำนวนคำระหว่างการเว้นวรรคที่กำลังพิจารณากับการเว้นวรรคใกล้เคียงทางด้านซ้าย) มีค่าเท่ากับ 7 และ NWrd_R (จำนวนคำระหว่างการเว้นวรรคที่กำลังพิจารณากับการเว้นวรรคใกล้เคียงทางด้านขวา) มีค่าเท่ากับ 6 ดังนั้นทั้ง NWrd_L และ NWrd_R จึงเป็นลักษณะสำคัญที่ 4 คือจำนวนคำระหว่างการเว้นวรรคที่พิจารณากับการเว้นวรรคใกล้เคียง เมื่อพิจารณา NWrd_SB ซึ่งเป็นลักษณะสำคัญที่แสดงจำนวนคำระหว่างการเว้นวรรคที่กำลังพิจารณากับจุดแบ่งประโยคก่อนหน้านี้หรือในกรณีที่ยังไม่มีการแบ่งประโยคก็ถือได้ว่าจุดเริ่มต้นประโยคเป็นจุดแบ่งประโยคก่อนหน้า มีค่าเท่ากับ 13 และสุดท้ายคือ NWrd_End ซึ่งเป็นลักษณะสำคัญที่แสดงจำนวนคำระหว่างการเว้นวรรคที่กำลังพิจารณากับจุดสิ้นสุดของข้อความ มีค่าเท่ากับ 10

การได้มาของลักษณะสำคัญที่แสดงถึงการปรากฏของคำกริยาระหว่างการเว้นวรรคที่พิจารณากับการเว้นวรรคก่อนหน้าที่เป็นการแบ่งประโยคจำเป็นที่จะต้องใช้ข้อมูลที่มีการกำกับชนิดของคำแล้วซึ่งในการวิจัยนี้มีตัวเลือกอยู่สองประเภทคือ ใช้การกำกับชนิดของคำหรือใช้การกำกับแคททิกอเรียลแกรมม่าซึ่งทั้งสองมีการแบ่งลักษณะย่อยที่แตกต่างกันแต่หากพิจารณาโดยละเอียดแล้วพบว่าการแบ่งประเภทตามหลักภาษาแบบคร่าวๆ เช่น นาม สรรพนาม กริยา ส่วนขยาย นั้นมีความใกล้เคียงกันมาก ดังนั้นการวิจัยนี้จึงเลือกใช้การกำกับประเภทของคำในระดับของหลักภาษาจากการกำกับของแคททิกอเรียลแกรมม่าด้วยเหตุผลที่ว่าการวิจัยนี้ต้องการนำเสนอการใช้แคททิกอเรียลแกรมม่าเพื่อแบ่งประโยคภาษาไทยแทนที่การใช้ชนิดของคำ ดังนั้นเพื่อความเหมาะสมต่อการนำไปใช้หรือพัฒนาต่อในอนาคตการเลือกใช้ลักษณะสำคัญที่ได้จากรูปแบบการกำกับที่สอดคล้องกันจึงมีความเหมาะสมมากกว่าการเลือกใช้จากรูปแบบการกำกับที่แตกต่างกัน สำหรับแคททิกอเรียลแกรมม่าที่แสดงถึงกริยามีทั้งสิ้น 21 ชนิด ดังแสดงในตารางที่ 3.1 ซึ่งอ้างอิงจาก [13]

ตารางที่ 3.1 แสดงแคททิกอเรียลแกรมม่าที่มีชนิดเป็นคำกริยา

CG Category	Type
s/(s/np)	Auxiliary verb
np/(np/np)	Auxiliary verb
(s\np)\(s\np)	Auxiliary verb
(s\np)/ws	Verb
(s\np)/pp	Verb
(s\np)/np	Verb
(s\np)/(s\np)	Verb, Auxiliary verb
(np\np)/((np\np)/np)	Auxiliary verb
((s\np)/ws)/pp	Verb
((s\np)/ws)/np	Verb
((s\np)/pp)\((s\np)/pp)	Auxiliary verb
((s\np)/pp)/np	Verb
((s\np)/np)\((s\np)/np)	Auxiliary verb
((s\np)/np)/np	Verb
((s\np)/np)/(s\np)	Verb
((np\np)/pp)\((np\np)/pp)	Auxiliary verb
((np\np)/np)\((np\np)/np)	Auxiliary verb
(s\np)\np	Verb
((s\np)/ws)/pp)/np	Verb
((s\np)/pp)/np)/np	Verb
((s/np)/np)/(s\np)/pp	Verb

1.2.3 แนวคิดการออกแบบและใช้ลักษณะสำคัญ

ในการใช้ระเบียบวิธีทางสถิติเพื่อตัดสินใจนั้นลักษณะสำคัญถือเป็นปัจจัยสำคัญที่จะส่งผลต่อความถูกต้องแม่นยำหรือความเหมาะสมที่ได้ของผลลัพธ์ ทั้งนี้งานวิจัยนี้ได้ให้ความสำคัญกับการเลือกใช้ลักษณะสำคัญที่สามารถแสดงถึงการใช้หรือรูปแบบที่ปรากฏของภาษาไทยที่มีต่อการแบ่งประโยค ดังนั้นลักษณะสำคัญที่จะนำมาทดลองจึงควรจะต้องมีความเป็นไปได้สูงที่จะให้

ผลลัพธ์ที่ดี ด้วยเหตุผลที่เหมาะสม หรือกล่าวได้ว่า ทุกลักษณะสำคัญที่เสนอล้วนแล้วแต่มีข้อมูลหรือสมมติฐานที่มีแนวโน้มที่แสดงให้เห็นว่าสามารถเพิ่มความถูกต้องแม่นยำของการแบ่งประโยคภาษาไทย ดังรายละเอียดที่แสดงต่อไปนี้

- ชนิดของคำ

  จากที่ได้กล่าวมาแล้วในบทที่ 2 ชนิดของคำถูกนำมาใช้เป็นลักษณะสำคัญแรกๆ ในแทบทุกการศึกษาวิจัยที่เกี่ยวข้องกับการวิเคราะห์ข้อความทั้งในงานวิจัยภาษาไทยและภาษาอื่นๆ ดังเห็นได้จากการศึกษาเรื่องการแบ่งประโยคภาษาไทย [9, 10] ซึ่งใช้ชนิดของคำเป็นลักษณะสำคัญหลัก ดังนั้นในการศึกษานี้จึงยังคงให้ความสำคัญกับชนิดของคำเพื่อดูผลลัพธ์ที่ได้หากใช้ร่วมกับลักษณะสำคัญอื่นๆที่เสนอ แต่อย่างไรก็ดีจากการศึกษางานวิจัยก่อนหน้านี้ [21, 22] ชนิดของคำมีแนวโน้มที่จะให้ผลลัพธ์ที่ด้อยกว่าการใช้แคททิกอเรียลแกรมม่า

- แคททิกอเรียลแกรมม่า

  แคททิกอเรียลแกรมม่าได้ถูกใช้งานมากขึ้นในหลากหลายปัญหาทางด้านการวิเคราะห์ข้อความ ตัวอย่างเช่น การใช้แคททิกอเรียลแกรมม่าเป็นลักษณะสำคัญหลักเปรียบเทียบกับการใช้ชนิดของคำในการทำนายการแบ่งวลีเพื่อใช้สำหรับการอ่านโดยระบบอัตโนมัติ [21] ซึ่งการศึกษานี้แสดงให้เห็นว่าแคททิกอเรียลแกรมม่าสามารถสะท้อนการเกิดขึ้นหรือธรรมชาติของภาษาไทยได้ดีกว่าการใช้ชนิดของคำ นอกจากนี้แคททิกอเรียลแกรมม่ายังให้ผลลัพธ์ที่ดีในการแบ่งข้อความยาวๆออกเป็นส่วนย่อยๆเพื่อเพิ่มประสิทธิภาพของระบบการแปลงต้นไม้ไวยากรณ์ [22] นอกจากภาษาไทยแล้วในภาษาต่างประเทศ แคททิกอเรียลแกรมม่ายังสามารถให้ผลลัพธ์ที่ดีในการแสดงถึงความสัมพันธ์เชิงความหมายของข้อความอีกด้วย [28] สาเหตุที่ทำให้แคททิกอเรียลแกรมม่ามีความโดดเด่นอย่างมากในการแก้ปัญหาด้านการประมวลผลข้อความคือ แคททิกอเรียลแกรมม่าเป็นการแทนคำต่างๆด้วยสัญลักษณ์ที่ไม่เพียงแต่บ่งบอกถึงหน้าที่ของคำคำนั้นที่มีต่อคำรอบข้องแต่ยังสามารถบอกได้ว่าบริบทรอบๆควรที่จะมีลักษณะอย่างไรและยังบอกถึงการลักษณะที่เหมาะสมสำหรับการรวมกันเพื่อให้เป็นส่วนประกอบที่ใหญ่ขึ้น ซึ่งตามหลักภาษาแล้วคือวลีหรือประโยค ดังนั้นในงานวิจัยนี้จึงเสนอให้ทดลองใช้แคททิกอเรียลแกรมม่าเป็นลักษณะสำคัญหลักแทนการใช้ชนิดของคำ และนอกจากนี้ยังมีการทดลองรวมการใช้งานของทั้งแคททิกอเรียลแกรมม่าและชนิดของคำเข้าด้วยกันเพื่อดูผลลัพธ์ที่เกิดขึ้น

  จากที่กล่าวว่าแคททิกอเรียลแกรมม่าสามารถบอกถึงการรวมกันเป็นองค์ประกอบทางภาษาที่ใหญ่ขึ้น ซึ่งตามหลักภาษาศาสตร์แล้วคือวลีหรือประโยค แต่หากพิจารณาคุณสมบัตินี้กับการแบ่งประโยคภาษาไทยจะพบว่า คุณสมบัติดังกล่าวไม่สามารถแสดงขอบเขตของประโยค

ภาษาไทยได้เสมอ เนื่องจากประโยคภาษาไทยมีความกำกวมและเกิดจากการนำคำต่าง ๆ มาต่อกัน ดังนั้นประโยคที่ถูกความหมายและถูกตรรกะจึงสามารถเขียนได้หลายรูปแบบจึงเป็นผลให้แคททิกอเรียลแกรมม่าที่มีการนิยามไว้ไม่สามารถครอบคลุมได้ทุกรูปแบบที่เกิดขึ้น ยกตัวอย่างเช่นประโยค “ชาวนาใช้วัวควายไถนา” ซึ่งเป็นประโยคจากฐานข้อมูลที่มีแบ่งคำ ประโยค และกำกับแคททิกอเรียลแกรมม่าไว้โดยนักภาษาศาสตร์ ประโยคนี้ถูกตัดคำออกเป็น “|ชาวนา|ใช้|วัว|ควาย|ไถนา|” และมีการกำกับแคททิกอเรียลแกรมม่าดังนี้ “|np|s\np/np|np|np|s\np|” เมื่อพิจารณาตามหลักการของแคททิกอเรียลแกรมม่า ข้อความดังกล่าวจะถูกแบ่งออกเป็น 2 ประโยค คือ “ชาวนาใช้วัว” และ “ควายไถนา” ซึ่งเป็นประโยคที่ไม่ถูกต้องตามความหมายที่แท้จริง นอกจากการบอกขอบเขตของประโยคที่ไม่ถูกต้องจากการพิจารณาเฉพาะขอบเขตที่ได้จากแคททิกอเรียลแกรมม่าเพียงอย่างเดียวแล้วยังมีกรณีที่หากพิจารณาตามหลักการของแคททิกอเรียลแกรมม่าแล้วไม่สามารถรวมองค์ประกอบให้เป็นประโยคได้ ดังตัวอย่างในภาพที่ 3.4

```
SEN:2
WRDTEXT:|เรา|จะ|ไป|โรงเรียน|
CG:|np|(s\np)/(s\np)|s\np/np|np|
POS:|PPRS|XVBM|VACT|NCMN|
SEN:3
WRDTEXT:|ไป|เที่ยว|
CG:|s\np/np|np|
POS:|VACT|VACT|
```

ภาพที่ 3.4 แสดงตัวอย่างของประโยคและการกำกับแคททิกอเรียลแกรมม่าในฐานข้อมูล

จากภาพที่ 3.4 ข้อความว่า “ไปเที่ยว” ถูกพิจารณาเป็นประโยคตามหลักภาษาศาสตร์ เนื่องจากในทางภาษาศาสตร์ ประโยคสามารถละประธานของประโยคไว้ได้ แต่หากพิจารณาตามหลักการของแคททิกอเรียลแกรมม่าแล้วจะพบว่า เมื่อรวมคำว่า “ไป” กับ “เที่ยว” ซึ่งมีการกำกับแคททิกอเรียลแกรมม่าเป็น “s\np/np” และ “np” ตามลำดับ แล้วจะได้ “ไปเที่ยว” ซึ่งมีแคททิกอเรียลแกรมม่าคือ “s\np” ซึ่งสำหรับแคททิกอเรียลแกรมม่า “s\np” ไม่ถูกจัดเป็นประโยค ดังนั้นตัวอย่างนี้จึงแสดงให้เห็นว่า ไม่สามารถพิจารณาหาขอบเขตของประโยคที่ถูกต้องได้จากการดูเพียงหลักการรวมกันของแคททิกอเรียลแกรม่าเท่านั้น เนื่องจากการเกิดขึ้นจริงของภาษาในทุกรูปแบบการใช้งานไม่สามารถถูกครอบคลุมได้ด้วยแคททิกอเรียลแกรมม่า ดังนั้นแคททิกอเรียลแกรมม่าจึงถูกใช้งานเป็นเพียงลักษณะสำคัญหนึ่งสำหรับการพิจารณาการแบ่งประโยคภาษาไทย

- คำที่อยู่โดยรอบ

คำที่อยู่โดยรอบคือบริบทจริงๆรอบจุดหรือตำแหน่งที่กำลังพิจารณา ยกตัวอย่างเช่น “เด็ก|นักเรียน|ชอบ|ไป|โรงเรียน|สาย| |และ|ต้อง|ถูก|ลงโทษ” หากพิจารณาจุดที่มีการเว้นวรรคโดยใช้ขนาดของหน้าต่างคือ 2 จะได้ว่า ลักษณะสำคัญที่เป็นคำที่อยู่โดยรอบคือ “โรงเรียน”, “สาย”,” และ”,”ต้อง” จากตัวอย่างแสดงให้อย่างชัดเจนว่าลักษณะสำคัญที่เป็นคำที่อยู่โดยรอบสามารถได้มาง่ายมากหากมีการแบ่งคำมาแล้ว แต่เนื่องจากเป็นการใช้คำจริงๆที่เกิดขึ้นในการใช้ภาษาจึงมีแน้วโน้มสูงมากที่จะมีความหลากหลายและแตกต่างกันเป็นอย่างมาก ซึ่งเหตุการณ์เช่นนี้ย่อมส่งผลเสียต่อการทำนายโดยใช้ระเบียบวิธีทางสถิติและทำให้ผลลัพธ์ที่ได้ขาดความแม่นยำ เนื่องจากหลักการของการทำนายโดยระเบียบวิธีทางสถิตินั้นคล้ายกับการเรียนรู้และหาคำตอบให้ใกล้เคียงกับสิ่งที่ระบบได้เรียนรู้มา แต่หากการเรียนรู้เป็นการเรียนรู้ที่กระจัดกระจายขาดซึ่งทิศทางที่แน่นอนย่อมส่งผลให้การทำนายได้ผลลัพธ์ที่ขาดความแน่นอนแม่นยำตามไปด้วย ดังนั้นลักษณะสำคัญนี้จึงเป็นลักษณะสำคัญที่มีแนวโน้มที่จะให้ผลลัพธ์ที่มีความแม่นยำที่ไม่สูงเมื่อเทียบกับการใช้ชนิดของคำหรือแคททิกอเรียลแกรมม่า ดังนั้นการวิจัยนี้จึงเลือกคำที่อยู่โดยรอบเป็นลักษณะสำคัญที่เป็นเส้นหลักล่าง (Baseline) เพื่อใช้เปรียบเทียบกับลักษณะสำคัญอื่นๆ แต่ทั้งนี้ในข้อเสียของคำที่อยู่โดยรอบที่ก่อให้เกิดการการกระจายตัวของข้อมูลอย่างมาก ก็ยังมีข้อดีที่เห็นได้ชัดเจนคือ การได้มาซึ่งลักษณะสำคัญที่ง่ายและรวดเร็วและยังสะท้อนความจริงที่เกิดขึ้นของภาษานั้นๆอย่างที่เป็น จุดนี้เองที่ทำให้มีการนำลักษณะสำคัญนี้ไปใช้ ด้วยเหตุผลด้านของการประมวลผลที่รวดเร็ว การแบ่งประโยคภาษาไทยเพื่อเป็นสิ่งเข้าของระบบแปลภาษาอัตโนมัติ [11] เป็นตัวอย่างที่ดีของการใช้ลักษณะสำคัญนี้ด้วยเหตุผลด้านความรวดเร็วของการได้มาซึ่งลักณะสำคัญ แต่ด้วยข้อด้อยของความแม่นยำ การศึกษาดังกล่าวจึงชดเชยด้วยการใช้ข้อมูลฝึกหัด (Training data) จำนวนมหาศาลเพื่อให้ระบบเรียนรู้จดทำทุกรูปแบบและมีแนวโน้มที่จะเจอรูปแบบที่คล้ายหรือเหมือนรูปแบบที่เคยเรียนรู้มา ดังนั้นผลลัพธ์ที่ได้จึงมีความถูกต้องแม่นยำในระดับที่สูง แต่การแก้ปัญหาด้วยวิธีดังกล่าวไม่สามารถทำได้กับทุกงานเนื่องจากข้อจำกัดด้านของข้อมูลฝึกหัดที่ส่วนใหญ่แล้วมักมีจำกัดรวมถึงการวิจัยนี้ด้วย ดังนั้นงานวิจัยโดยทั่วไปจึงเน้นการพัฒนาบนพื้นฐานที่ว่าด้วยทรัพยากรมีอยู่อย่างจำกัด การวิจัยนี้จึงมีสมมติฐานที่ว่าการใช้คำที่อยู่โดยรอบเป็นลักษณะสำคัญมีแนวโน้มที่จะให้ผลลัพธ์ที่มีความถูกต้องแม่นยำไม่สูงมาก แต่ยังเป็นลักษณะสำคัญที่มีความเป็นไปได้ที่จะให้ผลลัพธ์ที่ดีหากข้อมูลที่ใช้มีความใกล้เคียงกันมาก

- จำนวนคำระหว่างการเว้นวรรคที่กำลังพิจารณากับการเว้นวรรคใกล้เคียง

ลักษณะสำคัญนี้มีการใช้ในงานด้านการประมวลผลข้อความหลายงาน [19, 21, 29] และให้ผลลัพธ์ที่ค่อนดี แต่หากสังเกตให้ดีแล้วจะพบว่างานที่ใช้ลักษณะสำคัญนี้ล้วนแล้วแต่เป็นงานที่ศึกษาบนภาษาที่มีการเขียนที่ต่อเนื่องหรือกล่าวได้ว่าภาษาที่ไม่มีการเว้นวรรคระหว่างคำ เนื่องจากหากใช้ลักษณะสำคัญดังกล่าวกับภาษาที่มีการเว้นวรรคระหว่างคำ เช่น ภาษาอังกฤษ ภาษาฝรั่งเศส หรือภาษาเยอรมัน ลักษณะสำคัญดังกล่าวจะไม่สามารถแสดงออกถึงบริบทต่างๆ ได้เลย เนื่องจากการเว้นวรรคเป็นสิ่งที่เกิดขึ้นโดยปกติของการใช้งานภาษา แต่ในทางกลับกันภาษาที่ไม่มีการเว้นวรรคระหว่างคำ เช่นภาษาไทย การเว้นวรรคถือได้ว่ามีความหมายทั้งในแง่ของความต้องการสื่อสารของผู้ส่งสารและด้านไวยากรณ์ของภาษา ดังที่ได้กล่าวในบทที่ 2 นอกจากในแง่มุมของการใช้งานที่แสดงให้เห็นว่าลักษณะสำคัญนี้มีแนวโน้มที่ดีที่ช่วยให้ผลลัพธ์ของการแบ่งประมวลผลข้อความมีความถูกต้องแม่นยำมากขึ้นแล้ว เมื่อพิจารณาถึงการเกิดขึ้นของการเว้นวรรคในภาษาไทยมีในหลายกรณีที่ จำนวนคำระหว่างการเว้นวรรคที่กำลังพิจารณากับการเว้นวรรคใกล้เคียง มีแนวโน้มที่ดีในการบอกถึงลักษณะการเกิดขึ้น ยกตัวอย่าง เช่น การกล่าวถึงรายการที่ต่อเนื่องกัน เช่น “ในห้องเรียนมีสิ่งของมากมาย เช่น โต๊ะ เก้าอี้ กระดานดำ ชอล์ก ดินสอ และปากกา เป็นต้น” แน่นอนว่าการเว้นวรรคที่เห็นจากประโยคดังกล่าวมีความชัดเจนที่เป็นการเว้นวรรคที่ไม่แสดงถึงการจบประโยค และเมื่อพิจารณาโดยใช้จำนวนคำระหว่างการเว้นวรรคที่กำลังพิจารณากับการเว้นวรรคใกล้เคียง จะพบว่า จากประโยคดังกล่าว ลักษณะสำคัญนี้จะมีค่าเพียง 1 หรือ 2 เท่านั้น เมื่อพิจารณาเฉพาะรายการหลังคำว่า “เช่น” เหตุการณ์เช่นนี้สามารถพบได้บ่อยมากในข้อความภาษาไทยและไวยากรณ์ภาษาไทยได้มีการกำหนดว่าวรรคต่างๆระหว่างรายการคือวรรคเล็ก ซึ่งหมายความว่าเป็นการเว้นวรรคที่ไม่แสดงถึงการแบ่งประโยค แต่หากจะใช้กฎในการพิจารณาแล้วย่อมทำได้ยาก เนื่องจากกฎมีความแน่นอนและอาจส่งผลกระทบที่ทำให้เกิดความผิดพลาดกับการเกิดขึ้นในรูปแบบอื่นที่อาจเป็นวรรคตอนที่แสดงถึงการแบ่งประโยค จะสังเกตได้ว่า จำนวนคำระหว่างการเว้นวรรคที่กำลังพิจารณากับการเว้นวรรคใกล้เคียง จะมีจำนวนที่น้อยและมักเกิดขึ้นพร้อมๆกับแคททิกอเรียลแกรมม่าหรือชนิดของคำที่แสดงออกถึงคำนามหรือสรรพนาม (แสดงออกถึงสิ่งของหรือบุคคล) เนื่องจากเป็นการเขียนเพื่อแสดงรายการ ดังนั้นลักษณะสำคัญนี้จึงมีแนวโน้มสูงมากที่จะแก้ปัญหาเรื่องของการเว้นวรรคตอนในการแสดงรายการได้อย่างดี นอกจากนี้ในกรณีที่เป็นตัวเลขและตามด้วยหน่วยเช่น “ครูกำลังสอนนักเรียน 10 คน” เมื่อพิจารณาการเว้นวรรคหลังตัวเลขพบว่า จำนวนคำระหว่างการเว้นวรรคที่พิจารณากับการเว้นวรรคใกล้เคียงมีค่าเท่ากับ 1 (ทางซ้าย) และ 1 (ทางขวา) เมื่อประกอบกับการกำกับทั้งแคททิกอเรียลแกรมม่าและชนิดของ

คำของ “10” ซึ่งแสดงให้เห็นว่าเป็นตัวเลข ดังนั้นในกรณีเช่นนี้มีแนวโน้มสูงที่ลักษณะสำคัญนี้จะช่วยให้การทำนายการแบ่งประโยคมีความแม่นยำมากขึ้น

- จำนวนคำระหว่างการเว้นวรรคที่กำลังพิจารณากับการเว้นวรรคก่อนหน้าที่เป็นการแบ่งประโยค

งานวิจัยทางด้านภาษาศาสตร์ภาษาไทย [15] ได้แสดงให้เห็นว่า การจบใจความในหนึ่งประโยคมีลักษณะที่สำคัญประการหนึ่งคือ การที่จะเป็นประโยคได้ต้องมีความยาวที่เหมาะสม ซึ่งนิยามแบบชัดเจนของความยาวที่เหมาะสมนั้นไม่สามารถระบุได้อย่างแน่นอนซึ่งเป็นลักษณะที่เกิดขึ้นเหมือนกันในทุกๆภาษา แต่จากการศึกษาสิ่งหนึ่งที่สังเกตได้คือ ประโยคถึงแม้จะมีความแยวที่ไม่แน่นอนแต่ส่วนใหญ่แล้วการสร้างประโยคขึ้นมาจะมีความยาวที่ใกล้เคียงกัน หรือกล่าวได้ว่า ประโยคที่ซับซ้อนน้อยจะมีความยาวประมาณค่าหนึ่งและประโยคที่มีความซับซ้อนมากจะมีความยาวที่มากขึ้น เมื่อดูโดยรวมแล้วจะมีความยาวอยู่ในช่วงหนึ่งเป็นจำนวนมาก จากข้อสังเกตที่ได้จากงานวิจัยดังกล่าวทำให้เกิดความคิดว่า ความยาวของประโยคควรที่จะเป็นหนึ่งในลักษณะสำคัญที่ต้องพิจารณา และหากดูในมุมมองของบุคคลทั่วไปที่ต้องแบ่งประโยคด้วยการตัดสินใจของตัวเองจากการอ่านและทำความเข้าใจข้อความทั้งหมดแล้วโดยทั่วไปแล้วก็จะพยายามแบ่งประโยคให้มีความยาวที่เหมาะสม ยกตัวอย่างเช่น หากจุดที่พิจารณาห่างจากจุดสิ้นสุดของประโยคก่อนหน้าเพียงเล็กน้อยและเนื้อความที่ได้มีจำนวนไม่มาก มีแนวโน้มสูงที่จุดนั้นจะถูกพิจารณาเป็นวรรคเล็กหรือไม่เป็นจุดที่แสดงถึงการแบ่งประโยค นอกจากงานวิจัยทางด้านภาษาศาสตร์และข้อสังเกตจากการใช้ภาษาแล้ว ยังมีงานวิจัยหลายงาน แม้ว่าจะไม่ได้เป็นงานวิจัยด้านการแบ่งประโยคที่ใช้ลักษณะสำคัญนี้เพื่อแก้ปัญหาด้านการวิเคราะห์ข้อความและได้ผลลัพธ์ที่มีความถูกต้องแม่นยำมากขึ้น [19, 29, 30] ดังนั้นการวิจัยนี้จึงมีแนวคิดว่า จำนวนคำระหว่างการเว้นวรรคที่กำลังพิจารณากับการเว้นวรรคก่อนหน้าที่เป็นการแบ่งประโยคมีแนวโน้มสูงที่จะเป็นลักษณะสำคัญที่ช่วยให้ผลลัพธ์ของการแบ่งประโยคภาษาไทยมีความถูกต้องแม่นยำ

นอกจากงานวิจัยด้านภาษาศาสตร์ การศึกษาจากลักษณะของการใช้งานภาษา และงานวิจัยก่อนหน้าที่ให้ผลลัพธ์ที่ดีแล้ว เพื่อให้เกิดความมั่นใจมากขึ้นต่อการใช้ลักษณะสำคัญนี้ ผู้วิจัยได้ทำการตรวจสอบความยาวของประโยคจากฐานข้อมูลข้อความภาษาไทยที่มีการตรวจสอบความถูกต้องและกำกับรายละเอียดที่สำคัญโดยนักภาษาศาสตร์แล้วซึ่งถูกใช้เป็นข้อมูลสำหรับทำการทดลอง พบว่า มีการกระจายตัวระหว่างจำนวนคำในประโยคและจำนวนครั้งที่เกิดขึ้นดังภาพที่ 3.5 แสดงให้เห็นว่า ประโยคส่วนใหญ่ (จำนวนครั้งที่เกิดขึ้นเกิน 800 ครั้ง) มีจำนวนคำในประโยคอยู่ระหว่าง 6 ถึง 17 จากข้อมูลนี้เป็นการยืนยันที่ดีต่อข้อสรุปจากงานวิจัยด้าน

ภาษาศาสตร์ [15] และข้อสังเกตจากการใช้ภาษาว่า การจะพิจารณากลุ่มของข้อความใดๆว่าเป็นประโยคหรือไม่นั้น ความยาวหรือจำนวนคำภายในกลุ่มข้อความมีผลต่อการพิจารณาความเป็นประโยค จากรูปที่ 3.3 เมื่อพิจารณาช่วงที่มีความถี่น้อยของการเกิดขึ้นน้อยมากจะพบว่าเป็นช่วงที่ประโยคมีความสั้นมากหรือยาวมากเมื่อเทียบกับความยาวโดยเฉลี่ยทั่วไปของประโยค เมื่อศึกษาดูประโยคที่มีความยาวน้อยพบว่า ส่วนใหญ่แล้วเป็นประโยคความเดียวหรือประโยคที่มีลักษณะเป็นส่วนขยายของประโยคหลัก เช่น “|มี|ผลงาน|เป็น|รูปธรรม|” ซึ่งมีความยาวเพียง 4 คำเท่านั้น จากผลลัพธ์ที่ได้แสดงให้เห็นแนวโน้มว่าลักษณะพิเศษดังกล่าวสามารถเพิ่มความถูกต้องแม่นยำต่อการแบ่งประโยคภาษาไทยมากขึ้น เนื่องจากระเบียบวิธีเรียนรู้ทางสถิติจะพยายามแบ่งประโยคเมื่อจำนวนคำในข้อความมีความยาวอยู่ในช่วงที่มีความเป็นไปได้สูงที่จะเป็นประโยค และในทางกลับกันในช่วงที่มีความยาวน้อยมากหรือยาวมากระเบียบวิธีทางสถิติมีแนวโน้มที่จะไม่ระบุว่าการเว้นวรรคนั้นเป็นการเว้นวรรคที่แสดงการจบประโยค ซึ่งสำหรับข้อความที่มีความยาวน้อยย่อมเป็นช่วงให้มีความแม่นยำมากขึ้นเนื่องจากแนวโน้มของประโยคสั้นมีน้อย แต่ในทางกลับกันในช่วงที่มีความยาวของข้อความมากซึ่งจริงๆแล้วอาจเป็นข้อความที่ไม่ยาวมากและต้องการการจบประโยค ลักษณะสำคัญนี้อาจสร้างความผิดพลาดขึ้นได้ แต่อย่างไรก็ตามลักษณะสำคัญนี้ใช้เป็นลักษณะสำคัญที่เสริมให้มีความแม่นยำมาขึ้น ดังนั้นความผิดพลาดที่เกิดจากส่วนนี้อาจชดเชยได้โดยลักษณะสำคัญอื่นๆที่ใช้ร่วมกัน

ภาพที่ 3.5 แสดงการกระจายตัวของจำนวนคำในประโยค

- จำนวนคำระหว่างการเว้นวรรคที่กำลังพิจารณากับจุดสิ้นสุดของข้อความ (ในการทดลองคือจุดสิ้นสุดของย่อหน้า)

  ลักษณะสำคัญนี้ถูกสร้างมาจากเหตุผลที่ต้องการลดความผิดพลาดของการแบ่งประโยคภาษาไทย ที่เกิดขึ้นบริเวณที่ใกล้กับส่วนท้ายของข้อความ โดยมีสมมติฐานที่ว่า หากการเว้นวรรคที่พิจารณาอยู่ใกล้กับการสิ้นสุดของข้อความจะส่งผลให้มีแนวโน้มสูงที่จะเป็นการเว้นวรรคที่ไม่แสดงการจบประโยค เนื่องจากหากเป็นการแสดงการจบประโยคข้อความที่ตามมาภายหลังจะมีความยาวที่น้อยมากและมีแนวโน้มที่จะไม่เป็นประโยค ดังที่ได้กล่าวไว้ในหัวข้อที่ผ่านมา

- การปรากฏของคำกริยาระหว่างการเว้นวรรคที่พิจารณากับการเว้นวรรคก่อนหน้าที่เป็นการแบ่งประโยค

  ลักษณะสำคัญนี้เกิดจากหลักการของภาษาศาสตร์ที่ว่า ข้อความใดจะเรียกว่าเป็นประโยคหรือมีใจความที่สมบูรณ์ได้นั้นจะต้องประกอบด้วยคำกริยาอย่างน้อย 1 คำ [15] และจากงานวิจัยที่ต้องการระบุประโยคความซ้อนภาษาไทยโดยการใช้กฎ [24] ซึ่งหากพิจารณากฎจากงานวิจัยดังกล่าวจะพบว่าทุกกฎที่ได้เสนอมาจะต้องประกอบด้วยคำกริยาอย่างน้อย 1 คำในกฎ ดังนั้นลักษณะสำคัญนี้จึงมีแนวโน้มที่ดีที่จะช่วยให้การแบ่งประโยคมีความถูกต้องมากขึ้น

1.3 การตัดสินโดยใช้หลักไวยากรณ์ทางภาษาศาสตร์

จากที่กล่าวมาแล้วในบทที่ 2 ว่าไวยากรณ์ทางภาษาศาสตร์ที่ถูกกำหนดขึ้นโดยราชบัณฑิตย์นั้นมีทั้งส่วนที่แสดงถึงการแบ่งประโยค (วรรคใหญ่) และแสดงถึงการไม่แบ่งประโยค (วรรคเล็ก) เมื่อพิจารณาในส่วนของวรรคใหญ่พบว่า ไม่มีกฎที่ชัดเจนและเกี่ยวข้องกับลักษณะของข้อความที่จะบอกได้ว่าเป็นวรรคใหญ่ หรือกล่าวได้ว่า วรรคใหญ่เกิดจากความเหมาะสมของผู้ใช้งานที่จะแสดงการเว้นวรรค ซึ่งในทางการประมวลผลด้วยคอมพิวเตอร์แล้วไม่สามารถกระทำได้ และอีกส่วนหนึ่งคือวรรคเล็กซึ่งส่วนนี้มีหลายกฎที่มีความเกี่ยวข้องกับข้อความ ดังนั้นในส่วนนี้เองจึงสามารถนำมาเป็นกฎที่ช่วยเพิ่มความถูกต้องแม่นยำของการแบ่งประโยคภาษาไทยให้มากขึ้น แต่ด้วยข้อจำกัดของวรรคเล็กที่บอกเพียงว่าเป็นการเว้นวรรคตอนที่ไม่แสดงถึงการจบประโยค ดังนั้นการใช้กฎจึงช่วยลดการแบ่งประโยคที่ผิดพลาดแต่ไม่ได้ช่วยเพิ่มจำนวนของการแบ่งประโยค

กฎต่างๆที่จะถูกนำมาใช้นั้นมีบางกฎที่สามารถใช้ได้และบางกฎที่ไม่สามารถใช้ได้เนื่องจากเหตุผลสองประการคือ

1. ข้อจำกัดด้านการประยุกต์ใช้งานโดยคอมพิวเตอร์

การแก้ปัญหาโดยคอมพิวเตอร์คือการแก้ปัญหาที่มีการนิยามที่ชัดเจนหรือสามารถบอกได้ว่าเป็นสิ่งใดสิ่งหนึ่ง แต่การตัดสินใจของกฎบางประเภทไม่สามารถตัดสินใจได้อย่างชัดเจนด้วยเนื้อความที่เป็นสิ่งเข้า จำเป็นต้องอาศัยการตัดสินใจโดยประสบการณ์จากผู้ใช้ภาษา ยกตัวอย่างเช่น ระหว่างชื่อและนามสกุลก้มีการเว้นวรรคเล็ก ซึ่งหากเป็นการตัดสินใจโดยผู้ใช้ภาษาจะสามารถรู้ได้ว่ากลุ่มความแรกเป็นชื่อและอีกกลุ่มข้อความเป็นนามสกุล แต่ด้วยคอมพิวเตอร์แล้วปัญหาเช่นนี้ไม่สามารถที่จะแก้ได้ด้วยกรอบงานของการวิจัยนี้

2. ข้อยกเว้นของหลักไวยากรณ์

กฎหลายๆกฎสามารถมีการใช้งานที่หลากหลายและมีข้อจำกัดรวมทั้งข้อยกเว้นต่างๆ ดังนั้นไม่ใช่ทุกกฎที่จะสามารถนำมาใช้งานและได้ความถูกต้องที่แน่นอนเสมอ ยกตัวอย่างเช่น คำว่า “และ”, “หรือ” กรณีที่อยู่ในรายการที่มีมากกว่า 2 สิ่ง ให้เว้นวรรคเล็กหน้าคำว่า “และ”, “หรือ” ซึ่งกฎนี้กล่าวถึงการใช้คำทั้งสองในลักษณะที่เป็นคำบุพบท ซึ่งหากพิจารณาหลักไวยากรณ์ให้ละเอียดแล้วจะพบว่า ทั้งสองคำสามารถเป็นคำสันธานได้ด้วย และทั้งสองคำนี้เมื่อใช้เป็นคำสันธาน การเว้นวรรคหน้าคำทั้งสองจะคือการเว้นวรรคใหญ่ซึ่งแสดงการสิ้นสุดของประโยคก่อนหน้า

เนื่องด้วยเหตุผลสองประการที่กล่าวมา จึงจำเป็นต้องศึกษาความเป็นไปได้ของหลักเกณฑ์ไวยากรณ์ที่จะนำมาใช้ในการแบ่งประโยคภาษาไทย และเพื่อเป็นการยืนยันว่าหลักไวยากรณ์ที่จะนำมาใช้ครอบคลุมและมีแนวโน้มที่ช่วยลดความผิดพลาดที่เกิดขึ้นจากกระบวนการตัดสินโดยระเบียบวิธีทางสถิติ จึงมีการศึกษาอัตราส่วนการเกิดขึ้นของหลักไวยากรณ์แต่ละข้อในฐานข้อมูลที่ใช้เทียบกับจำนวนของการเว้นวรรคที่ไม่แสดงการแบ่งประโยคซึ่งมีทั้งสิ้น 68,264 การเว้นวรรครวมศึกษาความเป็นไปได้ของการนำมาใช้งาน ดังแสดงในตารางที่ 3.2โดยการปรากฏจะแสดงเป็นเปอร์เซ็นต์ของกฎที่ปรากฏเทียบกับการเว้นวรรคที่ไม่แสดงการแบ่งประโยค

ตารางที่ 3.2 แสดงรายละเอียดหลักไวยากรณ์เพื่อนำมาประยุกต์ใช้กับการแบ่งประโยคภาษาไทย

หลักไวยากรณ์	**การนำมาใช้**	**การปรากฏ (%)**	**หมายเหตุ**
1. เว้นวรรคเล็กระหว่างชื่อและนามสกุล	✗	-	ไม่สามารถระบุได้ว่าข้อความที่เป็นชื่อหรือนามสกุลได้

หลักไวยากรณ์	การนำมาใช้	การปรากฏ (%)	หมายเหตุ
3. ระหว่างชื่อของธนาคาร บริษัท กับคำว่า จำกัด ให้มีการเว้นวรรคเล็ก	✗	-	ไม่สามารถระบุชื่อเฉพาะดังกล่าวได้
4. เว้นวรรคเล็กระหว่างคำว่า “ห้างหุ้นส่วนจำกัด” และ “ห้างหุ้นส่วนสามัญนิติบุคคล” กับชื่อที่ตามมา	✓	2.92×10^{-3}	พบเพียงสองคำเท่านั้นในฐานข้อมูล
5. ระหว่างชื่อสถานที่ต่างๆให้มีการเว้นวรรคเล็ก	✗	-	ไม่สามารถระบุชื่อเฉพาะดังกล่าวได้
6. ระหว่างคำนำหน้านามกับนามให้มีการเว้นวรรคเล็ก	✗	-	คำนำหน้านามมีหลายรูปแบบและไม่มีการกำหนดเป็นมาตรฐาน
7. ระหว่างกลุ่มอักษรย่อต้องมีการเว้นวรรคเล็ก	✓	3.76	ใช้เครื่องหมายมหัพภาค ในการระบุว่าเป็นอักษรย่อ
8. เว้นวรรคเล็กระหว่างการเขียนตัวหนังสือกับตัวเลข	✓	5.40	-
9. เว้นวรรคเล็กระหว่างการเขียนวันและเวลา	✓	0	ไม่พบการใช้งาน
11. เว้นวรรคเล็กหลังข้อความที่เป็นหน่วยมาตรา	✗	4.43	พบว่าอัตราส่วนของการเว้นวรรคหลังหน่วยมาตรา ที่เป็นการวรรคเล็กต่อวรรคใหญ่เท่ากับ 3.5 : 1 ดังนั้นจึงไม่เหมาะสมที่จะใช้กฎข้อนี้
12. ระหว่างตัวหนังสือภาษาอื่นกับตัวหนังสือภาษาไทยให้มีการเว้นวรรคเล็ก	✓	6.12	มี 7 ตำแหน่งที่เป็นการเริ่มต้นของประโยค
13. เว้นวรรคเล็กระหว่างรายการที่กล่าวต่อกัน หรือลำดับของเลข เพื่อแยกรายการออกเป็นรายการย่อยๆ	✗	-	ไม่สามารถระบุส่วนของรายการได้ แต่ลำดับของตัวเลขครอบคลุมโดยกฎข้อ 8

หลักไวยากรณ์	การนำมาใช้	การปรากฏ (%)	หมายเหตุ
14. เว้นวรรคเล็กหน้าและหลังเครื่องหมายไปยาลใหญ่ ไม้ยมก เสมอภาค ทวิภาค (:) วิภัชภาค (:-)	✓	4.24	-นับเฉพาะส่วนของไม้ยมก -ไปยาลใหญ่นับรวมกับไปยาลน้อยในข้อ 16 -ไม่พบเครื่องหมายวิภัชภาค
15. หน้าเครื่องหมายอัญประกาศเปิดและวงเล็บเปิดให้เว้นวรรคเล็ก	✓	8.24	-
16. หลังเครื่องหมาย จุลภาค ไปยาลน้อย อัฒภาค วงเล็บปิด และอัญประกาศปิดให้เว้นวรรคเล็ก	✓	7.12	-
17. หลังข้อความที่เป็นหัวข้อให้เว้นวรรคเล็ก	✗	-	ไม่สามารถระบุข้อความที่เป็นหัวข้อได้
18. หน้าและหลังคำว่า ฉ และ ธ ให้เว้นวรรคเล็ก	✓	0.221	-ไม่ประกฏการใช้ ธ - ปรากฏการใช้ ฉ ที่ต้นประโยค 7 ตำแหน่ง
19. หน้าและหลังคำว่า “ได้แก่” ที่ตามมาด้วยรายการที่มากกว่าหนึ่งรายการ ให้เว้นวรรคเล็ก	✓	0.533	-
20. เว้นวรรคเล็กหน้าและหลังคำว่า “เช่น” ในการยกตัวอย่าง	✗	-	ไม่สามารถระบุข้อความที่แสดงการยกตัวอย่างได้
21. หน้าคำว่า “และ”, “หรือ” ในกรณีที่ใช้อยู่ในรายการให้เว้นวรรคเล็ก	✗	-	ไม่สามารถระบุได้ว่าคำเหล่านี้ใช้เพื่อแสดงรายการ
22. เว้นวรรคเล็กหน้าคำว่า “เป็นต้น” ที่อยู่หลังรายการ	✓	0.464	-
23. หลังคำว่า “ว่า” ถ้าส่วนที่ตามต่อมาคือประโยค ให้เว้นวรรคเล็กหลังคำว่า “ว่า”	✗	-	ไม่สามารถระบุข้อความที่ตามมาได้

นอกจากกฎที่บัญญัติโดยราชบัณฑิตยสถานดังแสดงในตารางที่ 3.2 แล้ว ยังมีข้อสังเกตอีกหลายข้อที่ปรากฏจากการใช้งาน ซึ่งในหลายข้อเป็นการใช้งานที่ไม่ถูกตามไวยากรณ์ที่ถูกกำหนดไว้ แต่ปรากฏในการใช้งานจริง ดังนั้นเพื่อให้การวิจัยสามารถนำไปประยุกต์ใช้งานได้จริงและสอดคล้องกับฐานข้อมูลซึ่งมีความผิดพลาดที่อาจเรียกได้ว่าเป็นความผิดพลาดด้านการใช้ภาษาที่เสมือนเป็นการใช้งานที่ถูกต้อง ซึ่งแสดงในตารางที่ 3.3

ตารางที่ 3.3 แสดงหลักเกณฑ์ของการใช้วรรคเล็กที่ได้จากการสังเกต

หลักเกฑณ์ที่ได้จากการสังเกต	**การนำมาใช้**	**การปรากฏ (%)**	**หมายเหตุ**
1. เว้นวรรคเล็กหลังคำว่า นาย, นาง, นางสาว, คุณ, ครู, อาจารย์	✓	0.051	ตามหลักไวยากรณ์หลังคำเหล่านี้ตามด้วยชื่อโดยไม่ต้องเว้นวรรค เช่น นายสมใจนึก ครูลัดดาวัลย์ เป็นต้น
2. เว้นวรรคเล็กก่อนและหลังเครื่องหมายทางคณิตศาสตร์	✓	6.638	-เครื่องหมาย “-” นับรวมที่ไม่ได้แสดงถึงการลบทางคณิตศาสตร์ด้วย
3. เว้นวรรคเล็กหลังคำว่า “เช่น” และ “ได้แก่” เพื่อยกตัวอย่างรายการ	✓	2.202	ไม่พบการใช้ทั้งสองคำที่ไม่แสดงดังการยกตัวอย่างรายการ
4. เว้นวรรคเล็กก่อนประพันธสรรพนาม (ที่, ซึ่ง,อัน) ซึ่งนำหน้าอนุประโยคเป็นประโยคความซ้อน	✓	2.612	ในฐานข้อมูลประโยคความซ้อนจัดเป็น 1 ประโยค
5. เว้นวรรคเล็กหน้าคำว่า “คือ”	✗	-	พบว่าอัตราส่วนของการเว้นวรรคหน้าคำว่า “คือ” ที่เป็นการวรรคเล็กต่อวรรคใหญ่เท่ากับ 2 : 1 ดังนั้นจึงไม่เหมาะสมที่จะใช้กฎข้อนี้
6. เว้นวรรคเล็กหลังเครื่องหมายอัศเจรีย์	✓	-	พบเพียง 1 ตำแหน่งเท่านั้น

จากตารางที่ 3.2 สรุปได้ว่า หลักไวยากรณ์ที่อ้างอิงจากราชบัณฑิยสถาน [14] ที่เกี่ยวข้องกับการเว้นวรรคเล็กทั้งหมด 23 ข้อสามารถนำมาใช้เพื่อเพิ่มความแม่นยำของการแบ่งประโยคภาษาไทยได้ 11 ข้อ คิดเป็น 40.528% ของการเว้นวรรคที่ไม่แสดงการแบ่งประโยค (วรรคเล็ก) และจากตารางที่ 3.3 สรุปได้ว่า หลักเกณฑ์ที่ได้จากการสังเกตการใช้ภาษาไทยที่สามารถนำมาใช้เพื่อเพิ่มความแม่นยำของการแบ่งประโยคภาษาไทยได้มีทั้งสิ้น 5 ข้อ คิดเป็น 11.503% ของการเว้นวรรคที่ไม่แสดงการแบ่งประโยค ดังนั้นสรุปได้ว่า หลักเกณฑ์ที่สามารถนำมาใช้ได้มีทั้งสิ้น 28 ข้อคิดเป็น 52.031% ของการเว้นวรรคที่ไม่แสดงการแบ่งประโยค และคิดเป็น 38.09% ของการเว้นวรรคทั้งหมด และมี 14 ตำแหน่งที่เกิดความผิดพลาดขึ้นเมื่อใช้กฎ คิดเป็น 0.02% ของการเว้นวรรคที่ไม่แสดงการแบ่งประโยค เนื่องจากการเว้นวรรคที่แสดงการสิ้นสุดของประโยค เห็นได้ว่าข้อยกเว้นที่เกิดขึ้นนั้นถือได้ว่าน้อยมากเมื่อเทียบกับผลลัพธ์ที่ดีขึ้น

1.4 การบูรณาการระเบียบวิธีทางสถิติร่วมกับหลักเกณฑ์การเว้นวรรค

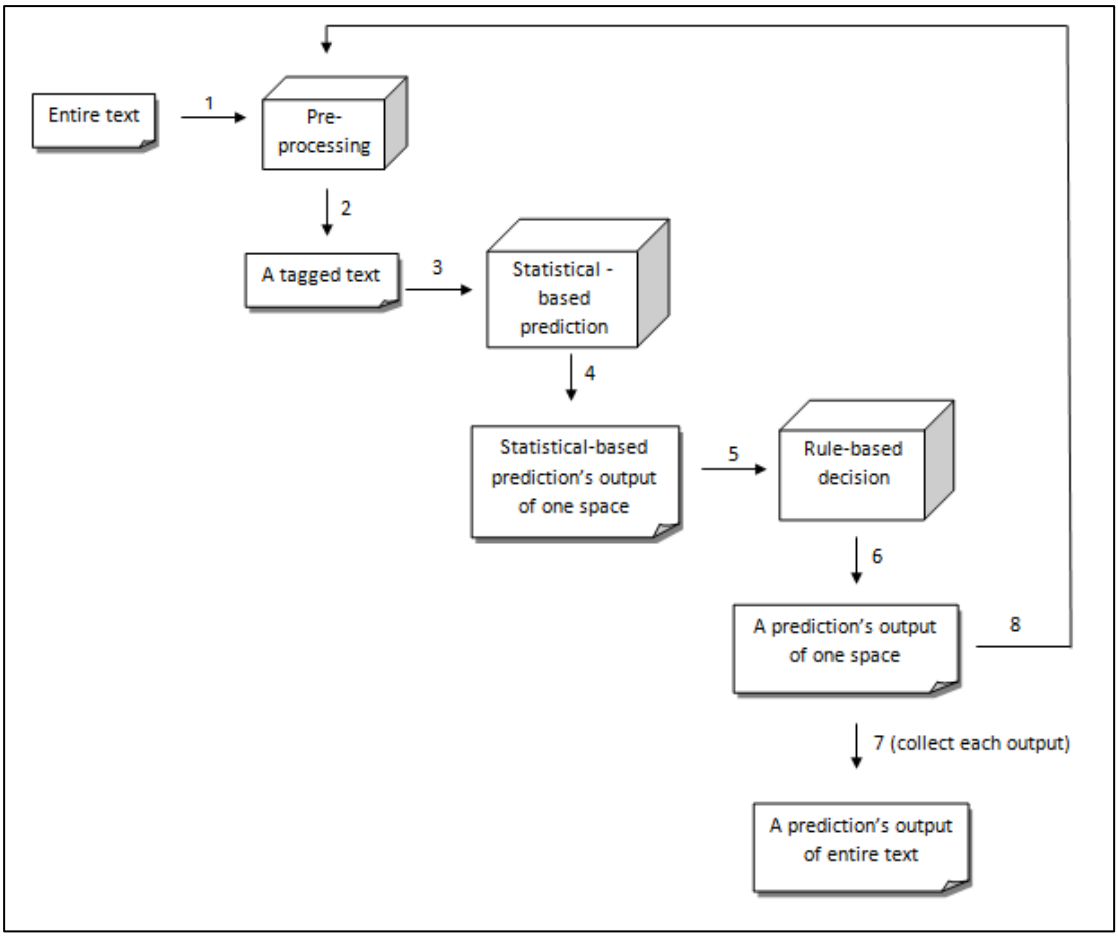
จากหัวข้อ 1.3 ที่แสดงให้เห็นว่าหลักเกณฑ์การเว้นวรรคตอนสามารถช่วยลดความผิดพลาดที่เกิดจากการแบ่งประโยคที่ผิดพลาดได้ ซึ่งผลลัพธ์ที่ได้จากระเบียบวิธีทางสถิติมีสองประเภทคือ เป็นการเว้นวรรคที่แสดงการแบ่งประโยคและการเว้นวรรคที่ไม่แสดงการแบ่งประโยค ดังนั้น หลักเกณฑ์ดังกล่าวจึงถูกประยุกต์ใช้กับผลลัพธ์ที่บ่งบอกถึงการเว้นวรรคที่แสดงการแบ่งประโยค หรืออาจกล่าวได้ว่า หลักเกณฑ์สามารถช่วยลดจำนวนของการแบ่งประโยคที่เกินได้ เมื่อพิจารณา รูปที่ 3.1 ซึ่งแสดงขั้นตอนโดยสังเขปของการแบ่งประโยคภาษาไทยที่นำเสนอพบว่า หากไม่มีการใช้ลักษณะสำคัญที่มีความเกี่ยวเนื่องกับผลของการตัดสินก่อนหน้าแล้ว กระบวนการที่นำเสนอจะยังคงมีประสิทธิภาพที่ดี เนื่องจากผลลัพธ์ทั้งหมดที่ได้จากระเบียบวิธีทางสถิติจะถูกนำไปประมวลผลต่อด้วยหลักเกณฑ์ดังตารางที่ 3.2 และตารางที่ 3.3 แต่หากมีการใช้ลักษณะสำคัญที่เกี่ยวเนื่องกับการตัดสินใจก่อนหน้า ผลการตัดสินใจที่ได้จากระเบียบวิธีทางสถิติจำเป็นต้องถูกนำไปประมวลผลต่อด้วยหลักเกณฑ์ เพื่อให้ได้ผลลัพธ์ที่ถูกต้องซึ่งจำเป็นต่อการนำไปใช้ในการประมวลผลต่อไป ดังนั้นเพื่อให้ทุกลักษณะสำคัญมีความถูกต้องจึงมีการตรวจสอบผลของความเกี่ยวเนื่องของลักษณะสำคัญกับผลลัพธ์จากการตัดสินก่อนหน้า ดังแสดงในตารางที่ 3.4

ตารางที่ 3.4 แสดงการความเกี่ยวข้องระหว่างลักษณะสำคัญและผลลัพธ์ก่อนหน้า

ลักษณะสำคัญ	**ความเกี่ยวข้องกับผลลัพธ์ก่อนหน้า**	**การแก้ไข**
1. ชนิดของคำ	✕	-
2. แคททิกอเรียลแกรมม่า	✕	-

ลักษณะสำคัญ	ความเกี่ยวข้องกับผลลัพธ์ก่อนหน้า	การแก้ไข
3. คำที่อยู่โดยรอบ	✗	-
4. จำนวนคำระหว่างการเว้นวรรคที่พิจารณากับการเว้นวรรคใกล้เคียง	✗	-
5. จำนวนคำระหว่างการเว้นวรรคที่กำลังพิจารณากับการเว้นวรรคก่อนหน้าที่เป็นการแบ่งประโยค	✓	ประมวลผลด้วยหลักเกณฑ์หลังจากได้ผลลัพธ์จากการประมวลผลด้วยระเบียบวิธีทางสถิติ เพื่อให้ได้ผลลัพธ์ที่ถูกต้องก่อนนำไปประมวลผลในขั้นตอนต่อไป
6. จำนวนคำระหว่างการเว้นวรรคที่กำลังพิจารณากับจุดสิ้นสุดของข้อความ (ในการทดลองคือจุดสิ้นสุดของย่อหน้า)	✗	-
7. การปรากฏของคำกริยาระหว่างการเว้นวรรคที่พิจารณากับการเว้นวรรคก่อนหน้าที่เป็นการแบ่งประโยค	✓	ประมวลผลด้วยหลักเกณฑ์หลังจากได้ผลลัพธ์จากการประมวลผลด้วยระเบียบวิธีทางสถิติ เพื่อให้ได้ผลลัพธ์ที่ถูกต้องก่อนนำไปประมวลผลในขั้นตอนต่อไป

จากตารางที่ 3.4 พบว่ามีลักษณะสำคัญสองข้อที่ต้องการความถูกต้องจากผลลัพธ์ก่อนหน้า คือ จำนวนคำระหว่างการเว้นวรรคที่กำลังพิจารณากับการเว้นวรรคก่อนหน้าที่เป็นการแบ่งประโยค และการปรากฏของคำกริยาระหว่างการเว้นวรรคที่พิจารณากับการเว้นวรรคก่อนหน้าที่เป็นการแบ่งประโยค ดังนั้นจึงมีปรับเปลี่ยนกระบวนการประมวลผลเล็กน้อย จากภาพที่ 3.1 เป็น ภาพที่ 3.4 เพื่อให้สอดคล้องกับการใช้ลักษณะสำคัญ



ภาพที่ 3.6 แสดงขั้นตอนโดยสังเขปของการแบ่งประโยคภาษาไทยที่สอดคล้องกับลักษณะสำคัญ

ภาพที่ 3.6 แสดงให้เห็นขั้นตอนของการแบ่งประโยคภาษาไทยที่แก้ไขแล้ว โดยเริ่มต้นข้อความทั้งหมดจะถูกประมวลผลด้วยขั้นตอนการเตรียมข้อมูลเบื้องต้น ผลที่ได้รับจากขั้นตอนนี้คือข้อความที่อยู่ในรูปแบบที่เหมาะสมต่อการประมวลผล ซึ่งขึ้นอยู่กับลักษณะสำคัญที่ต้องการใช้ โดยข้อความที่ได้จากขั้นตอนที่ 2 คือรูปแบบเฉพาะของการเว้นวรรคหนึ่งๆเท่านั้น สำหรับการประมวลผลจะใช้รูปแบบที่ใช้กันโดยทั่วไปคือจากซ้ายไปขวา ต่อจากนั้นจะเข้าสู่การทำนายด้วยระเบียบวิธีทางสถิติ ผลลัพธ์ที่ได้จากขั้นตอนที่ 4 คือผลลัพธ์ของหนึ่งการเว้นวรรค ต่อจากนั้นจะเข้าสู่การตัดสินใจด้วยหลักเกณฑ์ ผลลัพธ์ที่ได้จากขั้นตอนที่ 6 คือผลลัพธ์สุดท้ายของการเว้นวรรคที่พิจารณาซึ่งมีเพียงสองค่าคือ การเว้นวรรคนี้เป็นการเว้นวรรคที่แสดงถึงการจบประโยค (วรรคใหญ่) หรือไม่แสดงถึงการจบประโยค (วรรคเล็ก) ผลลัพธ์ที่ได้จะถูกส่งกลับไปที่ขั้นตอนของการเตรียมข้อมูลเพื่อเป็นสิ่งเข้าสำหรับการประมวลผลการเว้นวรรคต่อไป ทั้งนี้ผลลัพธ์ของแต่ละการเว้นวรรคจะถูกเก็บรวบรวมไว้จนได้ผลลัพธ์ครบทุกการเว้นวรรคที่ปรากฏในข้อความ

2. ข้อมูลที่ใช้ในการวิจัย

ข้อมูลที่จะใช้จำเป็นที่จะต้องมีการแบ่งคำ แบ่งประโยค ระบุชนิดของคำ และระบุประเภทของวากยสัมพันธ์ของแคททิกอเรียลแกรมม่า ดังนั้นจึงเลือกข้อความมาจาก Thai speech corpus for speech synthesis [31] ซึ่งมีคุณสมบัติดังกล่าว ลักษณะของข้อมูลแสดงดังภาพที่ 3.7

SEN:7
WRDTEXT:|ใน|โครงการ|ดังกล่าว|ต้อง|อาศัย|ความ|รู้|พื้นฐาน|ทาง|ด้าน|ซีเอ็นซี|ไฟฟ้า|และ|เครื่องกล|
CG:|pp/np|[np\np]/[np\np]|np\np|[np\np]\[np\np]|np|np/[s\np]|s\np/np|np|[np\np]/np|np|np|np|[np
\np]/np|np|
POS:|RPRE|NCMN|DDAC|XVMM|VSTA|FIXN|VSTA|NCMN|RPRE|NCMN|NCMN|NCMN|
JCRG|NCMN|
SEN:8
WRDTEXT:|ใน|ส่วน|ของ|การ|สร้าง|และ|พัฒนา|โปรแกรม|จะ|อ้างอิง|กับ|การ|ใช้
งาน|ที่|สะดวก|และ|คำสั่ง|มาตรฐานISO|
CG:|s\np|np|[np\np]/np|np/[s\np]|np|[np\np]/np|s\np/np|np|[s\np]/[s\np]|s\np/np|np|np/[s\np]|s\np/np|
np|np\np|[np\np]/np|np|np|
POS:|RPRE|NCMN|RPRE|FIXN|VACT|JCRG|VACT|NCMN|XVBM|VACT|RPRE|FIXN|VACT|
PREL|VATT|JCRG|NCMN|NCMN|
SEN:9
WRDTEXT:|ดังนั้น|ประโยชน์|ของ|การ|วิจัย|และ|พัฒนา|จะ|ได้|ต้นแบบ|ส่วนควบคุม|เครื่องกัด|แนว
ดิ่ง|ซีเอ็นซี|และ|ชุดฝึก|การ|เรียน|การ|สอน|เครื่องกัด|แนวดิ่ง|และ|เครื่องกลึง|ซีเอ็นซี|สำหรับ|กอง
วิทยาลัยเทคนิค| |กรมอาชีวศึกษา| |ที่|มี|ราคา|ถูก|กว่า|ต่างประเทศ|และ|ดูแลรักษา|ได้|ง่าย|
CG:|[s/s]/s|np|[np\np]/np|np/[s\np]|np|[np\np]/np|s\np|[s\np]/[s\np]|s\np/np|np|np|np|np|np|[np\np]/np|
np|np/[s\np]|np\np|np/[s\np]|s\np/np|np|np|[np\np]/np|np|np|[np\np]/np|np|np|np|np|[np\np]/[s\np]|s

ภาพที่ 3.7 แสดงตัวอย่างข้อมูลของฐานข้อมูล TSynC

3. เครื่องมือที่ใช้ในการวิจัย

1. Wagon CART Tool [32] เป็นเครื่องมือสำหรับใช้งานระเบียบวิธีเรียนรู้แบบ CART

Wagon CART Tool เป็นเครื่องมือหนึ่งใน Edinburgh Speech Tools Library ซึ่งเป็นเครื่องมือที่ใช้เฉพาะสำหรับการศึกษาวิจัยที่เกี่ยวข้องกับการประมวลผลภาษาธรรมชาติรวมถึงการประมวลผลข้อความ อีกทั้งเครื่องมือนี้ยังสามารถประยุกต์ใช้งานในงานด้านอื่นๆได้ เครื่องมือนี้ถูกออกแบบมาให้เหมาะสำหรับการใช้งานในการศึกษาวิจัยดังนั้นส่วนต่อประสาน

(Interface) จึงมีลักษณะที่เหมาะสำหรับการประมวลผลแบบกลุ่ม (Batch processing) เครื่องมือนี้มีคำสั่งในการใช้งานรวมถึงตัวเลือกสำหรับปรับแต่งค่าพารามิเตอร์ดังแสดงดังภาพ 3.8

```
      -desc ifile Field description file
      -data ifile Datafile, one vector per line
      -stop int" {50}" Minimum number of examples for leaf nodes
       -test ifile Datafile to test tree on
        -frs float" {10}" Float range split, number of partitions to split a float feature range into
      -dlist Build a decision list (rather than tree)
     -dtree Build a decision tree (rather than list) default
    -output ofile
         -o ofile File to save output tree in
-distmatrix ifile A distance matrix for clustering
     -quiet No questions printed during building
   -verbose Lost of information printing during build
 -predictee string name of field to predict (default is first field)
    -ignore string Filename or bracket list of fields to ignore
 -stepwise Incrementally find best features
   -swlimit float" {0.0}" Percentage necessary improvement for stepwise
     -swopt string Parameter to optimize for stepwise, for classification options are correct or entropy for regression options
            are rmse or correlation correct and correlation are the defaults
  -balance float For derived stop size, if dataset at node, divided by balance is greater than stop it is used as stop if balance is
            0 (default) always use stop as is.
 -held_out int Percent to hold out for pruning
      -heap int" {210000}" Set size of Lisp heap, should not normally need to be changed from its default, only with *very*
            large description files (> 1M)
 -noprune No (same class) pruning required
```

ภาพที่ 3.8 แสดงค่าพารามิเตอร์ของ Wagon CART tool (อ้างอิงจาก [32])

2. AutoIT Script Tools [33]

AutoIT Script Tools คือโปรแกรมสำหรับใช้ในสร้างบทคำสั่ง (Script) ซึ่งอยู่ในรูปแบบของภาษาทางคอมพิวเตอร์ที่เรียกว่า AutoIT ซึ่งมีลักษณะที่ไม่แตกต่างจากบทคำสั่งอื่นๆ เช่น VB หรือ Python แต่ด้วยเหตุผลที่ AutoIT มีความยืดหยุ่นในการใช้งานและประมวลผลที่สูงอีกทั้งยังใช้ทรัพยากรและเวลาในการประมวลผลที่ต่ำ นอกจากนี้ยังถูกออกแบบมาให้มีคำสั่งที่ทำงานได้ดีกับสิ่งเข้าที่อยู่ในรูปของข้อความรวมทั้งมีคำสั่งที่เกี่ยวข้องกับนิพจน์ทั่วไป (Regular expression) ซึ่งจำเป็นอย่างมากสำหรับการประมวลผลหลักเกณฑ์ ตัวอย่างของคำสั่ง AutoIT ที่ใช้งานมากในการประมวลผลหลักเกณฑ์ แสดงดังรูปที่ 3.9

Function Reference

StringRegExp

Check if a string fits a given regular expression pattern.

```
StringRegExp ( "test", "pattern" [, flag [, offset]] )
```

Parameters

test	The string to check
pattern	The regular expression to compare.
flag	**[optional]** A number to indicate how the function behaves. See below for details. The default is 0.
offset	**[optional]** The string position to start the match (starts at 1) The default is 1.

Flag	**Values**
0	Returns 1 (matched) or 0 (no match)
1	Return array of matches.
2	Return array of matches including the full match (Perl / PHP style).
3	Return array of global matches.
4	Return an array of arrays containing global matches including the full match (Perl / PHP style).

Character Classes

[:alnum:]	letters and digits
[:alpha:]	letters
[:ascii:]	character codes 0 - 127
[:blank:]	space or tab only
[:cntrl:]	control characters
[:digit:]	decimal digits (same as \d)
[:graph:]	printing characters, excluding space
[:lower:]	lower case letters
[:print:]	printing characters, including space
[:punct:]	printing characters, excluding letters and digits
[:space:]	white space (not quite the same as \s, it include VT: chr(11))
[:upper:]	upper case letters
[:word:]	"word" characters (same as \w)
[:xdigit:]	hexadecimal digits

ภาพที่ 3.9 แสดงคำสั่งการทำงานของ AutoIT Script

4 การวัดผล

งานวิจัยนี้ใช้หลักเกณฑ์การวัดผลโดยอ้างอิงจากการศึกษาที่เกี่ยวข้องกับการแบ่งกลุ่มวลีในภาษาอังกฤษ [34] ซึ่งงานวิจัยด้านการแบ่งประโยคภาษาไทยที่ผ่านมา [9 - 11] ได้เลือกใช้

หลักเกณฑ์นี้ในการวัดผลโดยตลอด หลักเกณฑ์ดังกล่าวประกอบด้วยการวัดผลใน 3 ดัชนีวัดผล คือ Sentence-break-correct, Space-correct และ False-break

$$\text{Sentence-break-recall} = (CB / RB) \times 100\% \quad (1)$$

$$\text{Space-correct} = (CS / RS) \times 100\% \quad (2)$$

$$\text{False-break} = (FB / RS) \times 100\% \quad (3)$$

โดยที่

- CB คือ จำนวนของการแบ่งประโยคที่ถูกต้องของข้อมูลทดสอบ
- FB คือ จำนวนของการแบ่งประโยคที่ผิดพลาดของข้อมูลทดสอบ
- CS คือ จำนวนของการแบ่งประโยคและไม่แบ่งประโยคที่ถูกต้องของข้อมูลทดสอบ
- RB คือ จำนวนของการแบ่งประโยคในข้อมูลอ้างอิง
- RS คือ จำนวนของการแบ่งประโยคและไม่แบ่งประโยคในข้อมูลอ้างอิง

เพื่อความเป็นระเบียบในการแสดงผลในตาราง ต่อไปจะแทน Sentence-break-recall ด้วย sb-recall

บทที่ 4

การทดลองและอภิปรายผล

การทดลองถูกออกแบบตามจุดประสงค์ของการศึกษาคือ ศึกษาผลของลักษณะสำคัญที่มีต่อการแบ่งประโยคภาษาไทย และผลของกฎซึ่งได้จากหลักเกณฑ์ไวยากรณ์ที่มีผลต่อการเพิ่มความถูกต้องแม่นยำของการแบ่งประโยคภาษาไทย ดังนั้นการทดลองจึงถูกแบ่งออกเป็นสองส่วนหลักคือ การทดลองเพื่อศึกษาลักษณะสำคัญและการทดลองเพื่อศึกษาผลของการใช้กฎ โดยแบ่งข้อมูลออกเป็นสองส่วน ส่วนแรกใช้เป็นข้อมูลฝึกฝน และอีกส่วนใช้เป็นข้อมูลสำหรับการวัดผลลัพธ์ที่ได้จากการฝึกฝนด้วยระเบียบวิธีต่าง ๆ ที่เสนอ

1. การทดลองเพื่อศึกษาลักษณะสำคัญ

การวิจัยนี้ศึกษาลักษณะสำคัญทั้งสิ้น 7 รายการ และใช้อักษรย่อแทนแต่ละลักษณะสำคัญเพื่อความเหมาะสมสำหรับการแสดงผลในตาราง ดังแสดงในตารางที่ 4.1

ตารางที่ 4.1 แสดงลักษณะสำคัญที่ใช้ทั้งหมดในการทดลองและอักษรย่อที่ใช้แทน

ลักษณะสำคัญ	อักษรย่อ
1. ชนิดของคำ	POS
2. แคททิกอเรียลแกรมม่า	CG
3. คำที่อยู่โดยรอบ	Word
4. จำนวนคำระหว่างการเว้นวรรคที่พิจารณากับการเว้นวรรคใกล้เคียง	NWrd
5. จำนวนคำระหว่างการเว้นวรรคที่กำลังพิจารณากับการเว้นวรรคก่อนหน้าที่เป็นการแบ่งประโยค	NWrd_SB
6. จำนวนคำระหว่างการเว้นวรรคที่กำลังพิจารณากับจุดสิ้นสุดของข้อความ (ในการทดลองคือจุดสิ้นสุดของย่อหน้า)	NWrd_End
7. การปรากฏของคำกริยาระหว่างการเว้นวรรคที่พิจารณากับการเว้นวรรคก่อนหน้าที่เป็นการแบ่งประโยค	V

จากตารางที่ 4.1 สามารถแบ่งลักษณะสำคัญได้เป็น 2 กลุ่มหลักคือ

1. ลักษณะสำคัญที่เป็นลักษณะสำคัญหลัก ประกอบด้วย ชนิดของคำ, แคททิกอเรียลแกรมม่า, คำที่อยู่โดยรอบ

2. ลักษณะสำคัญที่เป็นลักษณะสำคัญเสริม ทำหน้าที่ช่วยแสดงการปรากฏของภาษาให้ครอบคลุมมากยิ่งขึ้น ประกอบด้วย จำนวนคำระหว่างการเว้นวรรคที่กำลังพิจารณากับการเว้นวรรคก่อนหน้าที่เป็นการแบ่งประโยค จำนวนคำระหว่างการเว้นวรรคที่กำลังพิจารณากับจุดสิ้นสุดของข้อความ (ในการทดลองคือจุดสิ้นสุดของย่อหน้า) จำนวนคำระหว่างการเว้นวรรคที่พิจารณากับการเว้นวรรคใกล้เคียง และการปรากฏของคำกริยาระหว่างการเว้นวรรคที่พิจารณากับการเว้นวรรคก่อนหน้าที่เป็นการแบ่งประโยค

ดังนั้นในการทดลองเพื่อศึกษาหาลักษณะสำคัญที่ช่วยเพิ่มความแม่นยำให้กับการแบ่งวรรคตอนภาษาไทยจึงแบ่งเป็น 2 ส่วนตามการแบ่งกลุ่มข้างต้น

1.1 การทดลองที่ 1 ศึกษาเปรียบเทียบผลของลักษณะสำคัญหลัก

การทดลองนี้ ศึกษาเปรียบเทียบลักษณะสำคัญทั้งสิ้น 3 ลักษณะคือ ชนิดของคำ, แคททิกอเรียลแกรมม่า และคำที่อยู่โดยรอบ โดยลักษณะสำคัญหนึ่งที่มีในทุกชุดทดสอบคือ จำนวนคำระหว่างการเว้นวรรคที่พิจารณากับการเว้นวรรคใกล้เคียง เนื่องจากลักษณะสำคัญนี้ปรากฏแน่ชัดว่าช่วยส่งผลให้เกิดความแม่นยำเพิ่มมากขึ้นและสามารถคาดเดารูปแบบการเกิดขึ้นที่จำเพาะในหลายๆ ประเภทได้ ดังรายละเอียดในบทที่ 3 นอกจากลักษณะสำคัญทั้ง 3 แล้ว การทดลองนี้ยังนำเอาลักษณะสำคัญต่างๆมารวมกันเพื่อดูผลลัพธ์ที่เกิดขึ้น ภายใต้สมมติฐานว่า แคททิกอเรียลแกรมม่ามีแนวโน้มที่จะเป็นลักษณะสำคัญหลักที่ส่งผลดีต่อการแบ่งประโยคภาษาไทย ดังนั้นจึงมีชุดลักษณะสำคัญ (Feature set) ทั้งสิ้น 6 ชุด และการผลการทดลองแสดงดังตารางที่ 4.2 โดยตั้งค่าการทดลองดังนี้

- ทำการทดลองลองในลักษณะของการทวนสอบข้าม 10 รอบ บนฐานข้อมูล TSynC (10 fold cross-validation) หมายความว่า ข้อมูลที่ใช้ถูกแบ่งออกเป็น 10 ส่วน โดยแต่ละส่วนจะมีจำนวนย่อหน้าที่ใกล้เคียงกัน ทำการทดลองทั้งสิ้น 10 ครั้ง โดยแต่ละครั้ง เลือกข้อมูล 9 ส่วนเป็นข้อมูลเรียนรู้ (Learning data) และข้อมูลอีก 1 ส่วนเป็นข้อมูลทดสอบ โดยในแต่ละครั้งข้อมูลทดสอบต้องไม่ซ้ำกับข้อมูลทดสอบที่ใช้ไปแล้ว

- นำผลที่ได้แต่ละครั้งไปคำนวณตามดัชนีความแม่นยำ
- หาค่าเฉลี่ยของทั้ง 10 ครั้งและเปรียบเทียบ

ผลการทดลองที่ 1

ตารางที่ 4.2 แสดงผลการทดลองเปรียบเทียบผลของลักษณะสำคัญหลักต่อการแบ่งประโยค

ชุดลักษณะสำคัญ	Test 1 (%)	Test 2 (%)	Test 3 (%)	Test 4 (%)	Test 5 (%)	Test 6 (%)
ลักษณะสำคัญ	Word NWrd	POS NWrd	POS Word NWrd	CG NWrd	CG Word NWrd	CG POS NWrd
sb-recall	69.34	76.84	79.04	**83.54**	83.08	76.14
space-correct	72.13	88.58	90.24	**91.71**	90.62	88.54
false-break	8.03	3.79	1.94	1.88	**1.85**	3.83

วิเคราะห์ผลการทดลองที่ 1

จากผลการทดลองในตารางที่ 4.2 Test 1 ให้ผลลัพธ์ความแม่นยำที่ต่ำที่สุดในทุกดัชนีวัดผล มีสาเหตุมาจาก คำที่อยู่โดยรอบการเว้นวรรคตอนไม่สามารถแสดงบริบทของการใช้การเว้นวรรคได้อย่างคลอบคลุม อันเนื่องมาจากข้อมูลที่ใช้เป็นข้อมูลเรียนรู้มีจำนวนที่ไม่มากพอที่จะแสดงการเกิดขึ้นของคำซึ่งมีหลากหลายมากในภาษาและการเว้นวรรคตอนในภาษาไทยได้ครอบคลุมทุกกรณี ผลลัพธ์เช่นนี้เป็นสิ่งที่สามารถคาดเดาได้หากมีการใช้ลักษณะสำคัญที่ต้องการข้อมูลเรียนรู้จำนวนมากแต่เมื่อใช้งานจริงแล้วมีข้อมูลเรียนรู้ไม่เพียงพอที่จะครอบคลุมการเกิดขึ้นของคำได้ทั้งหมด แต่อย่างไรก็ดี การใช้คำที่เกิดขึ้นโดยรอบเป็นลักษณะสำคัญมีแนวโน้มที่จะให้ผลลัพธ์ที่ดีขึ้นหากมีข้อมูลเรียนรู้จำนวนมากและครอบคลุมคำศัพท์และกรณีต่างๆที่เกิดขึ้น ดังนั้นในการประมวลผลที่มีข้อมูลทดสอบจำกัดแล้ว การเลือกใช้การกำกับคำด้วยสัญลักษณ์ที่แทนความหมายเชิงการไวยากรณ์หรือแสดงลักษณะโดยทั่วไปของคำนั้นๆ จะส่งผลให้การประมวลผลมีความถูกต้องแม่นยำมากขึ้น ดังเห็นได้จาก Test 2 – Test 6 ซึ่งใช้ลักษณะสำคัญที่เป็นการแสดงความหมายเชิงไวยากรณ์ของคำ

เมื่อเปรียบเทียบผลระหว่าง Test 2 และ Test 3 พบว่า การใช้ชนิดของคำร่วมกับคำที่อยู่โดยรอบเป็นลักษณะสำคัญให้ผลลัพธ์ที่ดีกว่าการใช้ชนิดของคำในทุกดัชนีวัดผล เหตุผลที่เป็นเช่นนี้อาจเนื่องมาจาก ในกระบวนการของระเบียบวิธีเรียนรู้แบบ CART ซึ่งเป็นต้นไม้ตัดสินใจแบบทวิภาค โดยส่วนใหญ่แล้วลักษณะสำคัญที่มีความถี่จากการเรียนรู้มากจะถูกพิจารณาก่อน ต่อจากนั้นจึงเป็นลักษณะสำคัญที่มีความถี่จากการเรียนรู้ที่น้อยกว่าหรือแทบไม่พบ จึงทำให้ Test 3 มีลักษณะที่ใช้การตัดสินใจโดยใช้ชนิดของคำก่อนแล้วจึงตัดสินใจต่อด้วยการเรียนรู้จดจำจากคำที่ปรากฏซึ่งช่วยเพิ่มการตัดสินใจให้มีความถูกต้องมากยิ่งขึ้น

หากดูโดยภาพรวมแล้วพบว่าการใช้แคททิกอเรียลแกรมม่าเป็นลักษณะสำคัญหลักให้ผลลัพธ์ที่ดีที่สุดในการทดลองดังเห็นได้จาก ค่าดัชนี sentence-break-recall และ space-correct ที่มีค่าสูงที่สุดในขณะที่ค่าความผิดพลาดอยู่ในระดับที่ต่ำ เนื่องมาจาก แคททิกอเรียลแกรมม่าถูกสร้างขึ้นเพื่อแสดงรูปแบบการใช้คำในภาษาให้มีความครอบคลุมมากที่สุด โดยมีการนิยามประเภทของแคททิกอเรียลแกรมม่าไว้ทั้งสิ้น 120 ประเภทเมื่อเทียบกับชนิดของคำที่มีเพียง 47 ประเภท ไม่เพียงแต่ความครอบคลุมรูปแบบการเกิดขึ้นของภาษาเท่านั้นแคททิกอเรียลแกรมม่ายังถูกออกแบบมาเพื่อให้แสดงความเชื่อมโยงของคำในบริบท

เมื่อพิจารณา Test 5 ซึ่งใช้แคททิกอเรียลแกรมม่าร่วมกับคำที่อยู่โดยรอบ ผลของการแบ่งประโยคมีความแม่นยำลดลงเล็กน้อยเมื่อพิจารณาจากดัชนี space-correct ซึ่งแสดงความแม่นยำโดยรวมของระบบ อาจเกิดเนื่องจาก แคททิกอเรียลแกรมม่าเองสามารถแสดงรูปแบบที่เกิดขึ้นของการเว้นวรรคได้ครอบคลุมเป็นส่วนใหญ่แล้ว ดังนั้นเมื่อใช้ร่วมกับคำที่อยู่โดยรอบซึ่งมีความหลากหลายของการเกิดขึ้นที่มากจึงส่งผลให้เกิดการกระจายตัวของลักษณะสำคัญที่ใช้ในการเรียนรู้และส่งผลต่อความแม่นยำที่ลดลง

เมื่อพิจารณา Test 6 ซึ่งใช้แคททิกอเรียลแกรม่าร่วมกับชนิดของคำเป็นลักษณะสำคัญ พบว่า ความแม่นยำโดยรวมแล้วลดลงอย่างชัดเจน อาจด้วยเหตุผลของความซ้ำซ้อนของการใช้ลักษณะสำคัญ ทั้งแคททิกอเรียลแกรมม่าและชนิดของคำล้วนแต่ถูกออกแบบมาเพื่อแสดงหลักไวยากรณ์ของภาษา

จากการทดลองนี้สรุปได้ว่า การใช้แคททิกอเรียลแกรมม่าเป็นลักษณะสำคัญหลักร่วมกับจำนวนคำระหว่างการเว้นวรรคที่พิจารณากับการเว้นวรรคใกล้เคียงให้ผลลัพธ์ที่ดีที่สุดในการแบ่งประโยคภาษาไทย

หากพิจารณาข้อผิดพลาดจากข้อความที่ได้จากการแบ่งประโยคพบว่า ข้อผิดพลาดส่วนใหญ่เกิดขึ้นใกล้กับคำเชื่อม (เชื่อมคำและเชื่อมประโยค) และข้อผิดพลาดมักเกิดมากขึ้นเมื่อใช้ชนิดของคำเป็นลักษณะสำคัญหลักเมื่อเทียบกับการใช้แคททิกอเรียลแกรมม่า ยกตัวอย่างเช่น “JSBR” ถูกใช้กำกับทุกคำเชื่อมที่แสดงการเชื่อมอนุประโยค แต่เมื่อพิจารณาแล้วแต่ละคำเชื่อมไม่ได้แสดงการใช้งานที่เหมือนกัน ยกตัวอย่างเช่นคำว่า “ว่า” ซึ่งใช้เชื่อมประโยคและอนุประโยคที่ตามมา โดยอนุประโยคสามารถมีได้หลายลักษณะ เช่น ประโยคความซ้อน อนุประโยคที่มีใจความสมบูรณ์นามวลี หรือกริยาวลี ในขณะที่แคททิกอเรียลแกรมม่ามีการนิยามคำเชื่อมที่แสดงลักษณะการเกิดขึ้นที่แตกต่างกันในแต่ละประเภทของอนุประโยค ดังนั้นจึงช่วยให้การแสดงกำกับคำมีความใกล้เคียงกับลักษณะที่ปรากฏของการใช้ภาษา

1.2 การทดลองที่ 2 ศึกษาเปรียบเทียบผลของลักษณะสำคัญเสริม

การทดลองนี้ ศึกษาเปรียบเทียบลักษณะสำคัญทั้งสิ้น 3 ลักษณะคือ จำนวนคำระหว่างการเว้นวรรคที่กำลังพิจารณากับการเว้นวรรคก่อนหน้าที่เป็นการแบ่งประโยค จำนวนคำระหว่างการเว้นวรรคที่กำลังพิจารณากับจุดสิ้นสุดของข้อความ การปรากฏของคำกริยาระหว่างการเว้นวรรคที่พิจารณากับการเว้นวรรคก่อนหน้าที่เป็นการแบ่งประโยค เพื่อเปรียบเทียบดูผลของการใช้ลักษณะสำคัญเสริมเทียบกับการใช้เพียงลักษณะสำคัญหลัก ดังนั้นการทดลองจึงมีชุดลักษณะสำคัญ ทั้งสิ้น 4 ชุด และการผลการทดลองแสดงดังตารางที่ 4.3 โดยตั้งค่าการทดลองดังนี้

- ใช้ผลลัพธ์ที่ดีที่สุดจากการทดลองเพื่อหาลักษณะสำคัญหลักเป็นเส้นหลักล่าง
- ทำลองในลักษณะของ 10 fold cross-validation บนฐานข้อมูล TSynC หมายความว่า ข้อมูลที่ใช้ถูกแบ่งออกเป็น 10 ส่วน โดยแต่ละส่วนจะมีจำนวนย่อหน้าที่ใกล้เคียงกัน ทำการทดลองทั้งสิ้น 10 ครั้ง โดยแต่ละครั้ง เลือกข้อมูล 9 ส่วนเป็นข้อมูลเรียนรู้ (Learning data) และข้อมูลอีก 1 ส่วนเป็นข้อมูลทดสอบ โดยในแต่ละครั้งข้อมุลทดสอบต้องไม่ซ้ำกับข้อมูลทดสอบที่ใช้ไปแล้ว
- นำผลที่ได้แต่ละครั้งไปคำนวณตามดัชนีความแม่นยำ

- หาค่าเฉลี่ยของทั้ง 10 ครั้งและเปรียบเทียบ

ผลการทดลองที่ 2

ตารางที่ 4.3 แสดงลการทดลองเปรียบเทียบผลของลักษณะสำคัญเสริมต่อการแบ่งประโยค

ชุดลักษณะสำคัญ	Baseline (%)	Test 1 (%)	Test 2(%)	Test 3(%)
ลักษณะสำคัญ	CG Nwrd	CG NWrd NWrd_SB	CG NWrd NWrd_End	CG NWrd V
sb-recall	83.54	84.11	**91.14**	84.48
space-correct	91.71	**92.36**	79.95	91.14
false-break	**1.88**	4.16	7.51	2.01

วิเคราะห์ผลการทดลองที่ 2

จากผลการทดลองดังแสดงในตารางที่ 4.3 Test 1 มีความแม่นยำโดยรวมที่สูงที่สุดเมื่อวัดจากดัชนีวัดผล space-correct ซึ่งแสดงภาพรวมความถูกต้องเนื่องจากดูทั้งการแบ่งวรรคตอนที่เป็นการแสดงการจบประโยคและไม่เป็นการแสดงการจบประโยค จากผลการทดลองของ Test 1 เทียบกับ Baseline แสดงให้เห็นว่าถึงแม้โดยรวมจะมีความแม่นยำมากขึ้น แต่หากพิจารณาเฉพาะ false-break การเพิ่มลักษณะสำคัญที่แสดงจำนวนคำระหว่างการเว้นวรรคที่กำลังพิจารณากับการเว้นวรรคก่อนหน้าที่เป็นการแบ่งประโยคส่งผลให้เกิดความผิดพลาดของการแบ่งมากขึ้นอย่างชัดเจน อาจเนื่องจาก ลักษณะสำคัญดังกล่าวมีแนวโน้มที่จะแบ่งประโยคให้มีความยาวที่ไม่มาก ดังนั้นประโยคความที่ซับซ้อนและยาวจึงมีความเป็นเป็นไปได้สูงที่จะเกิดความผิดพลาดขึ้น เหตุผลที่เป็นเช่นนี้เนื่องจาก ข้อมูลส่วนใหญ่มีความยาวของประโยคอยู่ในช่วงที่ไม่ยาวมาก ดังรายละเอียดที่แสดงในบทที่ 3

เมื่อพิจารณาการ Test 2 ซึ่งใช้ลักษณะสำคัญที่แสดงจำนวนคำระหว่างการเว้นวรรคที่กำลังพิจารณากับจุดสิ้นสุดประโยคเป็นลักษณะสำคัญเสริม พบว่าเกิดความผิดพลาดที่สูงมากและส่งผลให้ความถูกต้องโดยรวมน้อยกว่า Baseline มาก ซึ่งผิดไปจากสมมติฐานที่ตั้งไว้ว่าลักษณะสำคัญนี้

น่าจะช่วยให้การพิจารณาการแบ่งประโยคบริเวณที่ใกล้กับจุดสิ้นสุดข้อความมีความแม่นยำมากขึ้น เมื่อวิเคราะห์ข้อมูลที่ผ่านการประมวลผลเบื้องต้นให้อยู่ในรูปแบบสำหรับการประมวลผลด้วยระเบียบวิธีทางสถิติแล้วพบว่า ลักษณะสำคัญดังกล่าวมีค่าอยู่ในช่วงที่ที่กว้างมากคือ 1 – 67 และมีการกระจายตัวที่สูงมากและขาดรูปแบบซึ่งแตกต่างจากลักษณะเฉพาะที่แสดงความยาวของประโยคซึ่งมีการกระจายตัวในลักษณะที่เกาะกลุ่ม ด้วยเหตุผลนี้เองจึงทำให้ Test 2 มีความถูกต้องน้อยลงเมื่อเทียบกับ Baseline

เมื่อพิจารณาผลของลักษณะสำคัญที่แสดงการปรากฏของคำกริยาระหว่างการเว้นวรรคที่พิจารณากับการเว้นวรรคก่อนหน้าที่เป็นการแบ่งประโยคดังแสดงใน Test3 เปรียบเทียบกับ Baseline พบว่า ความถูกต้องลดลงเล็กน้อยในทุกดัชนีการวัดผล เหตุผลที่เป็นเช่นนี้อาจเกิดจากการที่ประโยคภาษาไทยมีความซับซ้อนและในหนึ่งประโยคมักมีคำกริยาปรากฏมากกว่า 1 คำและยิ่งไปกว่านั้นการใช้คำในภาษาไทยส่วนใหญ่มักใช้คำนามที่เกิดจากการผสมระหว่างคำนามกับคำกริยาจึงอาจก่อให้เกิดความกำกวมเมื่อประมวลผล ยกตัวอย่างเช่น คำว่า “ธุรกิจนำเข้า” ในฐานข้อมูลถูกแบ่งออกเป็น 3 คำคือ “|ธุรกิจ|นำ|เข้า|” ซึ่งกำกับชนิดของคำเป็น คำนาม คำกริยา และคำกริยาตามลำดับ แต่หากพิจารณาตามความหมายแล้วพบว่า “ธุรกิจนำเข้า” คือคำนามหนึ่งคำ สืบเนื่องจากเหตุผลข้างต้นจึงทำให้การใช้ลักษณะสำคัญนี้ให้ผลลัพธ์ที่มีความถูกต้องน้อยลงเมื่อเทียบกับการไม่ใช้ลักษณะสำคัญดังกล่าว

2. การทดลองเพื่อศึกษาเปรียบเทียบผลของการใช้กฎ

การทดลองที่ 3 ศึกษาเปรียบเทียบผลของการใช้กฎ

การทดลองนี้ศึกษาผลของการใช้กฎที่มีผลต่อการลดความผิดพลาดของการแบ่งประโยคด้วยระเบียบวิธีทางสถิติ โดยตั้งค่าการทดลองดังนี้

- ตั้งค่าการทดลองส่วนของการแบ่งประโยคโดยระเบียบวิธีทางสถิติเช่นเดียวกับการทดลองที่ 2
- นำผลลัพธ์ที่ได้จากระเบียบวิธีทางสถิติมาตรวจสอบด้วยกฎดังแสดงในตารางที่ 3.3 และตารางที่ 3.4
- นำผลลัพธ์ที่ได้ไปคำนวณตามดัชนีความแม่นยำ

ผลการทดลองที่ 3

ตารางที่ 4.4 แสดงผลลัพธ์ที่ได้จากการใช้กฎเทียบกับการใช้เฉพาะระเบียบวิธีทางสถิติ

ชุดลักษณะสำคัญ	Test 1 (%)	Test 2 (%)	Test 3 (%)	Test 4 (%)	Test 5 (%)	Test 6 (%)	Test 7 (%)	Test 8 (%)
ลักษณะสำคัญ	CG Nwrd	CG Nwrd	CG NWrd NWrd_SB	CG NWrd NWrd_SB	CG NWrd NWrd_End	CG NWrd NWrd_End	CG NWrd V	CG NWrd V
การใช้กฎ	×	✓	×	✓	×	✓	×	✓
sb-recall	83.54	83.54	84.11	84.11	**91.14**	**91.14**	84.48	84.48
space-correct	91.71	91.77	92.36	**93.54**	79.95	84.72	91.14	91.23
false-break	1.88	**1.83**	4.16	2.99	7.51	2.74	2.01	1.92

วิเคราะห์ผลการทดลองที่ 3

ผลการทดลองที่ 3 พบว่า ทุกชุดลักษณะสำคัญที่มีการใช้กฎร่วมด้วยมีความถูกต้องมากขึ้นเมื่อเทียบกับการใช้เพียงระเบียบวิธีทางสถิติ แต่กฎที่นำมาใช้ส่งผลมากหรือน้อยแตกต่างกันต่อแต่ละชุดลักษณะสำคัญ เนื่องมาจาก การใช้กฎคือการลดข้อผิดพลาดในลักษณะของการแบ่งประโยคเกิน (false-break) ที่เกิดจากการแบ่งประโยคด้วยระเบียบวิธีทางสถิติ ดังนั้นหากผลลัพธ์จากระเบียบวิธีทางสถิติมีความผิดพลาดที่น้อยหรือความผิดพลาดที่เกิดขึ้นไม่ตรงกับกฎ ผลของการใช้กฎจะไม่ปรากฏเด่นชัด Test 2 และ Test 8 เทียบกับ Test 1 และ Test 7 ตามลำดับ แต่หากผลลัพธ์ที่ได้จากระเบียบวิธีทางสถิติมีความผิดพลาดมาก กฎจะสามารถช่วยลดความผิดพลาดที่เกิดขึ้นได้มากด้วยเช่นกัน ทั้งนี้ขึ้นอยู่กับชนิดของความผิดพลาดที่เกิดขึ้น ดังตัวอย่างเช่น Test 3 เทียบกับ Test 4 และ Test 5 เทียบกับ Test 6

เมื่อพิจารณาจากตารางที่ 4.4 ซึ่งแสดงผลการศึกษาที่รวมผลลัพธ์ของการศึกษาลักษณะสำคัญหลัก (การทดลองที่ 1) ลักษณะสำคัญเสริม (การทดลองที่ 2) และการใช้กฎเพื่อเพิ่มความถูกต้อง สามารถสรุปได้ว่า หากพิจารณาด้วยดัชนี space-correct ซึ่งแสดงความถูกต้องโดยรวมของการแบ่งประโยค การใช้แคททิกอเรียลแกรมม่า จำนวนคำระหว่างการเว้นวรรคที่พิจารณากับการเว้นวรรคใกล้เคียง จำนวนคำระหว่างการเว้นวรรคที่กำลังพิจารณากับการเว้นวรรคก่อนเป็นลักษณะ

สำคัญร่วมกับการใช้กฎให้ผลลัพธ์ที่ดีที่สุด คือ space-correct เท่ากับ 93.54% sentence-break-recall เท่ากับ 84.11% และ false-break เท่ากับ 2.99% และเมื่อเปรียบเทียบกับผลการศึกษาที่ผ่านมาซึ่งใช้ข้อมูลที่มีเนื้อความและมีรูปแบบการทดลองเช่นเดียวกับกับงานวิจัยนี้พบว่า ลักษณะสำคัญและกฎที่เสนอในงานวิจัยนี้ให้ผลลัพธ์ที่ดีกว่าการศึกษาในอดีต ดังแสดงในตารางที่ 4.5

ตารางที่ 4.5 แสดงความถูกต้องของการแบ่งประโยคด้วยวิธีที่เสนอเทียบกับการศึกษาในอดีต

	Result of [9]	Result of [10]	Result of [11]	This research 1	This research 2
Learning algorithm	Trigram	Winnow	Maximum Entropy	CART	CART
Feature	POS	POS Word	Word NWrd	CG NWrd	CG NWrd NWrd_SB Rules
sb-recall	79.82	77.27	83.5	83.54	**84.11**
space-correct	85.26	89.13	91.19	91.71	**93.54**
false-break	8.75	**1.74**	3.91	1.88	2.99

ตารางที่ 4.5 แสดงให้เห็นว่าการใช้ระเบียบวิธีทางสถิติแบบ CART โดยมีลักษณะสำคัญที่เหมาะสมร่วมกับกฎให้ความถูกต้องที่สูงกว่างานวิจัยที่ผ่านมาทั้งสองดัชนีวัดผลคือ sentence-break และ space-correct แต่เมื่อพิจารณาดัชนีวัดผล false-break แล้วงานวิจัยในอดีต [10] ยังคงมีค่าที่ต่ำมาเนื่องจากแนวทางการประมวลผลดังกล่าวมีแนวโน้มที่จะไม่แบ่งประโยค สังเกตได้จากค่าดัชนีวัดผล sentence-break-recall ที่ต่ำมากด้วยเช่นกัน

บทที่ 5

สรุปผลการวิจัยและข้อเสนอแนะ

1. สรุปผลการวิจัย

การศึกษานี้ได้เสนอการแก้ปัญหาการแบ่งประโยคภาษาไทยด้วยแคททิกอเรียลแกรมม่าและหลักเกณฑ์ทางไวยากรณ์ ภายใต้ชุดข้อมูลที่มีการกำกับรายละเอียดแล้ว นอกจากแคททิกอเรียลแกรมม่าซึ่งถูกใช้เป็นลักษณะสำคัญหลักแล้วการศึกษานี้ยังครอบคลุมถึงลักษณะสำคัญชนิดอื่น ทั้งนี้การศึกษานี้ครอบคลุมลักษณะสำคัญทั้งสิ้น 7 ชนิด รายละเอียดและผลของการใช้ลักษณะสำคัญแสดงดังตารางที่ 5.1

ตารางที่ 5.1 แสดงรายละเอียดและผลของการใช้ลักษณะสำคัญ

ลักษณะสำคัญ	**ความถูกต้องที่เพิ่มขึ้นสำหรับการแบ่งประโยคภาษาไทย**
1. ชนิดของคำ	✗
2. แคททิกอเรียลแกรมม่า	✓
3. คำที่อยู่โดยรอบ	✗
4. จำนวนคำระหว่างการเว้นวรรคที่พิจารณากับการเว้นวรรคใกล้เคียง	✓
5. จำนวนคำระหว่างการเว้นวรรคที่กำลังพิจารณากับการเว้นวรรคก่อนหน้าที่เป็นการแบ่งประโยค	✓
6. จำนวนคำระหว่างการเว้นวรรคที่กำลังพิจารณากับจุดสิ้นสุดของข้อความ (ในการทดลองคือจุดสิ้นสุดของย่อหน้า)	✗
7. การปรากฏของคำกริยาระหว่างการเว้นวรรคที่พิจารณากับการเว้นวรรคก่อนหน้าที่เป็นการแบ่งประโยค	✗

หากพิจารณาเฉพาะการใช้ลักษณะสำคัญที่เหมาะสมสำหรับการแบ่งประโยคภาษาไทยแล้วสามารถสรุปได้ว่า แคททิกอเรียลแกรมม่า, จำนวนคำระหว่างการเว้นวรรคที่พิจารณากับการเว้นวรรคใกล้เคียง และจำนวนคำระหว่างการเว้นวรรคที่กำลังพิจารณากับการเว้นวรรคก่อนหน้าที่เป็นการแบ่งประโยค เป็นลักษณะสำคัญที่เหมาะสม โดยให้ค่าความถูกต้องโดยรวม (space-correct) เท่ากับ 92.36% และมีความผิดพลาด (false-break) เท่ากับ 4.16% จากการศึกษาพบว่าจำนวนคำระหว่างการเว้นวรรคที่พิจารณากับการเว้นวรรคใกล้เคียงเป็นลักษณะสำคัญที่แสดงธรรมชาติของการเกิดขึ้นของประโยคที่ส่วนใหญ่แล้วมีความยาวที่อยู่ในช่วงที่เหมาะสมและเหตุปัจจัยที่ส่งผลให้แคททิกอ-เรียลแกรมม่าให้ผลลัพท์ที่ดีเนื่องจาก แคททิกอเรียลแกรมม่ามีการนิยามที่แสดงให้เห็นรายละเอียดของการเกิดขึ้นของภาษาที่มากกว่าลักษณะสำคัญชนิดอื่น นอกจากนี้ยังมีการศึกษาความเป็นไปได้ของการใช้หลักเกณฑ์ไวยากรณ์เพื่อเพิ่มความถูกต้องของการแบ่งประโยคพบว่า เมื่อใช้หลักเกณฑ์ไวยากรณ์ร่วมกับลักษณะสำคัญทั้งสามสามารถเพิ่มความถูกต้องโดยรวมขึ้นเป็น 93.54% และมีความผิดพลาดที่ 2.99% ซึ่งเมื่อเทียบกับการศึกษาในอดีตแล้ว จึงสรุปได้ว่าการใช้ลักษณะสำคัญดังกล่าวร่วมกับหลักเกณฑ์ไวยากรณ์เป็นระเบียบวิธีที่มีประสิทธิภาพสำหรับการแบ่งประโยคภาษาไทยและให้ค่าความถูกต้องที่ดีกว่าการศึกษาในอดีตที่ผ่านมา

2. ข้อเสนอแนะ

งานวิจัยนี้ศึกษาวิธีการแบ่งข้อความภาษาไทยให้ได้ผลลัพธ์เป็นประโยคโดยโดยมีแนวคิดจากหลักไวยากรณ์ทางภาษาไทยที่กำหนดว่า เมื่อจบใจความของประโยคให้เว้นวรรค (วรรคใหญ่) ดังนั้นงานวิจัยนี้จึงศึกษาการแบ่งประเภทของการเว้นวรรคว่าเป็นการแสดงการจบประโยคหรือไม่ แต่การใช้ภาษาในปัจจุบันได้มีการปรับเปลี่ยนรูปแบบการเขียนให้มีความง่ายมากขึ้น ดังนั้นเมื่อจบประโยคผู้เขียนอาจจะไม่แสดงการเว้นวรรค งานวิจัยนี้จึงไม่ครอบคลุมการแบ่งประโยคในทุกกรณี ดังนั้นจึงควรที่จะมีการศึกษาวิจัยเพิ่มเติมเกี่ยวกับการแบ่งประโยคโดยไม่ใช้สัญลักษณ์บ่งชี้ (งานวิจัยนี้ใช้การเว้นวรรคเป็นสัญลักษณ์บ่งชี้) หรือศึกษาเพิ่มเติมเพื่อหาระเบียบวิธีที่จะพิจารณาผลลัพธ์ที่ได้จากระเบียบวิธีที่ได้นำเสนอ

งานวิจัยนี้ศึกษาภายใต้ชุดข้อมูลที่มีการกำกับรายละเอียดบางส่วนแล้ว ซึ่งในอนาคตหากจะประยุกต์ใช้งานในลักษณะของการใช้งานโดยทั่วไป จำเป็นที่จะต้องมีการสร้างระเบียบวิธีสำหรับการกำกับรายละเอียดแคททิกอเรียลแกรมม่าที่มีความถูกต้องสูง

รายการอ้างอิง

[1] Walker, D., Clements, D., Darwin, M., and Amtrup J. Sentence Boundary Detection A Comparison of Paradigms for Improving MT Quality. Proceedings of MT Summit VIII, pp. 18-22. 2001

[2] Xu, J., Rechard Z., and Hermann N. Sentence Segmentation Using IBM Word Alignment Model 1. Proceedings of The 10th EAMT conference Practical application of machine translation, pp. 280-287. 2005.

[3] Cooke, J.R. Thai sentence particles and other topics. Department of Linguistics, Research School of Pacific Studies, Australian National University. 1989.

[4] Nantana Danvivathana. The Thai Writing System (Forum Phoneticum). Helmut Buske Verlag, 1987.

[5] Sayang Tepdang, Choochart Haruechaiyasak, and Rachada Kongkachandra. Improving Thai Word Segmentation with Named Entity Recognition. Proceedings of International Symposium on Communications and Information Technologies, pp. 940-945. 2010.

[6] Wigrai Thanadechteemapat, and Fung C. Thai Word Segmentation for Visualization of Thai Web Sites. Proceedings of International Conference on Machine Learning and Cybernetics, pp. 1544-1549. 2011.

[7] Choochart Haruechaiyasak, Sarawoot Kongyoung, and Dailey, M.N. A Comparative Study on Thai Word Segmentation Approaches. Proceedings of ECTI-CON, pp. 125-128. 2008.

[8] Longchupole Sungkornsarun. Thai Syntactical Analysis System by Method of Splitting Sentence from Paragraph for Machine Translation. Master Thesis, Department of Computer Engineering, Faculty of Engineering, King Monkut's Institute of Technology Ladkrabang, 1995.

[9] Pradit Mittrapiyanuruk and Virach Sornlertlamvanich. The Automatic Thai Sentence Extraction. Proceedings of the Fourth Symposium on Natural Language Processing, pp. 23-28. 2000.

[10] Paisarn Charoenpornsawat and Virach Sornlertlamvanich. Automatic Sentence Break Disambiguation for Thai. Proceedings of In International Conference on Computer Processing of Oriental Languages (ICCPOL), pp. 231-235. 2001.

[11] Slayden, G., Hwang, M. and Schwartz, L. Thai Sentence-Breaking for Large-Scale SMT. Proceedings of the 1st Workshop on South and Southeast Asian Natural Language Processing (WSSANLP), pp. 8-16. 2010.

[12] Breiman, L., Friedman, J., and Stone, C.J., and Olshen, R.A. Classification and Regression Tree. CRC Press, 1998.

[13] Taneth Ruangrajitpakorn, Kanokorn Trakultaweekoon, and Thepchai Supnithi. A Syntactic Resource for Thai: CG Treebank. Proceedings of the 7th Workshop on Asian Language Resource (ACL-IJCNPL), pp. 96-102. 2009.

[14] ราชบัณฑิตยสถาน. หลักเกณฑ์การใช้เครื่องหมายวรรคตอนและเครื่องหมายอื่นๆ หลักเกณฑ์การเว้นวรรค หลักเกณฑ์การเขียนคำย่อ ฉบับราชบัณฑิตยสถาน. ราชบัณฑิตยสถาน, 2548

[15] Wirote Aroonmanakun. Thoughts on Word and Sentence Segmentation in Thai. Proceedings of the Seventh International Symposium on Natural Language Processing, pp. 85-90. 2007.

[16] Virach. Sornlertlamvanich, Takahashi N., and Isahara H. Building a Thai Part-of-Speech Tagged Corpus (ORCHID). Journal of the Acoustical of Society of Japan, 20 (August 1999) : 189-198.

[17] C. Haruechaiyasak, S. Kongyoung and M. N. Dailey, “A Comparative Study on Thai Word Segmentation Approaches,” Proceedings of ECTI-CON, 2008, pp. 125-128

[18] Wigrai Thanadechteemapat and Fung C. Thai Word Segmentation for Visualization of Thai Web Sites. Proceedings of International Conference on Machine Learning and Cybernetics, pp. 1544-1549. 2011.

[19] Chatchawarn Hansakunbuntheung, Ausdang Thangthai, Chai Wutiwiwatchai, and Rungkarn Siricharoenchai. Learning Methods and Features for Coupus-based Phrase Break Prediction on Thai. Proceedings of Interspeech, pp. 325-328. 2005.

[20] Virongrong. Tesprasit, Paisarn Charoenpornsawat, and Virach Sornlertlamvanich. Learning phrase break detection in Thai text-to-speech. Proceedings of Interspeech, pp. 325-328. 2003.

[21] Sitipong Saychum and others. Categorial-Grammar-Based Phrase Break Prediction. Proceedings of ECTI-CON, pp. 954-957. 2011.

[22] Thepchai Supnithi, Peerachet Porkaew, Taneth Ruangrajitpakorn, Kanokorn Trakultaweekoon, Chanon Onman, and Asanee Kawtrakul. A Supervised Learning based Chuncking in Thai using Categorial Grammar. Proceedings of the Eighth Workshop on Asian Language Resources, pp 129-136. 2010.

[23] Sutheebanjard, P and Premchaiswadi, W. Thai Personal Name Entity Extraction without using Word Segmentation or POS Tagging. Proceedings of Eighth Intenational Symposium on Natural Language Processing, pp. 221-226. 2009.

[24] Yingyot Kanchina, Tipraporn Thanakulwarapas, Krit Kosawat and Sunant Anchaleenukoon. Recognizer for Thai Complex Sentences including Relative Pronouns. Proceedings of ECTI-CON, pp. 129-132. 2008.

[25] Chatchawarn Hansakunbuntheung, Ausdang Thangthai, Chai Wutiwiwatchai, and Rungkarn Siricharoenchai. Learning Methods and Features for Coupus-based Phrase Break Prediction on Thai. Proceedings of Interspeech, pp. 325-328. 2005.

[26] Parlikar, A. and Black, W.A. A Grammar Based Approch to Style Specific Phrase Prediction. Proceedings of Interspeech, pp. 2149-2152. 2011.

[27] Parlikar, *A. and* Black *W.A.* A Data-driven Phrasing for Speech Synthesis in Low-Resource Languages. Proceedings of ICASSP, pp. 4013-4016. 2012.

[28] Gildea, D. and Hockenmaier, J. Identifying Semantic Roles using Combinatory Categorial Grammar. Proceedings of the 2003 conference on Empirical methods in natural language processing, pp. pp, 57-64. 2003.

[29] Zhao, Z. and Zhu Y. Prediction of Prosodic Phrase Boundaries in Chinese TTS Based on Conditional Random Fields and Transformation Based Learning. Proceedings of Sixth International Conference on Fuzzy Systems and Knowledge Discovery. pp. 599-602. 2009.

[30] Navas, E., Hernaez, I., and Ezeiza, N. Assigning Phrase Breaks Using CARTs for Basque TTS. Proceedings of International Conference on Speech Prosody. pp. 527-531. 2002.

[31] Chatchawarn Hansakunbuntheung, Virong Tesprasit, and Virach Sornlertlamvanich. Thai Tagged Speech Corpus for Speech Synthesis. In Proceedings of Oriental COCOSDA, pp. 97-104. 2003.

[32] Taylor, P., Caley, R., Black, A.W., and King, S. Edinburgh Speech Tools Library System Documentation. University of Edinburgh, 1997.

[33] Bennett, J. and AutoIT Consulting Ltd. AutoIT Script Tools [online]. 2012. Available from: http://www.autoitscript.com/site/autoit [2012, May 19]

[34] Black, A.W. and Taylor, P. Assigning Phrase Breaks from Part-of-Speech Sequences. Computer Speech and Language, pp. 99-117. 1997.

ประวัติผู้เขียนวิทยานิพนธ์

นายณัฐชา ตังศิริรัตน์ เกิดเมื่อวันจันทร์ที่ 19 มิถุนายน พ.ศ.2532 ที่จังหวัดกรุงเทพมหานคร สำเร็จการศึกษาระดับมัธยมศึกษาจาก โรงเรียนหอวัง กรุงเทพมหานคร สำเร็จการศึกษาระดับปริญญาบัณฑิต จากคณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย กรุงเทพมหานคร ในปีการศึกษา 2554 มีผลงานตีพิมพ์คือ “Thai Sentence-Breaking using Categorial Grammar” จัดทำโดย “Nathacha Tangsirirat, Atiwong Suchato, Proadpran Punyabukkana” เผยแพร่ในเอกสารการประชุมวิชาการ “The Sixteenth International Computer Science and Engineering Conference: ICSEC'2012” หน้าที่ 139 - 144