การตัดคำและการกำกับหมวดคำภาษาไทยแบบเบ็ดเสร็จด้วยคอมพิวเตอร์

นาย นัฐวุฒิ ไชยเจริญ

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาอักษรศาสตรมหาบัณฑิต
สาขาวิชาภาษาศาสตร์ ภาควิชาภาษาศาสตร์
คณะอักษรศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย
ปีการศึกษา 2544
ISBN 974-17-0521-2

COMPUTERIZED INTEGRATED WORD SEGMENTATION AND PART-OF-SPEECH TAGGING OF THAI

Mr. Nattawut Chaicharoen

A Thesis Submitted in Partial Fulfillment of the Requirements
for the Degree of Master of Arts in Linguistics
Department of Linguistics
Faculty of Arts
Chulalongkorn University
Academic Year 2001

ISBN 974-17-0521-2

หัวข้อวิทยานิพนธ์	การตัดคำและการกำกับหมวดคำภาษาไทยแบบเบ็ดเสร็จด้วยคอมพิวเตอร์
โดย	นาย นัฐวุฒิ ไชยเจริญ
ภาควิชา	ภาษาศาสตร์
อาจารย์ที่ปรึกษา	อาจารย์ ดร.วิโรจน์ อรุณมานะกุล

คณะอักษรศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย อนุมัติให้นับวิทยานิพนธ์ฉบับนี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญามหาบัณฑิต

..... คณบดีคณะอักษรศาสตร์
(ผู้ช่วยศาสตราจารย์ ดร. ม.ร.ว. กัลยา ติงศภัทิย์)

คณะกรรมการสอบวิทยานิพนธ์

..... ประธานกรรมการ
(ผู้ช่วยศาสตราจารย์ ดร. สุดาพร ลักษณียนาวิน)

..... อาจารย์ที่ปรึกษา
(อาจารย์ ดร.วิโรจน์ อรุณมานะกุล)

..... กรรมการ
(ผู้ช่วยศาสตราจารย์ ดร. เพียรศิริ วงศ์วิภานนท์)

นัฐวุฒิ ไชยเจริญ : การตัดคำและการกำกับหมวดคำภาษาไทยแบบเบ็ดเสร็จด้วยคอมพิวเตอร์. (COMPUTERIZED INTEGRATED WORD SEGMENTATION AND PART-OF-SPEECH TAGGING OF THAI) อ. ที่ปรึกษา : อ. ดร. วิโรจน์ อรุณมานะกุล, 162 หน้า. ISBN 974-17-0521-2.

งานวิจัยนี้มีวัตถุประสงค์เพื่อสร้างโปรแกรมสำหรับตัดคำและกำกับหมวดคำแบบเบ็ดเสร็จด้วยคอมพิวเตอร์สำหรับภาษาไทย โดยใช้แบบจำลองไตรแกรมและชุดหมวดคำภาษาไทยที่ได้คัดสรรมา โดยมองว่าปัญหาการตัดคำและการกำกับหมวดคำเป็นส่วนงานเดียวกันซึ่งสามารถแก้ปัญหาไปพร้อมๆกันได้

ผู้วิจัยได้ทำการศึกษาเกณฑ์เรื่องคำ และนำเสนอชุดหมวดคำ เพื่อใช้สำหรับตัดคำและกำกับหมวดคำด้วยมือให้กับคลังข้อมูลซึ่งรวบรวมจากคลังข้อมูลของหนังสือพิมพ์กรุงเทพธุรกิจ ชุดหมวดคำภาษาไทยที่ใช้ในงานวิจัยนี้แบ่งเป็น 9 หมวดคำหลัก คือ นาม, กริยา, ตัวกำหนด, ตัวบอกปริมาณ, วิเศษณ์, คำหน้าหน่วยสร้างไร้ศูนย์, สันธาน, อนุภาค และเครื่องหมาย ตามเกณฑ์ทางวากยสัมพันธ์: การปรากฏร่วมของคำ และ การกระจายของคำ และแบ่งย่อยได้ทั้งหมด 26 หมวดคำสำหรับใช้เป็นป้ายกำกับหมวดคำในคลังข้อมูลและโปรแกรม

ในการทดลอง ให้โปรแกรมเรียนรู้ค่าสถิติจากคลังข้อมูลฝึกสอนที่ได้ทำการตัดคำและกำกับหมวดคำด้วยมือไว้ และทดสอบประสิทธิภาพกับข้อมูลทดสอบที่ไม่ได้มีการตัดคำ ผลการทดลองปรากฏว่า โปรแกรมสามารถกำกับหมวดคำและตัดคำได้ถูกต้อง 89.590% และ 96.087% ตามลำดับ ซึ่งแสดงให้เห็นว่าแบบจำลองไตรแกรมที่ใช้ปริบทหมวดคำข้างเคียงสามารถตัดคำและกำกับหมวดคำได้ประสิทธิภาพสูงในระดับหนึ่ง แต่เมื่อเทียบผลการตัดคำของแบบจำลองไตรแกรมที่ใช้หมวดคำข้างเคียงกับผลการตัดคำของแบบจำลองไตรแกรมที่ใช้รูปคำข้างเคียงแล้วพบว่า แบบจำลองที่ใช้หมวดคำข้างเคียงมีค่าความถูกต้องในการตัดคำต่ำกว่า ซึ่งแสดงให้เห็นว่า หากใช้แบบจำลองไตรแกรมเพื่อทำการตัดคำและกำกับหมวดคำภาษาไทย การแยกกระบวนการตัดคำและกระบวนการกำกับหมวดคำเป็นคนละกระบวนการน่าจะเหมาะสมมากกว่า โดยกระบวนการตัดคำควรเป็นกระบวนการขั้นต้นก่อนนำไปกำกับหมวดคำ

ภาควิชา.....ภาษาศาสตร์..... ลายมือชื่อนิสิต.....

สาขาวิชา.....ภาษาศาสตร์..... ลายมือชื่ออาจารย์ที่ปรึกษา.....

ปีการศึกษา 2544

4180137722 : MAJOR LINGUISTICS
KEY WORD: THAI / WORD / SEGMENTATION / PART OF SPEECH / TAGGING / TRIGRAM MODEL

NATTAWUT CHAICHAROEN : COMPUTERIZED INTEGRATED WORD SEGMENTATION AND PART-OF-SPEECH TAGGING OF THAI THESIS ADVISOR : WIROTE AROONMANAKUN, Ph.D., 162 pp. ISBN 974-17-0521-2.

This study aims at developing an integrated word segmentation and part-of-speech (POS) tagging program for Thai text, using trigram model and the selected POS tag set. The problem of word segmentation and POS tagging is treated as a single procedure in which those two problems are solved simultaneously.

We studied word criteria, and proposed a Thai POS set for using as a tool for manual segmentation and POS tagging on a corpus collected from Bangkok Business newspaper. The POS set in this study consists of 9 major categories, namely noun, verb, determiner, quantifier, adverb, exocentric marker, conjunction, particle, and punctuation, based on syntactic criteria: word co-occurrence, and word distribution. Major categories were further sub-categorized, yielding a total of 26 tags.

Training on manually segmented and tagged corpus, and testing on unsegmented test text, the result shows 89.590 % and 96.087 % accuracy for tagging and segmentation, respectively. This suggests that the POS trigram model can yield a fairly good result for tagging and segmentation in Thai. However, the segmentation accuracy is lower when compared with the result from the model that uses only word form trigram. This suggests that, when using a trigram model, it might be better to treat the word segmentation task and the POS tagging task as separated modules, i.e., the word segmentation task should precede the POS tagging task in Thai.

Department.....Linguistics..... Student's signature.....
Field of study....Linguistics..... . Advisor's signature.....
Academic year 2001

กิตติกรรมประกาศ

ผู้วิจัยขอขอบพระคุณ อ.ดร. วิโรจน์ อรุณมานะกุล อาจารย์ที่ปรึกษาวิทยานิพนธ์ เป็นอย่างสูงที่ได้ช่วยเหลือและให้คำปรึกษาแนะนำแนวทางในทุกขั้นตอน ตลอดจนแก้ไขและขัดเกลาวิทยานิพนธ์ฉบับนี้จนสำเร็จลุล่วงลงได้ และขอขอบพระคุณ ร.ศ. ดร. อมรา ประสิทธิ์รัฐสินธุ์ เป็นอย่างสูงเช่นกัน ที่ได้อนุเคราะห์งานวิจัยมาใช้เป็นแนวทางในวิทยานิพนธ์ฉบับนี้พร้อมทั้งให้คำปรึกษา ช่วยเหลือผู้วิจัยในการจัดทำชุดหมวดคำเป็นอย่างดีเสมอมา ซึ่งหากขาดความอนุเคราะห์จากอาจารย์ทั้ง 2 ท่านนี้แล้ว วิทยานิพนธ์ฉบับนี้คงมิอาจสำเร็จลงได้เลย และผู้วิจัยขอขอบพระคุณ ผ.ศ. ดร. สุดาพร ลักษณียนาวิน และ ผ.ศ. ดร. เพียรศิริ วงศ์วิภานนท์ กรรมการสอบวิทยานิพนธ์ที่ได้ให้คำปรึกษาและเสียสละเวลาเพื่อตรวจสอบแก้ไขวิทยานิพนธ์ฉบับนี้

นอกจากนี้ ผู้วิจัยขอขอบคุณคุณพิสิทธิ์ พรมจันทร์ สำหรับคำชี้แนะแนวทางในการทำวิทยานิพนธ์ และขอขอบคุณคณาจารย์ภาควิชาภาษาศาตร์ทุกท่านที่ได้ประสิทธิ์ประสาทความรู้ด้านภาษาศาสตร์ให้แก่ผู้วิจัย รวมทั้งขอขอบคุณเจ้าหน้าที่และเพื่อนๆภาควิชาภาษาศาสตร์ทุกคนที่ช่วยอำนวยความสะดวกและกระตุ้นเตือนผู้วิจัยในการทำวิทยานิพนธ์ฉบับนี้เสมอมา

สุดท้ายนี้ ผู้วิจัยขอขอบคุณคุณสิริธารินทร์ เจริญศิริ เป็นอย่างยิ่ง ที่ได้ช่วยเหลือและอำนวยความสะดวกในทุกๆด้าน พร้อมทั้งเป็นกำลังใจในการทำวิทยานิพนธ์อย่างดียิ่งตลอดมา และขอขอบพระคุณบิดามารดาสำหรับการสนับสนุนค่าใช้จ่ายระหว่างการศึกษา

สารบัญ

หน้า

หน้า

บทที่

หน้า

บทที่

หน้า

บทที่

สารบัญตาราง

สารบัญภาพ

บทที่ 1

บทนำ

1.1 ความเป็นมาและความสำคัญของปัญหา

ในปัจจุบันคอมพิวเตอร์มีบทบาทต่างๆในชีวิตมนุษย์เป็นอย่างมากในฐานะเป็นเครื่องมือช่วยอำนวยความสะดวกให้แก่มนุษย์ เนื่องจากคอมพิวเตอร์เป็นเครื่องจักรสมองกลที่มีสมรรถนะสามารถทำงานได้รวดเร็ว แม่นยำ สม่ำเสมอ สามารถจัดการกับปัญหาที่ซับซ้อนได้ และยังสามารถทำงานติดต่อกันได้เป็นเวลานาน จึงทำให้มนุษย์สนใจพัฒนาสมรรถนะของคอมพิวเตอร์ให้สามารถช่วยงานมนุษย์ในด้านต่างๆ โปรแกรมในด้านการประมวลผลภาษาธรรมชาติ (natural language processing - NLP) ต่างๆก็เป็นรูปแบบหนึ่งในการนำคอมพิวเตอร์มาช่วยอำนวยความสะดวกให้แก่มนุษย์ เช่น การประมวลผลคำ (word processing), การแปลภาษาด้วยเครื่อง (machine translation), การสังเคราะห์เสียงจากข้อความ (text-to-speech synthesis), การรู้จำเสียง (speech recognition) เป็นต้น ซึ่งเรื่องดังกล่าวเกิดจากความมุ่งหวังของมนุษย์ที่จะทำให้คอมพิวเตอร์สามารถรู้จำและเข้าใจภาษามนุษย์ได้ เพื่อที่มนุษย์จะสามารถติดต่อสื่อสารกับคอมพิวเตอร์ได้สะดวกขึ้นโดยใช้ภาษาธรรมชาติของมนุษย์เอง ดังนั้น การพัฒนาระบบการประมวลผลภาษาธรรมชาติด้วยคอมพิวเตอร์จึงนับได้ว่าเป็นความก้าวหน้าทางเทคโนโลยีที่จะช่วยให้เกิดความสะดวกในการติดต่อสื่อสารระหว่างมนุษย์กับเครื่องจักรได้ อันจะส่งผลให้เครื่องจักรสามารถทำงานได้ทัดเทียมมนุษย์มากขึ้น

อย่างไรก็ตาม การทำให้คอมพิวเตอร์สามารถเรียนรู้และเข้าใจภาษามนุษย์ได้นั้นไม่ใช่เรื่องง่าย การที่มนุษย์สามารถเรียนรู้ภาษาธรรมชาติได้ก็เพราะมนุษย์มีระบบประสาท ระบบสมองที่ซับซ้อน และมีกลไกการเรียนรู้ที่ลึกซึ้ง แต่คอมพิวเตอร์ซึ่งเป็นเครื่องจักรไม่ได้มีเครื่องมือต่างๆ ดังกล่าว รวมทั้งคอมพิวเตอร์ไม่มีสหัชญาณ* (intuition) ทางภาษา ดังนั้นปัญหาในการทำให้คอมพิวเตอร์เข้าใจภาษาในงานประมวลผลภาษาธรรมชาติจึงต้องอาศัยมนุษย์เป็นผู้กำหนดการทำงานให้แก่คอมพิวเตอร์เพื่อให้คอมพิวเตอร์สามารถเรียนรู้ภาษาธรรมชาติของมนุษย์ได้

---

* ศัพท์ภาษาไทยที่ อุดม วโรตม์สิกขดิตถ์ (2535: 142) ใช้

ในด้านภาษา ภาษาเป็นเครื่องมือสื่อสารอันมีประสิทธิภาพสูงของมนุษย์ มนุษย์ใช้ภาษาในการสืบทอด ส่งผ่าน และแลกเปลี่ยนความรู้ ความคิดกัน ภาษาที่มนุษย์ใช้สื่อสารมีทั้งภาษาพูดที่อยู่ในรูปเสียงพูด และภาษาเขียนที่อยู่ในรูปอักขระในภาษา สำหรับภาษาไทย ภาษาไทยสามารถสื่อสารได้ทั้งทางภาษาพูดและภาษาเขียนโดยมีระบบเสียงและระบบอักขระเป็นของตนเอง ภาษาไทยจัดเป็นประเภทภาษาคำโดดซึ่งมีรูปคำสำเร็จรูปสามารถนำไปใช้ในประโยคได้เลย แตกต่างจากประเภทภาษาคำควบซึ่งต้องมีคำประกอบเพื่อชี้หน้าที่หรือการกในประโยค เช่น ภาษามอญ เขมร มลายู และแตกต่างจากประเภทภาษาผันคำซึ่งมีวิภัติปัจจัยและวิธีการผันคำมากมาย เช่น ภาษาบาลี และภาษาตระกูลอินโดยูโรเปียนทั้งหลาย กำชัย ทองหล่อ (2515: 4) ได้นิยามภาษาคำโดดไว้ว่า *เป็นภาษาที่ไม่มีคำหรือเครื่องหมายแสดงความเกี่ยวข้องระหว่างคำที่เรียงร่วมกันอยู่ในประโยค และไม่มีการเปลี่ยนแปลงรูปคำให้เกิดความสัมพันธ์ในฐานเป็นการกต่างๆ เมื่อต้องการจะผูกประโยค ก็เอาคำแต่ละคำมาเรียงติดต่อกันเข้า* ลักษณะเช่นนี้ก็เป็นลักษณะหนึ่งที่ทำให้เกิดปัญหาต่อคอมพิวเตอร์ในการทำความเข้าใจภาษาไทย เพราะจะเกิดความกำกวมว่าคำที่ปรากฏในตำแหน่งต่างๆและมีรูปคำเหมือนกันจัดเป็นคำศัพท์เดียวกันหรือไม่

วิทยานิพนธ์ฉบับนี้สนใจศึกษาพัฒนาโปรแกรมเพื่อตัดคำและกำกับหมวดคำโดยอัตโนมัติให้กับข้อความภาษาไทยที่อยู่ในรูปตัวเขียน ซึ่งจำเป็นจะต้องแก้ไขปัญหาต่างๆ อันเกิดจากลักษณะภาษาเขียนของภาษาไทยที่ไม่เอื้ออำนวยต่อการพัฒนาโปรแกรมได้โดยง่าย อันได้แก่ การที่ภาษาเขียนสามารถเขียนคำเรียงติดต่อกันไปได้โดยไม่มีตัวบ่งขอบเขตของคำที่แน่ชัด (word boundary delimiter) เหมือนอย่างเช่นการเว้นช่องว่างระหว่างคำในภาษาอังกฤษ และการที่ภาษาไทยมีกลวิธีในการสร้างคำขึ้นใช้ใหม่ในภาษาโดยนำคำเดี่ยวที่มีอยู่เดิมมาประกอบกันเข้าเป็นคำผสม ทำให้เกิดปัญหาความกำกวมของการตัดแบ่งสายอักขระออกเป็นคำ นอกจากนี้ การที่ภาษาไทยเป็นภาษาคำโดด คำในภาษาไทยมีรูปคำสำเร็จ ไม่ต้องเปลี่ยนแปลงรูปคำเพื่อแสดงความสัมพันธ์ระหว่างคำในประโยค ทำให้รูปคำเดียวสามารถเป็นได้หลายหมวดคำ ก็ทำให้เกิดความกำกวมในการกำกับหมวดคำให้กับคำแต่ละคำที่ปรากฏในข้อเขียนภาษาไทย

การรู้จำคำโดยตัดคำในข้อเขียนภาษาไทยให้ถูกต้องนั้น เป็นเรื่องพื้นฐานของการประมวลผลภาษาไทย อันเป็นความจำเป็นขั้นต้นสำหรับคอมพิวเตอร์ที่จะต้องแก้ปัญหานี้ ซึ่งอาจกล่าวได้ว่าการตัดคำเป็นองค์ประกอบวิกฤติ (critical factor) ของการประมวลผลภาษาไทย (พิสิทธิ์ พรมจันทร์, 2540: 1) เนื่องจากมีงานต่างๆที่จำเป็นต้องเรียนรู้คำภาษาไทย และต้องการข้อมูลในรูปของคำ เช่น

1. การจัดรูปแบบเอกสารในงานประมวลผลคำ (word wrap)
2. การตรวจสอบตัวสะกดภาษาไทย (spelling check)
3. การวิเคราะห์วากยสัมพันธ์ (syntax analysis)
4. การแปลภาษาด้วยเครื่อง (machine translation)
5. การทำดรรชนีสำหรับเอกสาร (document indexing)
6. การทำอรรถาภิธาน (thesaurus)
7. การประมวลผลภาษาธรรมชาติ (natural language processing)
8. การสังเคราะห์เสียงพูดจากข้อความ (text-to-speech system)

ส่วนการกำกับหมวดคำให้กับแต่ละคำในข้อเขียนภาษาไทยนั้น เป็นเรื่องที่มีประโยชน์ต่อการวิเคราะห์วากยสัมพันธ์ของภาษาไทย มีระบบงานที่เมื่อกำกับหมวดคำน่าจะช่วยให้ทำงานได้สะดวกและถูกต้องยิ่งขึ้น เช่น

1. การแจงส่วนประโยค (sentence parsing)
2. การแปลภาษาด้วยเครื่อง (machine translation)

และยังช่วยเสริมประสิทธิภาพการทำงานของระบบงานต่างๆที่ต้องการข้อมูลในรูปของคำภาษาไทยดังที่กล่าวมาข้างต้นด้วย

การตัดคำและการกำกับหมวดคำภาษาไทยได้รับการพัฒนาอย่างต่อเนื่องมาเป็นลำดับจนถึงปัจจุบัน โดยเฉพาะงานด้านการตัดคำภาษาไทย มีงานวิจัยจำนวนไม่น้อยที่ได้เสนอวิธีการในการตัดคำภาษาไทยแบบต่างๆ โดยมุ่งหวังให้วิธีการที่เสนอมีประสิทธิภาพดีกว่าวิธีการที่ผ่านๆมา พิสิทธิ์ พรมจันทร์ (2540: 2) กล่าวว่า บางวิธีการจะได้ผลลัพธ์ทางเลือกในการตัดคำมากกว่าหนึ่งรูปแบบ เช่น งานวิจัยของรัตติกร วรากูลศิริพันธุ์ และคณะ (2538ก) และงานวิจัยของสมปรารถนา รัทยานนท์ (2535) เป็นต้น ซึ่งจำเป็นต้องเลือกทางเลือกใดทางเลือกหนึ่งที่ดีที่สุดโดยอาศัยกฎทางไวยากรณ์ (syntax) และความหมาย (semantic) มาช่วยตัดสิน งานวิจัยของรัตติกร วรากูลศิริพันธุ์ (2538ข) ได้เสนอวิธีการที่อาศัยความถี่ของการใช้คำภาษาไทยเพื่อเลือกประโยคที่มีการตัดคำที่ถูกต้อง เนื่องจากการใช้กฎไวยากรณ์ภาษาไทยในการเลือกจะทำให้ฐานความรู้มีขนาดใหญ่ เป็นต้น ถึงกระนั้นก็ตาม การตัดคำและการกำกับหมวดคำให้แก่ข้อเขียนภาษาไทยยังสามารถพัฒนาต่อไปได้อีก หน่วยงานต่างๆทั้งหน่วยงานของรัฐบาลและเอกชนต่างก็ให้ความสนใจพัฒนาวิธีการ

ตัดคำและการกำกับหมวดคำเพื่อเพิ่มประสิทธิภาพการประมวลผลภาษาไทยด้วยคอมพิวเตอร์ ซึ่งแต่ละงานจะมีความแตกต่างกันในด้านความถูกต้อง ความรวดเร็วของการทำงาน และปริมาณการใช้ทรัพยากรต่างๆ (พิสิทธิ์ พรมจันทร์, 2540: 1) เท่าที่ผ่านมา การแก้ปัญหาการกำกับหมวดคำภาษาไทยและปัญหาการตัดคำภาษาไทยถูกมองเป็น 2 ส่วนงานแยกจากกัน โดยบุญเสริม กิจศิริกุล (2541) มองว่า การตัดคำเป็นขั้นตอนแรกก่อนการกำกับหมวดคำ แต่วิทยานิพนธ์ฉบับนี้จะศึกษาพัฒนาวิธีการในการตัดคำและกำกับหมวดคำภาษาไทย โดยมองการแก้ปัญหาทั้งสองเรื่องเป็นส่วนงานเดียวกัน ผู้วิจัยเห็นว่า การตัดคำเป็นส่วนหนึ่งของการกำกับหมวดคำ กล่าวคือสามารถแก้ปัญหาทั้งสองไปพร้อมๆกันได้ เนื่องจากสามารถใช้ข้อมูลทางภาษาศาสตร์เรื่องหมวดคำมาช่วยตัดสินเลือกการตัดคำที่ถูกต้องของสายอักขระในปริบทหนึ่งๆได้ นอกจากนี้ ผู้วิจัยคาดว่าการกำกับหมวดคำจะช่วยให้ผลการตัดคำมีความถูกต้องมากกว่าการตัดคำโดยไม่พิจารณาเรื่องหมวดคำ

1.2 วัตถุประสงค์

เพื่อสร้างโปรแกรมสำหรับตัดคำและกำกับหมวดคำแบบเบ็ดเสร็จด้วยคอมพิวเตอร์โดยใช้แบบจำลองไตรแกรมและชุดหมวดคำที่คัดสรรมา

1.3 สมมติฐาน

1. กระบวนการตัดคำและกระบวนการกำกับหมวดคำสามารถรวมเป็นกระบวนการเดียวกันได้
2. แบบจำลองไตรแกรมซึ่งใช้ปริบทคำข้างเคียงในการคำนวณค่าความน่าจะเป็นของสายคำและสายหมวดคำเป็นกระบวนการที่เหมาะสมในการตัดคำและกำกับหมวดคำภาษาไทย

1.4 กรอบทฤษฎี

ประยุกต์ใช้วิธีการทางสถิติ (statistical techniques) โดยใช้แบบจำลองไตรแกรม (trigram model) เพื่อคำนวณหาค่าความน่าจะเป็นที่ใช้ในการตัดคำและกำกับหมวดคำให้กับข้อความภาษาไทย

1.5 เครื่องมือที่ใช้ในงานวิจัย

1. เครื่องไมโครคอมพิวเตอร์ส่วนบุคคล
2. โปรแกรมภาษา Perl ของบริษัท Active Perl

1.6 ขอบเขตงานวิจัย

1. ศึกษาการตัดคำและการกำกับหมวดคำให้กับข้อความภาษาไทยที่อยู่ในรูปตัวเขียน
2. ทดสอบโปรแกรมที่พัฒนาขึ้นมากับข้อมูลภาษาไทยจากคลังข้อมูลชุดทดสอบ (test set)

1.7 ขั้นตอนการวิจัย

1. ศึกษาทฤษฎี และพื้นฐานความรู้ทางภาษาศาสตร์ที่เกี่ยวข้อง ได้แก่ มโนทัศน์เรื่องคำ, การแบ่งวรรคตอนและเครื่องหมายต่างๆในภาษาไทย, การแบ่งคำภาษาไทยเป็นประเภทต่างๆ, การแบ่งหมวดคำในภาษาไทย
2. ศึกษาวิธีการตัดคำภาษาไทย และวิธีการกำกับหมวดคำในรูปแบบต่างๆที่ได้มีการพัฒนามาแล้ว
3. กำหนดชุดหมวดคำ (part-of-speech tag set) ที่จะใช้กำกับหมวดคำในวิทยานิพนธ์ฉบับนี้
4. กำหนดข้อมูลที่จะใช้เป็นคลังข้อมูล (corpus) โดยแบ่งเป็นคลังข้อมูลชุดฝึกสอน (training set) และคลังข้อมูลชุดทดสอบประสิทธิภาพของโปรแกรม (test set)
5. ทำการตัดคำและกำกับหมวดคำด้วยมือ (manual segmenting and tagging) ให้กับคลังข้อมูลฝึกสอน เพื่อใช้เป็นฐานความรู้ให้โปรแกรม
6. พัฒนาโปรแกรมตัดคำและกำกับหมวดคำภาษาไทยโดยอาศัยแบบจำลองไตรแกรม
7. พัฒนาโปรแกรมตัดคำภาษาไทยแบบที่ไม่นำข้อมูลเรื่องหมวดคำมาช่วย
8. ทดสอบการทำงานของโปรแกรมที่พัฒนาในขั้นตอน 6, 7
9. ประเมินผลการทำงานของโปรแกรม

1.8 ระเบียบและวิธีการวิจัย

1.8.1 การเก็บและจัดเตรียมข้อมูล

ผู้วิจัยรวบรวมข้อความภาษาไทยจากข้อมูลอิเล็กทรอนิกส์ที่เผยแพร่ไว้ในเวบไซต์ของหนังสือพิมพ์กรุงเทพธุรกิจ (http://www.bangkokbiznews.com) ทั้งหมดขนาดประมาณ 25,000 คำเพื่อนำมาใช้เป็นคลังข้อมูลในวิทยานิพนธ์ฉบับนี้ โดยได้เลือกใช้ข้อมูลของวันที่ 1 พฤษภาคม พ.ศ. 2543 ข้อมูลที่ใช้จัดเป็นข้อเขียนประเภทข่าวและบทความ ซึ่งมีการใช้ภาษาในรูปแบบภาษาเขียนที่ไม่เป็นทางการอันเป็นตัวอย่างของภาษาไทยที่พบได้ทั่วไปในชีวิตประจำวัน โดยผู้วิจัยทำการถ่ายโอนข้อมูลและแปลงให้อยู่ในรูปแฟ้มข้อมูลที่เป็นข้อความ (text file) จากนั้นผู้วิจัยจะเป็นผู้ทำการตัดคำและกำกับหมวดคำ (manual segmenting and tagging) ให้กับแต่ละคำในคลังข้อมูลตามเกณฑ์การตัดคำและชุดหมวดคำที่ใช้ในวิทยานิพนธ์ คลังข้อมูลทั้งหมดจะแบ่งเป็น 2 ส่วน คือ คลังข้อมูลฝึกสอน (training corpus) ซึ่งเป็นข้อความที่ตัดคำและกำกับหมวดคำด้วยมือแล้วขนาดประมาณ 80% เพื่อไว้สำหรับให้โปรแกรมเรียนรู้ค่าสถิติที่จะใช้ในการตัดคำและกำกับหมวดคำ ส่วนข้อมูลอีกส่วนหนึ่งขนาดประมาณ 20% เป็นข้อความที่ยังไม่ได้ทำการตัดคำและกำกับหมวดคำไว้ จะใช้เป็นคลังข้อมูลทดสอบ (test corpus) และคลังข้อมูลส่วน 20% นี้ ผู้วิจัยยังได้ทำการตัดคำและกำกับหมวดคำด้วยมือเพื่อใช้เป็นคำตอบสำหรับตรวจสอบความถูกต้องของผลลัพธ์จากโปรแกรม

1.8.2 การตัดคำและกำกับหมวดคำในคลังข้อมูลฝึกสอน

ผู้วิจัยจะเป็นผู้ตัดคำและกำกับหมวดคำภาษาไทยให้กับแต่ละคำในคลังข้อมูลฝึกสอนเพื่อใช้เป็นฐานความรู้ให้กับโปรแกรม (ดูรายละเอียดของเกณฑ์ในการตัดคำและชุดหมวดคำได้ในบทที่ 4) ข้อมูลภาษาไทยที่ทำการตัดคำและกำกับหมวดคำแล้วจะมีรูปแบบดังตัวอย่าง

ฉัน/NPRO_เดิน/V0_ไป/VN0_โรงเรียนNCM_

กล่าวคือ คำแต่ละคำในคลังข้อมูลจะถูกกำกับด้วยสัญลักษณ์หมวดคำโดยใช้เครื่องหมายทับ (/) คั่นระหว่างรูปคำและสัญลักษณ์หมวดคำ และแต่ละคำจะแยกจากกันด้วยเครื่องหมายขีดล่าง (_)

1.8.3 การทดสอบประสิทธิภาพของโปรแกรม

การทดสอบประสิทธิภาพจะกระทำหลังจากที่พัฒนาโปรแกรมแล้ว เพื่อทดสอบประสิทธิภาพการตัดคำและกำกับหมวดคำของโปรแกรมก่อนที่จะนำโปรแกรมไปใช้งานจริง ในการทดสอบ ผู้วิจัยจะให้โปรแกรมทดลองทำการตัดคำและกำกับหมวดคำให้กับข้อมูลชุดทดสอบซึ่งเป็นข้อความภาษาไทยที่ไม่ได้มีการตัดคำและกำกับหมวดคำ จากนั้นจึงประเมินผลโดยเปรียบเทียบผลการตัดคำและกำกับหมวดคำของโปรแกรมเทียบกับผลการตัดคำและกำกับหมวดคำให้กับข้อมูลชุดเดียวกันที่ทำโดยผู้วิจัย ตามที่อธิบายไว้ในหัวข้อ 1.8.4

1.8.4 การประเมินผลการตัดคำและกำกับหมวดคำ

การประเมินผลการตัดคำและการกำกับหมวดคำของโปรแกรมจะศึกษาผลที่ได้จากการทดสอบประสิทธิภาพ โดยประเมินผลเฉพาะประสิทธิภาพด้านความถูกต้องของการตัดคำและการกำกับหมวดคำเท่านั้น ไม่ประเมินผลประสิทธิภาพด้านความเร็วและปริมาณการใช้ทรัพยากร การประเมินผลความถูกต้องสามารถทำได้โดยวัดเป็นค่าร้อยละ F-measure (Van Rijsbergen, 1979 cited in Manning and Schutze, 1999) โดยแยกการประเมินผลเป็น 2 กรณี คือ ประเมินผลความถูกต้องของการตัดคำ และ ประเมินผลความถูกต้องของการกำกับหมวดคำ ในการประเมินผลประสิทธิภาพการตัดคำ ผู้วิจัยจะเปรียบเทียบผลการตัดคำของโปรแกรมแบบที่นำข้อมูลเรื่องหมวดคำมาพิจารณาไปพร้อมๆกัน กับผลการตัดคำของโปรแกรมแบบที่ไม่นำข้อมูลเรื่องหมวดคำมาช่วย โดยพิจารณาจากผลการตัดคำของคลังข้อมูลทดสอบชุดเดียวกัน เพื่อดูว่ามีความแตกต่างกันหรือไม่ อย่างไร และเปรียบเทียบประสิทธิภาพการตัดคำของโปรแกรมทั้ง 2 แบบ

ความกำกวมในการตัดคำและกำกับหมวดคำในภาษาไทยอาจส่งผลให้แต่ละคนตัดสินการตัดคำและกำกับหมวดคำแตกต่างกัน ดังนั้นในงานวิจัยนี้ ผู้วิจัยจะเป็นผู้ทำการตัดคำและกำกับหมวดคำให้กับคลังข้อมูลชุดทดสอบไว้ล่วงหน้า และใช้เป็นคำตอบสำหรับนำผลลัพธ์จากโปรแกรมมาตรวจสอบ

1.9 ประโยชน์ที่คาดว่าจะได้รับจากงานวิจัย

1. ชุดหมวดคำที่เหมาะกับการตัดคำและกำกับหมวดคำด้วยคอมพิวเตอร์

2. สามารถนำโปรแกรมไปประยุกต์ใช้ในการประมวลผลภาษาไทยด้วยคอมพิวเตอร์ด้านอื่นๆ เช่น การทำพจนานุกรม การแปลภาษาด้วยเครื่อง ฯลฯ

1.10 โครงร่างของบทต่าง ๆในวิทยานิพนธ์

ในบทที่ 2 ผู้วิจัยได้ศึกษาทบทวนว่า ภาษาไทยมีลักษณะเช่นใดที่ก่อให้เกิดปัญหาในการตัดคำและกำกับหมวดคำบ้าง (หัวข้อที่ 2.1 และหัวข้อที่ 2.2) รวมถึงทบทวนแนวคิดวิธีการตัดคำและวิธีการกำกับหมวดคำภาษาไทยที่ผ่านมา (หัวข้อที่ 2.3 และหัวข้อที่ 2.4) บทที่ 3 จะนำเสนอวิธีการจัดทำคลังข้อมูลภาษาสำหรับใช้ในการพัฒนาโปรแกรมตัดคำและกำกับหมวดคำ อันได้แก่ การรวบรวมชุดข้อมูล, รูปแบบของคลังข้อมูล, และขั้นตอนในการตัดคำและกำกับหมวดคำด้วยมือให้กับคลังข้อมูล บทที่ 4 ในหัวข้อที่ 4.1 จะนำเสนอและอภิปรายเกณฑ์ในการตัดคำที่คัดสรรมาใช้ตัดคำให้กับคลังข้อมูล และหัวข้อที่ 4.2 นำเสนอชุดหมวดคำภาษาไทยที่คัดสรรและนำมาปรับใช้ในวิทยานิพนธ์อย่างละเอียด บทที่ 5 กล่าวถึงลักษณะของปัญหาการตัดคำและกำกับหมวดคำจากมุมมองทางสถิติในหัวข้อที่ 5.1 และอธิบายแบบจำลองไตรแกรมที่นำมาใช้ในการพัฒนาโปรแกรมสำหรับตัดคำและกำกับหมวดคำแบบเบ็ดเสร็จในหัวข้อที่ 5.2 และกล่าวถึงแนวคิดของขั้นตอนวิธีวิเทอร์บิซึ่งนำมาช่วยเพิ่มประสิทธิภาพการทำงานของแบบจำลองไตรแกรมในหัวข้อที่ 5.3 บทที่ 6 ในหัวข้อที่ 6.1 กล่าวถึงการนำคลังข้อมูลที่จัดทำขึ้นมาใช้ประโยชน์เป็นคลังข้อมูลฝึกสอนและคลังข้อมูลทดสอบสำหรับโปรแกรม และอธิบายขั้นตอนการทำงานของส่วนประกอบต่างๆของโปรแกรมตัดคำและกำกับหมวดคำแบบเบ็ดเสร็จ ในหัวข้อที่ 6.2 อธิบายวิธีการประเมินผลประสิทธิภาพในการทำงานของโปรแกรม จากนั้นหัวข้อที่ 6.3 จะนำเสนอและอภิปรายผลการทดลองตัดคำและกำกับหมวดคำของโปรแกรมอย่างละเอียด และสุดท้าย บทที่ 7 กล่าวสรุปผลการทำงานของโปรแกรมตรวจสอบกับสมมติฐานที่ตั้งไว้ (หัวข้อที่ 7.1) และนำเสนอแนวทางในการพัฒนาการตัดคำและกำกับหมวดคำภาษาไทยต่อไป (หัวข้อที่ 7.2)

บทที่ 2

ทบทวนวรรณกรรม

บทนี้จะกล่าวถึงประเด็นต่างๆที่เกี่ยวข้องในการพัฒนาโปรแกรมเพื่อตัดคำและกำกับหมวดคำภาษาไทยซึ่งผู้วิจัยได้ทำการศึกษาทบทวนไว้ อันได้แก่ ในหัวข้อที่ 2.1 จะกล่าวถึงมโนทัศน์เรื่องคำและการรู้จำคำในภาษาไทย ในหัวข้อที่ 2.2 จะกล่าวถึงการจัดแบ่งหมวดคำภาษาไทยที่มีผู้นำเสนอไว้ก่อนหน้านี้ จากนั้น ในหัวข้อที่ 2.3 จะกล่าวถึงแนวคิดและวิธีการในการตัดคำภาษาไทยด้วยคอมพิวเตอร์ที่ผ่านมา และหัวข้อที่ 2.4 กล่าวถึงแนวคิดและวิธีการกำกับหมวดคำภาษาไทยด้วยคอมพิวเตอร์ที่ผ่านมา

2.1 มโนทัศน์เรื่องคำและการรู้จำคำในภาษาไทย

มโนทัศน์เรื่องคำเป็นเรื่องพื้นฐานสำหรับปัญหาการตัดคำในภาษาไทย ในส่วนนี้จะแสดงให้เห็นถึงลักษณะของภาษาไทยที่เป็นสาเหตุของปัญหาการตัดคำ ซึ่งเกิดจากการที่ภาษาไทยไม่มีการเขียนแยกคำที่ชัดเจน และนอกจากนี้ ปัญหาการตัดคำส่วนหนึ่งยังเกิดจากการที่ยังไม่มีเกณฑ์ที่ชัดเจนในการตัดสินขอบเขตของคำ ดังนั้น ผู้วิจัยจึงจะนำเสนอแนวคิดต่างๆที่เกี่ยวข้องกับเรื่องคำและการแยกประเภทของคำไทย เพื่อชี้ให้เห็นถึงปัญหาในเรื่องของการตัดคำนี้

หากพิจารณาระบบการเขียนของภาษาไทยจะพบว่ามีลักษณะเฉพาะตัว คือ คำจะเขียนเรียงติดต่อกันไปเป็นประโยคหรือข้อความได้โดยไม่มีการเว้นช่องว่างระหว่างคำเหมือนกับในภาษาอื่นๆหลายภาษา เช่น ภาษาอังกฤษ ภาษาฝรั่งเศส ฯลฯ ดังนั้นในข้อเขียนภาษาไทยจึงไม่ปรากฏตัวแบ่งขอบเขตของคำ (word boundary delimiter) ที่ชัดเจนแน่นอนว่า คำหนึ่งๆเริ่มต้นที่ใดและสิ้นสุดที่ใด ดังนั้นจึงทำให้เกิดความกำกวมในการตัดคำในภาษาไทย เช่น สายอักขระ (character string) ที่ว่า “ตากลม” สามารถตัดคำเป็น “ตา-กลม” (eye) (round) หรือ “ตาก-ลม” (to expose) (wind) หรือในกรณีของสายอักขระ “แม่น้ำ” อาจจะพิจารณาให้เป็นลำดับของคำ (word sequence) ที่ประกอบด้วยคำ 2 คำ คือ “แม่” (mother) และ “น้ำ” (water) หรืออาจพิจารณาให้เป็นคำประสมหนึ่งคำว่า “แม่น้ำ” (river) ก็ได้

มโนทัศน์เรื่องคำนี้ ตำราไวยากรณ์ต่างๆมักไม่ได้อธิบายหรือให้คำจำกัดความไว้ชัดเจนมากนัก โดยมักถือเอาว่า เจ้าของภาษารู้ดีอยู่แล้วว่าคำคืออะไร (อมรา ประสิทธิ์รัฐสินธุ์, 2544: 37) แต่ถึงแม้มโนทัศน์เรื่องคำจะยังไม่ชัดเจนนักก็ตาม ก็สามารถถือได้ว่าคำเป็นหน่วยที่มีความสำคัญในความคิดของผู้พูดภาษา ดังจะสามารถสังเกตได้จากการที่เด็กเริ่มเรียนรู้ภาษาก็เรียนเป็นคำ, พจนานุกรมในภาษาก็มักบรรจุคำเอาไว้, หรือการยืมระหว่างภาษาก็เป็นการยืมคำเสียเป็นส่วนใหญ่ (ปราณี กุลละวณิชย์ และคณะ, 2535: 92)

โดยคร่าวๆแล้ว คำ หมายถึง เสียงพูดหรือสายอักขระที่มีความหมายในภาษา โดย Leonard Bloomfield (1933: 178) ให้คำจำกัดความไว้ว่า คำ หมายถึง หน่วยที่เล็กที่สุดที่สามารถปรากฏตามลำพังได้ (minimum free form) ส่วน R.H. Robins (1964: 185) ได้อธิบายว่า สิ่งที่เป็นคำจะมีความคงตัว (stability) ไม่สามารถจะแยกย่อยให้เล็กลงไปได้อีก และจะจัดลำดับส่วนที่อยู่ในคำเสียใหม่ก็ไม่ได้ ส่วนตำราไวยากรณ์ไทยทั้งหลาย (พระยาอุปกิตศิลปสาร, 2514: 59, นววรรณ พันธุเมธา, 2527: 2, สมชาย ลำดวน, 2526: 102, เรืองเดช ปันเขื่อนขัติย์, 2541: 147) กล่าวไว้คล้ายคลึงกันว่า คำ หมายถึง กลุ่มของหน่วยเสียงหรือกลุ่มของตัวอักษรที่มีความหมายในภาษา คำจำกัดความเหล่านี้มีความแตกต่างกันในด้านมุมมองในการตัดสินคำ

ผู้วิจัยเห็นว่าความหมายเป็นสิ่งสำคัญในการรู้จำคำ กล่าวคือ มนุษย์สามารถรับรู้ได้ว่าสายอักขระใดถือเป็นคำก็ต่อเมื่อสายอักขระนั้นสื่อความหมายอย่างหนึ่งอย่างใดให้เป็นที่เข้าใจได้ แต่ความหมายก็เป็นสิ่งที่ซับซ้อน และไม่ปรากฏรูปออกมาให้สังเกตได้ชัดเจน เป็นเพียงมโนทัศน์ (concept) ที่ผู้ใช้ภาษานั้นๆมีอยู่ร่วมกัน ดังนั้น การจะรับรู้ความหมายได้ก็โดยอาศัยสหัชญาณ (intuition) ทางภาษา แม้หากจะยึดตามคำนิยามที่ปรากฏในพจนานุกรมฉบับราชบัณฑิตยสถาน (2525: 186) เป็นหลักแต่เพียงอย่างเดียวว่า *คำ หมายถึง เสียงพูดหรือลายลักษณ์อักษรที่เขียนหรือพิมพ์ขึ้นเพื่อแสดงความคิด โดยปรกติถือว่าเป็นหน่วยที่เล็กที่สุดซึ่งมีความหมายในตัว* ก็ยังมีปัญหา ไม่สามารถจะทราบได้แน่ชัดว่า คำนั้นมีความยาวแค่ไหนและมีขอบเขตอย่างไรอยู่นั่นเอง แม้แต่ในภาษาอังกฤษที่มีการเว้นช่องว่างระหว่างคำก็ยังมีปัญหาในกรณีของคำประสมที่บ้างก็มีขีดคั่นอยู่ตรงกลาง บ้างก็ไม่มี เช่น jack-of-all-trade, jackpot, jack rabbit (อุดม วโรตม์สิกขดิตถ์, 2535: 141) เป็นต้น ดังนั้นจะหาเกณฑ์ที่บอกขอบเขตของคำให้แน่นอนตายตัวคงเป็นเรื่องที่ลำบาก

ลักษณะอีกประการหนึ่งของภาษาไทยที่เป็นปัญหาต่อการตัดคำ คือ คำในภาษาไทยสามารถจัดได้เป็นหลายประเภทตามกลวิธีในการประกอบคำเพื่อให้ได้คำใหม่ๆมาใช้เพิ่มเติมในภาษาเพื่อให้เพียงพอต่อความจำเป็นในการสื่อสาร ตำราไวยากรณ์ภาษาไทยต่างๆ (เช่น กำชัย ทองหล่อ: 2515; บรรจบ พันธุเมธา: 2514; เรืองเดช ปันเขื่อนขัติย์: 2541; สุโขทัยธรรมาธิราช: 2533) ได้อธิบายชนิดของคำตามลักษณะการประกอบคำเป็นประเภทต่างๆ ได้แก่

(1) คำมูล (simple word) คือ คำเดี่ยว หน่วยคำเดียว มักเป็นคำหลักที่มีความหมายเดียว เช่น นอน อ่าน ช้าง คน สะพาย กะทิ อนามัย นาฬิกา เป็นต้น
(2) คำประสาน หรือ คำผสาน (complex word) คือ คำที่เกิดจากหน่วยคำไม่อิสระ (bound morpheme)* ประกอบเข้าด้วยกัน เรียกว่า “คำประสานแท้” หรือประกอบเข้ากับคำมูล เรียกว่า “คำประสานเทียม” เช่น ชดช้อย เกิดจากหน่วยคำไม่อิสระ 2 หน่วย คือ “ชด-“ และ “-ช้อย” ประกอบเข้าด้วยกัน ส่วน นักเรียน เกิดจากหน่วยคำไม่อิสระ “นัก-“ ประกอบเข้ากับคำมูล “เรียน”
(3) คำประสม (compound word) คือ คำที่เกิดจากการนำคำมูล 2 คำขึ้นไปมาประกอบกัน เช่น แม่น้ำ น้ำตก รถไฟ น้ำแข็ง เครื่องคิดเลข
(4) คำซ้อน หรือ คำไวพจน์ผสม (synonymous compound word) คือ คำที่เกิดจากการนำคำที่มีความหมายเหมือนหรือใกล้เคียงกันมาประกอบกัน เช่น บอกกล่าว เร็วไว รีบเร่ง แก่เฒ่า
(5) คำซ้ำ (reduplication) คือ คำที่เกิดจากการนำคำคำเดียวกันมาซ้ำกัน หรือนำคำที่มีเสียงคล้ายกันมาประกอบกันเข้า เช่น แดงๆ ดีๆ โฉ่งฉ่าง ตกอกตกใจ ออดๆแอดๆ ด๊ำดำ

อย่างไรก็ตาม คำแต่ละประเภทข้างต้นยังมีการแบ่งประเภทย่อยลงไปอีก และยังมีความลักลั่นกันอยู่ เช่น คำว่า “ดีเดอ” อาจจัดให้อยู่ในประเภทคำได้หลายประเภท โดยถ้าพิจารณาว่าเกิดจากหน่วยคำไม่อิสระ “-เดอ” ประกอบเข้ากับคำมูล “ดี” ก็จะจัดเป็นคำประสาน แต่ก็อาจพิจารณาเป็นคำซ้อนเพื่อเสียงได้ตามแนวคิดของบรรจบ พันธุเมธา (2514) ที่แบ่งประเภทคำซ้อนย่อยลงไปเป็นคำซ้อนเพื่อเสียงและคำซ้อนเพื่อความหมาย หรือหากพิจารณาว่าเกิดจากการซ้ำเสียงพยัญชนะต้นก็อาจจัดเป็นคำซ้ำได้เช่นกัน อย่างไรก็ดี คำแต่ละประเภทที่เกิดขึ้นใหม่นี้มีความหมายเฉพาะ ดังนั้นจึงถือเป็นคำหนึ่งคำทั้งสิ้น

---

* “หน่วยคำพันธะ” (อุดม วโรตม์สิกขดิตถ์: 2535: 152) หรือ “หน่วยคำประสาน” (เรืองเดช ปันเขื่อนขัติย์: 2541: 171)

ถึงแม้จะมีการอธิบายคำเป็นประเภทต่างๆตามลักษณะการประกอบคำอย่างที่กล่าวมา แต่ผู้วิจัยเห็นว่า ถ้าพิจารณาจากปัญหาการตัดคำ ประเภทของคำสามารถจัดแบ่งได้เป็น 2 กลุ่มใหญ่ๆ คือ

(1) คำเดี่ยว ได้แก่ คำมูล และคำประสานแท้ เนื่องจากคำประสานแท้นั้น เมื่อแยกหน่วยคำกันออกมาแล้วไม่มีส่วนใดที่สามารถปรากฏโดยลำพังเป็นคำได้ (คำว่า “นาฬิกา” ถึงแม้จะแยกเป็นคำว่า “นา” และ “กา” ได้ แต่คำว่า “นา” และ “กา” ไม่ได้มีความหมายเกี่ยวข้องใดๆกับคำว่า “นาฬิกา” เลย ดังนั้น “นา” และ “กา” ในตัวอย่างนี้จึงไม่จัดว่าเป็นหน่วยคำ เพราะฉะนั้นคำว่า “นาฬิกา” จึงมีหน่วยคำเดียวและจัดว่าเป็นคำเดี่ยว)
(2) คำประกอบ ได้แก่ คำที่เกิดจากการประกอบหน่วยคำตั้งแต่ 2 หน่วยคำขึ้นไปเข้าด้วยกัน และมีอย่างน้อย 1 หน่วยคำที่เป็นหน่วยคำอิสระได้ ซึ่งรวมคำประสม คำประสานเทียม คำซ้อน และคำซ้ำอยู่ในประเภทนี้

ปัญหาที่เกี่ยวข้องกับเรื่องประเภทของคำ คือ สายอักขระที่ปรากฏอาจจัดเป็นคำเดี่ยวหนึ่งคำ คำประกอบหนึ่งคำ หรือเป็นคำเดี่ยวหรือคำประกอบมากกว่าหนึ่งคำ ก็ได้ เช่น “น้ำแข็ง” (คำประกอบประเภทคำประสมหนึ่งคำ) กับ “น้ำ-แข็งตัว” (คำเดี่ยว 2 คำ), “ปากกาตก” (คำประกอบประเภทคำประสมหนึ่งคำ) กับ “ปาก-กายาวกว่าปากนก” (คำเดี่ยว 2 คำ), “เขาใช้ไม้เท้ายันตัวลุกขึ้น” (คำประกอบประเภทคำประสมหนึ่งคำ) กับ “เขาใช้ไม้-เท้าคางเวลานั่ง” (คำเดี่ยว 2 คำ), “เรือโคลงเพราะโค-ลงเรือ” (คำเดี่ยวหนึ่งคำ และคำเดี่ยว 2 คำตามลำดับ) เป็นต้น การตัดสินเลือกการตัดคำที่ถูกต้องสามารถทำได้โดยพิจารณาความหมายของคำในปริบท และดูตำแหน่งการปรากฏของคำซึ่งแสดงความสัมพันธ์ทางไวยากรณ์กับคำอื่นๆ เพื่อช่วยพิจารณาว่า สายอักขระดังกล่าวเป็นคำหนึ่งคำหรือไม่ เรื่องประเภทของคำในภาษาไทยนี้ แม้แต่คนไทยซึ่งเป็นเจ้าของภาษาก็ยังไม่สามารถตกลงเห็นพ้องกันได้เสมอไปว่า สายอักขระที่ปรากฏนั้นเป็นหนึ่งคำหรือไม่ เช่น “เครื่องพิมพ์ดีดไฟฟ้า” “โรงไฟฟ้าพลังน้ำ” “หม้อหุงข้าว” เป็นต้น แต่ละคนก็อาจตัดสินไม่เหมือนกันหากยึดเฉพาะความหมายเป็นเกณฑ์ในการตัดสินคำ เพราะความหมายเป็นเรื่องซับซ้อน แต่ละคนก็อาจมีมโนทัศน์ที่แตกต่างกัน ไม่สามารถกำหนดตายตัวลงไปได้ว่า คำหนึ่งๆจะต้องมีมโนทัศน์เช่นไร ดังนั้น ในวิทยานิพนธ์ฉบับนี้ นอกจากผู้วิจัยจะพิจารณาด้านความหมายแล้ว ยังศึกษาเกณฑ์ทางภาษาศาสตร์อื่นๆ ได้แก่ เกณฑ์ทางวากยสัมพันธ์, เกณฑ์ทางจิตวิทยา เพื่อคัดเลือกเกณฑ์ที่มี

ความลื่นไหลน้อยที่สุดเพื่อนำมาใช้พิจารณาในการตัดสินคำในภาษาไทยด้วย (ดังจะได้กล่าวอย่างละเอียดในบทที่ 4)

ในบางกรณี เมื่อสายอักขระภาษาไทยปรากฏอยู่โดดๆ แม้แต่เจ้าของภาษาก็เกิดความลำบากที่จะตัดคำให้ได้ถูกต้อง เช่น “ตากลม”, “โคลง” ก็จำเป็นต้องอาศัยปริบท (context) เพื่อช่วยพิจารณาว่า การตัดคำแบบใดมีความหมายเหมาะสมกับปริบทนั้นๆมากที่สุด โดยวิทยานิพนธ์ฉบับนี้มองว่า การแก้ไขความกำกวมในการตัดคำก็เปรียบเสมือนเป็นการแก้ไขความกำกวมในการเกิดร่วมกันของคำ (treat word segmentation disambiguation as word sequence disambiguation) แต่ในบางครั้งถึงแม้จะได้พิจารณาโดยอาศัยปริบทแล้ว ก็อาจจะไม่สามารถตัดสินได้แน่ชัดว่าควรจะเลือกการตัดคำแบบใดในปริบทนั้นๆ เช่น ในปริบทที่จำกัด “เขาตากลม” อาจหมายถึง “ตา-กลม” หรือ “ตาก-ลม” ก็ได้ แต่ละคนก็อาจตัดสินเลือกการตัดคำต่างกัน

นอกจากข้อเขียนภาษาไทยจะมีลักษณะเป็นคำเขียนเรียงติดต่อกันไปแล้ว ยังปรากฏการแบ่งวรรคตอนโดยใช้เครื่องหมายแบ่งวรรคตอนเป็นครั้งคราวด้วย การแบ่งวรรคตอนจึงอาจเป็นวิธีหนึ่งที่ช่วยแสดงขอบเขตของคำได้ กล่าวคือ เราสามารถทราบได้ว่าตำแหน่งที่อยู่หน้าเครื่องหมายแบ่งวรรคตอนเป็นตำแหน่งจบคำ และตำแหน่งที่อยู่หลังเครื่องหมายแบ่งวรรคตอนเป็นตำแหน่งเริ่มต้นคำเสมอ ซึ่งราชบัณฑิตยสถาน (2530) ได้กำหนดเกณฑ์ในการใช้เครื่องหมายวรรคตอนไว้ แต่แม้ว่าราชบัณฑิตยสถานจะได้กำหนดเกณฑ์การแบ่งวรรคตอนในภาษาไทยไว้ก็ตาม ข้อเขียนภาษาไทยโดยส่วนใหญ่ก็ไม่ได้แบ่งวรรคตอนตามเกณฑ์เสมอไป แต่มักแบ่งวรรคตอนเมื่อผู้เขียนเห็นว่าจบความหนึ่งๆโดยอาศัยความคุ้นเคย ผู้วิจัยจึงเห็นว่า การสร้างโปรแกรมตัดคำและกำกับหมวดคำภาษาไทยสามารถใช้ประโยชน์จากเครื่องหมายและการแบ่งวรรคตอนในภาษาไทยมาเพียงเพื่อช่วยตัดแบ่งข้อความออกเป็นส่วนๆเพื่อที่จะนำไปประมวลผลทีละส่วน ซึ่งจะเป็นการกำหนดขอบเขตของข้อความที่จะประมวลผลในรอบหนึ่งๆให้สั้นลง และน่าจะสามารถแก้ปัญหาความกำกวมในการตัดคำและปัญหาในการกำกับหมวดคำได้สะดวกและถูกต้องยิ่งขึ้น

2.2 การจัดแบ่งหมวดคำในภาษาไทย

หัวข้อนี้จะกล่าวถึงการจัดแบ่งหมวดคำในภาษาไทย โดยในหัวข้อ 2.2.1 กล่าวถึงความสำคัญของการจัดแบ่งหมวดคำเพื่อแสดงให้เห็นว่าการจัดแบ่งหมวดคำเป็นลักษณะพื้นฐานของระบบภาษาต่างๆ ซึ่งเกณฑ์ที่ใช้ในการจัดแบ่งหมวดคำก็มีได้หลายเกณฑ์ ส่วนหัวข้อที่ 2.2.2

กล่าวถึงชุดหมวดคำต่างๆในภาษาไทยที่ได้มีผู้นำเสนอไว้ก่อนหน้านี้ ซึ่งผู้วิจัยจะได้จัดเป็นกลุ่มตามเกณฑ์ที่ใช้ในการแบ่งหมวดคำ

2.2.1 ความสำคัญของการจัดแบ่งหมวดคำ

การจัดแบ่งหมวดคำเป็นลักษณะสากล (universality) ของภาษา คำในภาษาทุกภาษาสามารถจัดประเภทเป็นหมวดหมู่ต่างๆ โดยคำที่อยู่ในหมวดเดียวกันก็จะคล้ายคลึงกันตามเกณฑ์ที่ใช้ในการจัดหมวดหมู่ เช่น ถ้าใช้เกณฑ์ความหมาย คำที่มีความหมายทำนองเดียวกันก็จัดอยู่ในหมวดเดียวกัน ถ้าใช้เกณฑ์ตำแหน่งในการปรากฏของคำ คำที่มักปรากฏในตำแหน่งเดียวกันก็จัดอยู่ในหมวดเดียวกัน คำคำหนึ่งเมื่อใช้เกณฑ์หนึ่งอาจจัดอยู่ในหมวดเดียวกับอีกคำหนึ่ง แต่หากเปลี่ยนเกณฑ์ที่ใช้ในการแบ่งหมวดหมู่แล้วคำทั้งสองอาจแยกกันอยู่คนละหมวดก็ได้ แต่ละหมวดหมู่ของคำก็มีชื่อเรียกแตกต่างกัน เช่น คำนาม คำกริยา คำบุรพบท เป็นต้น ทั้งนี้เพื่อให้สะดวกในการนำคำเหล่านี้ไปใช้ในวลีและประโยคต่างๆ หมวดหมู่สำหรับคำต่างๆเหล่านี้ในภาษาไทยเรียกว่า “หมวดคำ” หรือ “ชนิดของคำ” ซึ่งตรงกับภาษาอังกฤษว่า word class หรือ part of speech หรือ grammatical category ทุกคำในภาษาต้องจัดอยู่ในหมวดคำใดหมวดคำหนึ่งอย่างน้อย 1 หมวดคำเสมอ Robins (1964: 218) กล่าวถึง การแบ่งหมวดคำไว้ว่า

> *In the grammatical analysis of languages words are assigned to word classes on the formal basis of its syntactic behavior, supplemented and reinforced by differences of morphological paradigms so that every word in a language is a member of word class*

การแบ่งหมวดคำเป็นสิ่งสำคัญสำหรับการวิเคราะห์ไวยากรณ์ภาษา เนื่องจากในระบบโครงสร้างภาษา ความสัมพันธ์ระหว่างคำในโครงสร้างไม่ใช่เป็นเพียงความสัมพันธ์ระหว่างคำดังกล่าวเท่านั้น แต่ยังแสดงให้เห็นถึงความสัมพันธ์ระหว่างหมวดคำของทั้งภาษาด้วย

2.2.2 ชุดหมวดคำในภาษาไทย

ลักษณะของภาษาไทยที่ทำให้เกิดความกำกวมในการกำกับหมวดคำ ได้แก่ การที่ภาษาไทยเป็นภาษาคำโดด คำในภาษาไทยมีรูปคำที่แน่นอนสำเร็จรูป สามารถนำไปใช้ในประโยคได้ทันที ไม่มีการเปลี่ยนแปลงรูปคำเพื่อแสดงความสัมพันธ์ระหว่างคำที่เรียงต่อกันในประโยค รูปคำหนึ่งๆจึงอาจเป็นได้หลายหมวดคำ ดังนั้น จึงทำให้เกิดปัญหาคำหลายหน้าที่ (polysemy) (สุโขทัยธรรมาธิราช: 2533, 312-313) ซึ่งทำให้เกิดความลำบากในการจัดแบ่งหมวดหมู่ของคำในภาษาไทย (ปัญหาเรื่องคำหลายหน้าที่นี้ จะกล่าวถึงโดยละเอียดอีกครั้งในบทที่ 4)

ในภาษาไทย นักไวยากรณ์ไทยหลายท่านได้เสนอการจัดแบ่งหมวดคำในภาษาไทยต่างๆกันไป โดยแต่ละท่านต่างก็ใช้เกณฑ์ในการจัดที่แตกต่างกัน แต่ในที่นี้ผู้วิจัยได้พิจารณาจากวิธีการและเกณฑ์ที่ใช้ในการจัดแบ่งหมวดคำแล้วเห็นว่า สามารถแบ่งได้เป็น 2 กลุ่มใหญ่ๆ ดังนี้

2.2.2.1 การจัดแบ่งหมวดคำโดยใช้ความรู้ทางภาษาของผู้จัดแบ่ง (intuition based approach)

การจัดแบ่งหมวดคำที่อยู่ในกลุ่มนี้ ผู้วิจัยใช้วิธีพิจารณาจากความรู้ทางภาษาของตนเองว่าควรมีหมวดคำอะไรบ้าง และแต่ละคำควรเป็นหมวดคำอะไร เกณฑ์หลักที่ใช้ในการวิเคราะห์โดยส่วนใหญ่เป็นเกณฑ์ความหมายเป็นสำคัญ งานวิจัยหรือตำราที่จัดอยู่ในกลุ่มนี้ ได้แก่

(1) พระยาอุปกิตศิลปสาร (2514) แบ่งคำเป็น 7 หมวดคำในปี พ.ศ. 2465 ได้แก่ คำนาม คำสรรพนาม คำกริยา คำวิเศษณ์ คำบุรพบท คำสันธาน และคำอุทาน โดยใช้ทั้งเกณฑ์ความหมาย เกณฑ์หน้าที่ และเกณฑ์ตำแหน่ง ผสมกัน การแบ่งหมวดคำตามแบบพระยาอุปกิตศิลปสารนี้ ในภายหลังมีผู้นำมาพัฒนาและอธิบายไว้อย่างละเอียดอีก ผลงานที่เด่นๆ คืองานของกำชัย ทองหล่อ (2515)
(2) กำชัย ทองหล่อ (2515) แบ่งคำเป็น 7 หมวดคำ โดยนำการแบ่งหมวดคำของพระยาอุปกิตศิลปสารมาพัฒนาและอธิบายไว้อย่างละเอียด
(3) บรรจบ พันธุเมธา (2514) แบ่งคำเป็น 8 หมวดคำโดยใช้เกณฑ์เหมือนกับพระยาอุปกิตศิลปสาร แต่ได้แยกหมวดคำลักษณนามจากหมวดคำย่อยของคำนามออกมาเป็นหมวดคำหลักอีก 1 หมวดคำ

(4) นววรรณ พันธุเมธา (2527) แบ่งคำตามหน้าที่ในการสื่อสารเป็น 6 หมวดคำ ได้แก่ คำเรียก-ร้อง คำหลัก (ได้แก่ คำนาม คำกริยา) คำแทน (สรรพนาม) คำขยาย (คุณศัพท์) คำเชื่อม (คำสันธาน) และคำเสริม (คำลงท้าย)

(5) อุดม วโรตม์สิกขดิตถ์ (2535) แบ่งคำตามหลักภาษาศาสตร์และปฏิบัติศาสตร์ได้เป็น 8 หมวดคำ ได้แก่ นาม (รวมคำนาม สรรพนาม ลักษณนาม บุพบทไว้ด้วยกัน) กริยา คำนำหน้ากริยา วิเศษณ์ สังขยา นิยมลักษณ์ คำท้ายประโยค และคำสันธาน

(6) เรืองเดช ปันเขื่อนขัติย์ (2541) แบ่งคำตามหน้าที่ของคำในประโยคเป็น 12 หมวดคำ ได้แก่ คำนาม คำสรรพนาม คำกริยา คำวิเศษณ์ คำคุณศัพท์ คำลักษณนาม คำเชื่อม คำนับ คำลงท้าย คำปฏิเสธ คำอุทาน คำกำหนด

2.2.2.2 **การจัดแบ่งหมวดคำจากการวิเคราะห์คลังข้อมูลหรือประโยคทดสอบ** (corpus based approach)

การจัดแบ่งหมวดคำที่อยู่ในกลุ่มนี้ ผู้วิจัยใช้วิธีวิเคราะห์ข้อมูลการใช้ภาษาจริงจากคลังข้อมูลหรือประโยคในภาษาไทย แล้วจึงสรุปเป็นชุดหมวดคำออกมา และเกณฑ์หลักที่ใช้ในการวิเคราะห์โดยส่วนใหญ่เป็นเกณฑ์ทางวากยสัมพันธ์โดยพิจารณาจากตำแหน่งในการปรากฏของคำ งานวิจัยหรือตำราที่จัดอยู่ในกลุ่มนี้ ได้แก่

(1) วิจินตน์ ภาณุพงศ์ (2532) แบ่งคำตามตำแหน่งในประโยคตามแนวคิดไวยากรณ์โครงสร้างเป็น 15 หมวดคำหลัก โดยอาศัยกรอบประโยคทดสอบ และแบ่งย่อยได้ทั้งหมด 26 หมวดคำ

(2) วิรัช ศรเลิศล้ำวาณิช และคณะ (Virach Sornlertlamvanich et al., 1997) แบ่งคำโดยใช้ทั้งเกณฑ์ความหมาย ตำแหน่งการปรากฏ และหน้าที่ของคำ แบ่งเป็น 14 หมวดคำหลักและแบ่งหมวดคำย่อยได้เป็น 47 หมวดคำ เพื่อใช้ในงานประมวลผลภาษาธรรมชาติของภาษาไทย

(3) อมรา ประสิทธิ์รัฐสินธุ์ (2543) แบ่งคำโดยใช้เกณฑ์ทางวากยสัมพันธ์ คือ เกณฑ์การปรากฏร่วม และเกณฑ์การกระจายของคำ โดยอิงจากแนวคิดของทฤษฎีไวยากรณ์พึ่งพาศัพทการกแบ่งได้เป็น 8 หมวดคำหลัก ได้แก่ กริยา, นาม, ตัวกำหนด, ตัวบอกจำนวน, วิเศษณ์, บุพบท, สันธาน และอนุภาค

ชุดหมวดคำต่างๆข้างต้นนี้ เมื่อนำหมวดคำหลักมาเปรียบเทียบกัน สามารถแสดงได้ดังตารางที่ 2-1

	กริยา	คุณศัพท์	วิเศษณ์	ตัวกำหนด	ตัวบอกปริมาณ	สรรพนาม	ลักษณนาม	นาม	บุพบท	สันธาน	คำลงท้าย	อุทาน	เครื่องหมาย
อุปกิต	กริยา	วิเศษณ์				สรรพนาม	นาม		บุพบท	สันธาน	วิเศษณ์	อุทาน	x
กำชัย	กริยา	วิเศษณ์				สรรพนาม	นาม		บุพบท	สันธาน	วิเศษณ์	อุทาน	x
บรรจบ	กริยา	วิเศษณ์				สรรพนาม	ลักษณนาม	นาม	บุพบท	สันธาน	วิเศษณ์	อุทาน	x
อุดม	กริยา	วิเศษณ์ / คำนำหน้ากริยา		นิยมลักษณ์	สังขยา	นาม				สันธาน	คำท้ายประโยค	นาม / กริยา	x
เรืองเดช	กริยา	คุณศัพท์	วิเศษณ์ / คำปฏิเสธ	คำกำหนด	คำนับ	สรรพนาม	ลักษณนาม	นาม	คำเชื่อม		คำลงท้าย	อุทาน	x
นววรรณ	กริยา		วิเศษณ์	คำขยาย		คำแทน	คำขยาย	นาม	คำเชื่อม		คำเสริม	คำเรียก-ร้อง	x
วิจินตน์	กริยา	คุณศัพท์	คำช่วยกริยา/คำปฏิเสธ/คำหน้าและหลังกริยา/คำกริยาวิเศษณ์/คำพิเศษ	คำบอกกำหนด	คำที่เกี่ยวกับจำนวน	สรรพนาม	ลักษณนาม	คำบอกเวลา นาม	บุพบท	คำเชื่อม	คำลงท้าย	x	x
วิรัช	verb	adverb	adverb/auxiliary/negator	determiner	(noun)	pronoun	classifier	noun	prep	conjunc	ending	interject	punctuation
อมรา	กริยา		วิเศษณ์	ตัวกำหนด	ตัวบอกปริมาณ	นาม			บุพบท	สันธาน	อนุภาค		x

ตารางที่ 2-1 ตารางเปรียบเทียบหมวดคำภาษาไทย

ตารางที่ 2-1 ได้แสดงเปรียบเทียบการจัดแบ่งหมวดคำหลักและชื่อชุดหมวดคำต่างๆโดยคร่าว ซึ่งจากตารางจะเห็นได้ว่า

(1) หมวดคำที่ทุกงานวิจัยจัดเป็นหมวดคำหลัก มีอยู่ 2 หมวด คือ นาม และ กริยา แม้ว่าคำจำกัดความจะมีความครอบคลุมแตกต่างกันไปในแต่ละงาน เช่น บางงานรวมเอาคำลักษณนามและคำสรรพนามไว้ในหมวดคำนาม แต่บางงานก็ไม่รวม บางงานรวมเอาคำคุณศัพท์ไว้ในหมวดคำกริยา แต่บางงานก็ไม่รวม ส่วนหมวดคำที่ทุกงานวิจัยจัดเป็นหมวดคำหลักแต่เรียกชื่อแตกต่างกัน คือ คำสันธานและคำวิเศษณ์ ซึ่งบางงานเรียกชื่อตามหน้าที่ของคำเป็น คำเชื่อมและคำขยาย ตามลำดับ

(2) หมวดคำที่บางงานวิจัยจัดแยกออกมาเป็นหมวดคำหลักแต่ก็มีบางงานที่จัดรวมไว้ในเป็นส่วนหนึ่งของหมวดคำอื่น ได้แก่ คำลักษณนาม, คำสรรพนาม, คำคุณศัพท์, ตัวกำหนด, ตัวบอกปริมาณ, บุพบท, คำลงท้าย, คำอุทาน

(3) หมวดคำที่มีงานวิจัยจำนวนน้อยได้แยกออกมาเป็นหมวดคำหลักด้วย ได้แก่ คำปฏิเสธ, คำช่วยกริยา, คำนำหน้ากริยา, คำหน้ากริยาและหลักกริยา, คำบอกเวลา

(4) หมวดคำเครื่องหมายนั้น งานวิจัยต่างๆไม่ค่อยได้กล่าวถึงไว้เพราะไม่ใช่คำในภาษา

ความแตกต่างที่ปรากฏนี้เป็นผลมาจากเกณฑ์การวิเคราะห์หมวดคำซึ่งแต่ละคนอาจใช้แตกต่างกัน จึงส่งผลให้ ชุดหมวดคำภาษาไทยมีการจัดแบ่งหมวดคำแบบต่างๆ บางส่วนก็สอดคล้องกัน บางส่วนก็แตกต่างกัน และสถานภาพของการจัดแบ่งหมวดคำภาษาไทยในปัจจุบันก็ยังไม่มีข้อยุติแน่ชัดว่าควรจะเป็นไปในรูปแบบใด แต่ละคนต่างก็เลือกใช้การแบ่งหมวดคำของนักไวยากรณ์หรือนักภาษาศาสตร์ไทยต่างกันไปตามที่ตนเห็นสมควร ทั้งยังมีผู้คิดค้นการจัดแบ่งหมวดคำภาษาไทยขึ้นใหม่อยู่เสมอ ผู้วิจัยเห็นว่า เนื่องจากประเด็นเรื่องการจัดแบ่งหมวดคำภาษาไทยยังคงเป็นข้อถกเถียงกันอยู่จนปัจจุบัน จึงน่าจะสามารถศึกษาและพัฒนาการจัดแบ่งหมวดคำในภาษาไทยเพิ่มเติมต่อไปได้ ดังนั้น วิทยานิพนธ์ฉบับนี้จึงเลือกทำการศึกษาและพัฒนาชุดหมวดคำภาษาไทยขึ้นมาใช้ในวิทยานิพนธ์นี้ด้วย

ในวิทยานิพนธ์ฉบับนี้ ผู้วิจัยได้เลือกใช้การแบ่งหมวดคำตามที่ อมรา ประสิทธิ์รัฐสินธุ์ (2543) ได้เสนอไว้ มาเป็นชุดหมวดคำต้นแบบสำหรับการพัฒนาชุดหมวดคำขึ้นใช้ในวิทยานิพนธ์ เนื่องจากเห็นว่าเป็นชุดหมวดคำที่ใช้เกณฑ์ทางวากยสัมพันธ์ซึ่งมีความคงที่ ไม่ลื่นไหล อันได้แก่ การปรากฏร่วม และ การกระจายของคำ และยังเป็นชุดหมวดคำที่ได้มาจากการวิเคราะห์คลังข้อมูลซึ่งเป็นการใช้ภาษาจริง อย่างไรก็ตาม ก็มีบางประเด็นที่ผู้วิจัยวิเคราะห์แตกต่างไปจาก

การวิเคราะห์ของอมรา ประสิทธิ์รัฐสินธุ์ และวิทยานิพนธ์เล่มนี้นอกจากจะได้นำเสนอชุดหมวดคำจากการวิเคราะห์ข้อมูลการใช้ภาษาจริงแล้ว ยังได้นำชุดหมวดคำที่วิเคราะห์ได้มาทดสอบกับคลังข้อมูลที่ใช้ในวิทยานิพนธ์นี้ แล้วพัฒนาเพิ่มเติมเพื่อใช้เป็นชุดหมวดคำที่เหมาะสม (รายละเอียดเรื่องการจัดหมวดคำที่ใช้ในวิทยานิพนธ์นี้แสดงไว้ในบทที่ 4)

2.3 วิธีการในการตัดคำภาษาไทยที่ผ่านมา

ที่ผ่านมาจนถึงปัจจุบัน งานด้านการตัดคำภาษาไทยได้รับการพัฒนาจากหน่วยงานวิจัยต่างๆทั้งของภาครัฐและภาคเอกชน โดยมีการพัฒนาแนวคิดและวิธีการต่างๆเพื่อใช้ในการตัดคำมาเป็นลำดับ แต่ละวิธีการต่างก็ให้ผลในด้านความถูกต้อง ความรวดเร็วของการทำงาน และปริมาณการใช้ทรัพยากรต่างๆ แตกต่างกัน วิธีการตัดคำภาษาไทยสามารถแบ่งได้เป็น 3 หลักการใหญ่ๆ คือ หลักการตัดคำโดยใช้กฎ (rule based approach), หลักการตัดคำโดยใช้พจนานุกรม (dictionary approach), และหลักการตัดคำโดยใช้คลังข้อมูล (corpus based approach) ดังนี้

2.3.1 หลักการตัดคำโดยใช้กฎ (rule based approach)

การตัดคำโดยใช้กฎเป็นความพยายามในขั้นเริ่มต้นของการพัฒนาระบบตัดคำภาษาไทย โดยใช้วิธีการตรวจสอบกฎเกณฑ์ทางอักขรวิธีที่กำหนดลักษณะของการประสมอักษร การเว้นวรรค และการขึ้นย่อหน้า เพื่อใช้เป็นเกณฑ์ในการบ่งชี้ขอบเขตของคำ ตัวอย่างเช่น

(1) การขึ้นย่อหน้าเป็นตัวบ่งชี้ถึงการสิ้นสุดความ
(2) การเว้นวรรคเป็นตัวบ่งชี้ถึงความเป็นไปได้ของการสิ้นสุดคำหรือประโยค
(3) กฎทางอักขรวิธีเป็นตัวบ่งชี้ถึงความเป็นไปได้ของการตัดคำในตำแหน่งนั้นๆ โดยได้แบ่งอักขระออกเป็น 5 กลุ่ม ได้แก่ (3.1) อักขระกลุ่ม Non-spacing character คือ รูปสระ วรรณยุกต์ และเครื่องหมาย ที่เมื่อประสมเข้ากับพยัญชนะแล้วไม่ทำให้มีการเคลื่อนขวาของตำแหน่ง อักขระในกลุ่มนี้ไม่สามารถปรากฏเดี่ยวได้ เช่น อั, อิ, อุ, อ็, อ่, อ์ (3.2) อักขระกลุ่มที่ต้องมีพยัญชนะตามเสมอ เช่น เ, แ, โ, ไ, ใ (3.3) อักขระกลุ่มที่ต้องมีพยัญชนะนำเสมอ เช่น ะ, า, อำ (3.4) อักขระกลุ่มที่เป็นตัวการันต์ที่มีทัณฑฆาตบังคับข้างบน เช่น ย์ เนื่องจากว่าตัวการันต์เป็นพยัญชนะสุดท้ายจึงไม่พิจารณาให้เป็นอักขระ

แรกของคำ (3.5) อักขระที่เหลือทั้งหมด วิรัช ศรเลิศล้ำวาณิช (2536) กล่าวว่า อักขระกลุ่มที่ 3.2, 3.3, 3.4 สามารถเป็นตัวจำกัด ไม่ให้มีการตัดคำในระหว่างอักขระนั้นๆ

วิรัช ศรเลิศล้ำวานิช (2536) กล่าวว่า กฎเกณฑ์ทั้ง 3 ลักษณะนี้สามารถใช้เป็นตัวช่วยในการบ่งชี้ขอบเขตของการพิจารณาในการเปรียบเทียบกับคำในพจนานุกรม

ผู้วิจัยเห็นว่า กฎทางอักขรวิธีที่นำมาใช้พิจารณาตัดคำข้างต้นไม่สามารถบ่งบอกตำแหน่งในการตัดคำได้อย่างถูกต้อง เนื่องจากอันที่จริงกฎดังกล่าวไม่ได้บ่งชี้ขอบเขตของคำ แต่เป็นตัวบ่งชี้ขอบเขตของพยางค์ ตัวอย่างเช่น “กระโดด” เป็นคำหนึ่งคำ ดังนั้นไม่สามารถตัดคำหลัง ะ (อักขระกลุ่มที่ 3.2) และไม่สามารถตัดคำหน้า โ (อักขระกลุ่มที่ 3.3) ได้ ส่วนตัวการันต์ (อักขระกลุ่มที่ 3.4) ก็ไม่ได้ปรากฏเป็นตำแหน่งจบคำเสมอไป ดังตัวอย่างเช่น กอล์ฟ เป็นต้น ดังนั้น กฎเกณฑ์ทางอักขรวิธีจึงเป็นกฎเกณฑ์ที่ใช้ในการตัดพยางค์มากกว่าที่จะเป็นกฎในการตัดคำ

พิสิทธิ์ พรมจันทร์ (2540: 10) กล่าวว่า วิธีการนี้มีข้อจำกัดมาก คือ ผลของการตัดคำอาจได้เป็นกลุ่มคำที่สามารถตัดคำแยกย่อยออกไปได้อีก ดังนั้นความถูกต้องของคำที่ได้จึงต่ำ แต่มีข้อดี คือ ความเร็วในการทำงานสูง ใช้ทรัพยากรน้อย ซึ่งวิธีการนี้ใช้ได้กับงานบางประเภทที่ไม่จำเป็นต้องตัดแยกคำให้ย่อยที่สุด เช่น งานจัดรูปแบบเอกสาร เป็นต้น

2.3.2 **หลักการตัดคำโดยใช้พจนานุกรม** (dictionary approach)

การตัดคำโดยใช้พจนานุกรมเป็นแนวคิดที่ได้รับการพัฒนาในยุคต่อมา โดยเก็บคำภาษาไทยไว้ในพจนานุกรม แล้วนำข้อความที่ป้อนเข้า (input) ไปค้นหาและเทียบสายอักขระกับคำในพจนานุกรม เพื่อหาว่าข้อความดังกล่าวควรตัดคำในบริเวณใด และประกอบด้วยคำใดบ้าง สมปรารถนา รัทยานนท์ (2535) ได้เสนอวิธีการใช้พจนานุกรมช่วยในการตัดคำภาษาไทย ใช้การค้นหาคำจากข้อความที่ป้อนเข้าไปเทียบกับคำในพจนานุกรมเพื่อนำแต่ละคำไปจัดเก็บไว้ในแถวลำดับหรืออะเรย์ (array) ชุดหนึ่ง โดยเริ่มค้นหาจากต้นข้อความ นำคำแรกที่เทียบเจอในพจนานุกรมไปจัดเก็บไว้ในอะเรย์ช่องที่หนึ่ง แล้วจึงตัดคำดังกล่าวออกไปจากข้อความ แล้วนำข้อความที่เหลือหลังจากตัดคำออกไป มาทำการเทียบคำกับพจนานุกรมเหมือนเดิม เพื่อนำแต่ละคำที่เทียบเจอไปจัดเก็บไว้ในอะเรย์ช่องต่อๆไปจนสิ้นสุดข้อความ (สมมติให้อะเรย์ช่องสุดท้ายเป็นช่องที่ n) แล้วจึงย้อนการทำงานกลับโดยทำการเทียบคำที่อยู่ในอะเรย์แต่ละช่องกับพจนานุกรม

ตั้งแต่ช่องที่ n, n-1, n-2,...ไปจนถึงอะเรย์ช่องที่ 1 เพื่อหาว่าในแต่ละช่องอะเรย์สามารถตัดคำในรูปแบบอื่นที่ต่างออกไปได้หรือไม่ หากอะเรย์ช่องใด (สมมติให้เป็นช่องที่ i ; i <= n) สามารถตัดคำในรูปแบบอื่นได้ก็จะตัดคำในรูปแบบใหม่นั้น แล้วจึงทำการเทียบคำต่อไปในช่องอะเรย์ถัดไป (ช่องที่ i+1) จนจบข้อความ การทำงานดังกล่าวจะทำย้อนกลับไปจนถึงอะเรย์ช่องที่ 1 แล้วจึงจบการทำงาน ดังนั้นผลที่ได้จะเป็นรูปแบบความเป็นไปได้ของการตัดคำทั้งหมดของข้อความที่ป้อนเข้า

ปัญหาที่พบในวิธีการตัดคำโดยใช้พจนานุกรม คือ เป็นไปไม่ได้ที่จะเก็บคำทุกคำในภาษาไทยลงในพจนานุกรมได้ โดยเฉพาะคำวิสามานยนาม เช่น ชื่อคน ชื่อสถานที่, ตัวเลข หรือคำที่เกิดขึ้นมาใหม่ (วิรัช ศรเลิศล้ำวาณิช, 2536) นอกจากนี้ วิธีการนี้จะสิ้นเปลืองทรัพยากรหน่วยความจำหลักค่อนข้างมาก เนื่องจากทั้งพจนานุกรมและอะเรย์ของคำที่ตัดได้จะเก็บไว้ในหน่วยความจำหลักทั้งหมด ประสิทธิภาพเชิงความถูกต้องของคำขึ้นอยู่กับปริมาณคำในพจนานุกรม ส่วนความเร็วขึ้นอยู่กับวิธีที่ใช้ในการค้นหาคำจากพจนานุกรม (พิสิทธิ์ พรมจันทร์, 2540: 11) และผลที่ได้จากการตัดคำวิธีนี้อาจมีได้มากกว่าหนึ่งทางเลือก ดังนั้นจึงต้องมีการเลือกทางเลือกของการตัดคำที่ถูกต้องต่อไปอีก

หลักการตัดคำโดยใช้พจนานุกรมนี้สามารถตัดคำได้ถูกต้องมากกว่าการใช้กฎ เพราะฉะนั้น จึงได้รับความนิยมและมีผู้พัฒนาวิธีการตัดคำภาษาไทยอื่นๆโดยใช้พจนานุกรมช่วยอีก เช่น วิธีการเทียบคำที่ยาวที่สุด (longest matching) และ วิธีการตัดคำให้ได้จำนวนคำและคำที่ไม่พบในพจนานุกรมน้อยที่สุด (maximal matching) เป็นต้น

2.3.2.1 **วิธีการเทียบคำที่ยาวที่สุด** (longest matching)

วิธีการเทียบคำที่ยาวที่สุดเป็นวิธีการตัดคำทางวิทยาการศึกษาสำนึก (heuristic) วิธีหนึ่งซึ่งต้องใช้พจนานุกรมช่วยรู้จำคำภาษาไทย โดยวิธีนี้จะทำการตรวจสอบหรือสแกนข้อความที่ป้อนเข้าจากซ้ายไปขวา นำไปเทียบกับพจนานุกรมดูว่า สายอักขระดังกล่าวเป็นหนึ่งคำหรือไม่ หากไม่พบว่าสายอักขระดังกล่าวสามารถเทียบเป็นคำได้ในพจนานุกรม ก็จะทำการลดความยาวของสายอักขระลงทีละตัว จนกว่าสายอักขระที่ตรวจสอบจะสามารถเทียบเป็นคำในพจนานุกรมได้ ก็จะทำเครื่องหมายเพื่อเป็นจุดย้อนกลับ จากนั้นก็จะเริ่มทำงานจากจุดย้อนกลับนั้นเพื่อตรวจสอบสายอักขระที่เหลือว่าจะสามารถตัดสายอักขระใดต่อไปให้เป็นคำได้ หากตัวเลือกในตอนแรกนี้สามารถทำให้ขั้นตอนวิธี (algorithm) ค้นหาคำที่เหลือได้ ตัวเลือกนี้ก็จะเป็นคำแรกของข้อความ

ได้จริง ไม่เช่นนั้นขั้นตอนวิธีก็จะกลับไปยังจุดย้อนกลับที่ทำเครื่องหมายไว้เพื่อแก้ไขคำแรกใหม่ จากนั้นก็จะเริ่มทำงานต่อไปโดยเริ่มจากจุดย้อนกลับ หากยังไม่สามารถเทียบสายอักขระกับคำในพจนานุกรมได้ก็จะทำการลดตัวอักษรลงทีละตัวจนกว่าจะเทียบคำในพจนานุกรมได้ และทำงานในรูปแบบนี้ต่อไปจนจบข้อความ ตัวอย่างเช่น ถ้าป้อนข้อความ “ความก้าวหน้าทางวิทยาศาสตร์มีบทบาทสำคัญ” เข้าไป ข้อความนี้เมื่อไม่สามารถเทียบคำกับพจนานุกรมได้ ก็จะลดลงเหลือ →“ความก้าวหน้าทางด้านวิทยาศาสตร์มีบทบาทสำคัั” → “ความก้าวหน้าทางด้านวิทยาศาสตร์มีบทบาทสำ” จนได้สายอักขระ “ความก้าวหน้า” ซึ่งสามารถเทียบคำในพจนานุกรมได้จึงตัดเป็นคำแรกของข้อความ และทำเครื่องหมายไว้เป็นจุดย้อนกลับ ผลของการตัดคำทั้งหมดจะเป็นดังตารางที่ 2-2

ส่วนของคำที่ยาวที่สุดที่ตัดได้	ส่วนที่เหลือหลังจุดย้อนกลับ
ความก้าวหน้า	ทางด้านวิทยาศาสตร์มีบทบาทสำคัญ
ทาง	ด้านวิทยาศาสตร์มีบทบาทสำคัญ
ด้าน	วิทยาศาสตร์มีบทบาทสำคัญ
วิทยาศาสตร์	มีบทบาทสำคัญ
มี	บทบาทสำคัญ
บทบาท	สำคัญ
สำคัญ	-

(วิรัช ศรเลิศล้ำวาณิช, 2536)

ตารางที่ 2-2 ผลการตัดคำด้วยวิธีเทียบคำที่ยาวที่สุด

วิธีการนี้ให้ความถูกต้องของการตัดคำได้ประมาณ 80% (วิรัช ศรเลิศล้ำวาณิช, 2536)

ข้อด้อยของวิธีการนี้เกิดจากลักษณะความพยายามที่จะตัดคำให้ได้ยาวที่สุดของอัลกอริธึม (greedy matching*) คือ การเลือกคำที่ยาวเกินไปตั้งแต่ต้นมีผลทำให้การตัดคำที่ตามมาผิดเพี้ยนไป ตัวอย่างเช่น ข้อความป้อนเข้า “ไปหามเหสี” (go to see the queen) จะตัดคำได้เป็น ไป (go)

* Surapant Meknavin and Boonserm Kijsirikul (2000)

หาม (carry) เห (deviate) สี (color) เสมอ เนื่องจากอัลกอริธึมจะพบคำว่า “หาม” ก่อนคำว่า “หา” เสมอ ทำให้การตัดคำสำหรับคำต่อๆไปผิดไปด้วย

2.3.2.2 วิธีการตัดคำให้ได้จำนวนคำและคำที่ไม่มีในพจนานุกรมน้อยที่สุด (maximal matching)

วิธีการตัดคำให้ได้จำนวนคำและคำที่ไม่มีในพจนานุกรม (unknown word) น้อยที่สุดก็เป็นวิธีการตัดคำทาง heuristic อีกวิธีหนึ่งที่ใช้พจนานุกรมช่วยรู้จำคำภาษาไทย วิธีการนี้พัฒนาโดย วิรัช ศรเลิศล้ำวาณิช (2536) เพื่อแก้ปัญหาที่ปรากฏในวิธีการเทียบคำที่ยาวที่สุด วิธีการนี้จะพัฒนาบนขั้นตอนวิธีของวิธีการเทียบคำที่ยาวที่สุด เริ่มจากการหาทางเลือกของรูปแบบการตัดคำทั้งหมดที่เป็นไปได้เสียก่อน โดยทำการย้อนกลับ (backtracking) ทีละคำหลังจากได้คำตอบจากวิธีการเทียบคำที่ยาวที่สุดแล้ว แล้วจึงเลือกทางเลือกที่มีจำนวนคำน้อยที่สุด Surapant Meknavin and Boonserm Kijsirikul (2000) กล่าวว่า การค้นหาทุกทางเลือกที่เป็นไปได้นี้ทำให้ต้องเสียเวลาในการคำนวณมาก แต่ก็สามารถลดเวลาลงได้โดยใช้โปรแกรมแบบพลวัติ (dynamic programming) ตัวอย่างเช่น หากป้อนข้อความ “ไปหามเหสี” เข้าไป ขั้นตอนวิธีนี้จะหาทางเลือกทั้งหมดของรูปแบบการตัดคำที่เป็นไปได้ ได้แก่

ไป (go) หาม (carry) เห (deviate) สี (color)
ไป (go) หา (see) [ม] เห (deviate) สี (color)
ไป (go) หา (see) มเหสี (queen)

โดยในขั้นแรก ขั้นตอนวิธีของวิธีการเทียบคำที่ยาวที่สุดจะได้คำตอบของการตัดคำเป็น “ไป-หาม-เห-สี” ก่อน หลังจากนั้นจึงเริ่มทำการย้อนกลับโดยเริ่มจากคำแรก พบว่า คำที่สอง “หาม” สามารถแบ่งได้เป็น “หา-ม” ได้ โดยเมื่อแบ่งเป็น “หา” แล้ว ทำให้สายอักขระที่เหลือ “มเหสี” สามารถตัดคำได้อีก 2 แบบ คือ “มเหสี” และ “[ม]-เห-สี” เมื่อทำการย้อนกลับกับทุกคำสิ้นสุดแล้ว ก็จะทำการคำนวณหาค่า cost ให้กับแต่ละทางเลือกที่เป็นไปได้ โดยบังคับให้มีการเกิดคำที่ไม่มีในพจนานุกรมน้อยที่สุด แล้วจัดเรียงผลลัพธ์โดยให้ทางเลือกที่น่าจะเป็นไปได้มากที่สุด (ค่า cost ต่ำที่สุด) มาเป็นอันดับแรก ตัวอย่างเช่น ข้อความที่ป้อนเข้าเป็น “กีฬาเป็นการออกกำลังกายอย่างหนึ่ง” จะได้ผลทางเลือกที่เป็นไปได้จัดเรียงตามค่า cost ดังนี้

ผลจากการทำการย้อนกลับ	ค่า cost
กีฬา/ เป็น/ การออกกำลังกาย/ อย่างหนึ่ง	4
กีฬา/ เป็นการ/ ออกกำลัง/ กาย/ อย่างหนึ่ง	5
กีฬา/ เป็นการ/ ออก/ กำลังกาย/ อย่างหนึ่ง	5
กีฬา/ เป็นการ/ ออกกำลัง/ กาย/ อย่าง/ หนึ่ง	6
กีฬา/ เป็นการ/ ออกกำลัง/ กา/ ยอ/ ย่าง/ หนึ่ง	7
กีฬา/ เป็นการ/ ออ/ [ก]/ กำลังกาย/ อย่างหนึ่ง	11
กีฬา/ เป็นการ/ ออกกำลัง/ กาย/ อย่า/ [ง]/ หนึ่ง	12
กีฬา/ เป็นการ/ ออ/ [ก]/ กำลังกาย/ อย่าง/ หนึ่ง	12
กีฬา/ เป็นการ/ ออ/ [ก]/ กำลังกาย/ อย่า/ [ง]/ หนึ่ง	18

(วิรัช ศรเลิศล้ำวาณิช, 2536)

ตารางที่ 2-3 ผลการตัดคำด้วยวิธีการตัดคำให้ได้จำนวนคำและคำที่ไม่มีในพจนานุกรมน้อยที่สุด

ขั้นตอนวิธีนี้จะเลือกทางเลือกการตัดคำที่มีค่า cost ต่ำที่สุด ซึ่งก็ขึ้นอยู่กับจำนวนคำที่ตัดอกมาได้และจำนวนคำที่ไม่มีในพจนานุกรม อย่างไรก็ตาม หากมีทางเลือกที่มีค่า cost เท่ากันและมีจำนวนคำเท่ากันมากกว่าหนึ่งทางเลือก ขั้นตอนวิธีนี้จะไม่สามารถตัดสินได้ว่าจะเลือกทางเลือกไหน ดังนั้นจึงต้องใช้ heuristic อื่นเข้ามาช่วย โดยส่วนใหญ่มักใช้วิธีการเทียบคำที่ยาวที่สุดเข้ามาช่วย ตัวอย่างเช่น ข้อความป้อนเข้า “ตากลม” จะมี 2 ทางเลือกที่มีจำนวนคำน้อยที่สุดเท่ากัน คือ

ตาก (expose) ลม (wind)
ตา (eye) กลม (round)

อัลกอริธึมจะเลือกทางเลือกแรกเป็นคำตอบ เนื่องจากคำแรกของทางเลือกแรก “ตาก” มีความยาวมากกว่าคำแรกของทางเลือกที่สอง “ตา”

Surapant Meknavin and Boonserm Kijsirikul (2000) กล่าวว่า วิธีการนี้มักจะมีประสิทธิภาพในการตัดคำสูงกว่าวิธีการเทียบคำที่ยาวที่สุด เนื่องจากวิธีการนี้ได้แก้ไขข้อจำกัดในการพยายามเลือกคำให้ยาวที่สุดตั้งแต่ต้น โดยใช้วิธีพิจารณาทางเลือกของการตัดคำที่เป็นไปได้ทั้งหมดก่อนที่จะเลือกการตัดคำ

2.3.3 **หลักการตัดคำโดยใช้คลังข้อมูล** (corpus based approach)

หลักการตัดคำโดยใช้คลังข้อมูลเป็นแนวคิดที่ได้รับการพัฒนาในยุคหลังๆซึ่งเป็นยุคที่มีผู้สนใจนำวิธีการทางสถิติ (statistical techniques) มาใช้ในการประมวลผลภาษาธรรมชาติมากขึ้น โดยใช้คลังข้อมูลทางภาษา (corpus) เป็นฐานความรู้สำหรับเก็บค่าความถี่ที่ใช้ในการตัดคำ

2.3.3.1 **วิธีการตัดคำโดยอาศัยค่าความน่าจะเป็น** (probabilistic word segmentation)

งานวิจัยของอัศนีย์ ก่อตระกูล (Asanee Kawtrakul et al., 1995 cited in Surapant Meknavin and Boonserm Kijsirikul, 2000) และงานของสุรพันธ์ เมฆนาวิน (Surapant Meknavin, 1995 cited in Surapant Meknavin and Boonserm Kijsirikul, 2000) รวมทั้งงานของ บุญเสริม กิจศิริกุล (2541) ได้ใช้แบบจำลองไตรแกรมของคำ (word trigram model) ร่วมกับแบบจำลองไตรแกรมของหมวดคำ (part-of-speech trigram model) เพื่อหารูปแบบการตัดคำที่เป็นไปได้มากที่สุด และลำดับหมวดคำ (tag sequence) ที่เป็นไปได้มากที่สุด ไปพร้อมๆกัน วิธีการนี้ต้องใช้คลังข้อมูลที่มีการตัดคำและกำกับหมวดคำเตรียมเอาไว้แล้ว โดยปัญหาการตัดคำภาษาไทยสามารถแสดงด้วยสมการที่ 2-1 ดังต่อไปนี้

(2-1)

$$
\begin{aligned}
\arg\max_{W_{1,n}} P(W_{1,n} \mid C_{1,m}) &= \arg\max_{W_{1,n}} \frac{P(C_{1,m} \mid W_{1,n}) * P(W_{1,n})}{P(C_{1,m})} \\
&= \arg\max_{W_{1,n}} P(W_{1,n}) \\
&= \arg\max_{W_{1,n}} \sum_{T_{1,n}} P(W_{1,n}, T_{1,n})
\end{aligned}
$$

(บุญเสริม กิจศิริกุล, 2541: 4)

กำหนดให้ $C_{1,m}$ หมายถึง สายอักขระที่ป้อนเข้าไปตั้งแต่ตัวที่ 1 ถึงตัวที่ m

$W_{1,n}$ หมายถึง สายคำที่สามารถตัดออกมาได้ ตั้งแต่คำแรกถึงคำที่ n

$T_{1,n}$ หมายถึง สายหมวดคำที่กำกับอยู่กับแต่ละคำ ทั้งหมด n คำ

ปัญหาของการตัดคำที่ คือ ต้องการหาสายคำ $W_{1,n}$ ที่ทำให้ค่าความน่าจะเป็นของ $P(W_{1,n})$ มีค่าสูงที่สุด โดยเมื่อนำหมวดคำเข้ามาช่วยคำนวณด้วยแล้ว จะสามารถหาค่า $P(W_{1,n})$ ได้ดังสมการ

สมการดังกล่าว มีความหมายว่า จากสายอักขระ $C_{1,m}$ ที่กำหนดให้ซึ่งเป็นข้อความที่ป้อนเข้าไป เราต้องการแบ่งสายอักขระนี้ออกเป็นคำ $W_1, W_2,\ldots,W_n$ (ซึ่งเขียนสั้นๆได้ว่า $W_{1,n}$) และมีหมวดคำเป็น $T_1, T_2,\ldots,T_n$ (หรือ $T_{1,n}$) เราต้องการตัดคำให้ได้ค่าของ $P(W_{1,n} \mid C_{1,m})$ ที่สูงที่สุด ซึ่งสามารถคำนวณได้จาก $P(W_{1,n}, C_{1,m}) / P(C_{1,m})$ และเราสามารถแปลงการหาค่าความน่าจะเป็นแบบมีเงื่อนไข (conditional probability) ตรงนี้ได้โดยนำกฎของเบย์ส (Bayes' rule) มาใช้ ดังนั้นจึงสามารถแปลงสมการได้เป็น $P(C_{1,m} \mid W_{1,n}) * P(W_{1,n}) / P(C_{1,m})$ (ดังแสดงในสมการ) ซึ่งสามารถลดเหลือเพียงการหาค่า $P(W_{1,n})$ (ดังแสดงในสมการ) เนื่องจาก $P(C_{1,m} \mid W_{1,n})$ มีค่าเท่ากับ 1 และ $P(C_{1,m})$ สามารถละไปได้ เนื่องจากเป็นตัวหารสำหรับทุกทางเลือกของการตัดคำที่จะนำมาเปรียบเทียบกัน จากนั้น จึงได้นำหมวดคำมาช่วยพิจารณาด้วย จึงได้สมการเป็นการหาค่า $\sum T_{1,n} P(W_{1,n}, T_{1,n})$ (ดังแสดงในสมการ) ซึ่งสามารถแปลงสมการได้เป็น $P(W_{1,n} \mid T_{1,n}) * P(T_{1,n})$ แล้วจึงประยุกต์ใช้แนวคิดของแบบจำลองไตรแกรม (trigram model) ที่มีสมมติฐานว่า:

(1) คำหนึ่งๆสามารถปรากฏ ณ ตำแหน่งใดๆในประโยคได้ โดยไม่ขึ้นกับคำหรือหมวดคำที่อยู่ก่อนหน้าหรือตามหลัง กล่าวคือ $P(W_{1,n} \mid T_{1,n})$ มีค่าประมาณเท่ากับ ผลคูณรวมของ $P(W_i \mid T_i)$ ของทุกคำ

(2) ความน่าจะเป็นที่หมวดคำหนึ่งจะปรากฏ ณ ตำแหน่งใดๆในประโยคจะขึ้นอยู่กับหมวดคำที่ปรากฏก่อนหน้า 2 หมวดคำเท่านั้น (trigram model) กล่าวคือ $P(T_{1,n})$ คำนวณแบบประมาณค่าได้จาก $P(T_i \mid T_{i-1}, T_{i-2})$

ดังนั้น สมการที่ 2-1 จะสามารถคำนวณโดยการประมาณค่าตามแบบจำลองไตรแกรมได้ดังสมการที่ 2-2

(2-2)

$$\underset{W_{1,n}}{\arg\max} \sum_{T_{1,n}} \prod_{I=1,n} P(W_i \mid T_i) * P(T_i \mid T_{i-1}, T_{i-2})$$

(บุญเสริม กิจศิริกุล, 2541: 5)

จากคลังข้อมูลภาษาไทยที่มีอยู่ เราสามารถหาค่าของ $P(W_i \mid T_i)$ ได้โดยนับจำนวนของคำ W_i ที่มีหมวดคำเป็น T_i หารด้วยจำนวนของ T_i ที่เป็นหมวดคำของคำใดๆ (จำนวนของ T_i ที่ปรากฏทั้งหมดในคลังข้อมูล) ส่วน $P(T_i \mid T_{i-1}, T_{i-2})$ เราสามารถหาได้โดยนับจำนวนหมวดคำ T_i ที่มีหมวดคำ T_{i-2} และ T_{i-1} นำหน้า หารด้วยจำนวนสายหมวดคำ T_{i-2} และ T_{i-1} ที่ปรากฏติดกันทั้งหมดในคลังข้อมูล ขั้นตอนการคำนวณค่าความน่าจะเป็นที่ใช้ในการตัดคำวิธีนี้สามารถสรุปได้ดังตารางที่ 2-5

1) คำนวณ conditional prob	$\arg\max_{W_{1,n}} P(W_{1,n} \mid C_{1,m})$	=	$\arg\max_{W_{1,n}} P(W_{1,n}, C_{1,m}) / P(C_{1,m})$
2) แปลงสมการโดยใช้ Bay's rule		=	$\arg\max_{W_{1,n}} P(C_{1,m} \mid W_{1,n}) * P(W_{1,n}) / P(C_{1,m})$
3) $P(C_{1,m} \mid W_{1,n})$ มีค่าเท่ากับ 1 และ $P(C_{1,m})$ เป็นตัวหารที่เท่ากันทุกทางเลือกสามารถตัดทิ้งได้		=	$\arg\max_{W_{1,n}} P(W_{1,n})$
4) นำ T มาช่วย		=	$\arg\max_{W_{1,n}} \sum_{T_{1,n}} P(W_{1,n}, T_{1,n})$
5) แปลง joint prob เป็น conditional prob		=	$\arg\max_{W_{1,n}} \sum_{T_{1,n}} P(W_{1,n} \mid T_{1,n}) * P(T_{1,n})$
6) นำแนวคิดไตรแกรมมาใช้	(1) lexical generation prob: คำหนึ่งๆสามารถปรากฏ ณ ตำแหน่งใดๆในประโยคได้โดยไม่ขึ้นกับสิ่งอื่น $P(W_{1,n} \mid T_{1,n}) \cong \prod_{i=1..n} P(W_i \mid T_i)$		
	(2) tag sequence prob: ความน่าจะเป็นที่หมวดคำหนึ่งๆจะปรากฏ ณ ตำแหน่งใดๆในประโยคจะขึ้นอยู่กับหมวดคำที่ปรากฏก่อนหน้า 2 หมวดคำเท่านั้น (trigram model) $P(T_{1,n}) \cong \prod_{i=1..n} P(T_i \mid T_{i-1}, T_{i-2})$		
7)		=	$\arg\max_{W_{1,n}} \sum_{T_{1,n}} \prod_{i=1..n} P(W_i \mid T_i) * P(T_i \mid T_{i-1}, T_{i-2})$

ตารางที่ 2-4 ขั้นตอนการตัดคำโดยอาศัยค่าความน่าจะเป็นตามแบบจำลองไตรแกรม

อย่างไรก็ตาม หากพิจารณาตามสมการที่ใช้ดังกล่าว จะพบว่าการนำหมวดคำมาคำนวณไม่ได้มีผลโดยตรงใดๆต่อการตัดคำ เพราะเป็นการหาเฉพาะสายคำที่ให้ค่าความน่าจะเป็นสูงที่สุด และค่าความน่าจะเป็นนี้ได้มาจากการนำค่าความน่าจะเป็นทั้งหมดของสายคำหนึ่งๆซึ่งมีการกำกับสายหมวดคำแบบต่างๆกันมารวมกัน ซึ่งจะเท่ากับการหาค่าความน่าจะเป็นของสายคำนั้นๆ

โดยไม่พิจารณาเรื่องหมวดคำ ตัวอย่างเช่น จากสายอักขระ “เขาเดินตากลมไปหามเหสี” ค่าความน่าจะเป็นของการตัดคำและกำกับหมวดคำที่เป็นไปได้ทั้งหมดสมมติว่าเป็นดังตารางที่ 2-5

การตัดคำและหมวดคำแบบต่างๆ	ค่าความน่าจะเป็นของสายคำและสายหมวดคำ	ค่าความน่าจะเป็นรวมของสายคำ
เขา/PN เดิน/VI ตาก/VT ลม/N ไป/VI หา/VT มเหสี/N	0.17	
เขา/PN เดิน/VI ตาก/NPrp ลม/N ไป/VI หา/VT มเหสี/N	0.04	0.3
เขา/N เดิน/VI ตาก/VT ลม/N ไป/VI หา/VT มเหสี/N	0.06	
เขา/N เดิน/VI ตาก/NPrp ลม/N ไป/VI หา/VT มเหสี/N	0.03	
เขา/PN เดิน/VI ตาก/VT ลม/N ไป/VI หาม/VT เห/VI สี/N	0.08	
เขา/PN เดิน/VI ตาก/VT ลม/N ไป/VI หาม/VT เห/VI สี/VT	0.1	
เขา/PN เดิน/VI ตาก/NPrp ลม/N ไป/VI หาม/VT เห/VI สี/N	0.05	
เขา/PN เดิน/VI ตาก/NPrp ลม/N ไป/VI หาม/VT เห/VI สี/VT	0.03	0.4
เขา/N เดิน/VI ตาก/VT ลม/N ไป/VI หาม/VT เห/VI สี/N	0.09	
เขา/N เดิน/VI ตาก/VT ลม/N ไป/VI หาม/VT เห/VI สี/VT	0.025	
เขา/N เดิน/VI ตาก/NPrp ลม/N ไป/VI หาม/VT เห/VI สี/N	0.02	
เขา/N เดิน/VI ตาก/NPrp ลม/N ไป/VI หาม/VT เห/VI สี/VT	0.005	
เขา/PN เดิน/VI ตา/N กลม/Adv ไป/VI หา/VT มเหสี/N	0.055	0.1
เขา/N เดิน/VI ตา/N กลม/Adv ไป/VI หา/VT มเหสี/N	0.045	
เขา/PN เดิน/VI ตา/N กลม/Adv ไป/VI หาม/VT เห/VI สี/N	0.07	
เขา/PN เดิน/VI ตา/N กลม/Adv ไป/VI หาม/VT เห/VI สี/VT	0.04	0.2
เขา/N เดิน/VI ตา/N กลม/Adv ไป/VI หาม/VT เห/VI สี/N	0.07	
เขา/N เดิน/VI ตา/N กลม/Adv ไป/VI หาม/VT เห/VI สี/VT	0.02	
รวม18 รูปแบบสายคำและสายหมวดคำ จัดเป็น 4 รูปแบบสายคำ	1.0	1.0

ตารางที่ 2-5 ค่าความน่าจะเป็นของสายคำและสายหมวดคำของประโยคตัวอย่าง

จากตารางที่ 2-5 ซึ่งแสดงรูปแบบการตัดคำและกำกับหมวดคำทั้งหมด 18 รูปแบบ ซึ่งมีการตัดคำ 4 รูปแบบ หากพิจารณาตามสมการที่ใช้ ซึ่งต้องการหาสายคำที่ให้ค่าความน่าจะเป็นสูงสุดเท่านั้น จะได้คำตอบเป็นการตัดคำแบบที่ 2 (จาก 4 รูปแบบ) ซึ่งมีค่าความน่าจะเป็นของสายคำเท่ากับ 0.4 ที่คำนวณได้จากการนำค่าความน่าจะเป็นของสายหมวดคำแบบต่างๆรวมกัน ทั้งที่ในความจริงแล้ว คำตอบที่ถูกต้องน่าจะเป็นการตัดคำแบบที่ 1 เพราะมีรูปแบบสายคำพร้อมสายหมวดคำที่ให้ค่าความน่าจะเป็นสูงสุด (จากทั้งหมด 18 รูปแบบ) คือ รูปแบบของสายคำและสายหมวดคำแบบที่ 1 (แสดงไว้ด้วยตัวหนา) แต่เนื่องจากเมื่อรวมค่าความน่าจะเป็นของสายหมวดคำแบบต่างๆในรูปแบบสายคำแบบที่ 1 แล้ว ได้เท่ากับ 0.3 ซึ่งน้อยกว่าค่าความน่าจะเป็นของรูปแบบสายคำแบบที่ 2 ทำให้คำตอบที่ถูกต้องไม่ได้ถูกเลือก ซึ่งก็เท่ากับเป็นการหาค่าความน่าจะเป็นของสายคำเท่านั้นโดยไม่ได้พิจารณาเรื่องหมวดคำ

ในประเด็นนี้ Surapant Meknavin and Boonserm Kijsirikul (2000) ได้กล่าวไว้ว่า มีผู้แย้งว่าควรจะคำนวณหาการตัดคำที่มีค่าความน่าจะเป็นร่วมของสายคำและสายหมวดคำ (joint probability of word sequence and tag sequence) ที่สูงที่สุด คือ $P(W_i , T_i)$ แทนที่จะเป็นเพียงแค่ $P(W_i)$ เท่านั้น โดยให้เหตุผลว่า เมื่อเวลามนุษย์คิดถึงประโยคหนึ่งๆ เขาจะคิดถึงในความหมายเดียวเท่านั้น ซึ่งสุรพันธ์ เมฆนาวิน และบุญเสริม กิจศิริกุล กล่าวว่า ได้ทำการคำนวณทั้ง 2 แบบแล้ว พบว่าไม่มีความแตกต่างที่มีนัยสำคัญระหว่าง 2 แบบนี้ แต่อย่างไรก็ตาม ผู้วิจัยไม่พบรายละเอียดใดๆของการคำนวณแบบที่หาค่าความน่าจะเป็นร่วมของสายคำและสายหมวดคำในบทความดังกล่าว ผู้วิจัยคาดว่า ผลจากการคำนวณ 2 แบบนี้น่าจะให้ผลลัพธ์แตกต่างกัน หากมองกระบวนการตัดคำเป็นส่วนเดียวกับกระบวนการกำกับหมวดคำ ไม่ใช่กระบวนการที่ต้องกระทำก่อนการกำกับหมวดคำ

2.3.3.2 วิธีการตัดคำโดยอาศัยคุณลักษณะของคำ (feature-based word segmentation)

วิธีการตัดคำโดยอาศัยคุณลักษณะของคำได้รับการพัฒนาโดยสุรพันธ์ เมฆนาวิน และคณะ (Surapant Meknavin et al., 1997) สุรพันธ์ได้นำวิธีการอาศัยคุณลักษณะ (feature-based approach) ซึ่งเป็นวิธีการที่ใช้ในงานด้านอื่นๆของการประมวลผลภาษาธรรมชาติหลายๆด้าน มาใช้เพื่อช่วยแก้ปัญหาในการตัดคำภาษาไทย เนื่องจากเห็นว่า วิธีการตัดคำโดยใช้ค่าความน่าจะเป็นได้นำเพียงข้อมูลเรื่องหมวดคำในปริบทที่จำกัดมาพิจารณาเท่านั้น อีกทั้งไม่ได้นำเรื่องการ

ปรากฏร่วมกันของคำที่ไม่ได้เกิดต่อเนื่องกัน (unordered long distance specific word collocations) มาพิจารณาด้วย ซึ่งหากจะพัฒนาวิธีการตัดคำโดยอาศัยค่าความน่าจะเป็นให้สามารถใช้ประโยชน์จากปริบทได้มากขึ้นก็จำเป็นต้องใช้คลังข้อมูลที่มีขนาดใหญ่มากและสิ้นเปลืองทรัพยากรหน่วยความจำในการเก็บตารางค่าสถิติมาก จึงเสนอวิธีการตัดคำโดยอาศัยคุณลักษณะของคำ

วิธีการตัดคำโดยอาศัยคุณลักษณะของคำเป็นวิธีการแบบผสม (hybrid approach) ที่นำคุณลักษณะ (feature) ของคำที่เกิดความกำกวมในการตัดคำเข้ามาช่วยเลือกการตัดคำที่ถูกต้องได้ โดย Surapant Meknavin et al (1997) ได้นำคุณลักษณะของคำมาช่วยเลือกการตัดคำที่ถูกต้องหลังจากได้ผลทางเลือกการตัดคำที่เป็นไปได้ทั้งหมดจากวิธีการตัดคำให้ได้จำนวนคำและคำที่ไม่มีในพจนานุกรมน้อยที่สุด (maximal matching) และกำกับหมวดคำให้กับทุกทางเลือกแล้ว คุณลักษณะอาจเป็นอะไรก็ตามที่สามารถบ่งชี้ลักษณะปริบทข้างเคียงของคำที่เกิดความกำกวม เช่น การเกิดร่วมกันของคำ (collocation), คำปริบท (context word) เป็นต้น แนวคิดของวิธีการนี้คือ พยายามเรียนรู้คุณลักษณะต่างๆจากคลังข้อมูลที่บ่งชี้ลักษณะของปริบทที่คำหนึ่งๆสามารถจะปรากฏได้ ซึ่งทำให้สามารถใช้คุณลักษณะต่างๆนั้นร่วมกันเพื่อแก้ปัญหาความกำกวมในการตัดคำ Surapant Meknavin et al. (1997) ได้เลือกใช้ RIPPER และ WINNOW ซึ่งเป็นอัลกอริธึมเรียนรู้ (learning algorithm) เพื่อเลือกคุณลักษณะต่างๆที่จะนำมาใช้ในการตัดคำภาษาไทยโดยอัตโนมัติจากคลังข้อมูลและเพื่อใช้คุณลักษณะเหล่านั้นร่วมกันแก้ปัญหาความกำกวม ซึ่งงานวิจัยดังกล่าวได้ลองใช้คุณลักษณะ 2 ประเภท ตามที่ Golding (1995 cited in Surapant Meknavin et.al, 1997) เคยทำไว้:

(1) การเกิดร่วมกันของคำ (collocation) เพื่อพิจารณารูปแบบภายใน L คำหรือหมวดคำที่เกิดต่อเนื่องกับคำเป้าหมาย (target word = คำที่เกิดความกำกวมที่เราสนใจ)
(2) คำปริบท (context word) เพื่อพิจารณาคำเฉพาะ (particular word) ที่ปรากฏภายในขอบเขต + / - K คำจากคำเป้าหมาย

ตัวอย่างเช่น ในการแก้ปัญหาความกำกวมของสายอักขระ “มากว่า” สามารถทำได้โดยหา เซตสับสน (confusion set =C) ซึ่งก็คือ รูปแบบความเป็นไปได้ทั้งหมดของการตัดคำ ได้เป็น C = {มาก ว่า, มา กว่า} จากนั้นจึงใช้ RIPPER และ WINNOW เพื่อเรียนรู้คุณลักษณะจากคลังข้อมูลฝึกสอนที่จะ

สามารถแยกการตัดคำทั้งสองแบบออกจากกันได้ ตัวอย่างของคุณลักษณะที่ใช้แก้ความกำกวมของ “มากว่า” ได้แก่

(1) มากว่า number (collocation ที่บ่งปริบทว่าควรตัดคำเป็น มา กว่า)

(2) พูด within –10 words (context word ที่บ่งปริบทว่าควรตัดคำเป็น มาก ว่า)

ผู้วิจัยเห็นว่า วิธีการตัดคำโดยอาศัยคุณลักษณะของคำเป็นรูปแบบหนึ่งของความพยายามที่จะใช้ประโยชน์จากปริบทในขอบเขตที่กว้างขึ้นกว่าที่ใช้ในวิธีการตัดคำโดยอาศัยค่าความน่าจะเป็น ซึ่งอันที่จริงสามารถเพิ่มข้อมูลปริบทเพื่อใช้แก้ปัญหาความกำกวมในการตัดคำได้หลายวิธี เช่น การใช้ข้อมูลโครงสร้างระดับวลี เป็นต้น

2.3.4 การเลือกประโยคที่ถูกต้องหลังการตัดคำ

วิธีการตัดคำบางวิธีจะให้ผลลัพธ์เป็นรูปแบบการตัดคำทั้งหมดที่เป็นไปได้ หรือมีทางเลือกของการตัดคำที่ได้มากกว่า 1 ทางเลือก ดังนั้นจึงจำเป็นต้องมีการเลือกรูปแบบการตัดคำที่คาดว่าถูกต้องที่สุด งานวิจัยของ รัตติกร วรากูลศิริพันธุ์ และคณะ (2538ข) เสนอการใช้ข้อมูลความถี่ของการใช้คำภาษาไทย (word usage frequency) เข้ามาช่วยในการเลือกประโยคที่ถูกต้องหลังการตัดคำ โดยคำนวณหาค่าความน่าจะเป็นของการนำคำนั้นไปใช้ในภาษาไทย (probability of usage) ซึ่งอาศัยแหล่งข้อมูลภาษาไทยต่างๆ เพื่อเป็นฐานข้อมูลตัวอย่าง การคำนวณหาค่าความน่าจะเป็นของการนำคำนั้นไปใช้ในภาษาไทยสามารถหาได้จากสมการที่ 2-3

(2-3)

$$Pu(W_1) = f_1 / N$$

โดยกำหนดให้

$Pu(W_1)$ หมายถึง ความน่าจะเป็นที่คำภาษาไทย W_1 จะถูกใช้ในภาษา

f_1 หมายถึง ความถี่หรือจำนวนครั้งของการใช้คำภาษาไทย W_1 ที่ปรากฏในคลังข้อมูล

N หมายถึง จำนวนคำทั้งหมดที่ปรากฏในคลังข้อมูล (คือ ความถี่หรือจำนวนครั้งที่ปรากฏของทุกคำในคลังข้อมูลรวมกัน)

การเลือกรูปแบบการตัดคำที่ถูกต้องสามารถทำได้โดย เปรียบเทียบค่า Pu ของคำที่อยู่ในลำดับเดียวกันในทางเลือกการตัดคำต่างๆที่ได้ แล้วเลือกทางเลือกที่มีค่าสูงกว่า หากคำในลำดับต้นมีค่าความน่าจะเป็นของการใช้เท่ากัน ก็เปรียบเทียบคำในลำดับถัดไปเรื่อยๆ

ผู้วิจัยเห็นว่า การเลือกประโยคที่ถูกต้องหลังการตัดคำสามารถนำไปใช้เสริมกับวิธีการตัดคำที่ได้ผลลัพธ์มากกว่า 1 ทางเลือก อันได้แก่ หลักการตัดคำโดยใช้พจนานุกรม ส่วนการตัดคำวิธีอื่นซึ่งได้ผลลัพธ์การตัดคำรูปแบบเดียวก็ไม่จำเป็นต้องใช้วิธีการนี้อีก วิธีการเลือกประโยคที่ถูกต้องแบบข้างต้นนี้จะอาศัยความน่าจะเป็นแบบ simple probabilistic model คือ หาความน่าจะเป็นที่คำจะถูกใช้ในภาษาโดยไม่ได้นำปริบทในการปรากฏของคำนั้นมาช่วยในการคำนวณ และวิธีนี้ให้ค่าน้ำหนักแก่คำที่ปรากฏต้นประโยคมากกว่าคำที่ปรากฏท้ายประโยค เนื่องจากไม่ได้เปรียบเทียบค่าความน่าจะเป็นของทั้งสาย แต่เลือกเปรียบเทียบจากค่าความน่าจะเป็นของคำที่อยู่ต้นประโยคก่อน

2.4 วิธีการในการกำกับหมวดคำภาษาไทยที่ผ่านมา

การกำกับหมวดคำ คือ การกำหนดหมวดคำให้แก่ข้อความป้อนเข้า โดยคอมพิวเตอร์จะเป็นผู้กำกับหมวดคำ (part-of-speech tagger หรือ POS tagger) ให้กับคำแต่ละคำในข้อความที่ป้อนเข้ามา โดยหมวดคำที่ใช้กำกับอาจแตกต่างกันในแต่ละงาน ซึ่งขึ้นอยู่กับว่า ผู้พัฒนาโปรแกรมกำกับหมวดคำจะกำหนดให้มีหมวดคำใดบ้างในงานของตน การกำกับหมวดคำมีประโยชน์ช่วยให้การแจงส่วนประโยค (parsing) สามารถทำได้สะดวกขึ้นและถูกต้องยิ่งขึ้น เนื่องจากว่า แต่เดิมความกำกวมจากการที่รูปคำหนึ่งๆเป็นได้หลายหมวดคำทำให้การแจงส่วนประโยคประสบปัญหา (บุญเสริม กิจศิริกุล, 2541: 9) วิธีการกำกับหมวดคำสามารถแบ่งได้เป็น 2 หลักการใหญ่ๆ คือ หลักการกำกับหมวดคำโดยใช้กฎ (rule based approach) และ หลักการกำกับหมวดคำโดยใช้แบบจำลองไตรแกรม (trigram model approach)

2.4.1 หลักการกำกับหมวดคำโดยใช้กฎ (rule based approach)

การกำกับหมวดคำโดยใช้กฎสามารถทำได้โดยเขียนกฎทางภาษาขึ้นมาใช้กำกับหมวดคำให้กับคำหนึ่งๆ โดยพิจารณาจากรูปแบบและหมวดคำที่อยู่ก่อนหน้าและหลัง เช่น กฎ “คำปัจจุบันจะไม่เป็นคำกริยา ถ้าคำที่อยู่ก่อนหน้าเป็นหมวดคำบ่งชี้ (determiner)” เป็นต้น (บุญเสริม กิจศิริกุล,

2541: 9) งานวิจัยในช่วงต้นๆ จะใช้คนเขียนกฎทางภาษาขึ้นมา หลักการกำกับหมวดคำโดยใช้กฎมีข้อดีในด้านความเร็วของการทำงาน เนื่องจากกฎที่ได้มีขนาดเล็ก แต่มีข้อด้อยคือ ต้องอาศัยนักภาษาศาสตร์เพื่อช่วยเขียนกฎหรือรูปแบบ และการเขียนกฎที่สมบูรณ์ทำได้ยาก ต่อมา Brill (1993) เสนอวิธีการสร้างรูปแบบ (template) ของกฎขึ้นมาโดยไม่ระบุรายละเอียด จากนั้นให้โปรแกรมแก้ไขกฎให้ดีขึ้นโดยอาศัยการเรียนรู้จากข้อผิดพลาดในรอบก่อนหน้าเพื่อแก้ไขกฎให้ถูกต้องมากยิ่งขึ้น

2.4.2 หลักการกำกับหมวดคำโดยใช้แบบจำลองไตรแกรม (trigram model approach)

หลักการกำกับหมวดคำโดยใช้แบบจำลองไตรแกรมได้นำ hidden Markov model (HMM) ซึ่งเป็นวิธีการที่ใช้กันแพร่หลายในงานด้านการรู้จำเสียงมาประยุกต์ใช้กับการกำกับหมวดคำ โดยรวบรวมค่าสถิติของความน่าจะเป็นของคำและหมวดคำต่างๆจากคลังข้อมูลไว้สำหรับใช้ในการคำนวณค่าความน่าจะเป็นในการกำกับหมวดคำ ปัญหาของการกำกับหมวดคำสามารถแสดงได้ ดังสมการที่ 2-4

(2-4)

$$
\begin{aligned}
\arg\max_{T_{1,n}} P(T_{1,n} \mid W_{1,n}) &= \arg\max_{T_{1,n}} \frac{P(W_{1,n}, T_{1,n})}{P(W_{1,n})} \\
&= \arg\max_{T_{1,n}} P(W_{1,n}, T_{1,n}) \\
&= \arg\max_{T_{1,n}} \prod_{i=1..n} P(W_i \mid T_i) * P(T_i \mid T_{i-1}, T_{i-2})
\end{aligned}
$$

(บุญเสริม กิจศิริกุล, 2541: 10)

กำหนดให้

$W_{1,n}$ หมายถึง สายคำที่สามารถตัดออกมาได้ ตั้งแต่คำแรกถึงคำที่ n

$T_{1,n}$ หมายถึง สายหมวดคำที่อาจเป็นไปได้ซึ่งกำกับอยู่กับคำแต่ละคำทั้งหมด n คำ

สมการที่ 2-4 มีความหมายว่า จากสายคำ $W_{1,n}$ ที่กำหนดให้ซึ่งเป็นข้อความที่ป้อนเข้าไปนี้ (ข้อความที่ป้อนเข้าไปเป็นข้อความที่มีการตัดคำไว้แล้ว) ต้องกำกับหมวดคำ $T_{1,n}$ ให้กับแต่ละคำในข้อความ โดยต้องการกำกับหมวดคำให้ได้ค่าความน่าจะเป็นของ $P(T_{1,n} \mid W_{1,n})$ มีค่าสูงที่สุด ซึ่งสามารถคำนวณได้จาก $P(W_{1,n}, T_{1,n}) / P(W_{1,n})$ ซึ่งสามารถละการคิด $P(W_{1,n})$ ที่เป็นตัวหารของทุกทางเลือกไปได้ ดังนั้นจึงเหลือเพียงการหาค่าของ $P(W_{1,n}, T_{1,n})$ ค่าความน่าจะเป็นร่วมตรงนี้สามารถคำนวณได้จาก $P(W_{1,n} \mid T_{1,n}) * P(T_{1,n})$ และสามารถประมาณค่าความน่าจะเป็นได้ตามแนวคิดของแบบจำลองไตรแกรมที่มีสมมติฐานว่า:

(1) คำหนึ่งๆสามารถปรากฏ ณ ตำแหน่งใดๆในประโยคได้ โดยไม่ขึ้นกับคำหรือหมวดคำที่อยู่ก่อนหน้าหรือตามหลัง กล่าวคือ $P(W_{1,n} \mid T_{1,n})$ มีค่าประมาณเท่ากับ ผลคูณรวมของ $P(W_i \mid T_i)$ ของทุกคำ

(2) ความน่าจะเป็นที่หมวดคำหนึ่งจะปรากฏ ณ ตำแหน่งใดๆในประโยคจะขึ้นอยู่กับหมวดคำที่ปรากฏก่อนหน้า 2 หมวดคำเท่านั้น (trigram model) กล่าวคือ $P(T_{1,n})$ คำนวณแบบประมาณค่าได้จาก $P(T_i \mid T_{i-1}, T_{i-2})$

ดังนั้น สมการที่ใช้สำหรับกำกับหมวดคำตามแบบจำลองไตรแกรม คือ $\prod_{i=1..n} P(W_i \mid T_i) * P(T_i \mid T_{i-1}, T_{i-2})$ (ดังแสดงในสมการที่ 2-4) จากคลังข้อมูลภาษาไทย สามารถหาค่าของ $P(W_i \mid T_i)$ ได้โดยนับจำนวนของคำ W_i ที่มีหมวดคำเป็น T_i หารด้วยจำนวนของ T_i ที่เป็นหมวดคำของคำใดๆ (จำนวนของ T_i ทั้งหมดที่ปรากฏในคลังข้อมูล) ส่วน $P(T_i \mid T_{i-1}, T_{i-2})$ สามารถหาได้โดยนับจำนวนหมวดคำ T_i ที่มีสายหมวดคำต่อเนื่อง T_{i-2} และ T_{i-1} นำหน้า หารด้วยจำนวนสายหมวดคำต่อเนื่อง T_{i-2}, T_{i-1} ที่ปรากฏติดกันทั้งหมดในคลังข้อมูล

การกำกับหมวดคำโดยใช้แบบจำลองไตรแกรมนี้ได้รับการวิจัยอย่างกว้างขวาง และพบว่าสามารถกำกับหมวดคำได้ถูกต้องสูง โดยเฉพาะกับภาษาอังกฤษ (Church, 1988; Charniak et al., 1993 อ้างถึงใน บุญเสริม กิจศิริกุล, 2541: 11) ข้อดีของการกำกับหมวดคำโดยใช้แบบจำลองไตรแกรม คือ ให้ผลความถูกต้องที่สูง และไม่จำเป็นต้องอาศัยนักภาษาศาสตร์เขียนกฎทางไวยากรณ์ แต่มีข้อด้อยคือ ต้องใช้เนื้อที่หน่วยความจำมากในการเก็บตารางค่าความน่าจะเป็นต่างๆ และความเร็วในการคำนวณช้ากว่าวิธีการใช้กฎเกณฑ์ แต่เนื่องจากปัจจุบันหน่วยความจำมีราคาถูกลงมาก และความเร็วในการประมวลผลของคอมพิวเตอร์ได้รับการพัฒนาอย่างต่อเนื่อง จึงคิดว่าข้อด้อยเหล่านี้ไม่เป็นปัญหามากนัก (บุญเสริม กิจศิริกุล, 2541: 11)

จากหลักการและแนวคิดต่างๆในการตัดคำและกำกับหมวดคำภาษาไทยที่ยกมากล่าวไว้ข้างต้น วิทยานิพนธ์ฉบับนี้ได้เลือกใช้หลักการที่อาศัยค่าความน่าจะเป็นเพื่อนำมาแก้ปัญหาการตัดคำและกำกับหมวดคำภาษาไทย โดยประยุกต์ใช้แบบจำลองไตรแกรมเพื่อสร้างโปรแกรมสำหรับตัดคำและกำกับหมวดคำ เนื่องจากผู้วิจัยเห็นว่า แบบจำลองไตรแกรมเป็นแนวคิดที่อาศัยหลักการของค่าความน่าจะเป็นที่ได้รับการวิจัยอย่างกว้างขวางจากงานวิจัยต่างๆทั้งภาษาไทยและภาษาอังกฤษ และยังให้ผลความถูกต้องสูงในการตัดคำและกำกับหมวดคำ อย่างไรก็ตาม วิทยานิพนธ์ฉบับนี้มองการแก้ปัญหาการตัดคำและกำกับหมวดคำเป็นเรื่องเดียวกัน ซึ่งน่าจะแก้ปัญหาทั้งสองเรื่องไปพร้อมๆกันได้ ดังนั้น วิทยานิพนธ์ฉบับนี้จึงต้องปรับใช้แนวคิดไตรแกรมเพื่อทำการตัดคำและกำกับหมวดคำภาษาไทยแบบเบ็ดเสร็จในขั้นตอนเดียว (ดูรายละเอียดในการคำนวณของแบบจำลองไตรแกรมที่ใช้ในวิทยานิพนธ์นี้ในบทที่ 5)

บทที่ 3

การจัดทำคลังข้อมูลภาษา

บทนี้จะกล่าวถึงการจัดทำคลังข้อมูลภาษาเพื่อใช้เป็นฐานความรู้สำหรับโปรแกรมในการตัดคำและกำกับหมวดคำภาษาไทย เนื่องจากโปรแกรมจำเป็นต้องเรียนรู้ค่าสถิติที่จะนำไปใช้ทั้งสำหรับตัดคำและกำกับหมวดคำ คลังข้อมูลที่ใช้ในวิทยานิพนธ์ฉบับนี้จึงมีลักษณะเป็นคลังข้อมูลภาษาไทยที่มีการตัดคำและกำกับหมวดคำ โดยในบทนี้จะเริ่มจากการกล่าวถึงบทบาทของคลังข้อมูลภาษาในงานด้านประมวลผลภาษาธรรมชาติ จากนั้นในหัวข้อที่ 3.1 กล่าวถึงชุดข้อมูลที่รวบรวมมาใช้เป็นคลังข้อมูล หัวข้อที่ 3.2 กล่าวถึงรูปแบบของข้อความในคลังข้อมูล และหัวข้อที่ 3.3 กล่าวถึงขั้นตอนในการตัดคำและกำกับหมวดคำด้วยมือให้กับคลังข้อมูล

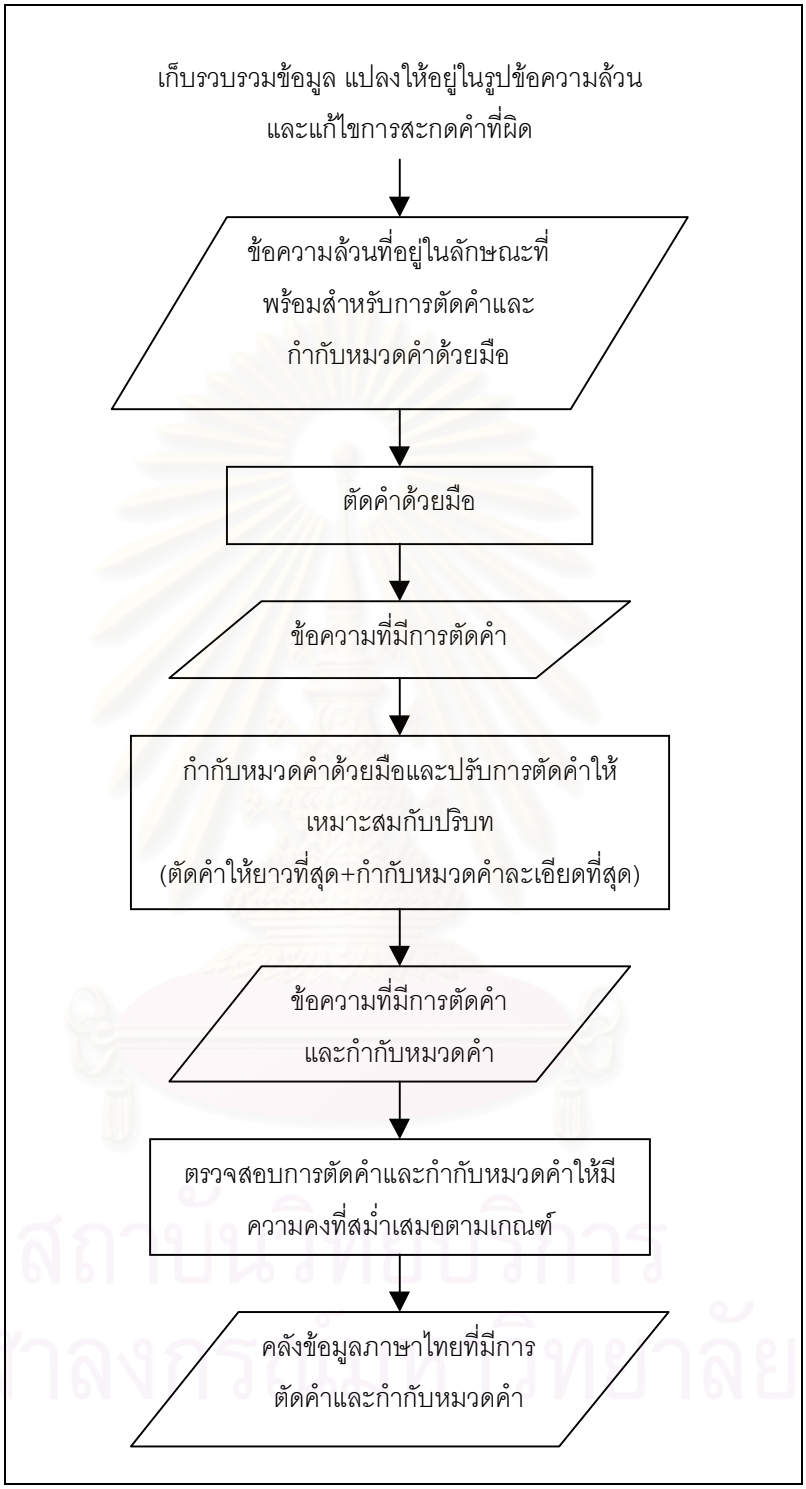
คลังข้อมูลภาษาถือเป็นทรัพยากรที่สำคัญสำหรับการวิเคราะห์ภาษาและการศึกษาทางภาษาศาสตร์ และในปัจจุบันได้ถูกนำมาใช้อย่างแพร่หลายในงานด้านประมวลผลภาษาธรรมชาติเพื่อเป็นฐานความรู้ด้านต่างๆให้กับระบบ โดยเฉพาะอย่างยิ่ง เมื่อยุคหลังๆการประมวลผลภาษาธรรมชาติมักประยุกต์วิธีการทางสถิติเข้ามาใช้ คลังข้อมูลก็จะเป็นแหล่งข้อมูลทางภาษาขนาดใหญ่สำหรับเก็บข้อมูลทางสถิติ เพื่อช่วยในการประมวลผลให้มีความถูกต้องแม่นยำและมีประสิทธิภาพสูงขึ้น หรือแม้แต่งานที่อาศัยกฎในปัจจุบันซึ่งได้รับการพัฒนาแก้ไขขึ้นมาใหม่ เช่น งานของ Brill (1993) ก็ยังให้ระบบเรียนรู้และสรุปกฎจากคลังข้อมูลเช่นกัน

คลังข้อมูลที่นำมาใช้ในการประมวลผลภาษาธรรมชาติได้รับการพัฒนารูปแบบและวิธีการเรื่อยมา ทำให้มีลักษณะแตกต่างกันไปทั้งในด้านรูปแบบและเนื้อหา มีทั้งที่เป็นคลังข้อมูลล้วน (plain corpus) ซึ่งประกอบด้วยข้อความอย่างเดียว และคลังข้อมูลที่มีการกำกับข้อมูล (annotated corpus) เช่น คลังข้อมูลที่กำกับหมวดคำ, คลังข้อมูลที่กำกับความหมาย หรือ บางงานต้องการข้อมูลที่เป็นตัวแทนของทั้งภาษาในขณะที่บางงานต้องการข้อมูลภาษาย่อย (sub-language) เฉพาะเรื่องเท่านั้น ทั้งนี้ขึ้นอยู่กับจุดประสงค์ วิธีการ และลักษณะของงานที่จะนำคลังข้อมูลไปใช้

สำหรับคลังข้อมูลภาษาไทยนั้น เพิ่งจะได้รับการพัฒนาในช่วงไม่กี่ปีที่ผ่านมา จึงทำให้ในปัจจุบันมีคลังข้อมูลภาษาไทยอยู่จำนวนน้อย คลังข้อมูลภาษาไทยที่ได้รับความนิยมนำไปใช้ในงานวิจัยต่างๆ ได้แก่ คลังข้อมูลออร์คิด (Orchid Corpus) (Virach Sornlertlamvanich et al., 1997) ซึ่งสร้างและพัฒนาโดยกลุ่มวิจัยภาษาและวิทยาการความรู้ (LINKS) แห่งศูนย์เทคโนโลยีอิเล็กทรอนิกส์และคอมพิวเตอร์แห่งชาติ (NECTEC) โดยมุ่งหวังที่จะให้บริการเป็นแหล่งข้อมูลภาษาไทยสำหรับการวิจัยทางภาษาและทางการประมวลผลภาษาธรรมชาติ คลังข้อมูลออร์คิดมีลักษณะเป็นคลังข้อความภาษาไทยที่มีการตัดแบ่งประโยคและคำและมีการกำกับหมวดคำ

การพัฒนาโปรแกรมซึ่งใช้แบบจำลองไตรแกรมสำหรับตัดคำและกำกับหมวดคำในวิทยานิพนธ์ฉบับนี้จำเป็นต้องเรียนรู้ข้อมูลทางสถิติจากคลังข้อมูลภาษาไทย ดังนั้น ผู้วิจัยจึงได้จัดทำคลังข้อมูลขึ้นมาสำหรับใช้ในวิทยานิพนธ์ สาเหตุที่ไม่ได้เลือกใช้คลังข้อมูลออร์คิด เนื่องจากว่าคลังข้อมูลออร์คิดที่เผยแพร่ไว้นั้นมีเนื้อหาค่อนข้างจำกัด เนื่องจากประกอบด้วยข้อมูลที่เป็นรายงานการประชุมทางวิชาการของทางเนคเทคเท่านั้น แต่วิทยานิพนธ์นี้สนใจศึกษาข้อมูลภาษาไทยที่ใช้กันทั่วไปในชีวิตประจำวัน ดังนั้นผู้วิจัยจึงได้รวบรวมข้อมูลภาษาไทยขึ้นมาเองและทำการตัดคำและกำกับหมวดคำด้วยมือให้กับคลังข้อมูล นอกจากนี้ เนื่องจากประเด็นเรื่องคำและหมวดคำภาษาไทยก็ยังคงเป็นประเด็นที่ยังไม่ลงตัวและน่าจะศึกษาต่อได้อีกมาก ผู้วิจัยจึงเลือกที่จะศึกษาเรื่องคำและหมวดคำในภาษาไทยไปด้วยเพื่อนำมาตัดคำและกำกับหมวดคำให้กับคลังข้อมูลที่จัดทำขึ้น (ดูรายละเอียดเรื่องเกณฑ์การตัดคำและการกำหนดชุดหมวดคำในบทที่ 4)

ส่วนถัดไปจะอธิบายรายละเอียดของการจัดทำคลังข้อมูลในวิทยานิพนธ์ฉบับนี้ ซึ่งโดยภาพรวมแล้ว การจัดทำคลังข้อมูลสำหรับวิทยานิพนธ์ฉบับนี้มีลำดับขั้นตอนดังรูปที่ 3-1



รูปที่ 3-1 ขั้นตอนกระบวนการสร้างคลังข้อมูล

3.1 ชุดข้อมูลที่ใช้เป็นคลังข้อมูล

ผู้วิจัยได้รวบรวมข้อความภาษาไทยซึ่งเป็นข้อมูลอิเล็กทรอนิกส์ที่เผยแพร่ไว้ในเวบไซต์ของหนังสือพิมพ์กรุงเทพธุรกิจ (http://www.bangkokbiznews.com) ขนาดประมาณ 25,000 คำเพื่อนำมาใช้เป็นคลังข้อมูล โดยเลือกข้อมูลของฉบับวันที่ 1 พฤษภาคม พ.ศ. 2543 ข้อมูลที่ใช้จัดเป็นข้อเขียนประเภทข่าวและบทความ ซึ่งประกอบด้วย รายงานข่าว บทความ คอลัมน์ต่างๆ และมีการใช้ภาษาในรูปแบบภาษาเขียนที่ไม่เป็นทางการ อันเป็นตัวอย่างของข้อความภาษาไทยที่คนไทยพบได้ทั่วไปในชีวิตประจำวัน

3.2 รูปแบบของข้อความในคลังข้อมูล

ข้อมูลภาษาไทยที่รวบรวมมาได้ ผู้วิจัยจะแปลงให้อยู่ในรูปของข้อความล้วน (plain text) โดยตัดส่วนที่เป็น HTML tag ออกไป แล้วจัดเก็บไว้เป็นแฟ้มข้อมูล พร้อมทั้งผู้วิจัยได้แก้ไขการสะกดคำที่ผิด และตัดข้อความในส่วนที่เป็นตารางออกไป ทั้งนี้พยายามคงข้อความเดิมให้มากที่สุดเพื่อให้เป็นตัวอย่างของการใช้ภาษาจริง หลังจากนั้น เนื่องจากข้อความในเวบไซต์มีเครื่องหมายขึ้นบรรทัดใหม่เพื่อแยกข้อความที่ต่อเนื่องกันออกเป็นบรรทัดๆ แม้ว่าจะยังไม่จบความก็ตาม ผู้วิจัยจึงได้ทำการจัดเนื้อความในย่อหน้าเดียวกันให้ต่อเนื่องกันไปโดยตัดเครื่องหมายขึ้นบรรทัดใหม่ที่แทรกอยู่ภายในข้อความของย่อหน้าเดียวกันออก และแทนที่การเว้นย่อหน้าด้วยสัญลักษณ์ระบุย่อหน้า <P> (ดูตารางที่ 3-1) เพื่อให้สะดวกต่อการพิจารณาในการตัดคำและกำกับหมวดคำด้วยมือ

คลังข้อมูลที่จัดทำขึ้นไม่มีการตัดแบ่งประโยค เนื่องจากว่าภาษาไทยยังมีความกำกวมในการตัดประโยคอยู่ ซึ่งเกิดจากการที่ภาษาไทยไม่มีตัวบ่งขอบเขตของประโยคที่แน่ชัดเหมือนอย่างเช่นภาษาอังกฤษที่ใช้เครื่องหมาย full-stop (.) ระบุการจบประโยค อีกทั้งเกณฑ์การตัดสินประโยคในภาษาไทยก็ยังไม่ลงตัว ตัวอย่างเช่น Matthews (1981: 27) กล่าวว่า ประโยคจะต้องประกอบด้วยส่วนหลักคือภาคแสดงซึ่งมีกริยาเป็นหัวใจของประโยค แต่จากการสังเกตในภาษาไทยก็มีสายอักขระที่ดูเหมือนจะเป็นประโยคแต่ไม่มีกริยาปรากฏอยู่ เช่น “*วันนี้วันเสาร์*” หรือสายอักขระ เช่น “*เขาชื่อสมชาย*” ซึ่งก็ยังถกเถียงกันได้ว่า “*ชื่อ*” ในที่นี้เป็นหมวดคำนามหรือหมวดคำกริยา หากจัดให้เป็นหมวดคำนามแล้วก็แสดงว่าไม่มีกริยาในสายอักขระที่ดูเหมือนจะเป็นประโยคนี้เลย ความกำกวมในเรื่องประโยคเช่นนี้เป็นอุปสรรคที่ทำให้ไม่สามารถตัดข้อความใน

คลังข้อมูลเป็นประโยคได้อย่างลงตัวเสมอไป ดังนั้น ในวิทยานิพนธ์ฉบับนี้ผู้วิจัยจะไม่ถือเอาประโยคเป็นหน่วยสำคัญ (ทั้งนี้ ไม่ได้หมายความว่าในภาษาไทยไม่มีประโยค) โดยจะถือว่าข้อเขียนภาษาไทยประกอบขึ้นจากถ้อยความ (utterance) ซึ่งได้ใจความ อาจเป็นประโยคหรือไม่ก็ได้ และมักมีการใส่เครื่องหมายวรรคตอนเมื่อจบถ้อยความหนึ่งๆ

สัญลักษณ์	คำอธิบาย
<T>	ชื่อเรื่อง, หัวข้อ (Title)
<P>	ย่อหน้า (Paragraph)
_	สัญลักษณ์แบ่งคำ (word delimiter)
/POS	กำกับหมวดคำ (Part-of-Speech)

ตารางที่ 3-1 สัญลักษณ์ในการกำกับคลังข้อมูล

จากนั้น ผู้วิจัยจะทำการตัดคำและกำกับหมวดคำให้กับคลังข้อมูลโดยใช้สัญลักษณ์ในตารางที่ 3-1 และสัญลักษณ์หมวดคำ (ดูรายละเอียดสัญลักษณ์หมวดคำในบทที่ 4) คลังข้อมูลที่ได้รับการตัดคำและกำกับหมวดคำแล้วจะมีรูปแบบดังข้างล่างนี้

ศูนย์/NCM_การศึกษา/NCM_ของ/PN_ประเทศ/NCM_

ซึ่งแสดงตัวอย่างข้อความว่า “ศูนย์การศึกษาของประเทศ” ที่ได้ทำการตัดคำและกำกับหมวดคำด้วยมือแล้ว โดยที่แต่ละบรรทัดจะเป็นข้อความที่ประกอบด้วยคำเรียงต่อกันไปโดยมีเครื่องหมายขีดล่าง (_) แบ่งคั่นระหว่างคำ และแต่ละคำประกอบด้วยรูปคำและหมวดคำโดยมีเครื่องหมายทับ (/) คั่นระหว่างรูปคำและหมวดคำ

3.3 ขั้นตอนในการตัดคำและกำกับหมวดคำด้วยมือให้กับคลังข้อมูล

ชุดข้อมูลที่แปลงเป็นข้อความล้วนและได้ทำการแก้ไขการสะกดคำที่ผิดและตัดข้อความส่วนที่เป็นตารางออก และได้จัดเนื้อความในย่อหน้าเดียวกันให้ต่อเนื่องกันไปแล้ว ผู้วิจัยจะได้นำมาตัดคำและกำกับหมวดคำด้วยมือ

การตัดคำและกำกับหมวดคำด้วยมือให้กับคลังข้อมูลเป็นขั้นตอนสำคัญในการจัดทำคลังข้อมูลฝึกสอน เพื่อใช้เป็นฐานความรู้ให้โปรแกรมได้เรียนรู้ข้อมูลทางสถิติที่จะใช้ในการแก้ปัญหาการตัดคำและกำกับหมวดคำ กระบวนการนี้เป็นกระบวนการที่ทำได้ลำบากและต้องใช้เวลามากเนื่องจากต้องทำการวิเคราะห์ข้อความที่อยู่ในคลังข้อมูลทั้งหมด ซึ่งข้อความดังกล่าวมีรูปแบบโครงสร้างและคำศัพท์ที่ใช้อยู่หลากหลาย ในการวิเคราะห์แต่ละข้อความผู้วิจัยต้องตัดสินใจว่าจะตัดคำที่ตำแหน่งใดและคำที่ตัดออกมาได้จะกำกับหมวดคำใด นอกจากนี้ ยังต้องพยายามให้คลังข้อมูลที่ตัดคำและกำกับหมวดคำแล้วมีความคงที่ (consistency) ในการตัดคำและกำกับหมวดคำด้วยมือมากที่สุด จากลักษณะปัญหาดังที่กล่าวมา ทำให้วิทยานิพนธ์นี้มีขั้นตอนในการจัดทำคลังข้อมูลดังจะอธิบายได้ดังนี้

ในขั้นแรก ผู้วิจัยจะทำการตัดคำด้วยมือเสียก่อนเพื่อแยกข้อความที่เขียนต่อเนื่องกันออกเป็นคำๆ เนื่องจากผู้วิจัยได้ทดลองทำการตัดคำและกำกับหมวดคำไปพร้อมกันแล้ว พบว่า การตัดคำและกำกับหมวดคำด้วยมือให้กับคลังข้อมูลไปพร้อมๆกันในทีเดียวทำให้ผู้วิจัยเกิดความสับสนทั้งในการตัดคำและกำกับหมวดคำเนื่องจากข้อความที่วิเคราะห์มีลักษณะที่หลากหลาย ซึ่งส่งผลให้ผู้วิจัยไม่สามารถตัดคำและกำกับหมวดคำให้คงที่ได้ ดังนั้น ผู้วิจัยจึงเลือกตัดคำเสียก่อน เพื่อให้เกิดความสะดวกต่อการอ่านในเวลาที่กำกับหมวดคำด้วยมือให้กับคลังข้อมูล อย่างไรก็ตาม ผลการตัดคำจากขั้นตอนนี้ยังคงมีข้อผิดพลาดอยู่บ้าง เนื่องจากพบว่า ในบางกรณีการพิจารณาตัดคำก็ขึ้นอยู่กับการตัดสินใจว่าจะกำกับหมวดคำแบบใดด้วย ซึ่งจะได้ทำการปรับการตัดคำในขั้นตอนต่อไป

ขั้นตอนต่อมา ผู้วิจัยตัดสินหมวดคำให้กับแต่ละคำและทำการกำกับหมวดคำลงในคลังข้อมูลที่มีการตัดคำไว้แล้ว พร้อมทั้งปรับการตัดคำไปพร้อมๆกันโดยพิจารณาจากปริบทข้างเคียงว่าในปริบทดังกล่าวควรตัดคำและกำกับหมวดคำแบบใด ซึ่งพบว่า ปัญหาการตัดคำและกำกับหมวดคำเกี่ยวข้องกันอย่างใกล้ชิด กล่าวคือ การเลือกตัดคำแบบใดแบบหนึ่งจะมีผลกระทบทำให้การกำกับหมวดคำเปลี่ยนแปลงไปเช่นกัน ในทางตรงกันข้าม การเลือกกำกับหมวดคำแบบใดแบบหนึ่งก็จะทำให้ต้องเปลี่ยนแปลงการตัดคำให้สอดคล้องกันด้วย ตัวอย่างข้างล่างแสดงให้เห็นความสัมพันธ์ระหว่างการตัดคำและกำกับหมวดคำ ซึ่งอาจเลือกตัดคำและกำกับหมวดคำได้มากกว่าหนึ่งแบบ (ดูรายละเอียดสัญลักษณ์หมวดคำได้ในบทที่ 4)

“ธนาคารแห่งประเทศไทย” “ธนาคารแห่งประเทศไทย/NPP_”

“ธนาคาร/NCM_ แห่ง/PN_ ประเทศไทย/NPP”

“ฝ่ายออกบัตร”	“ฝ่ายออกบัตร/NCM_”
	“ฝ่าย/NCM_ออก/VN0_บัตร/NCM”
“การคืนเงินต้น”	“การคืนเงินต้น/NCM_”
	“การคืน/NCM_เงินต้น/NCM_”
	“การ/PFX_คืน/VN0_เงินต้น/NCM”

อย่างไรก็ดี แม้ว่าผู้วิจัยได้พิจารณาปรับทเพื่อเลือกตัดคำและกำกับหมวดคำที่เหมาะสมให้กับข้อความแล้วก็ตาม ในหลายๆกรณีก็ยังเกิดความกำกวม ซึ่งผู้วิจัยยังไม่อาจตัดสินใจเด็ดขาดลงไปได้ว่าควรตัดคำและกำกับหมวดคำแบบใด ดังนั้น ในขั้นตอนนี้ ผู้วิจัยจึงพยายามตัดคำให้ได้คำยาวที่สุด และเลือกกำกับหมวดคำอย่างละเอียดที่สุดก่อน เพื่อให้สามารถตรวจสอบแก้ไขในภายหลังได้อย่างสะดวก การกำกับหมวดคำอย่างละเอียด หมายถึง ในกรณีที่ผู้วิจัยไม่แน่ใจว่าคำดังกล่าวควรเป็นหมวดคำใด ผู้วิจัยจะกำกับคำนั้นด้วยหมวดคำทั้งหมดเท่าที่น่าจะเป็นไปได้ในปริบทนั้น และในหลายกรณีผู้วิจัยได้เพิ่มเติมรายละเอียดของปริบทในการปรากฏของคำดังกล่าวด้วย ตัวอย่างเช่น

“มา/V0+AV=Post=direction_”	หมายถึง “มา” อาจเป็นคำกริยา(V0) หรือคำวิเศษณ์(AV) ซึ่งปรากฏหลังส่วนหลักที่มันขยาย (=Post) และแสดงความหมายเกี่ยวกับทิศทาง (=direction)
“เข้าใจ/VCV0[>ว่า]_ ว่า/PCOMP=VMod_”	หมายถึง “เข้าใจ” เป็นคำกริยาที่ตามหลังด้วยอนุพากย์ส่วนเติมเต็ม (complement clause) (VCV0) และตัวนำส่วนเติมเต็มที่ตามมานั้นคือคำว่า “ว่า”
	“ว่า” เป็นตัวนำส่วนเติมเต็ม (PCOMP) และเป็นหน่วยที่เกิดร่วมกับคำกริยา (=VMod)

ขั้นตอนถัดมา ผู้วิจัยได้ทำรายการคำและรายการหมวดคำทั้งหมดที่ได้จากขั้นตอนที่ผ่านมา เพื่อศึกษาว่า รูปคำและหมวดคำทั้งหมดมีอะไรบ้าง แต่ละรูปคำปรากฏเป็นหมวดคำใดใน

ปริบทใดได้บ้าง และแต่ละหมวดคำเป็นหมวดคำของรูปคำใดในปริบทใดบ้าง ซึ่งพบว่า ผลการตัดคำและกำกับหมวดคำจากขั้นตอนที่ผ่านมายังมีการตัดคำและกำกับหมวดคำที่ยังไม่คงที่อยู่เป็นจำนวนมาก ผู้วิจัยจึงได้ทำการตรวจสอบรูปคำและหมวดคำที่ยังตัดสินได้ไม่คงที่ลงตัว โดยค้นหาคำที่ยังมีความไม่คงที่ในคลังข้อมูลแล้วปรับการตัดคำและการกับหมวดคำให้เหมาะสมตามเกณฑ์ พร้อมกันนี้จากรายการคำและรายการหมวดคำจะสามารถพิจารณาได้ว่า หมวดคำชุดใดที่มีความคล้ายคลึงกันและน่าจะจับรวบเป็นหมวดคำเดียว และสามารถพิจาณาได้ว่ารูปคำหนึ่งๆซึ่งปรากฏในหลายปริบทน่าจะจัดเป็นหมวดคำเดียวกันหรือจัดเป็นคนละหมวดคำกัน ขั้นตอนนี้เป็นการตรวจสอบการตัดคำและการกำกับหมวดคำให้สม่ำเสมอตามเกณฑ์และมีความคงที่มากที่สุด และทำการตรวจสอบการตัดคำและกำกับหมวดคำไปจนกว่าทุกคำในคลังข้อมูลจะมีความคงที่สม่ำเสมอตามเกณฑ์ ในที่สุด จึงได้ผลลัพธ์เป็นคลังข้อมูลภาษาไทยที่มีการตัดคำและกำกับหมวดคำ ซึ่งพร้อมที่จะนำไปใช้ในการพัฒนาโปรแกรมตัดคำและกำกับหมวดคำ

โดยสรุปแล้ว ผู้วิจัยได้รวบรวมข้อมูลภาษาไทยจากเวบไซต์ของหนังสือพิมพ์กรุงเทพธุรกิจเพื่อนำมาใช้เป็นคลังข้อมูลในวิทยานิพนธ์ ซึ่งคลังข้อมูลดังกล่าวจำเป็นต้องมีการตัดคำและกำกับหมวดคำไว้เพื่อเป็นฐานความรู้ให้กับโปรแกรม ดังนั้น ผู้วิจัยจึงได้แปลงข้อมูลให้อยู่ในรูปข้อความล้วน แล้วทำการตัดคำและกำกับหมวดคำด้วยมือโดยพยายามให้มีความคงที่มากที่สุด ดังนั้นจึงต้องมีขั้นตอนต่างๆในการตัดคำและกำกับหมวดคำ และมีการตรวจสอบการตัดคำและกำกับหมวดคำให้คงที่สม่ำเสมอตามเกณฑ์ดังที่อธิบายไว้ บทต่อไปจะกล่าวถึง เกณฑ์ในการตัดสินคำและเกณฑ์การกำหนดชุดหมวดคำที่ได้นำมาใช้เพื่อตัดคำและกำกับหมวดคำให้กับคลังข้อมูลที่จัดทำขึ้นในบทนี้

บทที่ 4

การตัดคำและการกำหนดชุดหมวดคำ

การพัฒนาโปรแกรมสำหรับตัดคำและกำกับหมวดคำภาษาไทย โปรแกรมจำเป็นต้องเรียนรู้ข้อมูลทางสถิติจากคลังข้อมูลภาษาไทยที่จัดเตรียมไว้ อันได้แก่ การตัดคำและการกำกับหมวดคำ ซึ่งโปรแกรมจะเรียนรู้และเก็บข้อมูลไว้ในรูปของค่าสถิติในการปรากฏของคำและลำดับหมวดคำ ดังนั้น บทนี้จะได้กล่าวถึงเกณฑ์การตัดคำที่นำมาใช้ในวิทยานิพนธ์ในหัวข้อที่ 4.1 และกล่าวถึงการกำหนดชุดหมวดคำพร้อมทั้งนำเสนอชุดหมวดคำภาษาไทยที่จัดทำขึ้นในหัวข้อที่ 4.2

เกณฑ์การตัดคำและชุดหมวดคำถือได้ว่าเป็นเรื่องสำคัญทั้งในทางภาษาศาสตร์และในการพัฒนาคลังข้อมูลภาษาสำหรับการประมวลผลภาษาธรรมชาติ เนื่องจากว่าเกณฑ์ในการตัดคำและชุดหมวดคำที่นำมาใช้จะมีผลต่อการวิเคราะห์ภาษาเป็นอย่างมาก ในทางภาษาศาสตร์ เรื่องเหล่านี้ถือเป็นเรื่องพื้นฐานในการศึกษาภาษาอย่างเป็นระบบ ส่วนในทางการประมวลผลภาษาธรรมชาติ เกณฑ์ในการตัดคำและชุดหมวดคำเปรียบเสมือนความรู้ทางไวยากรณ์ภาษาที่แฝงตัวอยู่ในคลังข้อมูล คลังข้อมูลที่มีการตัดคำและการกำกับหมวดคำที่ต่างกันก็สะท้อนมุมมองทางไวยากรณ์ภาษาที่ต่างกันด้วย

4.1 การตัดคำภาษาไทย

ในส่วนนี้จะได้กล่าวถึงการตัดคำภาษาไทยในวิทยานิพนธ์นี้ โดยเริ่มจากกล่าวถึงลักษณะของภาษาไทยที่ทำให้เกิดปัญหาการตัดคำว่ามาจากการที่ภาษาไทยไม่ได้เขียนแยกคำไว้ และก็ยังไม่มีเกณฑ์ในการตัดสินขอบเขตของคำที่ชัดเจนอีกด้วย จากนั้นในหัวข้อที่ 4.1.1 จะได้กล่าวถึงเกณฑ์ต่างๆที่ได้เลือกใช้เพื่อช่วยในการตัดคำให้กับคลังข้อมูล อันได้แก่ เกณฑ์ทางความหมาย, เกณฑ์ทางวากยสัมพันธ์ และเกณฑ์ทางจิตวิทยา และจะได้กล่าวถึงแนวคิดและการนำเกณฑ์ต่างๆไปช่วยตัดสินความกำกวมในการตัดคำที่เป็นปัญหาในคลังข้อมูลในหัวข้อที่ 4.1.2 ดังนี้

จากลักษณะที่ระบบการเขียนของภาษาไทยสามารถเขียนคำเรียงติดต่อกันไปได้โดยไม่มีการเว้นช่องว่างระหว่างคำทำให้ภาษาไทยประสบปัญหาในการตัดคำ แม้แต่การรู้จำคำเดี่ยวซึ่ง

เป็นคำที่ไม่ค่อยมีปัญหานักก็ยังจำเป็นต้องอาศัยการตัดคำนั้นๆออกจากสายอักขระข้างเคียง การรู้จำคำยิ่งลำบากมากขึ้นในกรณีคำประกอบซึ่งเกิดจากการนำคำที่มีอยู่เดิมมาประกอบกันเข้าเป็นคำใหม่ เนื่องจากไม่ใช่เพียงต้องตัดคำออกจากสายอักขระข้างเคียงเท่านั้น แต่ยังต้องพิจารณาด้วยว่า สายอักขระที่ตัดออกมามีสถานะเป็นหนึ่งคำหรือไม่ ดังที่ได้กล่าวมาแล้วในบทที่ 2 ความพยายามที่จะอธิบายมโนทัศน์เรื่องคำในหลายๆภาษาทำให้มีงานวิจัยจำนวนไม่น้อย (Bloomfield, 1933; Brown and Miller, 1980; Crystal, 1971; Kramsky, 1969; Lehmann, 1983; Miller, 1991; Nida, 1949; Pike, 1967, 1977; Robins, 1964) ที่กล่าวถึงมโนทัศน์คำ และยกเกณฑ์ขึ้นมาสำหรับตัดสินว่า คำควรมีลักษณะและคำจำกัดความอย่างไร ถึงแม้ว่าเกณฑ์ต่างๆดังกล่าวยังไม่สามารถให้คำจำกัดความที่ลงตัวแน่นอนสำหรับคำได้ แต่ก็มีประโยชน์ช่วยชี้แนวทางในการตัดสินคำและช่วยให้เข้าใจลักษณะที่เป็นปัญหาได้ดียิ่งขึ้น

4.1.1 เกณฑ์การตัดคำภาษาไทย

หัวข้อนี้จะได้กล่าวถึงเกณฑ์และคำอธิบายที่งานวิจัยต่างๆพยายามใช้เพื่อกำหนดขอบเขตของคำ ซึ่งเกณฑ์การตัดคำที่กล่าวไว้ในงานวิจัยต่างๆสามารถแบ่งได้เป็น 4 ประเภทใหญ่ๆ คือ เกณฑ์ทางความหมาย, เกณฑ์ทางเสียง, เกณฑ์ทางวากยสัมพันธ์ และเกณฑ์ทางจิตวิทยา วิทยานิพนธ์ฉบับนี้จะตัดเกณฑ์ทางเสียงออกไป เนื่องจากข้อมูลที่ใช้ในวิทยานิพนธ์นี้อยู่ในรูปตัวเขียน ดังนั้นจึงจะพิจารณาจาก 3 เกณฑ์ที่เหลือเท่านั้น โดยที่ในหัวข้อ 4.1.1.1 จะได้กล่าวถึงเกณฑ์ทางความหมาย หัวข้อที่ 4.1.1.2 กล่าวถึงเกณฑ์ทางวากยสัมพันธ์ซึ่งแบ่งเป็นเกณฑ์ย่อยๆ อีก 6 เกณฑ์ และหัวข้อที่ 4.1.1.3 กล่าวถึงเกณฑ์ทางจิตวิทยา จากนั้นจะแสดงให้เห็นว่าวิทยานิพนธ์นี้เลือกใช้เกณฑ์เหล่านี้อย่างไรในการตัดสินหมวดคำให้กับคลังข้อมูล

จากคำอธิบายและเกณฑ์ต่างๆที่งานวิจัยจำนวนมากพยายามนำมาใช้อธิบายมโนทัศน์ของคำนั้น บางงานก็อธิบายคำในความคุ้นเคยของชาวตะวันตกเท่านั้น เช่น Miller (1991: 28) ยกตัวอย่างคำนิยามที่ว่า “คำเป็นสิ่งที่เขียนอยู่ระหว่างช่องว่างและไม่มีช่องว่างคั่นกลางคำ” แต่บางงานก็พยายามจำกัดความคำในฐานะเป็นหน่วยที่เป็นสากลทางภาษาของทุกภาษา ซึ่งผู้วิจัยได้คัดเลือกคำอธิบายและเกณฑ์ที่เห็นว่าเหมาะสมกับภาษาไทยเพื่อนำมาใช้อธิบายมโนทัศน์ของคำในวิทยานิพนธ์ฉบับนี้ ดังนี้

4.1.1.1 เกณฑ์ทางความหมาย

เกณฑ์นี้พิจารณาคำในแง่ของความหมาย ตามเกณฑ์นี้ คำจะถูกนิยามว่าเป็นหน่วยทางความหมาย หรือหน่วยที่แสดงความหมายหรือความคิดหรือมโนทัศน์เดี่ยว ซึ่งจะทำให้เกิดปัญหาเนื่องจากว่าการอธิบายความหมายหรือมโนทัศน์ยิ่งยากกว่าคำเสียอีกเพราะไม่ได้แสดงรูปอย่างชัดเจน ทำให้ไม่สามารถตัดสินได้แน่ชัดว่าสิ่งใดเป็นมโนทัศน์เดี่ยว เช่น ในภาษาไทย คำว่า “โรงไฟฟ้าพลังน้ำ”, “ชุดประจำชาติ” สามารถพิจารณาได้หลายแบบ ให้เป็นคำหนึ่งคำที่มีหนึ่งมโนทัศน์, เป็นคำหนึ่งคำที่มีหลายมโนทัศน์ หรือเป็นคำหลายคำและมีหลายมโนทัศน์ ก็ได้ นอกจากนี้เกณฑ์นี้ยังทำให้นิยามของคำไปเหลื่อมล้ำกับหน่วยคำด้วย เช่น “นัก-” ซึ่งเป็นหน่วยคำไม่อิสระก็แสดงความหมายว่า “ผู้กระทำ” ได้เช่นกัน ส่วนคำบางคำ เช่น “พูดกับคุณ”, “แม้จะไม่ค่อยชอบ” กลับไม่มีความหมายหรือแสดงความหมายไม่ชัดเจนนักในเชิงอ้างถึง ในภาษาของตะวันตกปรากฏการณ์เช่นนี้เกิดกับคำที่ Kramsky (1969: 19) เรียกว่า synsemantics ซึ่งแสดงความหมายทางไวยากรณ์ในประโยคเท่านั้น

จากเหตุผลดังกล่าว เห็นได้ว่า การนิยามคำด้วยเกณฑ์ทางความหมายเพียงอย่างเดียวนั้นยังไม่น่าเชื่อถือ เพราะไม่ครอบคลุมและลงตัวพอ อีกทั้งยังมีความลักลั่นระหว่างคนแต่ละคนได้ง่าย ดังนั้น วิทยานิพนธ์ฉบับนี้ไม่ได้อาศัยเกณฑ์ทางความหมายเพียงอย่างเดียวในการตัดสินคำ แต่ถือว่าความหมายเป็นคุณสมบัติพื้นฐานของคำ กล่าวคือ สิ่งที่เป็นคำจะต้องมีความหมาย อาจเป็นหนึ่งมโนทัศน์หรือหลายมโนทัศน์ก็ได้ ความหมายในที่นี้จะรวมถึง ความหมายในเชิงอ้างถึงสิ่งภายนอกภาษา ซึ่งอาจเป็นสิ่งรูปธรรม ได้แก่ สิ่งของ, การกระทำ, ลักษณะ หรือสิ่งนามธรรม ได้แก่ ความคิด ก็ได้ และยังรวมไปถึงความหมายในเชิงแสดงความสัมพันธ์ทางไวยากรณ์ในภาษา (grammatical meaning หรือ grammatical relation) ด้วย ซึ่งจะทำให้คำจำกัดความของคำครอบคลุมคำไวยากรณ์หรือคำที่ทำหน้าที่เชื่อมคำหรือเชื่อมความในภาษาไทยได้ด้วย

4.1.1.2 เกณฑ์ทางวากยสัมพันธ์

เกณฑ์ทางวากยสัมพันธ์พิจารณาคำในแง่การปรากฏของคำและส่วนประกอบภายในคำ แบ่งเป็นเกณฑ์ย่อย 6 เกณฑ์ ดังนี้

(1) เกณฑ์การปรากฏอิสระ (free, isolatability)

Bloomfield (1933: 178) เป็นผู้ให้นิยามคำ ว่า minimal free form ซึ่งหมายความว่า คำคือหน่วยที่เล็กที่สุดที่สามารถปรากฏเป็นประโยคได้ตามลำพัง เช่น คำว่า “น้อง” คำเดียวสามารถปรากฏเป็นประโยคที่ตอบคำถามว่า “นั่นของใคร” ได้

(2) เกณฑ์การแยกจากสิ่งข้างเคียง (separability)

คือ คำสามารถแยกออกจากสายอักขระข้างเคียงโดยการแทรกสิ่งอื่นลงไปได้ เช่น “รองเท้าสวย” สามารถแยกออกเป็น “รองเท้านั้นสวย” ได้

(3) เกณฑ์การไม่สามารถแทรกกลาง (uninterruptibility)

คือ คำไม่สามารถถูกแทรกกลางโดยสิ่งอื่นได้ เช่น “รองเท้า” ไม่สามารถแยกเป็น *“รองนั้นเท้า” ได้

(4) เกณฑ์การแทนที่ (replaceability)

คือ คำจะสามารถถูกแทนที่ด้วยคำอื่นได้ Kramsky (1969: 68) แสดงความเห็นว่า เกณฑ์นี้เป็นเกณฑ์ที่น่าเชื่อถือที่สุดและใช้ได้กับหลายๆภาษา อีกทั้งสามารถใช้เป็นเกณฑ์รองรับในกรณีที่ไม่สามารถใช้เกณฑ์การแยกจากสิ่งข้างเคียงได้ ภาษาไทยสามารถใช้เกณฑ์นี้ช่วยตัดสินคำที่เป็นบุพบท เช่น “บนโต๊ะ” ไม่สามารถแยกเป็น *“บนของโต๊ะ” หรือ *“บนจากโต๊ะ” ได้ แต่สามารถนำคำบุพบทอื่นมาแทนที่เป็น “ใต้โต๊ะ” ได้ เป็นต้น แต่ในทางตรงกันข้ามส่วนประกอบภายในคำไม่สามารถแทนที่ด้วยสิ่งอื่นได้ Pike (1977: 113) อธิบายไว้คล้ายคลึงกันว่า สิ่งที่ไม่สามารถแยกจากสิ่งข้างเคียงนั้นก็สามารถตัดสินให้เป็นคำได้ หากสิ่งนั้นสามารถปรากฏในสลอต (slot-filler role) เดียวกับคำอื่นได้ตามไวยากรณ์แทกมีมิค (Tagmemic)

(5) เกณฑ์การย้ายที่ (displaceability)

คือ คำจะมีความอิสระพอสมควรทำให้สามารถเคลื่อนย้ายไปในตำแหน่งต่างๆในประโยคได้ เช่น ในภาษาไทย “ครูตีนักเรียน” สามารถย้ายที่เป็น “นักเรียนครูตี”, “นักเรียนตีครู” ได้โดยไม่ผิดไวยากรณ์ ในขณะที่ส่วนประกอบภายในคำไม่สามารถย้ายที่ไปตำแหน่งต่างๆในประโยคได้โดยอิสระ เช่น *“ครูนักตีเรียน” หรือการย้ายที่ภายในคำเองก็ไม่สามารถทำได้ เช่น *“ครูตีเรียนนัก”

(6) เกณฑ์ความสัมพันธ์พิเศษภายในคำ (special relationship)

Pike (1967: 438) กล่าวว่า ความสัมพันธ์ระหว่างส่วนประกอบภายในคำอาจจะไม่เป็นไปตามรูปแบบทางวากยสัมพันธ์ (nonsyntactic patterns between morphemes may occur within words which are not found in phrases – the outcast VS to cast out) ในภาษาไทยพบว่า ในนามวลี คำที่มาขยายจะปรากฏตามหลังคำนามซึ่งเป็นส่วนหลัก เช่น “ผลการเรียนดี” แต่ก็มีคำจำนวนหนึ่งที่คำที่มาขยายปรากฏหน้าคำหลัก เช่น “มงคลสมัย”, “สุขภาพ” ซึ่งเป็นคำที่ได้รับอิทธิพลจากภาษาบาลี-สันสกฤตที่มีลักษณะโครงสร้างทางวากยสัมพันธ์คือ คำขยายจะอยู่หน้าคำนามที่เป็นส่วนหลักในนามวลี ดังนั้น คำเหล่านี้จึงจัดเป็นคำหนึ่งคำในภาษาไทยเพราะความสัมพันธ์ระหว่างส่วนประกอบภายในคำไม่เป็นไปตามรูปแบบทางวากยสัมพันธ์ของภาษาไทย

เกณฑ์ทางวากยสัมพันธ์เป็นเกณฑ์ที่ได้รับการยอมรับจากงานวิจัยต่างๆว่าน่าเชื่อถือมากที่สุดเพราะมีลักษณะที่เป็นวิทยาศาสตร์ สามารถตรวจสอบได้ อย่างไรก็ดี เกณฑ์นี้ยังมีข้อจำกัดในการตัดสินคำในภาษาไทยในบางกรณี ดังจะอธิบายได้ดังนี้ เมื่อยึดตามเกณฑ์การปรากฏอิสระ จะพบว่า แท้ที่จริงคำบางคำในภาษาไทยไม่ปรากฏเป็นประโยคโดยลำพัง เช่น “จาก”, “ถ้า”, “ซึ่ง” ซึ่งมีลักษณะเหมือนกับคำที่ Kramsky (1969) เรียกว่า synsemantic ของภาษาตะวันตก ซึ่งทำให้ต้องอาศัยเกณฑ์อื่น ได้แก่ เกณฑ์การแยกจากสิ่งข้างเคียง เช่น “ถ้าไม่มีอาจารย์ผู้สอน...” เป็น “ถ้าสมมติว่าไม่มีอาจารย์ผู้สอน...” ซึ่งเกณฑ์นี้ก็อาจใช้ในการแยกคำในหน่วยสร้างบุพบทวลีไม่ได้ เช่น “สถาบันการศึกษาของประเทศ” ไม่สามารถใช้วิธีแทรกคำที่ขยายคำนามเพื่อแยกคำออกจากกันเหมือนที่ทำในภาษาอังกฤษได้ เพราะในภาษาไทยคำที่มาขยายมักจะเกิดตามหลังคำหลัก เช่น “สถาบันการศึกษาของประเทศใหญ่” ดังนั้นทำให้ต้องอาศัยเกณฑ์อื่นอีก หรือเมื่ออาศัยเกณฑ์การแทนที่ เช่น “สถาบันการศึกษาของไทย” เป็น “สถาบันการศึกษาของประเทศที่อยู่ในภูมิภาคนี้” จะเห็นได้ว่า สิ่งที่มาปรากฏแทนที่คำในตำแหน่งดังกล่าวอาจเป็นหน่วยทางภาษาที่ใหญ่กว่าคำก็ได้ด้วย นอกจากนี้ การที่ภาษาไทยเป็นภาษาที่ลำดับของคำมีความสำคัญ จึงทำให้มีข้อจำกัดในย้ายที่ของคำ เช่น “ให้หนังสือแก่สมชาย” เมื่อย้ายที่แล้วก็อาจจะทำให้ผิดโครงสร้างทางวากยสัมพันธ์ในภาษาไทย เช่น “ให้หนังสือสมชายแก่” หรือการย้ายที่อาจจะทำให้ความหมายผิดเพี้ยนไป เช่น “ครูตีนักเรียน” เป็น “นักเรียนตีครู” นอกจากนี้ ยังพบว่าคำบางคำสามารถแทรกกลางได้ เช่น “เครื่องพิมพ์ดีดไฟฟ้า“ เป็น “เครื่องพิมพ์ดีดที่ใช้ไฟฟ้า” หรือ “เรือประมง” เป็น “เรือสำหรับประมง” ถึงแม้จะทำให้เอกภาพของคำน้อยลงไปบ้างแต่ก็ไม่ถึงกับทำให้ความหมายผิดเพี้ยนไปจากเดิมมากนัก

จากที่กล่าวมา ลักษณะของคำในทางวากยสัมพันธ์พอจะสรุปได้ว่า เมื่อเทียบกับหน่วยคำซึ่งมีขนาดเล็กกว่า คำมีความเป็นอิสระมากกว่า คือ คำสามารถแยกจากสิ่งข้างเคียงหรือปรากฏโดยลำพังได้ และมีความคล่องตัวในการย้ายที่หรือจะแทนที่คำด้วยคำอื่นก็ได้ แต่เมื่อเทียบกับวลีหรืออนุพากย์ซึ่งมีขนาดใหญ่กว่า คำจะมีความเหนี่ยวแน่นภายใน (internal cohesion หรือ internal stability) มากกว่า คือ คำไม่สามารถแทรกกลางด้วยสิ่งอื่นได้ และส่วนประกอบภายในคำก็ไม่สามารถย้ายที่หรือแทนที่ด้วยสิ่งอื่นได้ รวมทั้งความสัมพันธ์ระหว่างส่วนประกอบภายในคำอาจจะไม่เป็นไปตามรูปแบบความสัมพันธ์ทางวากยสัมพันธ์ด้วยก็ได้

4.1.1.3 เกณฑ์ทางจิตวิทยา

เกณฑ์นี้พิจารณาคำในแง่ความคิดของผู้พูดและผู้ฟังที่มีต่อคำ Kramsky (1969: 71) กล่าวว่า คำไม่สามารถอธิบายด้วยเกณฑ์ทางวากยสัมพันธ์เสมอไป โดยยกตัวอย่างคำว่า fathers-in-law, commanders-in-chief ซึ่งสามารถแทรกกลางคำด้วยหน่วยคำแสดงพหูพจน์ s ได้ และสามารถแทนที่ส่วนประกอบภายในคำเป็น mothers-in-law ก็ได้ Kramsky กล่าวว่า ในกรณีเหล่านี้ควรใช้เกณฑ์ทางจิตวิทยาพิจารณาว่า ในการสื่อสารผู้พูดและผู้ฟังตีความให้สายอักขระดังกล่าวมีสถานะเป็นหนึ่งคำหรือเป็นคำหลายคำ นอกจากนี้ ในภาษาอังกฤษระบบการเขียนที่มีการเว้นช่องว่างระหว่างส่วนประกอบภายในคำ, หรือมีการขีดคั่นระหว่างส่วนประกอบภายในคำ หรือการเขียนส่วนประกอบภายในคำติดกัน ก็มีอิทธิพลต่อความคิดในการตัดสินคำด้วยเกณฑ์ทางจิตวิทยานี้ถึงแม้จะไม่ใช่เกณฑ์ที่เป็นทางการ แต่ก็ช่วยให้รับรู้คำในมุมมองของผู้ใช้ภาษาได้ ดังนั้นเกณฑ์นี้จึงมักใช้โดยนักภาษาศาสตร์ที่วิเคราะห์ภาษาที่ยังไม่มีตัวเขียน

ในวิทยานิพนธ์ฉบับนี้ เกณฑ์ทางจิตวิทยาจะใช้เสริมในกรณีที่ไม่สามารถตัดสินคำด้วยเกณฑ์อื่นๆได้ ซึ่งโดยมากมักเป็นกรณีของคำประกอบประเภทคำประสมที่มีความกำกวมอย่างมากในการตัดสินว่าเป็นคำประสมหนึ่งคำหรือเป็นคำเดี่ยวๆหลายคำ เช่น โรงไฟฟ้าพลังน้ำ, หม้อหุงข้าว ผู้วิจัยเห็นว่าการตัดสินคำประสมมักขึ้นอยู่กับมุมมองของผู้ตัดสิน แต่ละคนก็อาจไม่เห็นพ้องกันในการตัดสิน ดังนั้นผู้วิจัยในฐานะเจ้าของภาษาและใช้ภาษาไทยในการสื่อสารอยู่สม่ำเสมอ จะเป็นผู้พิจารณาตัดสินว่าสายอักขระดังกล่าวจัดเป็นคำหนึ่งคำหรือเป็นคำหลายคำ

เกณฑ์ใหญ่ๆทั้ง 3 เกณฑ์นี้สามารถนำมาใช้ตัดสินมโนทัศน์เรื่องคำได้ โดยในวิทยานิพนธ์นี้ใช้เกณฑ์ทั้ง 3 ร่วมกันในการตัดคำด้วยมือให้กับคลังข้อมูล ถึงแม้ว่าแต่ละเกณฑ์จะมีข้อจำกัดของตนเอง ซึ่งทำให้การใช้เกณฑ์ใดเพียงเกณฑ์เดียวไม่สามารถตัดสินคำได้อย่างลงตัว แต่เมื่อใช้เกณฑ์ต่างๆร่วมกันแล้วก็สามารถลดข้อบกพร่องลงได้ ทำให้สามารถตัดสินคำได้ครอบคลุมและลงตัวยิ่งขึ้น โดยเกณฑ์หลักที่วิทยานิพนธ์นี้เลือกใช้ คือ เกณฑ์ทางวากยสัมพันธ์ต่างๆทั้งหมด เนื่องจากเห็นว่าเกณฑ์ทางวากยสัมพันธ์นี้สามารถสังเกตได้จากข้อมูลจริงและมีความลื่นไหลน้อยที่สุด และร่วมด้วยการใช้เกณฑ์ทางความหมายเพื่อเป็นตัวกำหนดว่าเกณฑ์ทางวากยสัมพันธ์นั้นสามารถนำมาใช้ได้หรือไม่ในแต่ละกรณี กล่าวคือ หากใช้เกณฑ์ทางวากยสัมพันธ์นั้นแล้วทำให้ความหมายผิดเพี้ยนไปหรือคำที่ได้ไม่มีความหมายก็จะถือว่าเกณฑ์ทางวากยสัมพันธ์ดังกล่าวไม่สามารถนำมาใช้ได้ และหากว่าเกณฑ์ทั้งสองยังไม่สามารถตัดสินได้ ผู้วิจัยก็จะเลือกใช้เกณฑ์ทางจิตวิทยาโดยผู้วิจัยในฐานะเจ้าของภาษาคนหนึ่งจะเป็นผู้พิจารณาว่า สายอักขระดังกล่าวจัดเป็นหนึ่งคำหรือเป็นหลายคำ

4.1.2 การตัดสินปัญหาความกำกวมในการตัดคำให้กับคลังข้อมูล

ในหัวข้อนี้จะได้กล่าวถึงการนำเกณฑ์ต่างๆที่ได้เสนอมาใช้ตัดสินความกำกวมลักษณะต่างๆในภาษาไทยที่ก่อให้เกิดปัญหาต่อการตัดคำให้กับคลังข้อมูล ในหัวข้อที่ 4.1.2.1 จะได้กล่าวถึงการตัดสินความกำกวมที่เกิดจากการที่คำในภาษาไทยเขียนติดกัน ในหัวข้อที่ 4.1.2.2 จะได้กล่าวถึงการตัดสินความกำกวมที่เกิดจากชื่อเฉพาะในภาษาไทยที่มักปรากฏร่วมกับคำนามสามัญอยู่เสมอ ในหัวข้อที่ 4.1.2.3 กล่าวถึงการตัดสินความกำกวมที่เกิดจากคำประกอบในภาษาไทยซึ่งเกิดจากในภาษาไทยยังไม่มีเกณฑ์ในการบ่งบอกขอบเขตของคำที่แน่ชัด จากนั้นในหัวข้อที่ 4.1.2.4 จะได้กล่าวถึงการพิจารณาตัดคำในลักษณะพิเศษอื่นๆในคลังข้อมูล ดังนี้

4.1.2.1 การตัดสินความกำกวมที่เกิดจากการที่คำในภาษาไทยเขียนติดกัน

ความกำกวมที่เกิดจากการที่คำในภาษาไทยเขียนติดกันนี้ ทำให้เกิดปัญหาว่าควรตัดคำที่ตำแหน่งใด ตัวอย่างเช่น

ตากลม	ตัดได้เป็น	ตา-กลม, ตาก-ลม
ภาพรอออกฉาก	ตัดได้เป็น	ภาพ-รอ-ออก-ฉาก, ภา-พร-ออ-อก-ฉาก

โคลง	ตัดได้เป็น	โคลง, โค-ลง
มากว่า	ตัดได้เป็น	มาก-ว่า, มา-กว่า

กล่าวคือ สายอักขระหนึ่งๆสามารถตัดคำได้หลายแบบ และแต่ละแบบก็มีความหมาย ในการแก้ปัญหาความกำกวมประเภทนี้ โดยทั่วไปสามารถอาศัยปริบทช่วยตัดสินเลือกการตัดคำที่ถูกต้องได้ เช่น “เขานั่งตากลม” ควรตัดเป็น “เขา-นั่ง-ตาก-ลม” ซึ่งมีความหมายเหมาะกับปริบท ในขณะที่ *“เขา-นั่ง-ตา-กลม” ไม่มีความหมายที่สมบูรณ์ ถึงแม้ว่าความกำกวมประเภทนี้ไม่พบในคลังข้อมูล แต่ได้นำมากล่าวไว้เนื่องจากเป็นความกำกวมที่โปรแกรมจะต้องแก้ปัญหาในการตัดคำด้วย

4.1.2.2 การตัดสินความกำกวมที่เกิดจากชื่อเฉพาะ

ความกำกวมประเภทนี้เกิดในคลังข้อมูล เนื่องจากในภาษาไทยสายอักขระที่เป็นชื่อเฉพาะมักปรากฏร่วมกับคำนามสามัญอยู่เสมอ ตัวอย่างในคลังข้อมูล เช่น

ชื่อเมือง ประเทศ สถานที่	เช่น	เมืองโฮจิมินห์, กรุงบรัสเซลส์, ประเทศมาเลเซีย
ชื่อสถาบัน บริษัท ธุรกิจ สิ่งประดิษฐ์	เช่น	กรมการประกันภัย, กรมประกันภัย, กรมส่งเสริมอุตสาหกรรม, กระทรวงพาณิชย์, กระทรวงคมนาคม, ธนาคารกรุงไทย, สำนักงานคณะกรรมาธิการสหภาพยุโรป, บริษัทจัสมิน อินเตอร์เนชั่นแนล จำกัด, บริษัทตลาดรองสินเชื่อเพื่อที่อยู่อาศัย, บริษัทผลิตไฟฟ้าจำกัด (มหาชน), ร้านเซเว่น-อีเลฟเว่น, สายการบินสิงคโปร์ แอร์ไลน์ส, คอมพิวเตอร์ไอบีเอ็ม, คณะกรรมการการเลือกตั้ง
ชื่อเชื้อชาติ, ภาษา ศาสนา	เช่น	เชื้อชาติจีน, คนจีน, คนไทย, ภาษาอังกฤษ, ศาสนาพุทธ
ชื่อเดือน วัน	เช่น	เดือนกันยายน, วันจันทร์
ชื่อโครงการ	เช่น	โครงการน้ำเทิน 2, โครงการ 14 สิงหาคม 2541
ชื่อคน กับคำนำหน้าชื่อหรือตำแหน่ง	เช่น	คุณชายจัตุมงคล, คุณทนง พิทยะ, นางรัตนาภรณ์ จึงสงวนสิทธิ์, นางสาวอาร์ลิน เดอ ลา ครูซ, นายชาญ อัศวโชค, นายกรัฐมนตรีฮุน เซน, รมช.วิชัย, รศ.ดร.ทองอินทร์ วงศ์โสธร

ความกำกวมที่เกิดขึ้นก็คือ สายอักขระเหล่านี้จะจัดเป็นคำหนึ่งคำ หรือเป็นคำหลายคำ ในวิทยานิพนธ์ฉบับนี้แยกคำนามสามัญกับชื่อเฉพาะที่เกิดร่วมกันนี้เป็นคนละคำกัน ด้วยเหตุผลดังต่อไปนี้

(1) ทั้งสองส่วนสามารถปรากฏโดยลำพังเป็นอิสระจากกัน เช่น “เดือน” กับ “กันยายน” หรือ “คน” กับ “จีน” สามารถปรากฏได้โดยลำพังโดยแต่ละส่วนยังมีความหมายคงเดิม
(2) ความกำกวมดังกล่าวสามารถใช้เกณฑ์การเกิดแทนที่ช่วยตัดสินได้ เช่น “คนจีน” กลายเป็น “ภาษาจีน” “ตัวอักษรจีน” หรือ “โครงการมิยาซาวา” เป็น “ทุนมิยาซาวา” เป็นต้น
(3) ความสัมพันธ์ทางความหมายระหว่างคำทั้งสองเป็นความสัมพันธ์แบบระบุประเภท คือ คำนามสามัญที่นำหน้าบอกสิ่งที่ชื่อเฉพาะที่ตามหลังเป็น กล่าวคือ “กันยายนเป็นเดือน”, “ไทยเป็นคน“ หรือ “ไทยเป็นภาษา” เป็นต้น
(4) ชื่อเฉพาะบางคำก็สร้างขึ้นจากคำนามสามัญ เช่น กรมส่งเสริมอุตสาหกรรม, กระทรวงพาณิชย์, บริษัทผลิตไฟฟ้าจำกัด กรณีเช่นนี้รูปภาษาไม่ได้บ่งบอกว่าสายอักขระดังกล่าวเป็นคำเดียวหรือไม่ จำเป็นต้องอาศัยความหมายในเชิงอ้างถึงสิ่งภายนอกภาษาเพื่อตัดสินว่าคำหรือกลุ่มคำนั้นเป็นชื่อของสถาบันหรือสิ่งเฉพาะเจาะจงหรือไม่ เช่น กระทรวงพาณิชย์ เป็นชื่อกระทรวงที่ทำหน้าที่ดูแลทางด้านการค้าของประเทศ, “ท่องไปในโลกกว้าง” เป็นชื่อหนังสือเป็นต้น และสังเกตได้ว่าคำเหล่านี้ในภาษาอังกฤษแม้ว่าเหมือนจะเป็นคำนามสามัญแต่มักเขียนเริ่มคำด้วยตัวอักษรใหญ่
(5) นามวลีอาจประกอบขึ้นจากคำนามสามัญคำเดียวประกอบกับหน่วยสร้างความรวม (coordinate construction) ที่มีชื่อเฉพาะหลายคำประกอบกันเป็นหน่วยสร้างได้ เช่น “ธนาคารกรุงเทพและกรุงไทย” ซึ่งหมายถึง “ธนาคารกรุงเทพ” และ “ธนาคารกรุงไทย”
(6) ชื่อคนที่เกิดร่วมกับคำนำหน้าชื่อหรือตำแหน่ง สามารถย้ายที่ได้ เช่น “พระยาอุปกิตศิลปสาร” เป็น “อุปกิตศิลปสาร, พระยา” บางคำแทรกกลางได้ (ด้วยช่องว่าง) เช่น “นายชาญ อัศวโชค” เป็น “นาย ชาญ อัศวโชค” นอกจากนี้ คำย่อก็จะย่อเฉพาะชื่อตำแหน่งหรือคุณวุฒิ เช่น “ร.ศ. ดร. ทองอินทร์” ย่อมาจาก “รองศาสตราจารย์ ดอกเตอร์ ทองอินทร์”

4.1.2.3 **การตัดสินความกำกวมที่เกิดจากคำประกอบ**

ความกำกวมของคำประกอบเป็นปัญหาที่พบมาก ซึ่งก่อให้เกิดความลำบากในการตัดคำภาษาไทยให้กับคลังข้อมูล เพราะนอกจากจะต้องตัดสายอักขระดังกล่าวออกจากสายอักขระ

ข้างเคียงแล้ว ยังต้องตัดสินด้วยว่าสายอักขระดังกล่าวมีสถานะเป็นคำหนึ่งคำหรือเป็นคำหลายคำ คำประกอบในวิทยานิพนธ์ฉบับนี้ หมายถึง คำใหม่ที่เกิดจากการประกอบหน่วยคำตั้งแต่ 2 หน่วยคำขึ้นไปเข้าด้วยกัน และมีหน่วยคำอย่างน้อย 1 หน่วยคำที่เป็นหน่วยคำอิสระได้ ซึ่งรวมเอาคำประสม คำประสานเทียม คำซ้อน และคำซ้ำอยู่ในประเภทนี้ ดังที่ได้กล่าวมาแล้วในบทที่ 2 ทั้งนี้ คำที่เกิดขึ้นมาใหม่นี้จะจัดเป็นคำใหม่หนึ่งคำมีความหมายเฉพาะเป็นของตนเอง ปัญหาของคำประกอบประเภทต่างๆนั้นมีข้อสังเกตและแนวทางการแก้ปัญหาต่างกัน ดังนี้

4.1.2.3.1 **การตัดสินความกำกวมของคำประสานเทียม** ซึ่งเกิดจากหน่วยคำไม่อิสระประกอบเข้ากับคำมูล เช่น “นักเรียน”, “ชาวบ้าน”, “โรงไฟฟ้า” สามารถแก้ปัญหาได้ โดยพิจารณาว่ามีส่วนประกอบที่ไม่สามารถปรากฏตามลำพังได้ จากตัวอย่าง “นัก-”, “ชาว-”, “โรง-” ไม่ปรากฏตามลำพัง ส่วนประกอบเหล่านี้จัดเป็นหน่วยคำไม่อิสระที่เกิดหน้าหรือคำอุปสรรค ดังนั้นจะไม่ประกอบเข้ากับส่วนที่นำมาข้างหน้า แต่จะประกอบเข้ากับส่วนที่ตามหลัง คือ “เรียน”, “บ้าน”, “ไฟฟ้า” เพื่อสร้างคำใหม่ขึ้นมา คำประกอบประเภทคำประสานเทียมนี้ถือว่าเป็นคำหนึ่งคำ

4.1.2.3.2 **การตัดสินความกำกวมของคำซ้ำ** ซึ่งเกิดจากคำเดียวกันเกิดซ้ำกันหรือคำที่มีเสียงคล้ายกันมาประกอบกันเข้า เช่น “แดงๆ”, “ดีๆ”, “โฉ่งฉ่าง”, “ด๊ำดำ”, “ตกอกตกใจ” สามารถแก้ปัญหาโดยสังเกตรูปแบบทางเสียงของส่วนประกอบทั้งสองจะคล้ายคลึงกัน คือมีเสียงเหมือนกันหรือมีสัมผัสสระหรือพยัญชนะระหว่างกัน คำประกอบประเภทคำซ้ำนี้ถือว่าเป็นคำหนึ่งคำ

4.1.2.3.3 **การตัดสินความกำกวมของคำซ้อน** ซึ่งเกิดจากคำที่มีความหมายใกล้เคียงหรือตรงข้ามกันมาประกอบเข้าด้วยกัน เช่น “บอกกล่าว”, “เร็วไว”, “แก่เฒ่า”, “เป็นตายร้ายดี” สามารถแก้ปัญหาได้โดยสังเกตรูปแบบทางความหมายของส่วนประกอบทั้งสองส่วนจะเกี่ยวเนื่องกัน ซึ่งอาจคล้ายคลึงหรือขัดแย้งกัน คำประกอบประเภทคำซ้อนนี้ถือว่าเป็นคำหนึ่งคำ

4.1.2.3.4 **การตัดสินความกำกวมของคำประสม** ซึ่งเกิดจากคำมูล 2 คำหรือมากกว่านั้นประกอบกันเข้าเป็นคำใหม่ เป็นความกำกวมที่มีจำนวนมากและตัดสินได้ยากที่สุด ในภาษาไทยไม่บ่งบอกด้วยการเว้นหรือไม่เว้นระหว่างส่วนประกอบภายในคำเหมือนอย่างเช่นภาษาอังกฤษ และไม่สามารถสังเกตร่องรอยในการประกอบเข้าด้วยกันได้ชัดเจนเท่ากับ

คำประกอบประเภทอื่น รูปแบบการประกอบเข้าด้วยกันของคำประสมก็มีลักษณะหลากหลาย ตัวอย่างในคลังข้อมูล เช่น

คำนาม-คำนาม	เช่น	พ่อแม่, เรือประมง, หลังมือ, หนี้ภาครัฐ, ฝ่ายการเงิน
คำนาม-คำกริยา	เช่น	ข้าวสวย, ท้องเดิน, ไก่เขี่ย, หุ้นอ้างอิง, ฝ่ายบริหาร
คำนาม-คำบุพบท	เช่น	บ้านนอก, วงใน
คำบุพบท-คำนาม	เช่น	ใต้ดิน
คำกริยา-คำนาม	เช่น	ล้วงลูก, ล้วงกระเป๋า, จับกบ, ทำงาน
คำกริยา-คำกริยา	เช่น	จำนำ, ลูบคม

ความกำกวมที่เกิดขึ้น ก็คือ สายอักขระดังกล่าวเป็นจะจัดเป็นคำประสมหนึ่งคำหรือเป็นคำหลายคำที่มีความสัมพันธ์ทางวากยสัมพันธ์ต่อกัน (syntactic group) หรือวลี อย่างไรก็ตาม วิทยานิพนธ์ฉบับนี้มีแนวทางที่ช่วยตัดสินความกำกวมของคำประสม ดังนี้

(1) ทางด้านความหมาย คำประสมโดยทั่วไปมีความหมายเฉพาะที่ไม่เท่ากับผลรวมทางความหมายของส่วนประกอบสองส่วน เช่น “กระดานดำ” ไม่เท่ากับ กระดานสีดำ และกระดานดำโดยทั่วไปก็เป็นสีเขียว, “ไฟฟ้า” ไม่ได้หมายถึง ไฟและฟ้า แต่เป็นคลื่นชนิดหนึ่งที่เดินทางโดยอาศัยตัวนำ เป็นต้น แตกต่างจากวลีที่มีความหมายเท่ากับผลรวมทางความหมายของแต่ละคำ เมื่อใช้เกณฑ์นี้ สำนวน (idiom) ก็จัดเป็นคำประสมด้วย เนื่องจากมีความหมายเฉพาะเช่นกัน เช่น “น้ำตาเช็ดหัวเข่า”, “กินน้ำใต้ศอก”

(2) คำประสมไม่มีความเป็นผลิตผลภาวะ[*] (productivity) คือ คำประสมไม่ใช่เกิดได้กับคำทุกคำ เช่น มีคำว่า “กระดานดำ” ในภาษาไทยในปัจจุบัน แต่ไม่มี *“กระดานแดง”, *“กระดานเขียว” ในขณะที่กลุ่มคำหรือวลีจะมีลักษณะเป็นผลิตผลภาวะ เช่น มีได้ทั้ง “ดอกไม้แดง”, “ดอกไม้เขียว”, “ดอกไม้เหลือง” (ในการใช้เกณฑ์นี้ ผู้วิจัยพิจารณาเฉพาะสภาวะของคำศัพท์ในภาษาในเวลาปัจจุบันเท่านั้น ไม่ได้ ดังนั้น จึงไม่ได้หมายความว่า คำศัพท์ เช่น “กระดานแดง” จะไม่มีโอกาสเกิดขึ้น เพราะคำศัพท์นี้อาจจะเกิดขึ้นในอนาคตก็ได้)

---

[*] ศัพท์ภาษาไทยที่ อมรา ประสิทธิ์รัฐสินธุ์, ยุพาพรรณ หุ่นจำลอง และสรัญญา เศวตมาลย์ (2544: 160) ใช้ ส่วนราชบัณฑิตยสถาน(http://www.royin.go.th) บัญญัติไว้ว่า “ผลิตภาพ”

(3) คำประสมมีระดับของความเหนี่ยวแน่น (cohesion) ต่างกัน Kramsky (1969: 54) กล่าวว่า ในภาษาอังกฤษระดับความเหนี่ยวแน่นภายในคำประสมจะแสดงออกมาจากการเขียนส่วนประกอบทั้งสองส่วนแยกกันหรือติดกัน คำประสมที่เขียนติดกันมีความเหนี่ยวแน่นมากที่สุด คำประสมที่เขียนเชื่อมแต่ละส่วนด้วยเครื่องหมายขีด (dash) มีความเหนี่ยวแน่นรองลงมา ส่วนคำประสมที่เขียนแยกกันมีความเหนี่ยวแน่นน้อยที่สุด แต่บางทีสายอักขระเดียวกัน พจนานุกรมบางเล่มก็เขียนแยกบางเล่มก็เขียนแบบมีขีดคั่น เช่น “sleeping car” กับ “sleeping-car”, “well known person” กับ “well-known person” หรือในภาษาเชค (Czech) คำประสมที่เป็นคำวิเศษณ์ เช่น “natolik”(in such a degree) ก็สามารถปรากฏแยกกันเป็นบุพบทวลี “no tolik”(so much) ได้ ซึ่งในกรณีนี้ Kramsky (1969) กล่าวว่าจะแสดงถึงระดับความเป็นเอกภาพทางมโนทัศน์ (conceptual unity or diversity) ของคำเหล่านี้ กล่าวคือ บุพบทวลีซึ่งเขียนคำแยกจากกัน มีเอกภาพทางมโนทัศน์น้อยกว่าคำวิเศษณ์ที่เขียนติดกัน สำหรับภาษาไทย คำอธิบายในแง่ระดับความเหนี่ยวแน่นนี้สามารถอธิบายได้ว่าทำไมคำหนึ่งๆจึงถูกตัดสินต่างกัน เช่น “โรงไฟฟ้าพลังน้ำ” บางครั้งถูกตัดสินเป็นหนึ่งคำ บางครั้งถูกตัดสินเป็น “โรงไฟฟ้า” กับ “พลังน้ำ” แยกกัน ในขณะที่ “ข้าวสวย” แทบจะไม่เคยถูกตัดสินให้เป็นคนละคำกันเลย และ “โรงไฟฟ้าพลังน้ำ” ก็ไม่น่าจะมีใครตัดสินแยกให้เป็น “โรงไฟฟ้าพลัง” กับ “น้ำ” เนื่องจากว่า “พลัง” กับ “น้ำ” มีความเหนี่ยวแน่นระหว่างกันมากกว่า “โรงไฟฟ้า” กับ “พลัง”

(4) กรณีสายอักขระมีความหมายในเชิงแสดงความเป็นเจ้าของ เช่น “นายกรัฐมนตรีอังกฤษ” “หนังสือห้องสมุด” จะไม่ถือว่าเป็นคำประสม แต่ถือว่าเป็นคำสองคำ เนื่องจากมีความหมายเท่ากับผลรวมของทั้งสองส่วน คือ “นายกรัฐมนตรีของอังกฤษ”

4.1.2.4 การตัดคำในลักษณะพิเศษอื่น ๆในวิทยานิพนธ์

(1) ตัดคำ “การ-” และ “ความ-” ที่นำหน้านามวลีแยกออกมามีสถานะเป็นหน่วยคำ เนื่องจากการสร้างคำนามหรือนามวลีด้วย “การ-” “ความ-” เป็นกระบวนการที่มีความเป็นผลิตผลภาวะ (productivity) และปรากฏเป็นจำนวนมากในภาษาไทย เช่น การแต่งงาน, การคิดไตร่ตรอง, ความรักชาติ, ความเป็นมนุษย์ เป็นต้น การแยกคำทั้งสองออกมาทำให้สามารถอธิบายตัวอย่าง “การพัฒนาและคิดสร้างสรรค์สิ่งใหม่...” ว่า “พัฒนา” และ “คิดสร้างสรรค์” ประกอบกันเป็น

หน่วยสร้างความรวม (coordinate construction) แล้วหน่วยคำ “การ-” จึงประกอบเข้ากับทั้งหน่วยสร้างอีกทีหนึ่ง ส่วนกรณีที่ไม่แยก “การ-“ และ “ความ-“ ออกมา ได้แก่ คำที่แสดงมาตรวัด เช่น ความยาว, ความสูง, ความถี่, ความสวย คำที่มีความหมายถึงเหตุการณ์ (event) เช่น การสัมมนา, การบรรยาย, การประชุม คำที่แสดงหน่วยของความรู้สึกนึกคิดไม่ได้แสดงการกระทำ(expression) เช่น ความคิด, ความรัก, ความรู้สึก คำที่หมายความว่า ” เรื่องหรือธุระที่เกี่ยวกับ” ที่ปรากฏหน้าคำนาม เช่น การเมือง, การเงิน, การธนาคาร และคำที่ประกอบกันขึ้นเป็นคำซ้อน เช่น การเรียนการสอน, การบ้านการเมือง ซึ่งถือว่าเป็นคำหนึ่งคำ

(2) ตัดคำในข้อความแสดงจำนวนนับโดยพิจารณาจากลักษณะของคำ กล่าวคือ การแสดงจำนวนนับสามารถแสดงได้ทั้งในรูปตัวเลขและตัวอักษร ใน 3 ลักษณะ ได้แก่ ตัวเลขอย่างเดียว“10,000,000” ตัดเป็นหนึ่งคำ, ตัวเลขผสมกับตัวอักษร “10 ล้าน” ตัดเป็นสองคำ คือ “10” กับ “ล้าน”, และ ตัวอักษรอย่างเดียว “สิบล้าน” ตัดเป็นสองคำคือ “สิบ” กับ “ล้าน” เนื่องจากเห็นว่าแต่ละส่วนสามารถปรากฏได้โดยลำพัง

(3) ชื่อกับนามสกุลตัดแยกเป็นคนละคำกัน เนื่องจากเห็นว่าแต่ละส่วนสามารถปรากฏได้โดยลำพัง

ผู้วิจัยมีความเห็นว่า ปัญหาเรื่อง “คำ” นี้ยังคงเป็นปัญหาที่ต้องการการศึกษาในแง่มุมต่างๆอีกมาก เกณฑ์และคำอธิบายเรื่องคำที่ใช้ในวิทยานิพนธ์ฉบับนี้จึงยังไม่ใช่จุดสิ้นสุดและจุดลงตัวเสียทีเดียว บางงานวิจัยเสนอว่าควรอธิบายคำในแง่ของระดับความเป็นคำ บางคำจัดเป็นสมาชิกหลักจากคำนิยามของคำ ส่วนบางคำเป็นสมาชิกที่อยู่ริมขอบ เช่น Ross (1973 อ้างถึงใน สุโขทัยธรรมาธิราช, 2533: 39-40) กล่าวว่าไม่อาจจะแบ่งแยกระหว่างคำนามกับประโยคได้อย่างเด็ดขาด บอกได้แต่เพียงว่าคำหรือกลุ่มคำใดมีลักษณะเป็นประโยคหรือเป็นคำนามมากน้อยกว่ากัน เป็นต้น นอกจากนี้ปัญหาในการตัดสินคำบางปัญหามีสาเหตุมาจากการเปลี่ยนแปลงของภาษาตามกาลเวลา ตัวอย่างเช่น ในบางภาษาบุพบทวลีเริ่มกลายเป็นคำ ซึ่งอาจสังเกตได้จากการไม่ปรากฏการเว้นช่องว่างระหว่างบุพบทกับคำนามที่ตามมา และคำบุพบทก็กลายเป็น clitic ไป ซึ่งถ้าได้รับการยอมรับในสังคมให้เป็นบรรทัดฐาน (norm) ในการใช้ภาษา บุพบทวลีก็จะกลายเป็นคำหนึ่งคำไปในที่สุด แต่ระหว่างช่วงการเปลี่ยนแปลงจากวลีกลายเป็นคำนี้ ทำให้เกิดความลำบากที่จะเห็นพ้องกันได้ว่าควรตัดสินว่าเป็นคำหนึ่งคำหรือเป็นวลี

4.2 การกำหนดชุดหมวดคำภาษาไทย

ในส่วนนี้จะได้กล่าวถึงการกำหนดชุดหมวดคำภาษาไทยในวิทยานิพนธ์นี้ โดยเริ่มจากกล่าวถึงการเลือกการจัดแบ่งหมวดคำที่นำมาใช้เป็นต้นแบบจากชุดหมวดคำภาษาไทยต่างๆที่มีผู้เสนอไว้ จากนั้น ในหัวข้อที่ 4.2.1 จะได้กล่าวถึงวิธีการกำหนดชุดหมวดคำขึ้นใช้ในวิทยาพนธ์นี้ ซึ่งได้ผลลัพธ์เป็นชุดหมวดคำภาษาไทยดังที่แสดงไว้โดยละเอียดในหัวข้อที่ 4.2.2 ชุดหมวดคำนี้เป็นชุดหมวดคำที่ได้นำไปใช้กำกับหมวดคำให้กับข้อมูลการใช้ภาษาจริง จึงเป็นชุดหมวดคำที่เหมาะสมและสามารถใช้งานได้จริงในระดับหนึ่ง ซึ่งในการใช้ชุดหมวดคำดังกล่าวกำกับคลังข้อมูลฝึกสอนได้พบปัญหาความกำกวมในการกำกับหมวดคำเป็นจำนวนมาก ดังนั้น หัวข้อ 4.2.3 จึงจะกล่าวถึงแนวทางที่ใช้ในการตัดสินปัญหาความกำกวมที่เกิดขึ้น ดังนี้

จากการทบทวนการจัดแบ่งหมวดคำภาษาไทยที่มีผู้เสนอไว้ดังแสดงในบทที่ 2 พบว่าภาษาไทยมีชุดหมวดคำอยู่ด้วยกันหลากหลายชุด การจัดแบ่งหมวดคำต่างๆที่มีอยู่แตกต่างกันทั้งในเรื่องหมวดคำที่ได้และในเรื่องวิธีการและเกณฑ์ที่ใช้ในการแบ่งหมวดคำ วิทยานิพนธ์ฉบับนี้ได้เลือกการจัดแบ่งหมวดคำที่ผู้วิจัยเห็นว่ามีความลื่นไหลน้อยที่สุด คือ การจัดแบ่งหมวดคำตามที่อมรา ประสิทธิ์รัฐสินธุ์ (2543) เสนอไว้ มาเป็นชุดหมวดคำต้นแบบ และนำมาทดลองใช้กับคลังข้อมูลในวิทยานิพนธ์นี้ แล้วจึงปรับชุดหมวดคำให้เหมาะสมสำหรับใช้กำกับหมวดคำให้กับคลังข้อมูลและเป็นชุดหมวดคำสำหรับโปรแกรมในการกำกับหมวดคำ และถึงแม้ว่ามีผู้ได้จัดทำชุดหมวดคำภาษาไทย (Virach Sornlertlamvanich et al, 1997) สำหรับใช้ในงานด้านการประมวลผลภาษาธรรมชาติมาก่อนแล้ว แต่จากการศึกษางานวิจัยดังกล่าว ผู้วิจัยเห็นว่าไม่ได้กล่าวถึงเกณฑ์และข้อถกเถียงในทางภาษาศาสตร์ที่ชัดเจนเพียงพอ จึงไม่ได้นำชุดหมวดคำดังกล่าวมาใช้ในวิทยานิพนธ์ฉบับนี้ เพราะผู้วิจัยอาจนำมาใช้ได้ไม่ตรงตามเกณฑ์และไม่มีความสม่ำเสมอคงที่ (consistency) ในการกำกับหมวดคำให้กับคลังข้อมูล

4.2.1 วิธีการกำหนดชุดหมวดคำภาษาไทย

ในหัวข้อนี้จะได้กล่าวถึงการนำการจัดแบ่งหมวดคำภาษาไทยของอมรา ประสิทธิ์รัฐสินธุ์ (2543) มาเป็นต้นแบบในการจัดแบ่งหมวดคำในวิทยานิพนธ์และพัฒนาต่อไปเป็นชุดหมวดคำที่ใช้ในวิทยานิพนธ์ฉบับนี้ โดยกล่าวถึงแนวคิด,หลักการและผลการจัดแบ่งหมวดคำของอมรา รวมทั้ง

สาเหตุที่เลือกใช้การจัดแบ่งหมวดคำดังกล่าว จากนั้น จะได้กล่าวถึงวิธีการที่ใช้ในการจัดแบ่งหมวดคำและหลักในการวิเคราะห์หมวดคำในวิทยานิพนธ์ฉบับนี้

วิทยานิพนธ์ฉบับนี้ได้เลือกใช้การจัดแบ่งหมวดคำของอมรา ประสิทธิ์รัฐสินธุ์ (2543) มาเป็นต้นแบบ ซึ่งงานวิจัยดังกล่าวได้ศึกษาวิจัยเรื่องหมวดคำจากการวิเคราะห์คลังข้อมูลภาษาไทยโดยใช้วิธีการทางวากยสัมพันธ์ เนื่องจากอมราเห็นว่าหมวดคำในภาษาไทยที่มีอยู่ยังมีข้อบกพร่องทั้งในเรื่องการใช้เกณฑ์ความหมายซึ่งเป็นอัตวิสัย ลื่นไหลได้ง่าย และเกณฑ์ที่ใช้ไม่สม่ำเสมอ ไม่คงที่และไม่ได้อยู่บนพื้นฐานของข้อมูลการใช้ภาษาจริง งานวิจัยดังกล่าวจึงยึดพื้นฐานของทฤษฎีไวยากรณ์พึ่งพาศัพทการก (Lexicase Dependency Grammar) (Starosta, 1988 อ้างถึงในอมรา ประสิทธิ์รัฐสินธุ์, 2543) ซึ่งอมราเห็นว่ามีจุดอ่อนน้อยที่สุดในการระบุหมวดคำ งานวิจัยดังกล่าวใช้หลักในการวิเคราะห์ คือ ใช้เกณฑ์ “การปรากฏร่วม”(co-occurrence) และ “การกระจายของคำ”(word distribution) เป็นเกณฑ์หลัก ไม่ใช้เกณฑ์ความหมายในการแบ่งหมวดคำ เพราะเห็นว่าความหมายเป็นสิ่งไม่แน่นอนและมีความลักลั่นระหว่างผู้พูดแต่ละคน และการวิเคราะห์จะยึดตามโครงสร้างที่เห็นในข้อมูลจริง โดยถือว่ามีโครงสร้างระดับเดียวในภาษา (ไม่แบ่งแยกเป็นโครงสร้างลึกและโครงสร้างผิว) งานวิจัยดังกล่าวสรุปว่า ในภาษาไทยมี 8 หมวดคำหลัก ได้แก่ กริยา, นาม, ตัวกำหนด, ตัวบอกปริมาณ, วิเศษณ์, บุพบท, สันธาน และอนุภาค และงานวิจัยนี้ยังมีลักษณะเด่นคือ การไม่แยก “คุณศัพท์” ออกมาเป็นหมวดคำต่างหากแต่ให้จัดรวมไว้เป็นกริยา และได้จัดแบ่งกริยาเป็น 16 หมวดคำย่อยตามเกณฑ์การปรากฏร่วมกับคำอื่นไว้ด้วย

สาเหตุที่วิทยานิพนธ์ฉบับนี้ได้เลือกการแบ่งหมวดคำตามที่อมรา ประสิทธิ์รัฐสินธุ์ (2543) เสนอไว้ มาเป็นต้นแบบสำหรับพัฒนาเป็นชุดหมวดคำภาษาไทยที่จะใช้กำกับหมวดคำให้กับคลังข้อมูลและเป็นชุดหมวดคำสำหรับโปรแกรม เพราะเห็นว่าการแบ่งหมวดคำของอมรามีจุดเด่นตรงที่ใช้เกณฑ์ทางวากยสัมพันธ์ที่มีความคงที่ สามารถสังเกตได้จากข้อมูลจริง นอกจากนั้น การแบ่งหมวดคำดังกล่าวยังได้มาจากการวิเคราะห์ข้อมูลภาษาไทยปัจจุบัน (ปี พ.ศ. 2543) ซึ่งเป็นข้อมูลที่ร่วมสมัยกับคลังข้อมูลที่จะใช้ในวิทยานิพนธ์ฉบับนี้ ซึ่งน่าจะสามารถนำเกณฑ์และหลักในการวิเคราะห์จากงานวิจัยนั้นมาเป็นต้นแบบและใช้จัดการกับข้อมูลในวิทยานิพนธ์ได้อย่างเหมาะสม

อย่างไรก็ดี งานวิจัยดังกล่าวได้ผลการวิเคราะห์เป็นหมวดคำหลักในภาษาไทย 8 หมวดคำ และแบ่งหมวดคำย่อยสำหรับกริยาไว้ 16 หมวดคำ วิทยานิพนธ์ฉบับนี้ได้ใช้หมวดคำดังกล่าวเป็นชุดหมวดคำตั้งต้นในการวิเคราะห์และจัดแบ่งหมวดคำภาษาไทย และยึดหลักในการวิเคราะห์และเกณฑ์หลักเรื่อง “การปรากฏร่วม” และ “การกระจายของคำ” เหมือนงานวิจัยต้นแบบ โดยได้นำชุด

หมวดคำและเกณฑ์หลักนั้นมาวิเคราะห์กับข้อมูลการใช้ภาษาจริงในวิทยานิพนธ์ เพื่อทดสอบว่าชุดหมวดคำดังกล่าวสามารถนำมาตัดสินหมวดคำให้กับข้อมูลในคลังข้อมูลนี้ได้หรือไม่ พบว่าหมวดคำหลัก 8 หมวดคำที่อมราเสนอไว้สามารถนำมาใช้ตัดสินข้อมูลในคลังข้อมูลที่ใช้ในวิทยานิพนธ์นี้ได้จริง แต่เนื่องจากผู้วิจัยเห็นว่าการกำกับหมวดคำในคลังข้อมูลเป็น 8 หมวดคำนั้นยังไม่ละเอียดเพียงพอสำหรับงานด้านต่างๆทางด้านการประมวลผลภาษาธรรมชาติที่ต้องการข้อมูลในรูปของคำที่มีการกำกับหมวดคำไว้ ดังนั้น จึงได้ทำการจัดแบ่งหมวดคำย่อยให้เหมาะสมเพื่อสร้างชุดหมวดคำสำหรับโปรแกรมตัดคำและกำกับหมวดคำที่จะพัฒนาขึ้น ซึ่งผู้วิจัยได้ทำการแบ่งหมวดคำย่อยของหมวดคำหลักบางหมวดคำ ได้แก่ นาม บุพบท วิเศษณ์ และได้ทำการปรับหมวดคำย่อยสำหรับกริยาที่อมราเสนอไว้ให้เหมาะสม ซึ่งทำให้จำนวนหมวดคำย่อยของกริยาลดลงเหลือเพียง 11 หมวด รวมทั้งได้เสนอให้มีหมวดคำเครื่องหมายเป็นหมวดคำหลักอีกหนึ่งหมวดคำ เนื่องจากเห็นว่าในคลังข้อมูลมีการใช้เครื่องหมายต่างๆอยู่เป็นจำนวนมาก ซึ่งเครื่องหมายเหล่านี้สมควรได้รับการกำกับหมวดคำเช่นเดียวกับคำอื่นๆในคลังข้อมูล ในการวิเคราะห์คลังข้อมูลและในการแบ่งหมวดคำย่อยนี้ ผู้วิจัยได้นำเกณฑ์ความหมายมาใช้ร่วมในการตัดสินหมวดคำด้วย เนื่องจากเห็นว่าจะละทิ้งในเรื่องความหมายไปเสียทีเดียวก็ไม่ได้ เนื่องจากความหมายเป็นสิ่งที่เป็นคุณสมบัติพื้นฐานของคำ ซึ่งสามารถช่วยตัดสินในกรณีที่รูปภาษาคลุมเครือและกรณีอปกติทั้งหลาย เช่น กรณีที่รูปภาษาที่ปรากฏซึ่งเป็นกฤตกรรมภาษา (performance) ไม่ได้ตรงไปตรงมากับความรู้ทางภาษาที่เป็นสามัตถิยะภาษา (competence) ตัวอย่างเช่น “พี่! เล็กน้ำ ไม่งอกชาม” คำว่า “งอก” ในที่นี้ไม่ได้มีความหมายเหมือนกับ กริยา “งอก” ใน “ต้นกล้ายังไม่<u>งอก</u>” แต่หมายถึง “ถั่วงอก” ซึ่งเป็นคำนาม แทบจะเป็นไปไม่ได้ที่คนไทยจะตัดสินคำว่า “ถั่วงอก” เป็นกริยา แต่จากตัวอย่าง คำนาม “งอก” ก็ปรากฏตามหลัง “ไม่” ได้ ฉะนั้นต้องพิจารณาความหมายด้วยว่าคำดังกล่าวมีความหมายอย่างไรในปริบท และตัวอย่างดังกล่าวยังแสดงให้เห็นว่า หากยึดติดอยู่กับโครงสร้างผิวแต่เพียงอย่างเดียว ก็จะไม่สามารถเห็นลักษณะธรรมชาติของภาษาได้ ดังนั้น ผู้วิจัยจึงได้เสนอหลักในการวิเคราะห์เพิ่มเติมขึ้นมาซึ่งแตกต่างจากหลักในการวิเคราะห์ของอมรา ได้แก่ อนุญาตให้มีการย้ายที่ของหน่วยสร้าง และอนุญาตให้มีสรรพนามไร้รูป ดังนั้น เมื่อประมวลหลักเกณฑ์ทั้งหมดที่ผู้วิจัยใช้ในการวิเคราะห์หมวดคำในวิทยานิพนธ์นี้ สามารถสรุปได้ดังนี้

(1) ใช้เกณฑ์ทางวากยสัมพันธ์ “การปรากฏร่วม” และ “การกระจายของคำ” เป็นเกณฑ์หลัก โดยพิจารณาเฉพาะคำที่เป็นส่วนหลักของหน่วยสร้าง
(2) พิจารณาความหมายในปริบทเพื่อช่วยเสริมในการตัดสินหมวดคำ

(3) โครงสร้างผิวเป็นโครงสร้างหลักในการตัดสินหมวดคำแต่ไม่ได้ยึดติดอยู่เพียงแค่โครงสร้างผิวอย่างเดียวเท่านั้น

(4) คำแต่ละคำสามารถเป็นได้เพียงหมวดคำเดียว ดังนั้น คำที่มีรูปเหมือนกันแต่มีการปรากฏต่างกัน จัดว่าเป็นคนละคำกัน เช่น คำว่า “ที่” ใน “คนที่กำลังเดิน” กับ “อยู่ที่บ้าน” จัดว่าเป็นคนละคำกัน ดังนั้นจึงมีคำว่า “ที่” 2 คำ

(5) ใช้ความรู้ภาษาแม่ (native intuition) ของผู้วิจัยเพื่อช่วยตัดสินในกรณีที่ข้อมูลไม่ชัดเจนหรือไม่ครอบคลุม

(6) ใช้ชื่อที่ใช้กันมาแพร่หลายในอดีตในการเรียกชื่อหมวดคำเพื่อให้เป็นที่คุ้นเคยและจำได้ง่าย แต่คำจำกัดความของหมวดคำแต่ละหมวดไม่จำเป็นต้องเหมือนกับที่ใช้มาในอดีตทุกประการ

(7) อนุญาตให้มีการย้ายที่ของหน่วยสร้าง เช่น “ดูปรากฏการณ์ที่ว่านี้ได้ทั่วไปรอบตัว” กับ “ปรากฏการณ์ที่ว่านี้ดูได้ทั่วไปรอบตัว” ในกรณีเช่นนี้ถือว่าหน่วยสร้างมีการย้ายที่ โดยที่ความสัมพันธ์ในการปรากฏร่วมยังคงเดิม นั่นคือ “ปรากฏการณ์ที่ว่านี้” ยังคงเป็นส่วนเติมเต็มของกริยา “ดู” เพียงแต่ได้ย้ายที่มาอยู่ข้างหน้าเท่านั้น

(8) อนุญาตให้มีสรรพนามไร้รูป (zero pronoun) ได้ กล่าวคือ กรณีที่คำดังกล่าวเป็นที่เข้าใจกันระหว่างผู้พูดและผู้ฟังแล้วก็อาจละคำนั้นไว้ก็ได้ เช่น “ผู้ชายที่เขาฆ่า” กริยา “ฆ่า” ต้องมีคำนามมาปรากฏร่วมเพื่อให้กริยาสมบูรณ์ แต่ก็มักพบว่าไม่มีคำนามปรากฏในตำแหน่งดังกล่าว ดังนั้นวิทยานิพนธ์นี้จึงพิจารณาให้ตำแหน่งดังกล่าวมีสรรพนามไร้รูปอยู่ ดังนี้ “ผู้ชายที่เขาฆ่า∅” โดยจะพิจารณาเป็นสรรพนามไร้รูปเฉพาะกรณีที่เป็นหน่วยจำเป็น (argument) เท่านั้น ส่วนกรณีที่คำนั้นเป็นส่วนเสริม (adjunct) ที่มีหรือไม่ก็ได้ จะถือว่า ส่วนเสริมเป็นหน่วยที่สามารถละไปได้อยู่แล้ว เช่น ตัวกำหนดที่ขยายคำนาม, คำวิเศษณ์ที่ขยายคำกริยา ทั้งนี้ การระบุว่าสรรพนามไร้รูปอ้างถึงสิ่งใดสามารถสืบค้นได้จากการพิจารณาปริบทเนื้อความข้างเคียง

4.2.2 ชุดหมวดคำภาษาไทยที่ใช้ในวิทยานิพนธ์

หัวข้อนี้จะได้กล่าวถึงผลของการวิเคราะห์หมวดคำโดยใช้วิธีการและหลักในการวิเคราะห์ดังที่กล่าวมา ซึ่งได้นำมาจัดทำเป็นชุดหมวดคำภาษาไทยสำหรับกำกับหมวดคำให้กับคลังข้อมูล และเป็นชุดหมวดคำสำหรับโปรแกรมในการกำกับหมวดคำ หมวดคำภาษาไทยที่ใช้ในวิทยานิพนธ์ฉบับนี้จำแนกได้เป็น 9 หมวดคำหลัก และเมื่อแบ่งหมวดคำย่อยแล้ว ได้เป็นหมวดคำทั้งหมด 26 หมวดคำสำหรับใช้กำกับในวิทยานิพนธ์ฉบับนี้ ซึ่งมีรายละเอียดดังนี้

(1) กริยา (Verb)

กริยา คือ หมวดคำที่ปรากฏหลังคำว่า “ไม่” ได้ เช่น ไม่ทราบ, ไม่สามารถหา, ไม่มีเงิน ในงานวิจัยต้นแบบให้เหตุผลสนับสนุนการใช้คำว่า “ไม่” เป็นเกณฑ์ในการระบุคำกริยาว่า เป็นเกณฑ์ที่น่าเชื่อถือที่สุดและมีผู้ใช้ได้อย่างมีประสิทธิภาพมาก่อนแล้ว เช่น Chao (1970), อุดม วโรตม์สิกขดิตถ์ (2538), จรัสดาว อินทรทัศน์ (2539 อ้างถึงในอมรา ประสิทธิ์รัฐสินธุ์, 2543) ส่วนหมวดคำอื่นๆไม่สามารถปรากฏหลังคำว่า “ไม่” หมวดคำกริยาในที่นี้รวมถึงคำที่บางงานวิจัย เรียกว่า คุณศัพท์* เช่น สวย ดี หนัก ร้อน เนื่องจากสามารถปรากฏหลังคำว่า “ไม่” ได้เหมือนกริยาทั่วไป (อมรา ประสิทธิ์รัฐสินธุ์, 2543) วิทยานิพนธ์นี้แบ่งกริยาเป็น 11 หมวดคำย่อย โดยพิจารณาจากเกณฑ์การปรากฏร่วมกับหน่วยที่ตามหลัง และพิจารณาเฉพาะหน่วยที่จำเป็นของกริยา (argument) ได้แก่

(1.1) **กริยาที่ปรากฏลำพัง** (V0) กริยาในหมวดนี้สามารถปรากฏโดยลำพังเป็นกริยาที่สมบูรณ์ได้ บางคำมักมีคำวิเศษณ์ปรากฏตามหลังด้วย ซึ่งถือว่าคำวิเศษณ์เป็นส่วนเสริมของกริยาเท่านั้น ตัวอย่างในคลังข้อมูล เช่น

ยังไม่ต้องกังวลใจในปีนี้และปีหน้า
เรื่องนี้จะต้องยุติลงเช่นไร
การหยุดการเสียหายแล้วยืนขึ้นกับเดินหน้าต่อไป
สร้างความเข้มแข็งให้เกิดขึ้น
หนี้ของประเทศแทนที่จะลดกลับเพิ่มสูงขึ้น

(1.2) **กริยาที่ตามด้วยคำนาม** (VN0) กริยาที่ตามด้วยคำนามในวิทยานิพนธ์ฉบับนี้หมายความรวมถึง คำที่เรียกกันว่า สกรรมกริยา ที่ตามด้วยคำนามซึ่งทำหน้าที่เป็นกรรม และ

---

* คำประเภทนี้งานวิจัยต่างๆจัดให้อยู่ในหมวดคำต่างๆกันและใช้ชื่อเรียกหลากหลาย เช่น พระยาอุปกิตศิลปสาร (2514) และ กำชัย ทองหล่อ (2515) เรียกว่า ลักษณวิเศษณ์ ซึ่งเป็นหมวดคำย่อยของวิเศษณ์, บรรจบ พันธุเมธา (2514) เรียกว่า คำคุณศัพท์ ซึ่งจัดเป็นประเภทย่อยของคำวิเศษณ์ที่ทำหน้าที่ขยายนาม, อุดม วโรตม์สิกขดิตถ์ (2535) เรียกว่า วิเศษณกริยา ซึ่งจัดเป็นประเภทย่อยของกริยา, วิจินตน์ ภาณุพงศ์ (2532) เรียกว่า คุณศัพท์ เมื่อใช้ขยายนาม และเรียกว่า กริยาอกรรมย่อย เมื่อใช้เป็นภาคแสดง, นววรรณ พันธุเมธา (2527) จัดให้เป็น กริยาแสดงสภาพ เมื่อใช้เป็นภาคแสดง และจัดให้เป็น คำขยาย เมื่อใช้ขยายคำหลัก

กริยาอื่นๆที่ตามด้วยคำนามซึ่งทำหน้าที่เป็นส่วนเติมเต็มด้วย เช่น “ผมเป็นนักเรียน”, “สมชายไปต่างประเทศ”, “เกิดอุบัติเหตุขึ้นบนทางด่วน” ตัวอย่างในคลังข้อมูล เช่น

แม้จะไม่ค่อยชอบตัวละครที่เล่นเป็นแมวคือเจ้าทอม
เกิดรายการ วิ่งไล่จับกันระหว่างคุณธารินทร์...กับคุณชายจัตุมงคล
ไม่สามารถหา "เงินตราต่างประเทศ" มาใช้หนี้
การรวมบัญชีจะมีการเดินหน้าต่อไป
การดึงเอาเงินลงทุนในหลักทรัพย์ต่างประเทศของฝ่ายออกบัตรมาไว้ที่ฝ่ายการธนาคาร
ธนาคารไทยพาณิชย์เป็นอีกแห่งหนึ่ง ที่ปรับโครงสร้างหนี้ได้ใกล้เคียงกัน
การขออนุญาตเดินทางไปต่างประเทศ
มีส่วนอยู่เบื้องหลังเรื่องนี้ด้วย

เมื่อพิจารณาจากคลังข้อมูลที่ใช้ในวิทยานิพนธ์และข้อมูลภาษาไทยทั่วไป พบว่า คำกริยาประเภทนี้มักนำไปใช้โดยไม่มีคำนามตามหลังอยู่บ่อยครั้ง ซึ่งสามารถอธิบายได้ว่า การที่คำนามไม่ปรากฏไม่ทำให้ความสมบูรณ์ของถ้อยความเสียไป เนื่องจากคำนามดังกล่าวสามารถสืบค้นได้จากข้อความข้างเคียง หรือเข้าใจได้จากปริบททางสถานการณ์ที่อยู่ภายนอกภาษา ซึ่งอาจเกิดจากกระบวนการทางวากยสัมพันธ์หรือทางสัมพันธสาร* (discourse) (Wirote Aroonmanakun, 1999) เช่น การละคำนามในคุณานุประโยคที่ซ้ำกับคำนามที่เป็นส่วนหลัก, การย้ายที่คำนามมาอยู่ต้นประโยคในประโยคกรรมวาจก (passive), การย้ายที่คำนามมาอยู่ต้นประโยคเพื่อเน้น, การละคำนามเมื่อคำนามดังกล่าวเคยกล่าวถึงไปแล้วในข้อความข้างหน้า หรือเมื่อคำนามดังกล่าวเป็นที่เข้าใจกันระหว่างผู้พูดและผู้ฟังแล้ว เป็นต้น กรณีเหล่านี้จะพิจารณาให้คำกริยาดังกล่าวยังเป็นกริยาที่ตามด้วยคำนาม โดยถือว่ามีสรรพนามไร้รูปปรากฏอยู่ เช่น “สิ่งที่หลายฝ่ายหวาดผวา∅“ ตัวอย่างในคลังข้อมูล เช่น

เป้าหมายที่ธนาคารกรุงเทพตั้งไว้ทั้งปีอยู่ที่ 1 แสนล้านบาท
สิ่งที่หลายฝ่ายหวาดผวาคือเชิงการเมืองของกกต.
ทั้ง 2 กรณียังเป็นประเด็นที่ต้องพิจารณาอย่างรอบคอบ

* คำศัพท์ภาษาไทยที่ อมรา ประสิทธิ์รัฐสินธุ์ (2543) ใช้เรียก discourse

การผิดพลาดเสียหายที่เกิดควรจะถูกสกัดหยุดลงได้
78 ส.ว.ซึ่งถูก กกต. แขวน

(1.3) **กริยาที่ตามด้วยบุพบทวลี** (VPN0) บุพบทวลีที่ปรากฏร่วมกับกริยาในหมวดนี้ถือว่าเป็นส่วนที่จำเป็นของกริยา และมักจะไม่ย้ายที่ไปอยู่ต้นประโยค เช่น “เบาะสีครีมกลมกลืนกับลายวอลเปเปอร์” ไม่ปรากฏเป็น “กับลายวอลเปเปอร์ เบาะสีครีมกลมกลืน” ในขณะที่หากบุพบทวลีเป็นเพียงส่วนเสริมเท่านั้น เช่น “เขากว้างขวางมากในวงการนี้” จะสามารถปรากฏเป็น “ในวงการนี้ เขากว้างขวางมาก” ได้ ในตัวอย่างหลังนี้ “กว้างขวาง” ถือว่าเป็นคำกริยาที่ปรากฏลำพัง(1.1) คำกริยาที่ปรากฏลำพังอื่นๆก็สามารถปรากฏร่วมกับบุพบทวลีที่เป็นส่วนเสริมในลักษณะเช่นนี้ได้ เช่น “เขานอนในรถ” “เขาทำงานอยู่บนโรงอาหาร” ซึ่งไม่ถือว่ากริยาดังกล่าวจัดอยู่ในหมวด 1.3 นี้ ตัวอย่างในคลังข้อมูล เช่น

ระบุความเห็นของเขาขัดแย้งกับแนวคิดดังกล่าว
เงินมีแต่ใช้ไม่ได้ ก็ไม่ต่างจากไม่มีเงิน
ที่บอกเงินก็มีคืออยู่ในบัญชีของฝ่ายออกบัตร
ประเทศนี้อยู่ภายใต้ระบอบปกครองของขุนทหาร
นอกเหนือจากประเด็นปัญหาเหล่านี้แล้ว
นักธุรกิจคนจีนที่มองหาช่องทางเข้าสู่ตลาดสินค้าในสหรัฐ
คลินตันกำลังพ่ายแพ้ต่อความเย้ายวน

(1.4) **กริยาที่ตามด้วยส่วนเติมเต็ม** (VCV0) ส่วนเติมเต็มในที่นี้คือ ส่วนเติมเต็มที่เป็นประโยคหรืออนุพากย์ส่วนเติมเต็ม (complement clause) ที่มีตัวนำส่วนเติมเต็ม (complementizer) เช่น “ที่”, “ว่า”, “ให้” ปรากฏนำหน้า ตัวอย่างในคลังข้อมูล เช่น

ผู้นำรัฐบาลเวียดนามยอมรับว่า ถึงแม้เวียดนามเหนือหรือเวียดนามคอมมิวนิสต์จะเป็นฝ่ายคว้าชัยชนะ
ผู้นำรัฐบาลเวียดนามกล่าวว่า ปัญหาหลักของเวียดนามคือปัญหาเศรษฐกิจ
ต้องถือว่าเป็นหนูตัวแสบ
คงจำกันได้ว่า รมต.คลังเอง คือ คนต้นคิด
สมกับที่จะไว้วางใจให้ตรวจสอบสถาบันการเงิน

มีแต่ดีใจที่คนไทยนำเงินออมไปซื้อรถยนต์ใหม่ใช้กัน
จึงไม่ค่อยนิยมที่จะไปศึกษาต่อกันที่ประเทศนั่น
รณรงค์ ส่งเสริม ชักชวนให้นักเรียนจากประเทศต่างๆทั่วโลกไปศึกษาต่อในสถาบันการศึกษาของประเทศตน
แต่อยากให้ประชาชนเข้าใจว่า การเลือกพรรคประชาธิปัตย์ เป็นทางเลือกที่ปลอดภัยที่สุด

(1.5) **กริยาที่ตามด้วยคำนามและบุพบทวลี** (VNPN0) บุพบทวลีที่ปรากฏร่วมกับกริยาในหมวดนี้ถือว่าเป็นส่วนที่สัมพันธ์กับกริยาโดยตรง ตัวอย่างในคลังข้อมูล เช่น

หน้าที่ของเด็กจะได้รับการดูแลเอาใจใส่จากผู้ใหญ่มากขึ้น
ศูนย์เทคโนโลยีทางการศึกษา ได้มีการให้ความรู้กับบุคลากรของศูนย์
นายชาญได้เสนอแผนปรับโครงสร้างหนี้ ให้กับเจ้าหนี้
การขายหุ้นแก่ผู้จำหน่ายเครื่องจักรที่จะนำมาติดตั้งในโรงงาน
มั่นใจภาคธนาคารจะนำประเทศสู่การฟื้นตัว
หุ้นเพิ่มทุนที่บริษัทมหาชนจำกัดออกให้เจ้าหนี้
การที่ปตท.ต้องจ่ายค่าก๊าซฯที่ซื้อจากแหล่งยาดานาและเยตะกุนให้กับพม่า

(1.6) **กริยาที่ตามด้วยคำนาม 2 ตัว** (VNN0) กริยาในหมวดนี้ต้องมีคำนาม 2 ตัวมารองรับจึงจะสมบูรณ์ กริยาหมวดนี้ตรงกับหมวดคำที่ วิจินตน์ ภาณุพงศ์ (2532) เรียกว่า กริยาทวิกรรม ตัวอย่างในคลังข้อมูล เช่น

โครงการให้คำปรึกษาผู้ประกอบการ
หนี้ครบกำหนดไม่มีปัญญาชำระหนี้เขาได้
การใช้หนี้ธนาคารชาติอื่นๆ
การให้โอกาสรัฐบาลแก้ไขปัญหามานาน โดยหวังว่า
แต่จะไม่ให้การคุ้มครองการบริหารงานที่เกิดจากการทุจริต

(1.7) **กริยาที่ตามด้วยคำนามและส่วนเติมเต็ม** (VNCV0) ส่วนเติมเต็มในที่นี้คืออนุพากย์ส่วนเติมเต็ม (complement clause) ที่มีตัวนำส่วนเติมเต็ม (complementizer) ปรากฏนำหน้า ซึ่งมีความสัมพันธ์กับกริยาโดยตรง ตัวอย่างในคลังข้อมูล เช่น

เปรียบการบริหารว่า คล้ายกับการขับเครื่องบิน
ไม่สามารถแก้จุดอ่อนที่มีอยู่เก่าก่อนให้หมดหรือเบาบางลงได้
การสร้างความเข้มแข็งให้เกิดขึ้น
สำหรับแบงก์ชาติที่ถูกสังคมตราหน้าว่า คือต้นเหตุแห่งความเสียหายทางเศรษฐกิจ
ทั้งสอนทั้งสั่งและชี้แนะ "เจ้าบ้าน" อย่างมากมายเกินงาม ว่าจะต้องบริหารเศรษฐกิจอย่างไร

(1.8) **กริยาที่ตามด้วยกริยาเติมเต็ม** (VV0) กริยาในหมวดนี้เป็นกริยาที่ต้องปรากฏร่วมกับกริยาอีกตัวหนึ่งจึงจะสมบูรณ์ กริยาหมวดนี้แตกต่างจากกริยาช่วย (1.10) ตรงที่คำกริยาที่ตามด้วยกริยาเติมเต็มยังคงมีความหมายด้านเนื้อหาอยู่ไม่ได้มีความหมายเป็นทั่วไปเหมือนกับกริยาช่วย และการที่คำกริยาในหมวดนี้ต้องการคำกริยาอีกตัวหนึ่งมาปรากฏตามหลังนั้นเป็นคุณสมบัติของคำนั้นๆเอง เช่น “เขาเผอิญเปิดไปช่องการ์ตูน” ซึ่งหากไม่มีคำกริยาที่ตามหลัง *“เขาเผอิญ” ก็จะไม่ได้ความ แตกต่างจากปรากฏการณ์กริยาเรียง (serial verb) ในภาษาไทย ซึ่งผู้วิจัยเห็นว่า ปรากฏการณ์กริยาเรียงเป็นเรื่องในระดับวากยสัมพันธ์ที่คำกริยาหลายคำสามารถปรากฏเรียงต่อเนื่องกันได้ เช่น “เขานั่งดูโทรทัศน์” ซึ่ง “นั่ง” และ “ดู” แม้จะปรากฏเรียงกันแต่ก็ไม่ได้จัดเป็นกริยาในหมวดนี้ วิทยานิพนธ์นี้ตัดสินว่า “นั่ง/V0” เป็นกริยาที่ปรากฏลำพัง(1.1) และ “ดู/VN0” เป็นกริยาที่ตามด้วยนาม(1.2) เพียงแต่ปรากฏในรูปแบบโครงสร้างกริยาเรียงในภาษาไทย ตัวอย่างในคลังข้อมูล เช่น

เผอิญเปิดไปช่องการ์ตูนทีเอ็นทีทางเคเบิลทีวี
ไม่อาจหามาตรการและวิธีทำ ที่จะช่วยสร้างความมั่นใจให้กับประชาชน
และทยอยจ่ายกันทุกสามเดือนที่ครบกำหนด
นักศึกษาจากในประเทศเอเชียเช่นมาเลเซีย สิงคโปร์ อินโดจีน รวมถึงประเทศไทยนิยมไปศึกษาต่อกันที่นั่น
ผมไม่ขอเอ่ยชื่อเพราะเกรงจะเป็นการโฆษณาเชียร์กัน

ส่วนเวียดนามเองนั้นมัวแต่รบทำสงครามต่อเนื่อง
สถาบันการศึกษาของไทยดังที่กล่าวมา ควรจะยอมลงทุนเพื่อคุณภาพ
หลายพรรคการเมืองเริ่มจัดกระบวนทัศน์
พร้อมจะนำบุคคลที่มีศักยภาพ หรือชื่อเสียงหอมหวานเข้าไปเสียบแทน
ความไม่ชัดเจนของผู้แทนฯที่เตรียมผละจากพรรคเดิมมาร่วมพรรคใหม่
เพราะนิสัยเดิมๆ ชอบหยอกล้อเล่นกับเด็กสาวๆ
แฟนๆการ์ตูนที่ต้องการพูดคุยเกี่ยวกับตัวการ์ตูนที่พวกเขาชื่นชอบ

(1.9) **กริยาที่ตามด้วยประโยค** (VS0) ตัวอย่างในคลังข้อมูล เช่น

เจอเขาฉายหนังการ์ตูนเรื่อง ทอมแอนด์เจอร์รี่
เบิกออกมาใช้โดยฝ่ายการธนาคารไม่ได้ก็เลยเหมือนไม่มีเงิน
เรามักได้ยินนักบริหารเปรียบการบริหารว่า คล้ายกับการขับเครื่องบิน
ไม่อาจเข้าใจการเปลี่ยนแปลงและขาดความรู้ใหม่ จนทำให้บริหารเอาชนะข้อจำกัดไม่ได้
ไม่มีแผนงานที่ชัดเจน ที่เชื่อว่าจะทำให้อนาคตของภาคเกษตรของไทยออกจากมุมมืด

(1.10) **กริยาช่วย** (VAUX) วิทยานิพนธ์นี้จัดกริยาช่วยแยกออกมาเป็นหมวดคำย่อยของกริยาหมวดหนึ่งด้วย เนื่องจากเห็นว่ากริยาช่วยมีลักษณะทางความหมายและการเกิดร่วมต่างจากกริยาอื่นๆอยู่พอสมควร ในด้านความหมาย กริยาช่วยมีความหมายทั่วไปมากขึ้น และช่วยเสริมความหมายทางไวยากรณ์ให้กับกริยาหลักในแง่การณ์ลักษณะ (aspect), มาลา (mood), หรือวาจก (voice) ในด้านการปรากฏ โดยทั่วไปกริยาช่วยจะเกิดร่วมกับกริยาหลัก ส่วนใหญ่กริยาช่วยมักเกิดหน้ากริยาหลัก เช่น (ไม่)ได้กิน, สามารถกิน, ควรกิน, ถูกกิน แต่ก็มีกริยาช่วยที่เกิดหลังกริยาหลักด้วย เช่น ไม่สามารถทำได้, ทำ(ไม่)ได้ เนื่องจากกริยาช่วยทุกตัวสามารถปรากฏหลังคำว่า “ไม่” ได้ ดังนั้นจึงจัดให้อยู่ในหมวดคำกริยา ตัวอย่างในคลังข้อมูล เช่น

ลูกหลานใครได้ไปเรียนต่อที่อเมริกาหรือที่อังกฤษ
ได้มีการพัฒนาความสามารถขึ้นมาใหม่หรือไม่
อนาคตของภาคเกษตรของไทยออกจากมุมมืด สู่ทางที่สว่างได้

เหตุผลเบื้องหน้าเบื้องหลัง ที่ประชาชนควรได้จากทั้งสองฝ่าย

สำหรับแบงก์ชาติที่ถูกสังคมตราหน้าว่า คือต้นเหตุแห่งความเสียหายทางเศรษฐกิจ

เราสามารถเป็นศูนย์การศึกษาระดับอุดมศึกษาในภูมิภาคเอเชียอาคเนย์ได้

ประเทศไทยซึ่งเคยมีศักดิ์ศรีและความพร้อมทุกด้าน แต่ทุกวันนี้กลับต้องมาอยู่ในสภาพ "กินน้ำใต้ศอก"

จึงไม่ค่อยนิยมที่จะไปศึกษาต่อกันที่ประเทศนั่น

ประเทศที่ได้เคยไปศึกษาร่ำเรียนมา

ส่วนกรณีของ “ไป” “มา” ที่เกิดร่วมกับกริยาหลัก เช่น เขาเดินไป, เขากินขนมชิ้นสุดท้ายไป, ซึ่งไม่สามารถปรากฏกับ “ไม่” ในปริบทของประโยคตัวอย่าง *เขาเดินไม่ไป, *เขากินขนมชิ้นสุดท้ายไม่ไป จะอาศัยความรู้ทางด้านความหมายของคำมาช่วย กล่าวคือ “ไป” ใน “เขาเดินไป” มีความหมายถึงการเคลื่อนที่เหมือนกับ “ไป” ที่เป็นกริยาหลัก ดังนั้นจึงจัดเป็นคำกริยา (เป็นกริยาประเภทหนึ่ง แต่ไม่ใช่กริยาช่วย) ส่วน “เขากินขนมชิ้นสุดท้ายไป” มีความหมายไม่เหมือน “ไป” ที่เป็นกริยา และไม่สามารถปรากฏกับ “ไม่” ด้วย วิทยานิพนธ์นี้จะจัดเป็นคำวิเศษณ์ ซึ่งได้แก่ “ไป” ที่มีความหมายบอกเวลาในอดีตที่ผ่านไปแล้ว และ “ไป” ที่บอกทิศทาง (ดังนั้น จึงไม่มีคำว่า “ไป” ที่เป็นกริยาช่วยในวิทยานิพนธ์นี้)

(1.11) กริยาคุณศัพท์ (VADJ) วิทยานิพนธ์นี้จัดกริยาคุณศัพท์แยกออกมาเป็นหมวดคำย่อยหมวดหนึ่งด้วย เนื่องจากเห็นว่ากริยาคุณศัพท์มีลักษณะทางความหมายที่แตกต่างจากกริยาทั่วไปอยู่พอสมควร กล่าวคือ กริยาคุณศัพท์มีความหมายระบุคุณสมบัติ สภาพ หรือลักษณะ และโดยทั่วไปสามารถใช้เปรียบเทียบระดับได้ (comparison) ทางด้านการปรากฏ กริยาคุณศัพท์สามารถปรากฏลำพังไม่ต้องการส่วนประกอบอื่นใดตามหลัง (เหมือนกริยาที่ปรากฏโดยลำพัง 1.1) ทางด้านหน้าที่ กริยาคุณศัพท์สามารถปรากฏเป็นภาคแสดง หรือเป็นส่วนขยายของคำนามหรือกริยาตัวอื่นได้ ตัวอย่างในคลังข้อมูล เช่น

ที่คนไม่ชอบเจ้าแมวทอมก็เพราะตัวมันใหญ่กว่า

ตรงนี้คุณทนง พิทยะน่าจะรู้ในรายละเอียดดี

ส่วนที่กู้และมีระยะเวลาคืนนานหน่อย

เพื่อให้คนดูได้อรรถรสไม่เงียบเหงา

ทำกิจกรรมที่ซ่อนเร้น ในสถานที่มิดชิด
เราอยู่ในสังคมที่มีทั้งคนที่ดีและเลวสะสมปนเปกันไป
เขามองหาหลักทรัพย์ที่น่าสนใจ

(2) **นาม** (Noun)

นาม คือ หมวดคำที่มีลักษณะสำคัญ 2 ประการ คือ (1) สามารถปรากฏหน้ากริยา, หลังกริยา หรือหลังบุพบทได้ (2) สามารถปรากฏหน้าตัวกำหนดได้ วิทยานิพนธ์นี้แบ่งหมวดคำนามเป็น 4 หมวดคำย่อย โดยพิจารณาจากทั้งความหมายและการปรากฏร่วม ได้แก่

(2.1) **นามสามัญ** (Common Noun) คือ คำนามที่สามารถปรากฏได้ตามเกณฑ์โดยทั่วไป คำนามสามัญมีความหมายเป็นชื่อทั่วไปไม่ได้บอกเจาะจงคนใดคนหนึ่งหรือสิ่งใดสิ่งหนึ่ง เช่น รัฐบาล, นายกรัฐมนตรี ตัวอย่างในคลังข้อมูล เช่น

ไซ่ง่อนอยู่ในกำมือพวกเขาตลอดเวลา
แต่ก็ยังมีศักยภาพและทรัพยากรทุกด้านพร้อมกว่าประเทศใดๆ
เรื่องนี้ ถ้าหากธนาคารแห่งประเทศไทยในฐานะคนรับผิดชอบจะต้องใช้หนี้ดังกล่าว
มีหน้าที่ก็ต้องทำ เพื่อปฏิรูปการเมือง
อาศัยการประมาณบางองค์ประกอบขึ้นมา
เงินมีแต่ใช้ไม่ได้

(2.2) **นามเฉพาะ** (Proper Noun) คือ คำนามที่โดยทั่วไปไม่ปรากฏหน้าตัวกำหนด แต่ก็สามารถปรากฏได้ในกรณีต้องการระบุแยกสองสิ่งที่มีชื่อเหมือนกัน เช่น “สมชายนี้ ไม่ใช่สมชายนั้น” และไม่ปรากฏกับกริยาคุณศัพท์ในหน่วยสร้างเข้าศูนย์ (endocentric construction) นอกจากนั้นคำนามเฉพาะมักปรากฏร่วมกับคำนามสามัญที่เป็นคำนำหน้าชื่อ เช่น มหาวิทยาลัยธรรมศาสตร์, ภาษาจีน คำนามเฉพาะมีความหมายชี้เฉพาะเจาะจง เป็นชื่อคนใดคนหนึ่งหรือสิ่งใดสิ่งหนึ่ง เช่น สมชาย, ธรรมศาสตร์ ตัวอย่างในคลังข้อมูล เช่น

ไซ่ง่อนอยู่ในกำมือพวกเขาตลอดเวลา
การบินไทยได้เชิญผู้บริหารสายการบินยูไนเต็ด แอร์ไลน์ และสายการบินลุฟท์ฮันซ่า

เจ้าหนี้ซับไมครอนถกแผนฟื้นฟูใหม่
คะแนนสงสารเต่ากับเกลียดชวน
ปตท.ต้องซื้อก๊าซฯจากพม่าในราคาที่สูงกว่าอ่าวไทย

(2.3) **ลักษณนาม** (Classifier Noun) คือ คำนามที่สามารถปรากฏในกรอบวลี [นาม ตัวบอกปริมาณ ___] และมักปรากฏในกรอบวลี [นามวลี ___ ตัวกำหนด] คำลักษณนามใช้บอกหน่วยนับหรือใช้เป็นหน่วยในการบอกปริมาณของคำนามทั่วไป เช่น คน, บาท, ครั้ง, ตัว, เล่ม ตัวอย่างในคลังข้อมูล เช่น

เงินจากธุรกิจผิดกฎหมายมีมูลค่าหลายหมื่นล้านบาท
ยกเว้นกรรมการผู้จัดการคนนั้น อาจเป็นคณะกรรมการบริหารอยู่ด้วย
ทำให้คณะกรรมการบริหารหลายบริษัท ต้องรับผิดชอบต่อความเสียหายจากปัญหาดังกล่าว
ทุกครั้งที่เราคิดถึงเรื่องชีวิตความเป็นอยู่ เราได้แต่กล้ำกลืนความรู้สึกไว้ลึกๆ

(2.4) **สรรพนาม** (Pronoun) คือ คำนามที่โดยทั่วไปไม่ปรากฏกับกริยาคุณศัพท์ที่ทำหน้าที่เป็นส่วนขยาย เช่น *“เธอสวยขับรถ” แต่เกิดกับกริยาคุณศัพท์ที่ทำหน้าที่เป็นภาคแสดงได้ เช่น “เธอสวย” คำสรรพนามใช้เรียกแทนคำนาม เช่น เขา, คุณ, ใคร, อะไร ตัวอย่างในคลังข้อมูล เช่น

คนอื่นที่เขาลงขันให้นั้นเงื่อนไขการชำระอาจจะยืดหยุ่นกว่า
อะไรที่เป็นไปไม่ได้หรือเกินจริง คนไทยจะบอกว่าพูดเป็นหนังการ์ตูนไปได้
ชอบปลอมพระองค์เพื่อไปถามทุกข์สุขความเดือดร้อนของราษฎรของพระองค์อยู่เสมอๆ
สภาพประเทศไทยคงไม่ต้องตกอยู่ในสภาพที่ต้องรอให้ใครมาช่วย

(3) **ตัวกำหนด** (Determiner)

ตัวกำหนด คือ หมวดคำที่สามารถปรากฏหลังคำนามได้ โดยตัวกำหนดจะเป็นส่วนขยายของคำนาม เช่น บ้านนี้, ชาติอื่น, สิ่งใด, ท่านทั้งหลาย, เงินทั้งหมด ตัวอย่างในคลังข้อมูล เช่น

ใช้หนี้ธนาคารชาติอื่นๆ ในแถบนี้
ยกเว้นกรรมการผู้จัดการคนนั้น อาจเป็นคณะกรรมการบริหารอยู่ด้วย
ผู้อ่านประมวลภาพเหล่านั้นออกมา
คนอื่นที่เขาลงขันให้นั้นเงื่อนไขการชำระอาจจะยืดหยุ่นกว่า
ชิพตัวไหนราคาเท่าไหร่
เรื่องของ"ทอมแอนด์เจอร์รี่"ก็เอวังด้วยประการฉะนี้แล
เสนอขายต่อประชาชนทั่วไป

(4) ตัวบอกปริมาณ (Quantifier)

ตัวบอกปริมาณ คือ หมวดคำที่ปรากฏอยู่หน้านาม และเป็นส่วนหนึ่งของหน่วยสร้างเข้าศูนย์ (endocentric construction) ที่ประกอบด้วยส่วนหลักและส่วนขยาย โดยตัวบอกปริมาณเป็นส่วนขยายของคำนาม ดังนั้นแม้ตัวบอกปริมาณจะปรากฏในตำแหน่งเดียวกับคำหน้าหน่วยสร้างไร้ศูนย์แต่อยู่ในหน่วยสร้างคนละประเภทกัน ตัวบอกปริมาณนี้รวมถึงจำนวนนับ (Numeral) ด้วย ตัวอย่างเช่น หนึ่งคน, กึ่งศตวรรษ, ครึ่งอัน, ปวงชน, สารพันปัญหา, ทุกที่, ทั้งปี ตัวอย่างในคลังข้อมูล เช่น

คิดเป็นเงินรวมกัน 2,500 ล้านเอสดีอาร์
สามารถหลบหนีไปได้ทุกครั้ง
บางคนอาจเก่งและชอบเรื่อง ขี่ม้า
หลายพรรคการเมืองเริ่มจัดกระบวนทัศน์
การจัดสรรพกำลังในแต่ละพื้นที่
บรรดาผู้ว่าส.ว.หน้าเดิมก็ยังเดินพาเหรดเข้าสภา

(5) วิเศษณ์ (Adverb)

วิเศษณ์ คือ หมวดคำที่ปรากฏหน้าหรือหลังกริยาได้เช่นเดียวกับคำนาม เช่น กิจกรรมนี้จะเป็นจุดเริ่ม, อย่าเพิ่งรีบกลับ, ผู้อ่านคงต้องให้ความสนใจ, ครูสวนขึ้นทันที, อิ่มตื้อ, จ้องเล่นงานรัฐบาลอยู่, ไม่กินเส้นกัน แต่ที่ทำให้วิเศษณ์มีเกณฑ์การปรากฏที่ต่างจากคำนามคือ คำวิเศษณ์ไม่สามารถปรากฏกับตัวกำหนด ในวิทยานิพนธ์ฉบับนี้แบ่งวิเศษณ์เป็น 2 หมวดคำย่อย ได้แก่

(5.1) **วิเศษณ์** (Adverb) คือ คำวิเศษณ์ทั่วไปตามเกณฑ์ที่กล่าวมาข้างต้น ตัวอย่างในคลังข้อมูล เช่น

เงื่อนไขการชำระอาจจะยืดหยุ่นกว่า
จะเริ่มจากลูกค้าของ ธอส. และเอสเอ็มซีก่อน
ดูกันต่อไปว่า กกต.จะว่าอย่างไร
คลินตันกำลังพ่ายแพ้ต่อความเย้ายวนเข้าอีกแล้ว
จะใช้ระบบเสียงข้างมาก หรือเสียงเกือบเป็นเอกฉันท์
ปัญหาคอรัปชั่นยิ่งรุนแรงขึ้นในช่วง 2-3 ปีหลัง
ประธานาธิบดีคลินตันคงจะแวะที่ปากีสถานด้วย
มีสัญญาณว่าจะค่อยๆ ขยับเพิ่มขึ้น
การไม่ไว้วางใจที่จะส่งลูกหลานไปเรียนต่อ
ที่ร้ายที่สุด พวกเขาอาจจะอุ้มคุณไปที่ไหนสักแห่ง
ปลอมเป็นคนขับแท็กซี่เพื่อฟังผู้โดยสารปรับทุกข์บ้าง ปลอมเป็นคนไข้ไปตามโรงพยาบาลบ้าง

(5.2) **วิเศษณ์บอกปฏิเสธ** (Negator) คือ คำที่ใช้แสดงปฏิเสธ ได้แก่ ไม่, มิ ในวิทยานิพนธ์นี้จัดให้คำเหล่านี้เป็นคำวิเศษณ์เพราะมีการปรากฏตรงตามเกณฑ์ของคำวิเศษณ์ คือ ปรากฏนำหน้ากริยา และไม่สามารถปรากฏหน้าตัวกำหนด แต่ที่แยกออกมาเป็นหมวดคำย่อยหมวดคำหนึ่งก็เนื่องจากเห็นว่า เป็นคำที่มีความสำคัญเพราะใช้แยกหมวดคำกริยาออกจากหมวดคำอื่นๆที่เหลือ ดังนั้น การกำหนดวิเศษณ์บอกปฏิเสธแยกออกมาจึงน่าจะช่วยระบุคำกริยาที่ปรากฏตามหลังได้ ตัวอย่างในคลังข้อมูล เช่น

เป็นยอดเงินต้นไม่รวมดอกเบี้ยถึง 1,655 ล้านดอลลาร์
แม้จะไม่ค่อยชอบตัวละคร
ประเทศไทยคงไม่ต้องตกอยู่ในสภาพที่ต้องรอให้ใครมาช่วย
การไม่ไว้วางใจที่จะส่งลูกหลานไปเรียนต่อ
เงินมีแต่ใช้ไม่ได้
แต่แรกนั้นคลินตันมิได้กำหนดไว้แน่ชัดว่า ...

(6) **คำหน้าหน่วยสร้างไร้ศูนย์** (Exocentric marker)*

คำหน้าหน่วยสร้างไร้ศูนย์ คือ หมวดคำที่ปรากฏหน้าคำนามและหน้ากริยาได้ และไม่สามารถปรากฏหลังคำว่า “ไม่” เช่น เธอซบไปกับเก้าอี้, เรียนรู้จากประสบการณ์ของตนเอง, เสียสละเพื่อคนอื่น, รีบกลับบ้านก่อนฝนตก, เกลี่ยให้เรียบ, ทำงานมากขึ้นกว่าเดิม และวลีนั้นจะต้องเป็นหน่วยสร้างไร้ศูนย์ (exocentric construction) ซึ่งหมายความว่า ทั้งคำหน้าหน่วยสร้างไร้ศูนย์และส่วนประกอบอีกส่วนเป็นส่วนหลักที่จำเป็นต้องปรากฏทั้งคู่ คำหน้าหน่วยสร้างไร้ศูนย์ในที่นี้รวมถึงคำที่รู้จักกันในชื่อของ subordinate conjunction ด้วย เช่น ก่อน, หลัง, เพราะ, ถ้าหาก เป็นต้น เหตุผลหลักที่พิจารณาคำพวกนี้เข้าเป็นหมวดคำหน้าหน่วยสร้างไร้ศูนย์ ก็คือ ลักษณะการประกอบร่วมกับส่วนที่ตามมาในลักษณะที่เป็นหน่วยสร้างไร้ศูนย์ คำในหมวดนี้แบ่งเป็น 4 หมวดคำย่อย ได้แก่

(6.1) **บุพบทนำหน้านาม** (Preposition preceding Noun) คือ คำหน้าหน่วยสร้างไร้ศูนย์ประเภทที่ปรากฏหน้าคำนามเสมอ ตัวอย่างในคลังข้อมูล เช่น

โดนกกต. จับแขวน ประจานกลางเมือง
การจ่ายค่าไฟฟ้าให้กับเอ็กโกจะต่ำกว่าการจ่ายค่าก๊าซฯให้กับพม่า
มีความสอดคล้องกับบทบาทภารกิจของทางศูนย์
รักษาส่วนแบ่งจำนวนผู้โดยสารในเส้นทางข้ามทวีป
การรองรับการปฏิรูปการศึกษา ด้วยการเชิญผู้เชี่ยวชาญเป็นกลุ่มตัวอย่างสัมภาษณ์
อดีตเอกอัครราชทูตฝรั่งเศสประจำสหภาพยุโรป
การชำระคืนเงินต้นพร้อมดอกเบี้ย

* หมวดคำนี้ อมรา ประสิทธิ์รัฐสินธุ์ (2543) เรียกรวมกันว่า บุพบท โดยไม่ได้แบ่งหมวดคำย่อยลงไป ส่วนอุปสรรคสร้างนาม ซึ่งวิทยานิพนธ์นี้จัดให้อยู่ในหมวดคำหน้าหน่วยสร้างไร้ศูนย์นี้ด้วย อมราไม่ได้กล่าวแยกออกมาเนื่องจากหมวดคำนี้มีสถานะเป็นเพียงหน่วยคำที่ทำหน้าที่สร้างนามวลีเท่านั้น

(6.2) **บุพบทนำหน้ากริยา** (Preposition preceding Verb) คือ คำหน้าหน่วยสร้างไร้ศูนย์ที่ปรากฏนำหน้ากริยาหรือกริยาวลี[**] อันได้แก่ คำที่เรียกว่า subordinate conjunction ทั้งหลายที่จัดคำประเภทนี้ไว้ในหมวดคำหน้าหน่วยสร้างไร้ศูนย์ด้วยเนื่องจากบุพบทนำหน้ากริยามีลักษณะการปรากฏตรงตามเกณฑ์ของคำหน้าหน่วยสร้างไร้ศูนย์ คือ ไม่สามารถปรากฏหลังคำว่า “ไม่” และมีความสัมพันธ์ร่วมกับส่วนประกอบที่ตามมาในลักษณะที่เป็นหน่วยสร้างไร้ศูนย์ ตัวอย่างในคลังข้อมูล เช่น

กวาดสายตามองด้านหน้าร้านอาหารของเขา ก่อนจะเปิดปากเล่า
ก่อนหน้าที่จะจับตัวประกัน... นักรบขบวนการอาบู เซยาฟได้จับ...
ถ้าเงินไม่มา ขาเราไม่ยอมเดินไปลงคะแนน
ถึงแม้ว่าเวียดนามจะมีการพัฒนาที่น่าประทับใจ
ไม่ต้องการให้มีการพิจารณาคดีนี้ เพราะเกรงว่า เจ้าหน้าที่..จะถูก...

(6.3) **ตัวนำส่วนเติมเต็ม** (Complementizer) คือ คำที่ปรากฏนำหน้าอนุพากย์ส่วนเติมเต็ม (complement clause) หรือปรากฏนำหน้าคุณานุประโยค (relative clause) เพื่อทำหน้าที่เป็นส่วนเติมเต็มของกริยาหรือนาม วิทยานิพนธ์นี้จัดให้ตัวนำส่วนเติมเต็มเป็นหมวดคำหน้าหน่วยสร้างไร้ศูนย์เนื่องจากมีลักษณะการปรากฏตรงตามเกณฑ์ของคำหน้าหน่วยสร้างไร้ศูนย์ คือ ไม่สามารถปรากฏหลังคำว่า “ไม่” และมีความสัมพันธ์ร่วมกับส่วนประกอบที่ตามมาในลักษณะที่เป็นหน่วยสร้างไร้ศูนย์เช่นกัน ตัวอย่างในคลังข้อมูล เช่น

ไม่ค่อยชอบตัวละครที่เล่นเป็นแมวคือเจ้าทอม
ในมงคลสมัยที่พระบาทสมเด็จพระเจ้าอยู่หัว เจริญพระชนมายุครบ 6 รอบ
ประสงค์ที่จะเข้ารับการศึกษาใหม่ ในหลักสูตรอื่นๆ
เลขาธิการ ก.ค.กล่าวอีกว่า ระเบียบกระทรวงศึกษาธิการฉบับดังกล่าวนี้ ได้นำหลักการ

[**] แนวคิดที่บุพบทสามารถปรากฏนำหน้ากริยาได้ในภาษาไทย ไม่ใช่แนวคิดใหม่ พระยาอุปกิตศิลปสาร (2514) ก็ให้คำจำกัดความบุพบทไว้ว่า เป็นคำที่ใช้นำหน้านาม สรรพนาม หรือกริยาประเภทสภาวมาลา และกำชัย ทองหล่อ (2515: 342) ได้อธิบายและยกตัวอย่างคำบุพบทที่ปรากฏนำหน้ากริยา เช่น เขากินเพื่ออยู่, เขาทำงานกระทั่งตาย บุพบทที่นำหน้าประโยค เช่น เขามาตั้งแต่ฉันตื่นนอน และบุพบทที่นำหน้าวิเศษณ์ เช่น เขาต้องมาหาฉันโดยเร็ว เป็นต้น

เป็นหลักประกันว่า นับแต่นี้ต่อไป เด็กและเยาวชนทุกคน
พวกข้ามีคะแนนออกมา ให้เห็นกันจะๆ อย่างนี้
เป็นผลให้เด็กและเยาวชน ได้รับการพัฒนาให้เป็นทรัพยากรบุคคล
สหายร่วมแนวของเสี่ยชัช ซึ่งเป็นระดับเดอะเฮียในวงการ

(6.4) **อุปสรรคสร้างนาม** (Prefix of Nominalization) ได้แก่ “การ-“ และ “ความ-“ ซึ่งเป็นหน่วยคำอุปสรรค (prefix) ที่ปรากฏนำหน้าคำหรือวลีเพื่อสร้างคำนามหรือนามวลี วิทยานิพนธ์ฉบับนี้แยกหน่วยคำเหล่านี้ออกมาจัดเป็นหมวดคำหนึ่งด้วย เนื่องจากพบว่า กระบวนการสร้างนามวลีด้วย “การ-” และ “ความ-” เป็นกระบวนการที่มีผลิตผลภาวะ และมีความถี่ในการปรากฏสูง สาเหตุที่จัดให้คำอุปสรรคสร้างนามอยู่ในหมวดคำหน้าหน่วยสร้างไร้ศูนย์เนื่องจากเมื่อพิจารณาการปรากฏร่วมกับส่วนประกอบที่ตามมาแล้วพบว่า มีลักษณะตรงตามเกณฑ์การปรากฏของคำหน้าหน่วยสร้างไร้ศูนย์ คือไม่สามารถปรากฏหลังคำว่า “ไม่” และมีความสัมพันธ์ร่วมกับส่วนประกอบที่ตามมาในลักษณะที่เป็นหน่วยสร้างไร้ศูนย์ ตัวอย่างในคลังข้อมูล เช่น

เห็นเงินในกระเป๋าคนอื่นแล้วตาโต แสดง[ความอยากได้]จนออกนอกหน้า
ในเส้นทางบินจากยุโรป-ออสเตรเลียนั้น การบินไทยจะมี[ความได้เปรียบ]
เรียกจิตวิญญาณแห่ง[ความรักชาติ]กลับคืนมาอีกครั้ง
ใน[การแถลงข่าวของข้อมูลเศรษฐกิจการเงิน] ของธนาคารแห่งประเทศไทย

(7) **สันธาน** (Conjunction)

สันธาน คือ หมวดคำที่อยู่ระหว่าง 2 ถ้อยความ (utterance) ไม่สัมพันธ์กับคำใดคำหนึ่ง และหน่วยทั้งสองที่ขนาบสันธานจะต้องมีสมดุลพอสมควร สันธานทำหน้าที่ทางสัมพันธสาร (discourse) สันธานแตกต่างจากคำหน้าหน่วยสร้างไร้ศูนย์ตรงที่ คำหน้าหน่วยสร้างไร้ศูนย์ต้องปรากฏในหน่วยสร้างจะละออกไม่ได้เพราะเป็นส่วนจำเป็น ส่วนสันธานละออกได้แม้จะทำให้ความกระชับของถ้อยความลดลงไปบ้าง เช่น นี่ คือ ธรรมชาติ, เขาพยายามสุดชีวิต แต่ไม่ประสบผลสำเร็จ, ถ้าคุณตั้งใจเรียน และทำการบ้านทุกครั้ง..., เปิดเครื่องสักครู่ แล้วเลือกโปรแกรม นอกจากนี้ คำที่เป็นคำเชื่อมหน้าถ้อยความ ก็จัดเป็นคำสันธานด้วย เช่น “...อย่างไรก็ตาม เขาก็ยังเป็นพ่อเราอยู่” ตัวอย่างในคลังข้อมูล เช่น

'ทอม' กับ 'เจอร์รี่'

อะไรที่เป็นไปไม่ได้หรือเกินจริง

เห็นเจอร์รี่คอยแกล้งทอมและสามารถหลบหนีไปได้ทุกครั้ง

เป้าหมายของทุกคนคือการเข้าสู่ตำแหน่งใหญ่

ที่บอกเงินก็มีคืออยู่ในบัญชีของฝ่ายออกบัตร แต่เบิกออกมาใช้โดยฝ่ายการธนาคารไม่ได้

บางทีอาจเป็นเรื่องไม่ลองไม่รู้ '...อย่างไรก็ดี การฟื้นคืนชีพอีกหนของ"ซัช เตาปูน"

เหมาะสมจะส่งลูกหลานไปเรียนต่อได้อย่างสะดวกสบาย อีกทั้งเชื้อชาติและภาษาวัฒนธรรมก็ใกล้เคียงกัน

(8) **อนุภาค** (Particle)

อนุภาค คือ หมวดคำที่ปรากฏตำแหน่งหน้าสุด หรือท้ายสุดของถ้อยความ ไม่สัมพันธ์กับคำใดคำหนึ่ง แต่ทำหน้าที่ทางสัมพันธสาร อนุภาคนี้รวมถึง คำลงท้ายประโยคและคำอุทาน เช่น แหม ไม่น่าเชื่อเลยนะ, เฮ้ย เป็นไง สบายดีหรือ, ไชโย ไทยชนะ 3 ต่อ 2, เร็วเข้าเถอะ และรวมถึงตัวบ่งชี้หัวเรื่อง (topic marker) ด้วย เช่น สิ่งที่เขาพูดนั้น เป็นตัวอย่างที่ดีของ.. ตัวอย่างในคลังข้อมูล เช่น

เจอเขาฉายหนังการ์ตูนเรื่อง ทอมแอนด์เจอร์รี่ แล้วละก้อ

จุ๊ จุ๊ นั่นไง

ข้องใจอีกมั้ย

เขาจะแวะปากีสถานหรือไม่

ที่ได้คะแนนหลักแสนนั้น นับเป็นกรณีศึกษาที่น่าสนใจ (“นั้น” ในที่นี้จัดให้เป็นอนุภาคที่ทำหน้าที่เป็น topic marker)

(9) **เครื่องหมาย** (Punctuation)

เครื่องหมาย หมายถึง เครื่องหมายวรรคตอนในภาษาไทย ถึงแม้ในงานวิจัยต้นแบบจะไม่ได้กล่าวถึงไว้ เนื่องจากไม่ใช่คำในภาษา แต่เนื่องจากในคลังข้อมูลปรากฏใช้เครื่องหมายเหล่านี้อย่างมากมาย ดังนั้น จึงจัดเครื่องหมายเป็นหมวดคำอีกหมวดคำหนึ่งด้วย ในแง่การปรากฏ เครื่องหมายสามารถปรากฏได้ในตำแหน่งต่างๆโดยไม่มีความสัมพันธ์ทางวากยสัมพันธ์กับคำใดคำหนึ่งในภาษา เครื่องหมายสามารถปรากฏภายในคำ เช่น อี-เมล์, พ.อ. หรือปรากฏระหว่างถ้อยความก็ได้

เช่น หมายเหตุตบท้าย : สำหรับข่าวลือ ... เป็นต้น ในแง่ความหมาย เครื่องหมายสามารถใช้สื่อความหมายหรือแสดงจุดประสงค์บางอย่างได้ เช่น ? ใช้แสดงความสงสัย, “ ” ใช้แสดงข้อความที่ยกมาหรือเพื่อเน้นคำ ตัวอย่างจากคลังข้อมูล เช่น

'ทอม' กับ 'เจอร์รี่'
คุณธารินทร์ (แสดงเป็นทอมเพราะชื่อคล้องโดยบังเอิญ)
หมายเหตุตบท้าย : สำหรับข่าวลือ ...
อยู่ในสภาพ "กินน้ำใต้ศอก”

หมวดคำหลักในภาษาไทย 8 หมวดคำ สามารถสรุปคำจำกัดความตามเกณฑ์ในการปรากฏได้โดยแสดงเป็นคุณลักษณ์ (feature) การปรากฏร่วมและการกระจายของคำ ดังตารางที่ 4-1 (หมวดคำเครื่องหมาย ไม่ได้แสดงคุณลักษณ์ในการปรากฏไว้ เพราะสามารถระบุได้จากลักษณะรูปร่างของตัวเครื่องหมายเองเนื่องจากไม่ใช่คำในภาษา) ส่วนสัญลักษณ์หมวดคำที่ใช้เป็นป้ายระบุหมวดคำในคลังข้อมูลและในการกำกับหมวดคำของโปรแกรมทั้งหมด 26 ป้ายหมวดคำจะแสดงไว้ในตารางที่ 4-2 ท้ายบทนี้

หมวดคำหลัก	คุณลักษณ์การปรากฏ
กริยา	+[ไม่_]
นาม	-[ไม่_] , { +[_V] , + [V_] } , +[_D]
ตัวกำหนด	-[ไม่_] , +[N_]
ตัวบอกปริมาณ	-[ไม่_] , +[_N]
วิเศษณ์	-[ไม่_] , { +[_V] , + [V_] } , -[_D]
คำหน้าหน่วยสร้างไร้ศูนย์	-[ไม่_] , { +[_N] , + [_V] }
สันธาน	-[ไม่_] , +[U_U]
อนุภาค	-[ไม่_] , { +[_U] , +[U_] }

โดยที่:
+ หมายถึง สามารถปรากฏในตำแหน่งนั้นได้
- หมายถึง ไม่สามารถปรากฏในตำแหน่งนั้น
{ } หมายถึง เลือกเกิดในตำแหน่งใดๆใน { } ได้
U หมายถึง ถ้อยความ (utterance)
N หมายถึง นาม (noun)
V หมายถึง กริยา (verb)
D หมายถึง ตัวกำหนด (determiner)

ตารางที่ 4-1 คุณลักษณ์แสดงเกณฑ์การปรากฏของหมวดคำหลักในภาษาไทย

4.2.3 การตัดสินปัญหาความกำกวมในการกำกับหมวดคำให้กับคลังข้อมูล

หัวข้อนี้จะได้กล่าวถึงการตัดสินปัญหาความกำกวมของหมวดคำให้กับคลังข้อมูลซึ่งเป็นปัญหาที่พบในระหว่างการจัดเตรียมคลังข้อมูลฝึกสอน โดยเริ่มจากการอธิบายลักษณะและสาเหตุของความกำกวมในการกำกับหมวดคำ แล้วจึงกล่าวถึงการนำชุดหมวดคำที่จัดแบ่งไว้ มาตัดสินความกำกวมในแต่ละกรณี

ลักษณะของภาษาไทยที่คำมีรูปคำสำเร็จรูป ไม่มีการผันวิภัติปัจจัย ทำให้รูปคำหนึ่งๆ สามารถปรากฏในตำแหน่งต่างๆและทำหน้าที่ต่างๆได้อย่างหลากหลาย อันเป็นปัญหาของคำหลายหน้าที่ (polysemy) (สุโขทัยธรรมาธิราช, 2533: 312-313) ซึ่งทำให้เกิดความกำกวมในการตัดสินว่าคำที่ปรากฏในตำแหน่งต่างๆกันนั้นเป็นคำเดียวกันหรือไม่ และเป็นหมวดคำใด เช่น “เขา<u>กอง</u>หนังสือไว้บนโต๊ะ” กับ “ไปหยิบหนังสือจาก<u>กอง</u>นั้น”, “เขา<u>จาก</u>บ้านไป 3 ปีแล้ว” กับ “เขาไป<u>จาก</u>บ้าน 3 ปีแล้ว” อมรา ประสิทธิ์รัฐสินธุ์ (2543:48) อธิบายว่า เกิดจากการเปลี่ยนแปลงของภาษาที่เรียกว่า “การกลายเป็นคำไวยากรณ์” (grammaticalization) ซึ่งหมายถึง กระบวนการที่คำในภาษาซึ่งแต่เดิมเกิดในตำแหน่งเดียวกลายเป็นคำที่สามารถเกิดในตำแหน่งอื่นๆเพิ่มขึ้น ส่วนใหญ่จะเป็นกรณีที่คำเนื้อหา (content word) เช่น คำกริยา, คำนาม ปรากฏหน้าที่เพิ่มขึ้นเป็นคำไวยากรณ์ (grammatical word) ที่มีความหมายทางเนื้อหาลดลง แต่เพิ่มความหมายและหน้าที่ทางไวยากรณ์ขึ้น เช่น คำวิเศษณ์, คำบุพบท, คำสันธาน, คำกริยาช่วย เป็นต้น

จากการวิเคราะห์คลังข้อมูล พบว่า ความกำกวมในการตัดสินหมวดคำให้กับคำหลายหน้าที่ที่ปรากฏในคลังข้อมูล สามารถแบ่งอธิบายได้เป็น 2 กลุ่มใหญ่ๆ คือ ความกำกวมระหว่างคำนามกับหมวดคำอื่นซึ่งจะกล่าวในหัวข้อที่ 4.2.3.1 และความกำกวมระหว่างคำกริยากับหมวดคำอื่นซึ่งจะกล่าวในหัวข้อที่ 4.2.3.2 นอกจากนั้นยังพบว่าคำบุพบทและสันธานก็อาจก่อให้เกิดความกำกวมในการตัดสินหมวดคำได้เช่นกัน จึงได้นำมากล่าวไว้ในหัวข้อที่ 4.2.3.3

4.2.3.1 การตัดสินความกำกวมระหว่างนามกับหมวดคำอื่น

มีรูปคำอยู่จำนวนหนึ่งที่ทำให้เกิดความกำกวมระหว่างคำนามและหมวดคำอื่นๆ เช่น กริยา สันธาน บุพบท วิเศษณ์ และความกำกวมระหว่างหมวดคำย่อยของคำนาม คือ นามสามัญกับลักษณนาม ดังนี้

4.2.3.1.1 การตัดสินความกำกวมระหว่างคำนามสามัญกับคำลักษณนาม

คำลักษณนามในภาษาไทย มีทั้งคำลักษณนามที่มีรูปคำแตกต่างจากคำนามสามัญตัวที่มันขยาย เช่น หนังสือ 1 เล่ม, ขลุ่ย 1 เลา, ช้าง 2 เชือก, บ้าน 2 หลัง ซึ่งการตัดสินว่าคำดังกล่าวเป็นคำนามสามัญหรือคำลักษณนามทำได้โดยพิจารณาความหมายของคำเมื่อใช้ในปริบท ตัวอย่างเช่น “คัน<u>หลัง</u>จังเลย” กับ “<u>หลัง</u>นี้สวยดีนะ” ในตัวอย่างแรก “หลัง” เป็นคำนามสามัญซึ่งหมายถึงอวัยวะของร่างกาย ส่วนตัวอย่างที่สองเป็นคำลักษณนามซึ่งหมายถึงบ้าน เป็นต้น และมีทั้งคำลักษณนามที่มีรูปคำเหมือนกับคำนามสามัญตัวที่มันขยาย เช่น คน, บริษัท, ประเทศ ซึ่งเกิดความกำกวมในการระบุว่าคำดังกล่าวเป็นคำนามสามัญหรือคำลักษณนาม อย่างไรก็ตาม แม้ว่าทั้งคำนามสามัญและคำลักษณนามจะสามารถปรากฏตามเกณฑ์หลักของคำนาม คือ สามารถปรากฏหน้ากริยา, หลังกริยา, หรือหลังบุพบทได้ และปรากฏหน้าตัวกำหนดได้ทั้งคู่ แต่มีรายละเอียดในการปรากฏต่างกัน ตัวอย่างเช่น

แมว	*ตัว
แมวนี้	<u>ตัว</u>นี้
*แมวแมวนี้	แมว<u>ตัว</u>นี้
แมวสวย	<u>ตัว</u>สวย

*แมวแมวสวย	แมว<u>ตัว</u>สวย
แมวที่สวย	ตัวที่สวย
*แมวแมวที่สวย	แมว<u>ตัว</u>ที่สวย
*แมวแมวที่สวยนี้	แมว<u>ตัว</u>ที่สวยนี้
*แมวที่สวยแมวนี้	แมวที่สวย<u>ตัว</u>นี้
*แมว 2 แมว	แมว 2 <u>ตัว</u>
*แมว 2 แมวนี้	แมว 2 <u>ตัว</u>นี้
2 แมว	2 <u>ตัว</u>
หมากัดแมว	*หมากัด<u>ตัว</u>
หมากัดแมวนี้	หมากัด<u>ตัว</u>นี้
ให้อาหารแก่แมว	*ให้อาหารแก่<u>ตัว</u>
ให้อาหารแก่แมวนี้	ให้อาหารแก่<u>ตัว</u>นี้

จากตัวอย่าง เห็นได้ว่า ตำแหน่งในหน่วยสร้างนามวลีที่คำลักษณนามปรากฏได้ แต่คำนามสามัญไม่สามารถปรากฏได้ คือ

(1) [นามสามัญ ตัวบอกปริมาณ ___] เช่น “แมว 2 ตัว”, *“แมว 2 แมว”

(2) [นามสามัญ ___ ตัวกำหนด] เช่น “แมวตัวนี้”, *“แมวแมวนี้”

(3) [นามสามัญ ___ ส่วนขยายที่เป็นกริยาหรือคุณานุประโยค] เช่น “แมวตัวสวย”, *“แมวแมวสวย”, “แมวตัวที่สวย”, *“แมวแมวที่สวย”

ในทางตรงกันข้าม ตำแหน่งในหน่วยสร้างนามวลีที่คำนามสามัญปรากฏได้ แต่คำลักษณนามไม่สามารถปรากฏได้ คือ

(4) [#_#] หมายถึง คำนามสามัญสามารถปรากฏตามลำพังในหน่วยสร้างนามวลีได้ แต่คำลักษณนามต้องปรากฏกับส่วนขยายเสมอ เช่น *“ตัว” ,“ตัวนี้”, “ตัวที่สวย”

ดังนั้น วิทยานิพนธ์นี้ใช้ตำแหน่งในการปรากฏในนามวลีเหล่านี้ช่วยแก้ปัญหาความกำกวมระหว่างคำนามสามัญและคำลักษณนามที่มีรูปคำเหมือนกันได้ ดังนี้

“คน” เป็นคำนามสามัญ เพราะคำลักษณนามไม่สามารถปรากฏลำพังได้

“คน 2 คน” ตัวแรกเป็นคำนามสามัญ ส่วนตัวหลังเป็นคำลักษณนาม

“คนคนนี้“ ตัวแรกเป็นคำนามสามัญ ส่วนตัวหลังเป็นคำลักษณนาม

ส่วนตำแหน่งที่ทั้งคำนามสามัญและคำลักษณนามสามารถปรากฏได้เหมือนกัน ได้แก่

(5) [ตัวบอกปริมาณ ___] เช่น 2 แมว, 2 ตัว

(6) [___ ตัวกำหนด] เช่น แมวนี้, ตัวนี้

(7) [___ ส่วนขยายที่เป็นกริยาหรือคุณานุประโยค] เช่น แมวสวย, แมวที่สวย, ตัวสวย, ตัวที่สวย

ในกรณีเช่นนี้ หากรูปคำที่ปรากฏในตำแหน่งดังกล่าวสามารถเป็นได้ทั้งคำนามสามัญและคำลักษณนาม เช่น “3 คน”, “คนนี้”, “คนสวย”, “คนที่สวย” ก็จะทำให้เกิดความกำกวม วิทยานิพนธ์นี้จะตัดสินให้ “คนนี้”, “คนสวย”, “คนที่สวย” เป็นคำนามสามัญ เนื่องจากคำนามสามัญเป็นคำที่มีความหมายหลัก มีความหมายด้านเนื้อหาที่ชัดเจนมากกว่าคำลักษณนาม และคำนามสามัญสามารถปรากฏร่วมกับส่วนขยายในรูปแบบเหล่านี้ได้โดยทั่วไป เช่น “ตุ๊กตานี้”, “รถที่สวย” เป็นต้น ส่วน “3 คน” จะตัดสินเป็นคำลักษณนามเนื่องจากเป็นรูปแบบที่มีความถี่ในการปรากฏมากกว่าและเป็นหน้าที่หลักของคำลักษณนามที่ใช้บอกจำนวนนับของคำนามสามัญ

4.2.3.1.2 การตัดสินความกำกวมระหว่างคำนามกับคำบุพบท

คำนามบางคำมีรูปคำเหมือนกับคำบุพบท เช่น “หน้า”, “หลัง”, “กลาง”, “ข้าง”, “ทาง”, “ของ”, “ที่” ตัวอย่างเช่น

ไม่รวมวัตถุดิบ และสินค้าขั้นกลาง/NCM	เคยโดนกกต.จับแขวน ประจานกลาง/PNเมือง
รอยยิ้มบนหน้า/NCMของนักธุรกิจหนุ่ม	เลี้ยงโต๊ะจีนพวกชุมนุมหน้า/PNทำเนียบ
เป็นไปในทาง/NCMตั้งรับและถอยร่น	การขยายตัวทาง/PNเศรษฐกิจ

ความแตกต่างระหว่างคำนามกับคำบุพบท คือ คำบุพบทต้องปรากฏร่วมกับส่วนประกอบที่ตามหลังอีกส่วนเพื่อสร้างบุพบทวลีซึ่งมีลักษณะเป็นหน่วยสร้างไร้ศูนย์ บุพบทวลีสามารถย้ายตำแหน่งได้แต่ส่วนประกอบทั้งสองส่วนต้องย้ายตำแหน่งไปด้วยกัน ไม่สามารถแยกจากกันได้ ส่วนคำนามนั้นสามารถปรากฏลำพังได้และส่วนขยายที่ตามหลังคำนามสามารถละไปได้ นอกจากนี้กรณีที่คำนามนั้นมีความหมายเกี่ยวกับตำแหน่งพื้นที่ซึ่งมักปรากฏอ้างอิงกับอีกคำหนึ่งอยู่เสมอ เช่น “อยู่หลังบ้าน” คำว่า “หลัง” ปรากฏโดยอ้างอิงตำแหน่งจากคำว่า “บ้าน” ก็เกิดความกำกวมในทางความหมายด้วยว่า หมายถึง “หลังบ้าน(นอกบ้าน)” หรือ “ส่วนหลังของบ้าน(ในบ้าน)” ซึ่ง

ต้องพิจารณาความหมายจากปริบทที่ปรากฏด้วย หากเป็นกรณีแรก คือ “หลังบ้าน(นอกบ้าน)” จะถือว่าเป็นคำบุพบทแสดงความสัมพันธ์ที่คำนามในบุพบทวลีมีต่อคำกริยาหรือคำนามอีกคำ ส่วนหากเป็นกรณีหลัง คือ “ส่วนหลังของบ้าน(ในบ้าน)” จะถือว่าเป็นคำนามเพราะแสดงถึงส่วนของพื้นที่ โดยใช้คำแสดงความเป็นเจ้าของ “ของ” ในการทดสอบได้ กล่าวคือ คำนามสามารถปรากฏในตำแหน่ง [___ ของ นาม] ได้ ส่วนคำบุพบทจะไม่สามารถปรากฏได้ การใช้คำว่า “ของ” ในการทดสอบนี้สามารถนำไปใช้แก้ปัญหาความกำกวมของคำว่า “ของ” ได้ด้วย เช่น “ของคนอื่น” สามารถพูดได้ว่า “ของของคนอื่น” ได้ แสดงว่าตัวอย่างนี้เป็นคำนามหมายถึงสิ่งของ แต่ “หนังสือของคนอื่น” ไม่สามารถพูดว่า “หนังสือของของคนอื่น” ดังนั้นตัวอย่างนี้เป็นคำบุพบทที่แสดงความสัมพันธ์แบบเป็นเจ้าของ

4.2.3.1.3 การตัดสินความกำกวมระหว่างคำนามกับคำสันธาน

ความกำกวมระหว่างคำนามกับคำสันธานที่พบในคลังข้อมูลในวิทยานิพนธ์ คือ คำว่า “ส่วน” ดังตัวอย่าง

หลังจากนั้นก็จะเบาลงใน<u>ส่วน/NCM</u>ของไอเอ็มเอฟ	ส่งลูกหลานมาเรียนที่ไทยกันเป็นส่วนใหญ่ <u>ส่วน/C</u>เวียดนามเองนั้นมัวแต่รบ
ใน<u>ส่วน/NCM</u>ของพรรคประชาธิปัตย์สนามเมืองกรุง	... <u>ส่วน/C</u>การชำระคืนเงินต้นพร้อมดอกเบี้ยที่กู้มา

ความแตกต่างระหว่างคำนามกับคำสันธานคือ คำสันธานเป็นหน่วยทางสัมพันธสารซึ่งอยู่ระหว่างถ้อยความที่มีความสมดุลกันพอควร และสามารถละได้ จากตัวอย่าง “ส่วน” ทำหน้าที่เชื่อมระหว่างถ้อยความที่มีเนื้อความไม่คล้อยตามกัน และสามารถละได้ ส่วนคำนามเป็นหน่วยหลักทางวากยสัมพันธ์ที่มีความสำคัญในหน่วยสร้าง จากตัวอย่าง “ใน<u>ส่วน</u>ของไอเอ็มเอฟ” คำว่า”ส่วน” เป็นส่วนหลักของหน่วยสร้างนามวลี “ส่วนของไอเอ็มเอฟ” ซึ่งก็เป็นส่วนประกอบหลักส่วนหนึ่งของบุพบทวลี “ในส่วนของไอเอ็มเอฟ” อีกทีหนึ่ง จึงไม่สามารถละคำนามไปได้เพราะจะทำให้บุพบทวลีไม่สมบูรณ์

4.2.3.1.4 การตัดสินความกำกวมระหว่างคำนามกับตัวกำหนดหรือคำวิเศษณ์

มีรูปคำบางรูปซึ่งเป็นได้ทั้งคำนามและเป็นได้ทั้งตัวกำหนดหรือคำวิเศษณ์ ดังตัวอย่าง

มีความหมายไม่ได้ยิ่งหย่อนไปกว่ากัน/NPRO	มาศึกษาต่อที่เมืองไทยกัน/AVเป็นจำนวนมาก
เป็นกำลัง/NCMในการพัฒนาบ้านเมือง	สภาพการแก้ไขเศรษฐกิจที่กำลัง/AVเป็นอยู่
หลังจากนั้น/NPROจะนำทรัพย์สิน	กำลังศึกษาต่อที่ประเทศนั้น/D

จากตัวอย่าง ตำแหน่งในการปรากฏของคำนามและตัวกำหนดหรือคำวิเศษณ์แตกต่างกันตามเกณฑ์หลักของแต่ละหมวดคำ คำนามเป็นคำที่มีความหมายหลักและเป็นส่วนสำคัญทางวากยสัมพันธ์ เช่น “กัน” และ “นั้น” ในตัวอย่างที่เป็นคำนามเป็นส่วนประกอบหลักในบุพบทวลี, “กำลัง” ในตัวอย่างที่เป็นคำนาม เป็นส่วนเติมเต็มของกริยา ส่วนหมวดคำอื่นๆใช้เป็นส่วนขยายเพื่อเพิ่มความชัดเจน เช่น “กัน” และ “กำลัง” ในตัวอย่างที่ใช้เป็นวิเศษณ์ เป็นส่วนขยายของกริยา และ “นั้น” ในตัวอย่างที่เป็นตัวกำหนดเป็นส่วนขยายของคำนาม ซึ่งสามารถละไปได้ นอกจากนี้คำว่า “นั้น” ยังปรากฏเป็นตัวบ่งชี้หัวเรื่อง (Topic marker) ได้อีกด้วย เช่น “ส่วนที่กู้จากคนอื่นที่เขาลงขันให้นั้น/PTเงื่อนไขการชำระอาจจะยืดหยุ่นกว่า”, “ส่วนเวียดนามเองนั้น/PTมัวแต่รบทำสงครามต่อเนื่อง” ตัวบ่งชี้หัวเรื่องในวิทยานิพนธ์นี้จัดให้เป็นคำอนุภาค (PT)

4.2.3.2 การตัดสินความกำกวมระหว่างกริยากับหมวดคำอื่น

การแก้ปัญหาความกำกวมระหว่างกริยากับหมวดคำอื่นๆนั้น จะใช้เกณฑ์หลักสำหรับคำกริยา คือ เกณฑ์การปรากฏกับคำว่า “ไม่” ได้ ส่วนหมวดคำอื่นๆจะไม่สามารถปรากฏกับคำว่า “ไม่” ได้ หมวดคำที่เกิดความกำกวมกับคำกริยามีดังนี้

4.2.3.2.1 การตัดสินความกำกวมระหว่างคำกริยากับคำบุพบท

คำบุพบทซึ่งจัดเป็นประเภทหนึ่งของหมวดคำหน้าหน่วยสร้างไร้ศูนย์ มีทั้งคำบุพบทที่นำหน้านามและคำบุพบทที่นำหน้ากริยา คำบุพบททั้งสองประเภทมีอยู่เป็นจำนวนมากที่มีรูปคำเหมือนกับคำกริยา เช่น จาก, ถึง, ให้, ว่า, ต่อ, ตาม, ข้าม, พอ, ประจำ ตัวอย่างเช่น

ไม่ว่าจะประจำ/VPN0อยู่ตามโรงพยาบาล	เอกอัครราชทูตฝรั่งเศสประจำ/PNสหภาพยุโรป
กระโดดข้าม/VN0กำแพงเครื่องกีดขวาง	จำนวนผู้โดยสารในเส้นทางข้าม/PNทวีป
แค่นี้พอ/V0มั้ย?	พอ/PVมาถึงพระเจ้าฮุสเซนผู้เพิ่งสิ้นพระชนม์ไป
ก็คงตาม/VN0เสี่ยชัชเข้ามา	ไม่ว่าจะประจำอยู่ตาม/PNโรงพยาบาล
แต่ผลที่ตาม/V0มาจะเสียหาย	
ดูกันต่อไปว่า กกต.จะว่า/V0อย่างไร	ดูกันต่อไปว่า/PCOMP กกต.จะว่าอย่างไร
อินเทอร์เน็ตช่วยต่อ /VN0ชีวิต	เบื่อหน่ายต่อ/PNแนวทางการทำงานของ กกต.
อุตสาหกรรม'การ์ตูน'	มีมูลค่าหลายหมื่นล้านบาทต่อ/PNปี

จรัสดาว อินทรทัศน์ (2539, อ้างถึงใน อมรา ประสิทธิ์รัฐสินธุ์, 2543: 32-34) ได้ทำการศึกษาการที่คำกริยากลายเป็นคำบุพบทไว้อย่างละเอียด และสรุปว่าในทางวากยสัมพันธ์ บุพบทไม่สามารถปรากฏตามหลังคำว่า “ไม่” และไม่สามารถแยกจากคำนามที่ตามหลังได้ การย้ายที่ต้องย้ายไปทั้งบุพบทและคำนามที่ตามหลัง สำหรับกริยาจะปรากฏกับคำว่า “ไม่” ได้ และสามารถแยกจากคำนามที่เป็นกรรมได้ ส่วนทางด้านอรรถศาสตร์ คำกริยามีความหมายเฉพาะ ส่วนบุพบทมีความหมายทั่วไปและความหมายจางกว่ากริยา (อมรา ประสิทธิ์รัฐสินธุ์, 2543: 34) (บุพบท ตามคำจำกัดความของ อมรา ประสิทธิ์รัฐสินธุ์ (2543) หมายถึง คำที่สามารถปรากฏนำหน้าคำนามหรือกริยาได้ โดยประกอบกับคำนามหรือกริยาที่ตามหลังเป็นหน่วยสร้างไร้ศูนย์ และไม่สามารถปรากฏหลังคำว่า “ไม่” ได้ ดังนั้นเมื่อเทียบกับหมวดคำในวิทยานิพนธ์นี้แล้ว หมวดคำบุพบทของอมราจึงเท่ากับหมวดคำหน้าหน่วยสร้างไร้ศูนย์ในวิทยานิพนธ์นี้ โดยที่อมราไม่ได้แบ่งหมวดคำย่อยลงไปอีก แต่วิทยานิพนธ์นี้แบ่งย่อยออกเป็น บุพบทนำหน้านาม, บุพบทนำหน้ากริยา, ตัวนำส่วนเติมเต็ม, และอุปสรรคสร้างนาม)

ผู้วิจัยเห็นว่าคำอธิบายดังกล่าวสามารถนำมาใช้แก้ปัญหาความกำกวมระหว่างคำกริยากับคำบุพบทในวิทยานิพนธ์นี้ได้ แต่เนื่องจากบุพบทในวิทยานิพนธ์นี้มีทั้งบุพบทที่นำหน้านามและ

บุพบทที่นำหน้ากริยา ดังนั้นต้องขยายคำอธิบายให้ครอบคลุมว่า การย้ายที่บุพบทวลีจะต้องย้ายไปทั้งบุพบทและส่วนประกอบหลักที่ตามหลังอีกส่วนหนึ่งซึ่งอาจเป็นคำนามหรือคำกริยา

4.2.3.2.2 การตัดสินความกำกวมระหว่างคำกริยากับคำวิเศษณ์

จากการวิเคราะห์คลังข้อมูล พบว่ามีคำจำนวนหนึ่งที่สามารถปรากฏเป็นกริยาหลัก และปรากฏร่วมกับกริยาหลักเป็นส่วนขยายได้ด้วย เช่น ไป, มา, เข้า, ออก, ไว้, อยู่ ในกรณีที่คำเหล่านี้ไม่ใช่คำกริยาหลัก แต่ปรากฏร่วมกับกริยาหลักอีกตัวหนึ่งจะทำให้เกิดความกำกวมว่า คำดังกล่าวจัดเป็นคำกริยาหรือเป็นคำวิเศษณ์ เนื่องจากคำวิเศษณ์ในภาษาไทยจะปรากฏหน้าหรือหลังกริยา และในภาษาไทยคำกริยาก็สามารถปรากฏหลังกริยาด้วยกันได้เช่นกัน ตัวอย่างเช่น

ไม่ยอมเปิดเผยตัวว่า ไป/VN0บ้านเสี่ยชัชบ่อยๆ
ปรับตัวลดลงไป/V0แตะที่ระดับต่ำสุด
กลับไป/VN0ประเทศบ้านเกิดเมืองนอน

พวกเขามา/V0กันทุกวันล่ะ
ก่อนจะมา/VPN0ที่ตอนจบเดียวกัน
ส่งลูกหลานมา/V0เรียนต่อ

เพื่อขึ้น/V0เป็นกลุ่มโฆษณาอันดับ 1 โลก
โดยให้เครื่องกลับบินขึ้น/V0ไปใหม่และบินคืบหน้าไปได้
อัตราดอกเบี้ยอยู่/VPN0ที่ 6%
ดึงเอาเงินลงทุนในหลักทรัพย์ต่างประเทศของฝ่ายออกบัตรมาไว้/VNPN0ที่ฝ่ายการธนาคาร
การเลือกตั้งอื่นๆ ภายใต้รัฐธรรมนูญใหม่

เลี้ยงโต๊ะจีนพวกชุมนุมหน้าทำเนียบจนพุงกาง เข็ดไป/AVตามๆกัน
ผู้ที่ซื้อใบสมัครเป็นนักศึกษา มสธ.ไป/AVแล้ว
กิจการบางประเภท อาจไป/AVเกี่ยวข้องกฎระเบียบหรือข้อบังคับของหน่วยงานอื่น
ดึงเอาเงินลงทุนในหลักทรัพย์ต่างประเทศของฝ่ายออกบัตรมา/AVไว้ที่ฝ่ายการธนาคาร
ร่างกฎหมายที่รัฐสภาบัญญัติขึ้นมา/AVในช่วงที่...

ตั้งซุ้มเก็บเงินขึ้น/AVในทำเนียบขาว
ราคาก๊าซฯในอนาคตจะสูงขึ้น/AV

ปัจจุบันก็ยังใช้กันอยู่/AV
ควรจะคิดอ่านทำอะไรรอท่าไว้/AV

ได้/VAUXผ่านพ้นไป
โดยให้เครื่องกลับบินขึ้นไปใหม่และบิน
คืบหน้าไปได้/VAUX
ยิ่งอยากได้มาก/VADJ
อาจจะได้น้อย/VADJกว่า 6%

การแก้ปัญหาความกำกวมระหว่างคำกริยากับคำวิเศษณ์นี้ ผู้วิจัยใช้วิธีทดสอบตามเกณฑ์ของคำกริยา คือ หากสามารถปรากฏหลังคำว่า “ไม่” ได้ก็จะจัดว่าเป็นคำกริยา ดังนั้นบางรูปคำอาจจัดเป็นคำกริยาอย่างเดียว เช่น “กินมาก”, “กินน้อย”, “วิ่งเร็ว” เพราะสามารถปรากฏตามหลังคำว่า “ไม่” ได้เสมอ แต่บางรูปคำมีทั้งที่เป็นคำกริยาและเป็นคำวิเศษณ์ เช่น “ซื้อใบสมัครไป/AVแล้ว” เป็นคำวิเศษณ์แสดงความหมายว่าเหตุการณ์ได้เกิดขึ้นในอดีต และความหมายของ “ไป” ในที่นี้แตกต่างจาก “ไป” ที่เป็นกริยาอยู่มาก ส่วน “เดินไป/V0” ถึงแม้ว่าในกรณีนี้จะไม่สามารถปรากฏเป็น *“เดินไม่ไป” แต่เมื่อดูที่ความหมายแล้วเห็นว่ายังมีความหมายเหมือน “ไป” ที่เป็นกริยา เช่น “ฉันไม่ไปโรงเรียน” ซึ่งหมายถึงการเคลื่อนที่ไปในทิศทางห่างจากตัวผู้พูด จึงจัดให้เป็นคำกริยาด้วย นอกจากนี้ยังใช้วิธีทดสอบได้ว่า ถ้าคำทั้งสองเป็นกริยาทั้งคู่ ก็น่าจะสามารถแยกสองคำดังกล่าวโดยใช้กับประธานตัวเดียวกันแล้ว ความหมายของข้อความจะยังคงเดิมอยู่ (อมรา ประสิทธิ์รัฐสินธุ์, 2543: 22; วิจินตน์ ภาณุพงศ์, 2532: 64) เช่น “ฉันเดินไป” สามารถแยกเป็น “ฉันเดิน” “ฉันไป” ได้โดยยังสื่อความหมายเหมือนเดิมอยู่ ดังนั้นจึงจัดว่า “ไป” ในที่นี้เป็นกริยา

คำที่เกิดความกำกวมเมื่อปรากฏร่วมกับคำกริยาอย่างที่กล่าวมานี้ บางคำ เช่น ไป, มา, ขึ้น, ลง, ไว้ ในงานวิจัยอื่นๆ ก็อธิบายไว้แตกต่างกัน เช่น วิจินตน์ ภาณุพงศ์ (2532) จัดไว้เป็นหมวดคำแยกออกมาต่างหาก เรียกว่า หมวดคำหน้ากริยา, หมวดคำหลังกริยา ตามตำแหน่งที่ปรากฏเมื่อเทียบกับกริยา นววรรณ พันธุเมธา (2527) จัดไว้เป็นหมวดคำขยายตามหน้าที่ที่ใช้ขยายคำกริยา ดังนั้น แนวคิดที่ใช้แก้ปัญหาความกำกวมประเภทนี้ในวิทยานิพนธ์นี้เป็นเพียงวิธีการหนึ่ง ประเด็นปัญหาความกำกวมเหล่านี้ยังสามารถศึกษาต่อไปได้อีก

4.2.3.2.3 การตัดสินความกำกวมระหว่างคำกริยากับคำกริยาคุณศัพท์

วิทยานิพนธ์นี้จัดคำที่เรียกกันว่าคุณศัพท์ (adjective) เป็นหมวดคำกริยาตามที่ อมรา ประสิทธิ์รัฐสินธุ์ (2543) เสนอไว้ เนื่องจากเห็นด้วยกับอมราว่า คำชนิดนี้สามารถปรากฏหลังคำว่า “ไม่” ได้เหมือนคำกริยาทั่วไป และยังสามารถปรากฏในตำแหน่งต่างๆได้เหมือนกริยา หรือสามารถทำหน้าที่เหมือนกับกริยาได้ (อมรา ประสิทธิ์รัฐสินธุ์, 2543: 25-26) แต่ได้จัดแยกออกมาเป็นหมวดคำย่อยต่างหาก เรียกว่า กริยาคุณศัพท์ (adjectival verb) ที่จัดแยกออกมาก็เนื่องจากเห็นว่าคำกริยาคุณศัพท์นี้มีลักษณะที่แตกต่างจากกริยาทั่วไปตรงที่ มีความหมายระบุคุณสมบัติ สภาพ หรือลักษณะ นอกจากนี้ สามารถใช้เปรียบเทียบระดับได้ (comparison) โดยทั่วไปกริยาคุณศัพท์สามารถปรากฏในตำแหน่ง [___ กว่า] และ [___ ที่สุด] ได้ทันที หรือปรากฏในตำแหน่ง [___ มากกว่า] และ [___ มากที่สุด] ก็ได้ ในขณะที่คำกริยาเมื่อใช้เปรียบเทียบจะปรากฏในตำแหน่ง [___ มากกว่า] และ [___ มากที่สุด] เท่านั้น (วิจินตน์ ภาณุพงศ์, 2532: 54-55) ตัวอย่างเช่น

ค่าใช้จ่ายในการศึกษาเล่าเรียนถูก/VADJกว่าไปเรียนต่อที่สหรัฐอเมริกา
การปรับยุทธวิธีใหม่ จะดี/VADJกว่าการเดินย่ำรอยเดิมหรือไม่
เรื่องนี้เสี่ยชัชเจ๋ง/VADJกว่าใคร
ที่ร้าย/VADJที่สุด พวกเขาอาจจะอุ้มคุณ...
คดีความทางกฎหมายชนิดหิน/VADJที่สุด
กำลังเป็นไปในทางตั้งรับ/V0และถอยร่น/V0 มากกว่า "การบริหารจัดการเชิงรุก"
ได้ผล/V0มากขึ้นกว่าเดิม
เป็นกลุ่มที่ได้เปรียบ/V0มากที่สุด

ดังนั้น คำว่า “มาก” และ “น้อย” ที่ปรากฏหลังกริยาก็จัดเป็นกริยาคุณศัพท์ด้วย เพราะสามารถปรากฏหลังคำว่า “ไม่” และปรากฏนำหน้า “กว่า” และ “ที่สุด” ได้ทันที

อย่างไรก็ตาม ถึงแม้ว่าเกณฑ์นี้จะใช้ได้ดีพอสมควร แต่ก็ยังมีปัญหาอยู่บ้าง เนื่องจากคำกริยาอื่นบางคำก็สามารถปรากฏหน้า “กว่า” และ “ที่สุด” ได้ทันทีเช่นกัน เช่น “ฉันชอบเขาที่สุด” “ฉันเกลียดเขามากกว่าใครๆ” กริยาที่ปรากฏในตำแหน่งนี้ได้จะเป็นกริยาที่มีความหมายเชิง

ความรู้สึก ส่วนคำกริยาที่มีความหมายในเชิงรูปธรรม เช่น การกระทำ, การเคลื่อนไหว ไม่สามารถปรากฏในตำแหน่งนี้ได้ เช่น *เขาทำงานกว่าผู้จัดการ” *”เขาทำงานที่สุด”

4.2.3.3 การตัดสินความกำกวมระหว่างบุพบทกับสันธาน

คำบุพบทในวิทยานิพนธ์ฉบับนี้มีทั้งบุพบทที่นำหน้าคำนามและบุพบทที่นำหน้าคำกริยา ในขณะที่สันธานจะปรากฏระหว่างถ้อยความ ซึ่งถ้อยความอาจประกอบขึ้นจากคำนามคำเดียวหรือนามวลี หรือประกอบขึ้นจากคำกริยาคำเดียวหรือกริยาวลีก็ได้เช่นกัน ดังนั้น จึงทำให้เกิดความกำกวมระหว่างคำบุพบทกับคำสันธานในบางกรณี ตัวอย่างเช่น

คุณปองพล อดิเรกสาร กับ/Cความสำเร็จในการเขียนหนังสือ
ต้องการทักษะที่ต่างกันกับ/PNที่หลายคนมีอยู่
คนอังกฤษนิยมใช้กันมาแต่/PNโบราณ
ล้อเลียนกันแรงไปหน่อย แต่/Cก็มิได้ห้าม
ประชาชนจำเป็นต้องยอมรับ หาก/PVต้องการที่จะอยู่รอด
ชาวต่างชาติแห่เข้ามาในเวียดนามตอนนี้ เพราะ/PVพวกเขาคิดว่า...
ในสมัยที่อเมริกันเรืองอำนาจ และ/Cในสมัยที่ฝรั่งเศสเป็นเจ้าอาณานิคม

ความแตกต่างระหว่างคำบุพบทและสันธาน คือ คำบุพบทเป็นหน่วยทางวากยสัมพันธ์เป็นส่วนที่จำเป็นจะละออกไปจากบุพบทวลีไม่ได้ และถ้าจะย้ายที่ต้องย้ายไปทั้งบุพบทวลี คือ ทั้งคำบุพบทและส่วนประกอบอีกส่วนที่ตามหลัง เนื่องจากบุพบทวลีเป็นหน่วยสร้างไร้ศูนย์ ส่วนคำสันธานเป็นหน่วยทางสัมพันธสาร ทำหน้าที่เชื่อมระหว่าง 2 ถ้อยความที่มีความสมดุลกันพอควรและสามารถละไปได้ ดังนั้น หากคำดังกล่าวสามารถละไปได้โดยความหมายยังเหมือนเดิมอยู่ก็จัดว่าเป็นคำสันธาน แต่หากละไปแล้วทำให้ประโยคไม่สมบูรณ์ก็จัดว่าเป็นคำบุพบท

4.3 สรุป

ในบทนี้ ผู้วิจัยได้นำเสนอเกณฑ์ในการพิจารณาตัดคำ, การกำหนดชุดหมวดคำ, การนำชุดหมวดคำไปใช้กำกับข้อมูลจริงในคลังข้อมูลฝึกสอน และปัญหาความกำกวมในการกำกับหมวดคำที่พบพร้อมทั้งการตัดสินกรณีความกำกวมต่างๆ ทั้งหมดนี้เป็นประเด็นปัญหาพื้นฐาน

ทางภาษาศาสตร์ที่ผู้วิจัยต้องทำการศึกษาวิเคราะห์เพื่อจัดเตรียมคลังข้อมูลฝึกสอนสำหรับใช้กับแบบจำลองไตรแกรมที่นำเสนอในวิทยานิพนธ์ฉบับนี้ โดยที่ประเด็นเรื่องคำและเรื่องหมวดคำดังที่กล่าวอภิปรายมานี้ เป็นประเด็นที่สัมพันธ์กันอย่างใกล้ชิด เนื่องจากมโนทัศน์เรื่องหมวดคำก็ตั้งอยู่บนพื้นฐานของมโนทัศน์เรื่องคำ ดังนั้น จึงต้องมีการกำหนดเกณฑ์สำหรับตัดสายอักขระออกเป็นคำเสียก่อน แล้วถึงพิจารณาหมวดคำได้ โดยที่ทั้งคำและหมวดคำต่างมีอิทธิพลต่อการพิจารณาตัดคำและกำกับหมวดคำในคลังข้อมูล และจากข้อจำกัดและประเด็นปัญหาความกำกวมทั้งหลายที่กล่าวมา สามารถเห็นได้ว่า เรื่องคำและเรื่องหมวดคำในภาษาไทยยังสามารถศึกษาหาวิธีการแก้ปัญหาต่อไปได้อีกมาก เกณฑ์ในการตัดสินคำ, ชุดหมวดคำ และวิธีการแก้ปัญหาความกำกวมต่างๆที่ผู้วิจัยเลือกใช้สามารถช่วยแก้ปัญหาในการตัดสินคำและหมวดคำในคลังข้อมูลได้ดีในระดับหนึ่ง เพราะเป็นเกณฑ์การตัดคำและชุดหมวดคำที่ได้นำไปใช้กับข้อมูลภาษาจริง แต่ก็อาจไม่ใช่วิธีแก้ปัญหาที่ดีที่สุดและยังอาจเป็นที่ถกเถียงในทางวิชาการต่อไปได้

	สัญลักษณ์	ความหมาย	ตัวอย่างคำ
1	V0	กริยาที่ปรากฏลำพัง	มา, พัฒนา, ฟื้นตัว, ลอยนวล
2	VN0	กริยาตามด้วยนาม	เกลียด, สะสาง, สกัด, สำรวจ
3	VPN0	กริยา+บุพบท+นาม	พิจารณา(จาก), รับผิดชอบ(ต่อ), สอดคล้อง(กับ)
4	VCV0	กริยา+ส่วนเติมเต็ม	เชื่อ(ว่า), เสนอ(ให้), สนใจ(ที่)
5	VNPN0	กริยา+นาม+บุพบท+นาม	คืน...(ยัง), เสนอขาย...(แก่), ได้รับ...(จาก)
6	VNN0	กริยา+นาม+นาม	ชำระ, ใช้, ให้
7	VNCV0	กริยา+นาม+ส่วนเติมเต็ม	ถาม..(ว่า) ,กล่าวหา..(ว่า), เตือน..(ให้)
8	VV0	กริยาตามด้วยกริยา	ต้องการ, ชอบ, พยายาม, ทยอย, สนใจ
9	VS0	กริยาตามด้วยประโยค	คาด, ทำให้, เจอ, ห้าม
10	VAUX	กริยาช่วย	ควรทำ, ต้องทำ, ถูกตี, เคยทำ
11	VADJ	กริยาคุณศัพท์	ดี, เลว, สูง, กะทัดรัด, น่าสนใจ, มาก, ยากจน
12	NCM	นามสามัญ	ศัพท์, ลูกหลาน, วิชาการ, ราคา, มนุษย์
13	NPP	นามเฉพาะ	สหรัฐ, ไทย, ซับไมครอน, กุมภาพันธ์
14	NCSF	ลักษณนาม	คน, แผ่น, แห่ง, ดอลลาร์
15	NPRO	สรรพนาม	เขา, นาง, ใคร, อะไร, นี่
16	D	ตัวกำหนด	ดังกล่าว, ทั่วไป, ทั้งหมด, คนนี้, คนนั้น, คนไหน
17	Q	ตัวบอกปริมาณ	หลาย, อีกคน, 2 คน, สิบล้านบาท
18	AV	วิเศษณ์	ทำค่อยๆ, กำลังทำ, ดีที่สุด, ไม่, เกือบ, เต็มทน
19	AVNEG	วิเศษณ์บอกปฏิเสธ	ไม่, มิ
20	PN	บุพบทนำหน้านาม	ใน, ที่, บน, แก่
21	PV	บุพบทนำหน้ากริยา	เพราะ, ถึงแม้, ถ้าหาก
22	PCOMP	ตัวนำส่วนเติมเต็ม	ที่, ว่า, ให้, ซึ่ง
23	PFX	อุปสรรคสร้างคำนาม	การ, ความ, การที่
24	C	สันธาน	แต่, ส่วน, หรือ, และ, อย่างไรก็ตาม
25	PT	อนุภาค	จ๊ะ, ครับ, มั๊ย, ฟะ, หรอก, หรือยัง, หรือไม่
26	PUNC	เครื่องหมาย	( ) , : ?

ตารางที่ 4-2 สัญลักษณ์หมวดคำสำหรับกำกับหมวดคำ

บทที่ 5

การตัดคำและกำกับหมวดคำภาษาไทยโดยใช้แบบจำลองไตรแกรม

ดังที่ได้อธิบายไปในบทที่ผ่านๆมาถึงปัญหาในการตัดคำและกำกับหมวดคำในภาษาไทย ว่า ความกำกวมในการตัดคำและกำกับหมวดคำภาษาไทยเกิดจากลักษณะเฉพาะตัวของภาษาไทย คือ การที่คำในภาษาไทยสามารถเขียนเรียงติดต่อกันไปเป็นข้อความได้โดยไม่มีการแบ่งช่องว่างระหว่างคำ ซึ่งทำให้เกิดความลำบากในการระบุว่าคำหนึ่งๆเริ่มและจบที่ตำแหน่งใด และภาษาไทยมีกระบวนการสร้างคำขึ้นใช้ใหม่ในภาษาโดยประกอบคำที่มีอยู่เดิมเข้าด้วยกันให้กลายเป็นคำประกอบ เช่น คำประสาน, คำประสม ฯลฯ ซึ่งก็ทำให้เกิดความกำกวมในการระบุว่าสายอักขระหนึ่งๆจัดเป็นคำหนึ่งคำหรือเป็นคำหลายคำ นอกจากนั้นคำในภาษาไทยซึ่งไม่มีการเปลี่ยนรูปเมื่อนำไปใช้ในตำแหน่งและหน้าที่ต่างๆยังทำให้เกิดความลำบากในการระบุว่า รูปคำดังกล่าวในตำแหน่งต่างๆเป็นคำศัพท์เดียวกันหรือคนละคำศัพท์กัน ประเด็นเหล่านี้เป็นปัญหาด้านภาษาที่ทำให้เกิดความลำบากเมื่อต้องการใช้คอมพิวเตอร์ประมวลผลภาษาไทย โดยเฉพาะในการประมวลผลระดับสูง เช่น การแปลภาษา, การแจงส่วนประโยค ฯลฯ จากอุปสรรคดังกล่าว วิทยานิพนธ์นี้จึงมุ่งสร้างโปรแกรมตัดคำและกำกับหมวดคำภาษาไทย โดยประยุกต์แนวคิดทางสถิติมาใช้ในการแก้ปัญหา ดังนั้น ในส่วนแรกของบทนี้จะอธิบายถึงลักษณะปัญหา และแนวคิดที่ใช้แก้ปัญหาความกำกวมในการตัดคำและกำกับหมวดคำภาษาไทย จากนั้น ในส่วนที่ 5.2 จะนำเสนอแบบจำลองไตรแกรมที่ใช้สำหรับตัดคำและกำกับหมวดคำแบบเบ็ดเสร็จ ซึ่งในการพัฒนาโปรแกรมตามแบบจำลองดังกล่าว สามารถเพิ่มประสิทธิภาพในการทำงานได้โดยใช้เทคนิคการโปรแกรมแบบพลวัติ จึงจะได้กล่าวถึงขั้นตอนวิธีวิเทอร์บีซึ่งเป็นการเทคนิคการโปรแกรมแบบพลวัติที่นำมาใช้ในที่นี้ไว้ในตอนที่ 5.3 ด้วย

5.1 ลักษณะปัญหาของการตัดคำและกำกับหมวดคำ

จากลักษณะที่คำในภาษาไทยไม่มีตัวบ่งขอบเขตของคำ ข้อความในภาษาไทยจึงอยู่ในรูปสายของอักขระปรากฏต่อๆกันไป ดังนั้น ปัญหาในการตัดคำภาษาไทยจึงเป็นปัญหาในเรื่องการตัดสายอักขระออกเป็นคำ ซึ่งสามารถแสดงในรูปของสมการทางคณิตศาสตร์ได้ดังสมการที่ 5-1 (ดัดแปลงจาก บุญเสริม กิจศิริกุล, 2541 และ Allen, 1995)

(5-1) $$W = \max_{w_1,...,w_n} \arg \text{PROB}(w_1,..., w_n \mid c_1,..., c_m)$$

โดยที่ W คือ สายคำ $w_1,...,w_n$ ที่ทำให้ค่าความน่าจะเป็นตามสมการ 5-1 มีค่าสูงที่สุด

$w_1,...,w_n$ คือ สายคำที่ตัดออกมา ตั้งแต่คำที่ 1 ถึงคำที่ n

$c_1,...,c_m$ คือ สายอักขระที่ป้อนเข้าไป ตั้งแต่ตัวที่ 1 ถึงตัวที่ m

และ $n \leq m$ คือ จำนวนคำที่ตัดออกมาได้จะน้อยกว่าหรือเท่ากับจำนวนตัวอักขระที่ป้อนเข้าไป

จากสมการที่ 5-1 หมายถึง ข้อความที่ป้อนเข้าประกอบไปด้วยลำดับของอักขระ $(c_1,c_2,...,c_m)$ ให้เลือกลำดับของคำที่ตัดออกมาได้ $(w_1,w_2,...,w_n)$ จากสายอักขระที่กำหนดให้ ที่จะทำให้ค่าความน่าจะเป็นตามสมการมีค่าสูงที่สุด (คือ W)

ส่วนปัญหาในการกำกับหมวดคำ เกิดจากการที่คำในภาษาไทยสามารถปรากฏในตำแหน่งต่างๆได้โดยไม่เปลี่ยนรูปคำ ดังนั้น ปัญหาในการกำกับหมวดคำก็คือปัญหาในเรื่องการเลือกหมวดคำที่ถูกต้องให้กับคำแต่ละคำ ซึ่งแสดงได้ดังสมการที่ 5-2

(5-2) $$T = \max_{t_1,...,t_n} \arg \text{PROB}(t_1,..., t_n \mid w_1,..., w_n)$$

โดยที่ T คือ สายหมวดคำ $t_1,...,t_n$ ที่ทำให้ค่าความน่าจะเป็นตามสมการ 5-2 มีค่าสูงที่สุด

$t_1,...,t_n$ คือ สายหมวดคำที่กำกับให้กับแต่ละคำ ตั้งแต่คำที่ 1 ถึงคำที่ n

จากสมการที่ 5-2 หมายถึง จากสายคำ $(w_1,w_2,...,w_n)$ ที่กำหนดให้ ให้เลือกสายหมวดคำที่กำกับให้แต่ละคำของสายคำดังกล่าว $(t_1,t_2,...,t_n)$ ที่จะทำให้ค่าความน่าจะเป็นตามสมการ 5-2 มีค่าสูงที่สุด (คือ T)

5.2 แบบจำลองไตรแกรมสำหรับแก้ปัญหาการตัดคำและกำกับหมวดคำภาษาไทย

จากลักษณะของทั้งสองปัญหาที่แสดงไว้ดังสมการที่ 5-1 และ 5-2 สามารถนำแนวคิดทางสถิติมาประยุกต์ใช้เพื่อแก้ปัญหาการตัดคำและกำกับหมวดคำ โดยใช้แนวคิดของแบบจำลองไตรแกรมซึ่งได้รับความนิยมอย่างแพร่หลายในการประมวลผลภาษาธรรมชาติ สำหรับปัญหาการตัดคำและกำกับหมวดคำภาษาไทยนั้น วิทยานิพนธ์นี้มองปัญหาทั้งสองเป็นส่วนงานเดียวกัน ซึ่งน่าจะสามารถแก้ปัญหาไปพร้อมๆกันได้ ดังนั้น ปัญหาของการตัดคำและกำกับหมวดคำภาษาไทยในวิทยานิพนธ์ฉบับนี้ ก็คือ ปัญหาในเรื่องการตัดสายอักขระออกเป็นคำและกำกับด้วยหมวดคำที่ถูกต้องไปพร้อมๆกัน ซึ่งสามารถแสดงได้ดังสมการที่ 5-3

(5-3) $$\max_{t_1,...,t_n} \arg \text{PROB}(w_1,..., w_n , t_1,..., t_n \mid c_1,..., c_m)$$

จากสมการที่ 5-3 หมายถึง ข้อความที่ป้อนเข้าประกอบไปด้วยลำดับของอักขระ $(c_1,c_2,...,c_m)$ ให้เลือกลำดับของคำที่ตัดออกมาได้ $(w_1,w_2,...,w_n)$ ที่มีการกำกับด้วยลำดับของหมวดคำ $(t_1,t_2,...,t_n)$ จากสายอักขระที่กำหนดให้ ที่จะทำให้ค่าความน่าจะเป็นตามสมการมีค่าสูงที่สุด ซึ่งเมื่อพิจารณาจากสมการแล้วจะเห็นว่า เท่ากับเป็นการหาสายคำและสายหมวดคำที่มีค่าความน่าจะเป็นสูงที่สุดเท่านั้น เนื่องจากสายอักขระจะถูกกำหนดจากข้อความที่ป้อนเข้าไปอยู่แล้ว ซึ่งแสดงได้ดังสมการที่ 5-4

(5-4)

$$\max_{t_1,...,t_n} \arg \text{PROB}(w_1,..., w_n , t_1,.., t_n \mid c_1,..., c_m) = \max_{t_1,...,t_n} \arg \text{PROB}(w_1,..., w_n, t_1,..., t_n)$$

จากสมการที่ 5-4 ซึ่งเป็นการหาค่าความน่าจะเป็นร่วม (joint probability) สามารถแปลงสมการดังกล่าวให้อยู่ในรูปของการหาค่าความน่าจะเป็นแบบมีเงื่อนไข (conditional probability) ได้ดังแสดงในสมการที่ 5-5

(5-5) $$\text{PROB}(t_1,..., t_n) \times \text{PROB}(w_1,...,w_n \mid t_1,..., t_n)$$

แต่การคำนวณค่าความน่าจะเป็นตามสมการที่ 5-5 โดยตรงต้องอาศัยคลังข้อมูลที่มีขนาดใหญ่มาก เนื่องจากจะต้องครอบคลุมสายหมวดคำทั้งสาย คือ $t_1,...,t_n$ และต้องครอบคลุมถึงกรณีที่สายหมวดคำดังกล่าวจะปรากฏเป็นสายคำ $w_1,...,w_n$ ด้วย ดังนั้น จึงทำการคำนวณสมการนี้โดยใช้การประมาณค่าตามแนวคิดของแบบจำลองไตรแกรมซึ่งจำกัดขอบเขตของปริบทที่ใช้ทำนายความน่าจะเป็นในการปรากฏของคำหนึ่งๆ จึงทำให้สามารถใช้คลังข้อมูลที่มีขนาดเล็กลงได้ แบบจำลองไตรแกรมมีสมมติฐานว่า ความน่าจะเป็นในการปรากฏของคำหนึ่งๆขึ้นอยู่กับคำที่ปรากฏก่อนหน้า 2 คำเท่านั้น

จากการใช้แนวคิดของแบบจำลองไตรแกรม สามารถประมาณค่าความน่าจะเป็นของหมวดคำหนึ่งๆ ได้โดยพิจารณาเฉพาะ 2 หมวดคำก่อนหน้าเท่านั้น ดังนั้น ค่าความน่าจะเป็นของสายหมวดคำในส่วนแรกของสมการที่ 5-5 คือ $PROB(t_1,...,t_n)$ ซึ่งหมายถึง ความน่าจะเป็นของลำดับหมวดคำ (tag sequence probability) สามารถคำนวณได้ดังสมการที่ 5-6

(5-6) $$PROB(t_1,..., t_n) \cong \prod_{i=1...n} PROB(t_i \mid t_{i-1}, t_{i-2})$$

ส่วน $PROB(w_1,...,w_n \mid t_1,...,t_n)$ ซึ่งหมายถึง ความน่าจะเป็นในการปรากฏของรูปคำ (lexical generation probability) กล่าวคือ ค่าความน่าจะเป็นของสายคำ $w_1,...,w_n$ เมื่อให้สายหมวดคำ $t_1,...,t_n$ สามารถประมาณค่าได้โดยตั้งสมมติฐานว่า ค่าความน่าจะเป็นที่จะปรากฏเป็นรูปคำนั้นๆ เมื่อกำหนดหมวดคำให้จะไม่ขึ้นกับคำหรือหมวดคำก่อนหน้าหรือตามหลัง ดังแสดงในสมการที่ 5-7

(5-7) $$PROB(w_1,..., w_n \mid t_1,..., t_n) \cong \prod_{i=1...n} PROB(w_i \mid t_i)$$

ดังนั้น จากสมการที่ 5-5 สามารถจะคำนวณความน่าจะเป็นของสายคำและสายหมวดคำโดยอาศัยการประมาณค่าได้ดังสมการที่ 5-8 ซึ่งสามารถนำไปใช้ได้จริง โดยค่าความน่าจะเป็นสามารถคำนวณได้จากการนับค่าความถี่ในการปรากฏในคลังข้อมูล

(5-8) $$\prod_{i=1...n} PROB(t_i \mid t_{i-1}, t_{i-2}) \times PROB(w_i \mid t_i)$$

ดังนั้น ในวิทยานิพนธ์ฉบับนี้จะนำสมการที่ 5-8 นี้ มาใช้แก้ปัญหาการตัดคำและกำกับหมวดคำภาษาไทย ซึ่งผลลัพธ์ที่ได้จากสมการนี้จะเป็นทั้งคำตอบของปัญหาการตัดคำและเป็นทั้งคำตอบของปัญหาการกำกับหมวดคำด้วย ซึ่งทำให้เห็นว่าการตัดคำและกำกับหมวดคำสามารถแก้ปัญหาไปพร้อมๆกันได้โดยมองเป็นส่วนงานเดียวกัน ต่างจากงานวิจัยที่ผ่านมาซึ่งมองว่า การตัดคำเป็นส่วนงานขั้นต้นที่ต้องแก้ปัญหาก่อน

อย่างไรก็ดี การคำนวณจากสมการที่ 5-8 โดยตรงจะเสียเวลาในการคำนวณเป็นอย่างมาก เนื่องจากจะทำการคำนวณเส้นทางทั้งสายทุกเส้นทางก่อนแล้วจึงเลือกเส้นทางที่มีค่าความน่าจะเป็นสูงที่สุด (brute force algorithm) ซึ่งทำให้โปรแกรมทำงานได้ช้า ดังนั้นจึงมีการนำเทคนิคการโปรแกรมแบบพลวัต (dynamic programming) เข้ามาใช้ช่วยคำนวณ ดังจะได้อธิบายในส่วนถัดไป

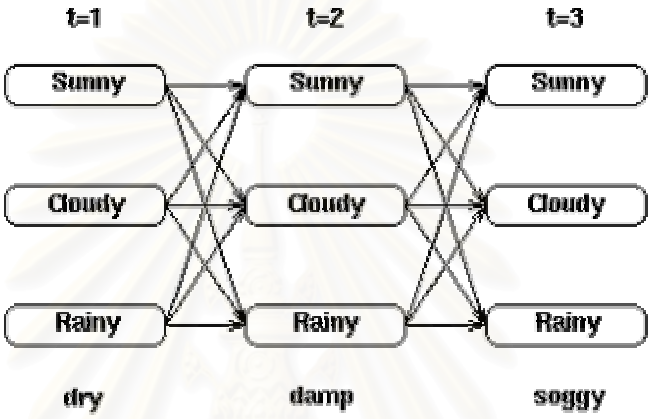
5.3 ขั้นตอนวิธีวิเทอร์บิ

ดังที่ได้กล่าวไปในตอนท้ายของส่วนที่แล้วว่า การคำนวณค่าความน่าจะเป็นโดยตรงจากสมการที่ 5-8 สามารถทำได้ช้ามาก เนื่องจากต้องเสียเวลาในการคำนวณทุกเส้นทางที่เป็นไปได้ ตัวอย่างเช่น สายคำสายหนึ่งที่มีจำนวนคำทั้งหมด N คำ และหมวดคำทั้งหมดในภาษาไทยมีจำนวน V หมวดคำ ในกรณีที่แย่ที่สุดคือทุกคำในสายคำดังกล่าวสามารถปรากฏเป็น V หมวดคำ กรณีดังกล่าวจะต้องเสียเวลาในการคำนวณ $k \times N^V$ ครั้งต่อสายคำเดียว โดย k คือค่าคงที่ (ไพศาล เจริญพรสวัสดิ์, 2541: 20) จะเห็นว่า เวลาที่ใช้จะขึ้นอยู่กับจำนวนคำและจำนวนหมวดคำที่เป็นไปได้ของแต่ละคำในสายนั้น ซึ่งจะทำให้เวลาที่ใช้เป็นสัดส่วนแบบเอกซ์โปเนนเชียล (Exponential) นอกจากนี้แล้ว เนื่องจากวิทยานิพนธ์ฉบับนี้ทำการตัดคำไปพร้อมๆกับกำกับหมวดคำ ดังนั้น สายอักขระหนึ่งๆที่ป้อนเข้าไปสามารถมีแบบการตัดคำได้มากกว่าหนึ่งแบบ จึงยิ่งทำให้เสียเวลาในการคำนวณสายอักขระหนึ่งๆมากกว่า $k \times N^V$ เสียอีก

เพื่อลดความสิ้นเปลืองในการคำนวณของโปรแกรม วิทยานิพนธ์นี้จึงได้นำเทคนิคการโปรแกรมแบบพลวัตเข้ามาช่วย ซึ่งเทคนิคที่เลือกใช้นี้มีชื่อว่า “ขั้นตอนวิธีวิเทอร์บิ” (Viterbi Algorithm) ซึ่งเป็นเทคนิคที่ได้รับความนิยมนำมาใช้กับแบบจำลองไตรแกรม และมีผู้นำมาใช้สำหรับการกำกับหมวดคำภาษาไทยแล้ว (ไพศาล เจริญพรสวัสดิ์, 2541) แนวคิดของขั้นตอนวิธีวิเทอร์บิสำหรับการกำกับหมวดคำ คือ ในแต่ละตำแหน่งของแต่ละเส้นทางให้จดจำเฉพาะค่าความน่าจะเป็นของเส้นทางก่อนหน้าที่ดีที่สุดที่จะนำไปสู่แต่ละหมวดคำ (store for each point in the

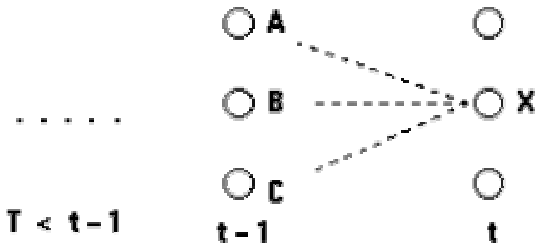
trellis the probability of the most probable path that leads to that node, Manning and Schutze, 1999: 308) ซึ่งสามารถแสดงได้จากตัวอย่าง ดังรูปที่ 5-1 ถึงรูปที่ 5-3

(http://www.comp.leeds.ac.uk/scs-only/teaching-materials/HiddenMarkovModels/html_dev/viterbi_algorithm/)



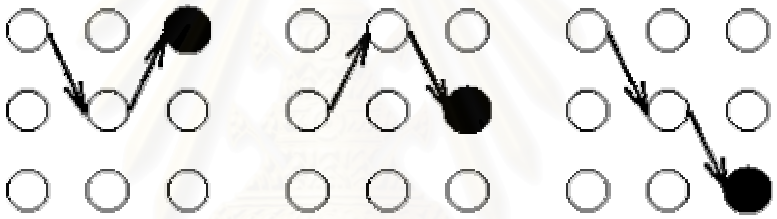
รูปที่ 5-1 เส้นทางการคำนวณที่เป็นไปได้ทั้งหมด 27 เส้นทาง

จากรูปที่ 5-1 แสดงตัวอย่างเส้นทางการเกิดของสถานการณ์การตากสาหร่าย ซึ่งจำลองแทนประโยคในภาษา โดยมีลักษณะที่สังเกตได้ (แทนรูปคำ) 3 ลักษณะ คือ dry, damp, soggy ในแต่ละช่วงเวลา (t = 1 ถึง 3 วัน) และมีสภาวะอากาศที่เป็นไปได้ทั้งหมด 3 สภาวะ(แทนหมวดคำ) ได้แก่ Sunny, Cloudy, Rainy การคำนวณโดยตรงต้องทำการสร้างเส้นทางจนจบทั้งสามช่วงเวลา ทั้งหมด 3*3 = 27 เส้นทาง แล้วจึงทำการเปรียบเทียบค่าความน่าจะเป็นของแต่ละเส้นทางเพื่อเลือกเส้นทางที่ดีที่สุด ส่วนขั้นตอนวิธีวิเทอร์บิจะทำไปทีละช่วงเวลา และจดจำเฉพาะเส้นทางที่ดีที่สุดที่มาจบลงที่แต่ละสภาวะ ณ ช่วงเวลานั้น ดังนั้นในช่วงเวลาแรกจะคำนวณค่าความน่าจะเป็นของ dry จากแต่ละสภาวะ (แทนการคำนวณความน่าจะเป็นในการปรากฏของรูปคำ) และจดจำ 3 เส้นทาง คือ เส้นทางที่ดีที่สุดที่มาจบที่สภาวะ Sunny, เส้นทางที่ดีที่สุดที่มาจบที่สภาวะ Cloudy และเส้นทางที่ดีที่สุดที่มาจบที่สภาวะ Rainy จากนั้นจึงขยายเส้นทางไปทีละหนึ่งช่วงเวลา โดย ณ ช่วงเวลาต่อไปทุกช่วง แต่ละสภาวะจะมีเส้นทางที่เข้ามา 3 เส้นทาง



รูปที่ 5-2 เส้นทางทั้งหมดที่ไปจบลงที่สภาวะ X

จากรูปที่ 5-2 เส้นทางที่มาจบลงที่สภาวะ X ณ ช่วงเวลาใดๆจะมีทั้งหมดแค่ 3 เส้นทาง วิธีวิเทอร์บิจะเลือกจดจำเฉพาะเส้นทางที่ดีที่สุดเส้นทางเดียวจาก 3 เส้นทางนั้น และทำเช่นนี้ต่อไปเรื่อยๆจนจบสาย ดังแสดงในรูปที่ 5-3



รูปที่ 5-3 เส้นทางที่ดีที่สุดของแต่ละสภาวะเมื่อจบสาย

จากรูปที่ 5-3 เมื่อจบสาย ณ ช่วงเวลาสุดท้าย แต่ละสภาวะจะเหลือเส้นทางที่เดินมาเพียงแค่เส้นทางที่ดีที่สุดเส้นทางเดียวเท่านั้น คือ เส้นทางที่มาจบที่สภาวะ Sunny ณ ช่วงเวลา t=3, เส้นทางที่มาจบที่สภาวะ Cloudy ณ ช่วงเวลา t=3, และเส้นทางที่มาจบที่สภาวะ Rainy ณ ช่วงเวลา t=3 ดังรูป แล้วจึงทำการเปรียบเทียบค่าความน่าจะเป็นระหว่าง 3 เส้นทางนี้ เพื่อเลือกเส้นทางที่ดีที่สุดเพียงเส้นทางเดียวเป็นคำตอบ สมมติว่า พบว่าเส้นทางที่ดีที่สุดคือ เส้นทางที่มาจบลงที่สภาวะ Cloudy ณ ช่วงเวลา t=3 ก็จะทำการย้อนรอย (backtrack) เส้นทางนั้นเพื่อหาว่า สายที่เป็นคำตอบคือ Cloudy-Sunny-Cloudy หรือในกรณีของการกำกับหมวดคำก็คือ สายหมวดคำที่มีความน่าจะเป็นสูงที่สุด

การใช้ขั้นตอนวิธีวิเทอร์บิจะช่วยลดเวลาในการคำนวณในกรณีของไตรแกรมลงเหลือ $k \times N^3T$ ทำให้เวลาที่ใช้ในการคำนวณจะลดลงอย่างมาก เนื่องจากขั้นตอนวิธีวิเทอร์บิจะคำนวณ

เฉพาะเส้นทางที่จำเป็นเท่านั้น ดังนั้นวิทยานิพนธ์ฉบับนี้จึงได้นำขั้นตอนวิธีวิเทอร์บิมาช่วยในการคำนวณความน่าจะเป็นของสายคำและสายหมวดคำด้วย

ในบทนี้ ผู้วิจัยได้นำเสนอแบบจำลองไตรแกรมที่นำมาใช้ในการแก้ปัญหาการตัดคำและการกำกับหมวดคำแบบเบ็ดเสร็จ พร้อมทั้งได้นำเสนอเทคนิคขั้นตอนวิธีวิเทอร์บีเพื่อเพิ่มประสิทธิภาพการทำงานของแบบจำลองไตรแกรม แบบจำลองนี้ได้พัฒนาขึ้นเพื่อทดสอบกับคลังข้อมูลที่จัดทำขึ้นมา โดยผู้วิจัยได้เลือกใช้ภาษา Perl ในการพัฒนาโปรแกรมดังกล่าว ซึ่งรายละเอียดของการทดสอบ ผลการทดสอบ และการประเมินผลที่ได้ จะได้กล่าวถึงในบทต่อไป

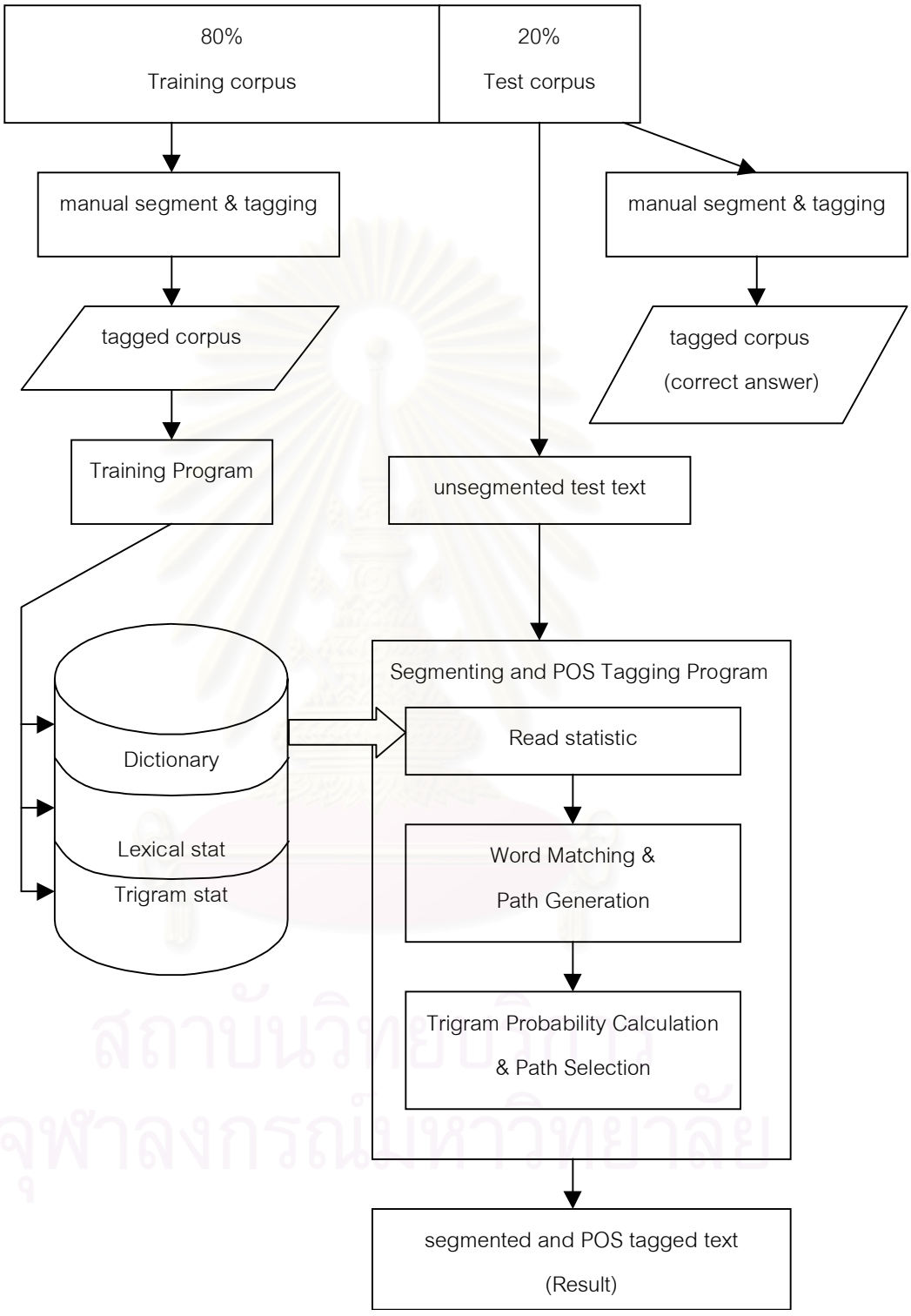
บทที่ 6

ผลการตัดคำและกำกับหมวดคำภาษาไทยแบบเบ็ดเสร็จ

บทที่ผ่านมาได้กล่าวถึงประเด็นทั้งหมดที่เกี่ยวข้องในการสร้างโปรแกรมตัดคำและกำกับหมวดคำภาษาไทยแบบเบ็ดเสร็จไปแล้ว บทนี้จะได้กล่าวถึงการนำโปรแกรมที่สร้างขึ้นมาทดลองทำการตัดคำและกำกับหมวดคำให้กับข้อความภาษาไทย โดยในหัวข้อที่ 6.1 จะกล่าวถึงขั้นตอนในการทดลองตัดคำและกำกับหมวดคำ อันประกอบด้วย การใช้ประโยชน์จากคลังข้อมูลที่จัดทำขึ้น และขั้นตอนในการทำงานของส่วนประกอบต่างๆของโปรแกรมตัดคำและกำกับหมวดคำแบบเบ็ดเสร็จ ในหัวข้อที่ 6.2 จะอธิบายถึงวิธีการประเมินผลประสิทธิภาพการตัดคำและการกำกับหมวดคำ และในหัวข้อที่ 6.3 จะได้นำเสนอและอภิปรายผลการทดลองตัดคำและผลการทดลองกำกับหมวดคำภาษาไทย รวมทั้งเปรียบเทียบผลการตัดคำแบบที่นำหมวดคำมาช่วยกับแบบที่ไม่ได้นำหมวดคำมาช่วย

6.1 ขั้นตอนในการทดลองตัดคำและกำกับหมวดคำ

ส่วนนี้จะอธิบายถึงขั้นตอนต่างๆในการนำโปรแกรมตัดคำและกำกับหมวดคำแบบเบ็ดเสร็จและคลังข้อมูลภาษาไทยที่จัดทำขึ้นมาทดลองทำการตัดคำและกำกับหมวดคำเพื่อทดสอบการทำงานของโปรแกรม โดยมีขั้นตอนทั้งหมดดังรูปที่ 6-1 แล้วจึงอธิบายรายละเอียดของขั้นตอนต่างๆ ดังนี้ ในหัวข้อที่ 6.1.1 จะอธิบายวิธีการใช้ประโยชน์จากคลังข้อมูลเพื่อเป็นฐานความรู้สำหรับเรียนรู้ค่าสถิติและเพื่อเป็นข้อความทดสอบของโปรแกรม ในหัวข้อที่ 6.1.2 กล่าวถึงส่วนประกอบต่างๆของโปรแกรมแบบเบ็ดเสร็จและขั้นตอนที่โปรแกรมทำการตัดคำและกำกับหมวดคำ และในหัวข้อที่ 6.1.3 จะแสดงให้เห็นลักษณะของผลลัพธ์ที่ได้จากโปรแกรม



รูปที่ 6-1 ขั้นตอนในการทดลองตัดคำและกำกับหมวดคำแบบเบ็ดเสร็จ

6.1.1 การใช้ประโยชน์จากคลังข้อมูล

หัวข้อนี้จะได้กล่าวถึง วิธีการนำคลังข้อมูลที่จัดทำขึ้นมาใช้ประโยชน์สำหรับการทดลองตัดคำและกำกับหมวดคำของโปรแกรม โดยจากรูปที่ 6-1 คลังข้อมูลภาษาที่รวบรวมมาจะแบ่งเป็น 2 ส่วน ได้แก่ (1) คลังข้อมูลฝึกสอน มีขนาดประมาณ 20,000 คำ (80%) และ (2) คลังข้อมูลทดสอบประสิทธิภาพ มีขนาดประมาณ 5,000 คำ (20%)

ในคลังข้อมูลฝึกสอนผู้วิจัยได้ทำการตัดคำและกำกับหมวดคำด้วยมือไว้เพื่อเป็นฐานความรู้สำหรับโปรแกรม (ดูรายละเอียดขั้นตอนในการตัดคำและกำกับหมวดคำด้วยมือในบทที่ 3) จากนั้นจึงใช้โปรแกรมฝึกสอน (training program) เพื่อเรียนรู้ค่าความถี่ทั้งหมดที่ปรากฏในคลังข้อมูลฝึกสอนนี้ กระบวนการนับค่าความถี่ในคลังข้อมูลฝึกสอน แสดงได้ดังตัวอย่างข้างล่างนี้

ต้อง/VAUX_ทำ/VN0_ /PUNC_เพื่อ/PV_ปฏิรูป/VN0_การเมือง/NCM

...

...

(1) [VAUX-VN0-PUNC]

(2) [VN0-PUNC-PV]

(3) [PUNC-PV-VN0]

(4) [PV0-VN0-NCM]

จากตัวอย่าง โปรแกรมฝึกสอนจะทำการนับค่าความถี่ของลำดับหมวดคำทีละ 3 คำ (ไตรแกรมของหมวดคำ) พร้อมกับนับค่าความถี่ของคำศัพท์ไปด้วย ในลำดับที่ (1) โปรแกรมจะนับค่าความถี่ของลำดับหมวดคำ [VAUX-VN0-PUNC] พร้อมกับนับค่าความถี่ของคำศัพท์ “ /PUNC” จากนั้นจะเลื่อนตำแหน่งไปทีละหนึ่งคำ ในลำดับที่ (2) จะนับค่าความถี่ของลำดับหมวดคำ [VN0-PUNC-PV] พร้อมกับนับค่าความถี่ของคำศัพท์ “เพื่อ/PV” แล้วจึงเลื่อนไปยังตำแหน่งถัดไปเรื่อยๆ จนนับค่าความถี่ได้ครบทุกคำในคลังข้อมูลฝึกสอน

ค่าความถี่ที่เรียนรู้จากกระบวนการข้างต้นจะแยกจัดเก็บไว้ในไฟล์ข้อมูล 3 ไฟล์ ได้แก่ ไฟล์พจนานุกรม, ไฟล์สถิติคำศัพท์, และไฟล์สถิติไตรแกรม แต่ละไฟล์มีรายละเอียดของข้อมูล ดังนี้

(1) **ไฟล์พจนานุกรม** จัดเก็บรายการคำศัพท์ทั้งหมดไว้สำหรับการเทียบคำ คำศัพท์แต่ละตัวในพจนานุกรมประกอบด้วยรูปคำและหมวดคำ หากรูปคำเดียวมีหลายหมวดคำก็จะจัดเก็บเป็นคำศัพท์คนละรายการกัน ตัวอย่างเช่น

หน้า/D
หน้า/NCM
หน้า/PN
หน้ากระดาษ/NCM
หน้าตา/NCM
หน้าต่าง/NCM
...
...

แต่เนื่องจากรายการคำศัพท์ในไฟล์พจนานุกรมนี้ได้มาจากการเรียนรู้คลังข้อมูลฝึกสอนเท่านั้น จึงอาจจะไม่สามารถครอบคลุมคำศัพท์ทุกคำในข้อความทดสอบได้ ดังนั้น ในข้อความทดสอบจึงมีคำศัพท์อยู่จำนวนหนึ่งที่ไม่ปรากฏในไฟล์พจนานุกรม หรือที่เรียกว่า คำที่ไม่รู้จัก (unknown word) ซึ่งจะทำให้โปรแกรมเกิดข้อผิดพลาดในการทำงานเพราะโปรแกรมไม่สามารถเทียบคำเจอได้ และเนื่องจากวิทยานิพนธ์นี้ไม่ได้มีจุดมุ่งหมายที่จะศึกษาหรือนำเสนอวิธีการแก้ปัญหาการระบุคำที่ไม่รู้จักโดยอัตโนมัติ ดังนั้นผู้วิจัยจึงได้ทำการเพิ่มเติมคำที่ไม่รู้จักที่ปรากฏในข้อความทดสอบลงในไฟล์พจนานุกรมด้วยเพื่อแก้ปัญหาการเทียบคำไม่เจอ (คำศัพท์ทั้งหมดในไฟล์พจนานุกรมแสดงไว้ในภาคผนวก ก) การเพิ่มคำที่ไม่รู้จักลงในไฟล์พจนานุกรมนี้จะไม่กระทบต่อค่าความถี่การปรากฏของคำศัพท์อื่นๆ เนื่องจากว่าไฟล์พจนานุกรมไม่ได้จัดเก็บค่าความถี่ของคำศัพท์ไว้ ค่าความถี่ในการปรากฏของคำศัพท์จะได้จัดเก็บแยกไว้ในไฟล์สถิติคำศัพท์

(2) **ไฟล์สถิติคำศัพท์** จัดเก็บค่าความถี่การปรากฏของคำศัพท์แต่ละคำที่นับได้จากคลังข้อมูลฝึกสอน ลักษณะของข้อมูลในไฟล์สถิติคำศัพท์นี้จะคล้ายคลึงกับข้อมูลในไฟล์พจนานุกรม แต่เพิ่มเติมข้อมูลค่าความถี่ในการปรากฏของคำศัพท์แต่ละตัวไว้ด้วย การแยกไฟล์พจนานุกรมและไฟล์สถิติคำศัพท์ออกจากกันทำให้สามารถเพิ่มเติมคำศัพท์ลงในไฟล์พจนานุกรมได้โดยไม่กระทบต่อค่าความถี่ในไฟล์สถิติคำศัพท์ ตัวอย่างข้อมูลที่เก็บไว้ในไฟล์สถิติคำศัพท์ มีลักษณะดังนี้

หน้า/D:4

หน้า/NCM:6

หน้า/PN:1

หน้ากระดาษ/NCM:2

หน้าตา/NCM:2

หน้าต่าง/NCM:1

...

...

ซึ่งมีความหมายว่า คำว่า “หน้า” ปรากฏเป็นตัวกำหนด (D) 4 ครั้ง, เป็นคำนามสามัญ (NCM) 6 ครั้ง, และเป็นคำบุพบทนำหน้านาม (PN) 1 ครั้งในคลังข้อมูลฝึกสอน และสายอักขระ “หน้า” ยังปรากฏเป็นส่วนย่อยภายในคำอื่นๆด้วย เช่น หน้ากระดาษ, หน้าตา, หน้าต่าง เป็นต้น

(3) **ไฟล์สถิติไตรแกรม** จัดเก็บค่าความถี่การปรากฏของลำดับหมวดคำในรูปไตรแกรมที่นับได้จากคลังข้อมูลฝึกสอน ตัวอย่างข้อมูลในไฟล์สถิติไตรแกรมมีลักษณะดังนี้

VNN0^NCM#NCM:3,NPRO:1,&PFX#V0:3,&

ซึ่งแสดงการเก็บค่าความถี่ไตรแกรมของลำดับหมวดคำ ดังนี้

trigram{VNN0}{NCM}{NCM} มีความถี่ 3

trigram{VNN0}{NCM}{NPRO} มีความถี่ 1

trigram{VNN0}{PFX}{V0} มีความถี่ 3

ค่าความถี่ไตรแกรมของหมวดคำจากตัวอย่าง มีความหมายว่า ในคลังข้อมูลฝึกสอนหมวดคำ VNN0 ปรากฏในลำดับหมวดคำ VNN0-NCM-NCM 3 ครั้ง (คือ VNN0 ปรากฏตามด้วย NCM และตามด้วย NCM เรียงกันเป็นลำดับ), ปรากฏในลำดับหมวดคำ VNN0-NCM-NPRO 1 ครั้ง, และปรากฏในลำดับหมวดคำ VNN0-PFX-V0 3 ครั้ง จากไตรแกรมทั้ง 3 ชุดแสดงให้เห็นว่า VNN0 ปรากฏทั้งหมด 7 ครั้งในคลังข้อมูลฝึกสอน

ส่วนคลังข้อมูลสำหรับทดสอบประสิทธิภาพมีลักษณะเป็นข้อความที่ไม่ได้มีการตัดคำและกำกับหมวดคำไว้เพื่อใช้เป็นข้อความทดสอบป้อนเข้าไปให้โปรแกรมได้ทำการทดลองตัดคำและกำกับหมวดคำ ซึ่งมีลักษณะดังรูปที่ 6-2 ข้อมูลชุดเดียวกันนี้ผู้วิจัยจะได้ทำการตัดคำและกำกับหมวดคำด้วยมือเตรียมไว้ล่วงหน้าเพื่อใช้เป็นคำตอบที่จะนำไปตรวจสอบกับผลลัพธ์ที่ได้จากโปรแกรม ซึ่งมีลักษณะดังรูปที่ 6-3

พนมเปญ - ยูเอ็นยังสงวนท่าที รอให้รัฐบาลพนมเปญยืนยันอีกครั้งเรื่องการยอมให้ใช้ระบบศาลระหว่างประเทศ พิจารณาคดีผู้นำเขมรแดง

หลังจากที่วุฒิสมาชิกจอห์น เคอร์รี ของสหรัฐ เปิดเผยเมื่อวันเสาร์ (29 เม.ย.) หลังพบหารือกับนายกรัฐมนตรีฮุน เซน ผู้นำกัมพูชาว่า รัฐบาลพนมเปญ ได้ยอมรับข้อเสนอของสหประชาชาติ (ยูเอ็น) เกี่ยวกับการตั้งศาลอาชญากรสงคราม เพื่อพิจารณาความผิดของผู้นำเขมรแดงแล้ว แต่ดูเหมือนว่า ยูเอ็นยังไม่มั่นใจในความสำเร็จของการเจรจาครั้งนี้นัก

รูปที่ 6-2 ลักษณะของข้อความทดสอบ

พนมเปญ/NPP_ /PUNC_-/PUNC_ /PUNC_ยูเอ็น/NPP_ยัง/AV_ สงวน/VN0_ท่าที/NCM_ /PUNC_รอ/VCV0_ให้/PCOMP_รัฐบาล/NCM_ พนมเปญ/NPP_ยืนยัน/VN0_อีก/Q_ครั้ง/NCSF_ /PUNC_เรื่อง/NCM_ การ/PFX_ยอม/VCV0_ให้/PCOMP_ใช้/VN0_ระบบ/NCM_ศาล/NCM_ ระหว่าง/PN_ประเทศ/NCM_ /PUNC_พิจารณา/VN0_คดี/NCM_ ผู้นำ/NCM_เขมรแดง/NPP_

หลังจากที่/PV_วุฒิสมาชิก/NCM_จอห์น/NPP_ /PUNC_เคอร์รี/NPP_ /PUNC_ของ/PN_สหรัฐ/NPP_ /PUNC_เปิดเผย/VCV0_เมื่อ/PN_วัน/NCM_ เสาร์/NPP_(/PUNC_29/NCM_ /PUNC_เม.ย./NPP_)/PUNC_ /PUNC_ หลัง/PV_พบ/VPN0_หารือ/VPN0_กับ/PN_นายกรัฐมนตรี/NCM_ฮุน/NPP_ /PUNC_เซน/NPP_ /PUNC_ผู้นำ/NCM_กัมพูชา/NPP_ว่า/PCOMP_ /PUNC_รัฐบาล/NCM_พนมเปญ/NPP_ /PUNC_ได้/VAUX_ยอมรับ/VN0_ ข้อเสนอ/NCM_ของ/PN_สหประชาชาติ/NPP_ /PUNC_(/PUNC_ยู เอ็น/NPP_)/PUNC_ /PUNC_เกี่ยวกับ/PN_การ/PFX_ตั้ง/VN0_ศาล/NCM_ อาชญากร/NCM_สงคราม/NCM_ /PUNC_เพื่อ/PV_พิจารณา/VN0_ ความผิด/NCM_ของ/PN_ผู้นำ/NCM_เขมรแดง/NPP_แล้ว/AV_ /PUNC_ แต่/C_ดูเหมือนว่า/AV_ /PUNC_ยูเอ็น/NPP_ยัง/AV_ไม่/AVNEG_ มั่นใจ/VPN0_ใน/PN_ความ/PFX_สำเร็จ/VADJ_ของ/PN_การ/PFX_ เจรจา/V0_ครั้ง/NCSF_นี้/D_นัก/PT_

รูปที่ 6-3 ลักษณะของคำตอบของการตัดคำและกำกับหมวดคำ

6.1.2 **ส่วนประกอบของโปรแกรมตัดคำและกำกับหมวดคำแบบเบ็ดเสร็จ**

หัวข้อนี้จะนำเสนอส่วนประกอบต่างๆของโปรแกรมแบบเบ็ดเสร็จและขั้นตอนการทำงานของโปรแกรมเพื่อทำการตัดคำและกำกับหมวดคำให้กับข้อความทดสอบ ในการทดลองตัดคำและกำกับหมวดคำให้กับข้อความทดสอบ ข้อความทดสอบจะถูกแบ่งเป็นส่วนๆตามเครื่องหมายวรรคตอนในภาษาไทย และนำไปประมวลผลทีละส่วน โดยข้อความทดสอบจะผ่านส่วนประกอบหลัก 3 ส่วนของโปรแกรม ดังรูปที่ 6-1 ได้แก่ ส่วนอ่านค่าสถิติ (read statistic), ส่วนเทียบคำและต่อสาย

คำ (word matching & path generation), และส่วนคำนวณค่าความน่าจะเป็นตามแบบจำลองไตรแกรม (trigram probability calculation & path selection) ซึ่งแต่ละส่วนมีหน้าที่ในการทำงาน ดังนี้

6.1.2.1 **ส่วนอ่านค่าสถิติ**

ส่วนอ่านค่าสถิติ มีหน้าที่ในการนำข้อมูลที่จัดเก็บไว้ในไฟล์พจนานุกรม, ไฟล์สถิติคำศัพท์, และไฟล์สถิติไตรแกรม เข้ามาเก็บไว้ในหน่วยความจำเพื่อการทำงานในขั้นตอนต่อไป โดยข้อมูลที่อ่านได้จากไฟล์พจนานุกรมนี้จะใช้สำหรับการเทียบคำและต่อสายคำ ส่วนข้อมูลความถี่ที่อ่านได้จากไฟล์สถิติคำศัพท์และไฟล์สถิติไตรแกรมจะใช้สำหรับการคำนวณค่าความน่าจะเป็นในการตัดคำและกำกับหมวดคำ

6.1.2.2 **ส่วนเทียบคำและต่อสายคำ**

ส่วนเทียบคำและต่อสายคำ มีหน้าที่เทียบสายอักขระในข้อความทดสอบที่ป้อนเข้าไปกับรายการคำศัพท์ในไฟล์พจนานุกรมเพื่อหาว่าสามารถเทียบเจอคำอะไรบ้างในข้อความทดสอบ โดยเริ่มจากการเทียบเจอทุกคำที่เป็นไปได้ในข้อความทดสอบ ตัวอย่างเช่น ข้อความ “นายกรัฐมนตรีของอังกฤษ” จะเทียบเจอคำว่า “นา”, “นาย”, “นายก”, “นายกรัฐมนตรี”, “ยก”, “รัฐ”, “รัฐมนตรี”, “มน”, “ตรี”, “ขอ”, “ของ”, “อังกฤษ”, “งก” เป็นต้น พร้อมกันนี้ ได้สร้าง chart สำหรับคำแต่ละคำที่เทียบเจอไว้เพื่อใช้สำหรับการต่อสายคำ โดยแต่ละ chart ระบุตำแหน่งเริ่มของคำ, ตำแหน่งจบของคำ และคำศัพท์ที่เทียบเจอ ดังตัวอย่างข้างล่างนี้

chart{0}{2}{นา/NCM}
chart{0}{3}{นาย/NCM}
chart{0}{4}{นายก/NCM}
chart{0}{12}{นายกรัฐมนตรี/NCM}
chart{2}{4}{ยก/NCSF}
chart{2}{4}{ยก/VN0}
chart{4}{7}{รัฐ/NCM}
chart{4}{12}{รัฐมนตรี/NCM}

chart{7}{9}{มน/VADJ}

chart{9}{12}{ตรี/NCM}

chart{12}{14}{ขอ/NCM}

chart{12}{14}{ขอ/VN0}

chart{12}{15}{ของ/NCM}

chart{12}{15}{ของ/P}

chart{14}{16}{งอ/VADJ}

chart{14}{16}{งอ/VN0}

chart{15}{21}{อังกฤษ/NPP}

chart{17}{19}{งก/VADJ}

chart ทั้งหมดที่ได้จากการเทียบคำจะนำมาใช้สำหรับการต่อสายคำ โดยโปรแกรมพยายามสร้างเส้นทางเชื่อมต่อระหว่าง chart เพื่อหาเส้นทางที่สามารถเชื่อมต่อตั้งแต่ต้นข้อความไปจนจบข้อความได้ ในระหว่างการสร้างเส้นทาง สามารถลดความสิ้นเปลืองในการคำนวณได้โดยจดจำเฉพาะเส้นทางที่ดีที่สุดที่มาจบลงที่ chart นั้นๆตามแนวคิดของขั้นตอนวิธีวิเทอร์บิ (Manning and Schutze, 1999) ดังที่ได้กล่าวไปในบทที่ 5 ผลลัพธ์จากขั้นตอนนี้ คือ เส้นทางที่ดีที่สุดที่มาจบลงที่คำสุดท้ายต่างๆกัน ดังนั้นผลลัพธ์จากขั้นตอนนี้อาจมีได้มากกว่า 1 เส้นทาง

6.1.2.3 **ส่วนคำนวณค่าความน่าจะเป็นตามแบบจำลองไตรแกรม**

ส่วนคำนวณค่าความน่าจะเป็นตามแบบจำลองไตรแกรม มีหน้าที่คำนวณค่าความน่าจะเป็นของเส้นทางต่างๆที่เป็นผลลัพธ์มาจากส่วนเทียบคำและต่อสายคำ แล้วเปรียบเทียบค่าความน่าจะเป็นของทุกเส้นทาง เพื่อเลือกเส้นทางที่มีค่าความน่าจะเป็นสูงสุดเป็นคำตอบของการตัดคำและกำกับหมวดคำแบบเบ็ดเสร็จ ส่วนนี้จะทำการคำนวณค่าความน่าจะเป็นในการตัดคำและกำกับหมวดคำตามแบบจำลองไตรแกรม ซึ่งแสดงได้ดังสมการที่ 6-1 (ดูรายละเอียดของแบบจำลองไตรแกรมได้ในบทที่ 5)

(6-1) $$\prod_{i=1...n} \mathrm{PROB}(t_i \mid t_{i-1}, t_{i-2}) \times \mathrm{PROB}(w_i \mid t_i)$$

กล่าวคือ สำหรับคำแต่ละคำ จะต้องทำการคำนวณค่าความน่าจะเป็น 2 ค่า ได้แก่

(1) ค่าความน่าจะเป็นของลำดับหมวดคำ (tag sequence probability) คือ ความน่าจะเป็นในการปรากฏของหมวดคำโดยพิจารณาเฉพาะ 2 หมวดคำก่อนหน้าเท่านั้น ค่าความน่าจะเป็นตรงนี้สามารถคำนวณได้โดยนับจำนวนของลำดับหมวดคำ t_{i-2} t_{i-1} t_i ที่ปรากฏในคลังข้อมูลเทียบกับจำนวนของลำดับหมวดคำ t_{i-2} t_{i-1} ทั้งหมดที่ปรากฏในคลังข้อมูล ดังสมการที่ 6-2

(6-2) $$\text{PROB}(t_i \mid t_{i-1}, t_{i-2}) = \frac{\text{count}\ (t_{i-2}\ t_{i-1}\ t_i)}{\text{count}\ (t_{i-2}\ t_{i-1})}$$

(2) ค่าความน่าจะเป็นในการปรากฏของคำ (lexical generation probability) คือ ความน่าจะเป็นในการปรากฏเป็นรูปคำนั้นๆเมื่อกำหนดหมวดคำให้ ค่าความน่าจะเป็นตรงนี้สามารถคำนวณได้โดยนับจำนวนของคำศัพท์ที่มีรูปคำเป็น w_i และมีหมวดคำเป็น t_i ที่ปรากฏในคลังข้อมูลเทียบกับจำนวนของหมวดคำ t_i ทั้งหมดที่ปรากฏในคลังข้อมูล ดังสมการที่ 6-3

(6-3) $$\text{PROB}(w_i \mid t_i) = \frac{\text{count}\ (w_i, t_i)}{\text{count}\ (t_i)}$$

ส่วนกรณีที่ลำดับหมวดคำ t_{i-2} t_{i-1} t_i ไม่ปรากฏในคลังข้อมูล ซึ่งจะส่งผลให้ค่าความน่าจะเป็นของทั้งเส้นทางเท่ากับศูนย์ไปด้วย จะใช้วิธี Witten-Bell Smoothing (Witten and Bell,1991 cited in Jurafsky and Martin, 2000) ช่วยปรับค่าความน่าจะเป็นไม่ให้เท่ากับศูนย์ ผลลัพธ์จากส่วนนี้จะได้เส้นทางของการตัดคำและกำกับหมวดคำที่มีค่าความน่าจะเป็นสูงสุดเพียงเส้นทางเดียวที่เลือกมาเป็นคำตอบสำหรับการตัดคำและกำกับหมวดคำของโปรแกรมแบบเบ็ดเสร็จ

นอกจากนี้ วิทยานิพนธ์นี้จะเปรียบเทียบผลการตัดคำของโปรแกรมแบบเบ็ดเสร็จที่นำหมวดคำข้างเคียงมาช่วยในการตัดคำกับผลการตัดคำแบบที่ไม่ได้นำหมวดคำข้างเคียงมาช่วย ซึ่งการตัดคำแบบที่ไม่ได้นำหมวดคำข้างเคียงมาช่วยนั้น มีลักษณะปัญหาในการตัดคำดังสมการที่ 6-4 (เหมือนกับสมการที่ 5-1)

(6-4) $$W = \max_{w_1,\ldots,w_n} \arg\ \text{PROB}(w_1,\ldots,w_n \mid c_1,\ldots,c_m)$$

ซึ่งหมายความว่า เป็นการหาสายคำที่มีค่าความน่าจะเป็นสูงที่สุดเมื่อกำหนดให้สายอักขระมา ซึ่งเท่ากับว่า เป็นการหาสายคำที่มีค่าความน่าจะเป็นสูงที่สุดเท่านั้น เนื่องจากสายอักขระถูกกำหนดจากข้อความที่ป้อนเข้าไปอยู่แล้ว ดังนั้น สมการที่ 6-4 สามารถแปลงเป็นสมการที่ 6-5

(6-5) $$W = \max_{w_1,...,w_n} \arg \; PROB(w_1,...,w_n)$$

จากสมการที่ 6-5 แสดงให้เห็นว่า ต้องการสายคำที่มีค่าความน่าจะเป็นสูงสุดเท่านั้น โดยไม่ได้พิจารณาค่าความน่าจะเป็นของสายหมวดคำพร้อมกันไปด้วย สมการที่ 6-5 นี้ได้มีงานวิจัยที่นำไปประยุกต์ใช้ในรูปแบบต่างๆ เช่น บุญเสริม กิจศิริกุล (2541), Asanee Kawtrakul et. al. (1995 cited in Surapant Meknavin and Boonserm Kijsirikul, 2000), Surapant Meknavin (1995 cited in Surapant Meknavin and Boonserm Kijsirikul, 2000) เป็นต้น

วิทยานิพนธ์ฉบับนี้ก็ได้นำสมการที่ 6-5 มาประยุกต์ใช้เพื่อเป็นตัวอย่างของแนวคิดแบบที่ไม่ได้นำหมวดคำข้างเคียงมาช่วย โดยได้นำแนวคิดของแบบจำลองไตรแกรมมาใช้กับรูปคำ กล่าวคือ ใช้รูปคำข้างเคียงเพื่อช่วยในการคำนวณค่าความน่าจะเป็นของสายคำเพื่อแก้ปัญหาการตัดคำ ซึ่งแสดงได้ดังสมการที่ 6-6

(6-6) $$\prod_{i=1...n} PROB(w_i \mid w_{i-1}, w_{i-2})$$

สมการที่ 6-6 เป็นการคำนวณหา word sequence probability คือ หาค่าความน่าจะเป็นในการปรากฏของรูปคำ w_i โดยพิจารณาเฉพาะรูปคำก่อนหน้า 2 คำเท่านั้น คือ w_{i-2} w_{i-1} ค่าความน่าจะเป็นนี้สามารถคำนวณได้โดยนับจำนวนของลำดับรูปคำ w_{i-2} w_{i-1} w_i ที่ปรากฏในคลังข้อมูล เทียบกับจำนวนของลำดับรูปคำ w_{i-2} w_{i-1} ทั้งหมดที่ปรากฏในคลังข้อมูล ดังสมการที่ 6-7

(6-7) $$PROB(w_i \mid w_{i-1}, w_{i-2}) = \frac{count\ (w_{i-2}\ w_{i-1}\ w_i)}{count\ (w_{i-2}\ w_{i-1})}$$

จากสมการที่ 6-7 แสดงให้เห็นว่า ในการเรียนรู้ค่าสถิติจากคลังข้อมูลฝึกสอน โปรแกรมฝึกสอนจำเป็นต้องมีการนับค่าความถี่ของลำดับรูปคำเพื่อใช้ในการคำนวณตามสมการ ดังนั้น จากตัวอย่างกระบวนการนับค่าความถี่ในคลังข้อมูลฝึกสอนข้างต้น (ในตอนต้นของหัวข้อ 6.1.1) คือ

ต้อง/VAUX_ทำ/VN0_ /PUNC_เพื่อ/PV_ปฏิรูป/VN0_การเมือง/NCM

ในที่นี้ต้องนับค่าความถี่ของลำดับรูปคำทีละ 3 คำ (ไตรแกรมของรูปคำ) ดังนี้ ... [ต้อง-ทำ-] , [ทำ- -เพื่อ] , [-เพื่อ-ปฏิรูป] , [เพื่อ-ปฏิรูป-การเมือง] ... ตามลำดับไปจนครบทุกคำในคลังข้อมูลฝึกสอน แล้วเก็บค่าความถี่นี้ไว้ในอีกไฟล์หนึ่ง คือ ไฟล์สถิติไตรแกรมของรูปคำ ส่วนขั้นตอนและส่วนประกอบอื่นๆของโปรแกรมนี้จะทำได้ในลักษณะเดียวกับขั้นตอนและส่วนประกอบของโปรแกรมแบบเบ็ดเสร็จ

ตารางที่ 6-1 แสดงสมการและค่าความถี่ที่ใช้ในการคำนวณค่าความน่าจะเป็นของโปรแกรมทั้งสองแบบเปรียบเทียบกัน ดังนี้

โปรแกรมแบบที่ใช้หมวดคำข้างเคียง (โปรแกรมแบบเบ็ดเสร็จ)	**โปรแกรมแบบที่ใช้รูปคำข้างเคียง (ไม่ได้นำหมวดคำข้างเคียงมาช่วย)**
$\max_{t_1,...,t_n} \arg \text{PROB}(w_1,..., w_n, t_1,..., t_n \mid c_1,..., c_m)$	$\max_{w_1,...,w_n} \arg \text{PROB}(w_1,..., w_n \mid c_1,..., c_m)$
$\max_{t_1,...,t_n} \arg \text{PROB}(w_1,..., w_n, t_1,..., t_n)$	$\max_{w_1,...,w_n} \arg \text{PROB}(w_1,..., w_n)$
$\text{PROB}(t_1,..., t_n) \times \text{PROB}(w_1,...,w_n \mid t_1,..., t_n)$	
$\prod_{i=1...n} P(t_i \mid t_{i-1}, t_{i-2}) \times P(w_i \mid t_i)$	$\prod_{i=1...n} P(w_i \mid w_{i-1}, w_{i-2})$
$\frac{\text{count}(t_{i-2}\ t_{i-1}\ t_i)}{\text{count}(t_{i-2}\ t_{i-1})} \times \frac{\text{count}(w_i, t_i)}{\text{count}(t_i)}$	$\frac{\text{count}(w_{i-2}\ w_{i-1}\ w_i)}{\text{count}(w_{i-2}\ w_{i-1})}$

ตารางที่ 6-1 ตารางเปรียบเทียบลำดับที่มาของสมการและค่าความถี่ที่ใช้ของโปรแกรมทั้งสองแบบ

6.1.3 ผลลัพธ์จากการตัดคำและกำกับหมวดคำอัตโนมัติด้วยโปรแกรมแบบเบ็ดเสร็จ

ผลลัพธ์จากการตัดคำและกำกับหมวดคำอัตโนมัติด้วยโปรแกรม คือ ข้อความทดสอบที่ได้รับการตัดคำและกำกับหมวดคำซึ่งมีค่าความน่าจะเป็นสูงที่สุดเพียงแบบเดียว แล้วเก็บแยกไว้ในไฟล์ผลลัพธ์สำหรับนำไปประเมินผลต่อไป ลักษณะของผลลัพธ์จากการตัดคำและกำกับหมวดคำด้วยโปรแกรมจะมีลักษณะเหมือนกับลักษณะของคำตอบที่ผู้วิจัยได้ตัดคำและกำกับหมวดคำด้วยมือไว้ล่วงหน้า ลักษณะของผลลัพธ์แสดงได้ดังรูปที่ 6-4

พนมเปญ/NPP_ /PUNC_-/PUNC_ /PUNC_ยูเอ็น/NPP_ยัง/AV_สงวน/VN0_ท่าที/NCM_ /PUNC_รอ/VCV0_ให้/PCOMP_รัฐบาล/NCM_พนมเปญ/NPP_ยืนยัน/VN0_อีก/Q_ครั้ง/NCSF_ /PUNC_เรื่อง/NCM_การ/PFX_ยอม/VCV0_ให้/PCOMP_ใช้/VN0_ระบบ/NCM_ศาล/NCM_ระหว่าง/PN_ประเทศ/NCM_ /PUNC_พิจารณา/VN0_คดี/NCM_ผู้นำ/NCM_เขมรแดง/NPP
หลังจากที่/PV_วุฒิสมาชิก/NCM_จอห์น/NPP_ /PUNC_เคอร์รี/NPP_ /PUNC_ของ/PN_สหรัฐ/NPP_ /PUNC_เปิดเผย/VPN0_เมื่อ/PN_วัน/NCM_เสาร์/NPP_(/PUNC_29/NCM_ /PUNC_เม.ย./NPP_)/PUNC_ /PUNC_หลัง/PV_พบ/VN0_หารือ/VPN0_กับ/PN_นายกรัฐมนตรี/NCM_ฮุน/NPP_ /PUNC_เซน/NPP_ /PUNC_ผู้นำ/NCM_กัมพูชา/NPP_ว่า/PCOMP_ /PUNC_รัฐบาล/NCM_พนมเปญ/NPP_ /PUNC_ได้/VAUX_ยอมรับ/VN0_ข้อเสนอ/NCM_ของ/PN_สหประชาชาติ/NPP_ /PUNC_(/PUNC_ยูเอ็น/NPP_)/PUNC_ /PUNC_เกี่ยวกับ/PN_การ/PFX_ตั้ง/VN0_ศาล/NCM_อาชญากร/NCM_สงคราม/NCM_ /PUNC_เพื่อ/PV_พิจารณา/VN0_ความ/PFX_ผิด/VADJ_ของ/PN_ผู้นำ/NCM_เขมรแดง/NPP_แล้ว/AV_ /PUNC_แต่/C_ดูเหมือน/AV_ว่า/PCOMP_ /PUNC_ยูเอ็น/NPP_ยัง/AV_ไม่/AVNEG_มั่นใจ/V0_ใน/PN_ความ/PFX_สำเร็จ/V0_ของ/PN_การ/PFX_เจรจา/V0_ครั้ง/NCSF_นี้/D_นัก/PT

รูปที่ 6-4 ลักษณะของผลลัพธ์จากการตัดคำและกำกับหมวดคำด้วยโปรแกรมแบบเบ็ดเสร็จ

6.2 วิธีการประเมินผล

การประเมินผลการตัดคำและกำกับหมวดคำในวิทยานิพนธ์ฉบับนี้จะประเมินผลที่ได้จากการทดลองกับข้อความทดสอบ โดยประเมินผลเฉพาะประสิทธิภาพด้านความถูกต้องของการตัดคำและกำกับหมวดคำเท่านั้น ไม่ประเมินผลประสิทธิภาพด้านความเร็วและปริมาณการใช้ทรัพยากร การประเมินผลความถูกต้องทำได้โดยตรวจสอบผลลัพธ์จากโปรแกรมกับคำตอบซึ่งได้จากการตัดคำและกำกับหมวดคำด้วยมือ แต่เนื่องจากในบางกรณีมนุษย์เองก็ไม่สามารถเห็นพ้องได้ว่าการตัดคำและกำกับหมวดคำแบบใดเป็นคำตอบที่ถูกต้อง แต่ละคนอาจเลือกรูปแบบการตัดคำและกำกับหมวดคำต่างกันไปได้ ดังนั้นในวิทยานิพนธ์นี้ ผู้วิจัยจะเป็นผู้ทำการตัดคำและกำกับหมวดคำให้กับข้อความทดสอบไว้ล่วงหน้า และถือว่าเป็นคำตอบที่มีความถูกต้องเท่ากับ 100% สำหรับนำไปตรวจสอบกับผลลัพธ์จากโปรแกรม

การวัดค่าความถูกต้องทำได้โดยใช้วิธีการวัดค่า F-measure (Van Rijsbergen, 1979 cited in Manning and Schutze, 1999) ซึ่งคำนวณได้ดังสมการที่ 6-8

(6-8)
$$F = \frac{1}{\alpha \frac{1}{P} + (1 - \alpha) \frac{1}{R}}$$

ค่า F-measure จะพิจารณาจากค่าความแม่นยำ (P = Precision) และ ค่าความครบถ้วน (R = Recall) ซึ่งในการวัดประสิทธิภาพนี้พิจารณาให้ค่าทั้งสองมีน้ำหนักเท่าๆกัน ($\alpha = 0.5$) ทำให้สามารถแปลงสมการในการวัดค่าให้อยู่ในรูปที่ง่ายขึ้นได้ คือ

(6-9)
$$F = \frac{2 \times P \times R}{P + R}$$

โดยแบ่งการประเมินผลเป็นการประเมินผลการตัดคำ และการประเมินผลการกำกับหมวดคำ ดังนี้

1. การประเมินผลความถูกต้องของการตัดคำ ทำได้โดย คำนวณค่าความแม่นยำ (P) จากจำนวนคำที่โปรแกรมตัดคำได้ถูกต้องเทียบกับจำนวนคำทั้งหมดที่โปรแกรมตัดคำได้ และคำนวณค่า

ความครบถ้วน (R) จากจำนวนคำที่โปรแกรมตัดคำได้ถูกต้องเทียบกับจำนวนคำทั้งหมดที่ผู้วิจัยตัดคำได้ แล้วนำไปคำนวณค่า F-measure ตามสมการที่ 6-9

2. การประเมินผลความถูกต้องของการกำกับหมวดคำ ทำได้โดย คำนวณค่าความแม่นยำและค่าความครบถ้วนเหมือนที่ใช้ในการประเมินผลการตัดคำ แต่ให้พิจารณาทั้งคำและหมวดคำไปด้วยกัน แล้วคำนวณค่า F-measure ตามสมการที่ 6-9

นอกจากนี้ วิทยานิพนธ์นี้จะได้เปรียบเทียบผลการตัดคำของโปรแกรมแบบที่นำข้อมูลเรื่องหมวดคำมาพิจารณาไปพร้อมๆกัน กับการตัดคำแบบธรรมดาที่ไม่นำข้อมูลเรื่องหมวดคำมาช่วย โดยพิจารณาจากผลการตัดคำของข้อความทดสอบชุดเดียวกัน เพื่อดูว่ามีความแตกต่างกันหรือไม่อย่างไร และเปรียบเทียบประสิทธิภาพการตัดคำของโปรแกรมทั้ง 2 แบบ

6.3 ผลการทดลองตัดคำและกำกับหมวดคำแบบเบ็ดเสร็จ

ส่วนนี้จะได้กล่าวถึงผลการทำงานของโปรแกรมแบบเบ็ดเสร็จที่พัฒนาขึ้นมาซึ่งได้นำไปทดลองตัดคำและกำกับหมวดคำให้กับคลังข้อมูลชุดทดสอบที่ได้จัดเตรียมไว้ และทำการประเมินผลความถูกต้องดังที่ได้กล่าวมาแล้ว โดยแยกเป็น การนำเสนอและอภิปรายผลการวัดประสิทธิภาพในการทดลองกำกับหมวดคำในหัวข้อที่ 6.3.1 และการนำเสนอและอภิปรายผลการวัดประสิทธิภาพในการทดลองตัดคำในหัวข้อที่ 6.3.2 โดยเปรียบเทียบกับประสิทธิภาพในการตัดคำของโปรแกรมแบบที่ใช้แบบจำลองไตรแกรมของรูปคำซึ่งไม่ได้นำหมวดคำมาช่วย เพื่อศึกษาว่าการนำปริบทหมวดคำของคำข้างเคียงมาช่วยในการตัดคำจะสามารถช่วยให้ตัดคำได้ถูกต้องมากกว่าปริบทรูปคำของคำข้างเคียงหรือไม่ อย่างไร ซึ่งผู้วิจัยตั้งสมมติฐานว่า การนำปริบทหมวดคำของคำข้างเคียงมาช่วยในการตัดคำจะช่วยให้สามารถตัดคำได้ถูกต้องมากยิ่งขึ้น จากนั้น หัวข้อที่ 6.3.3 จะได้สรุปผลการทดลองทั้งหมด

ในการทดสอบประสิทธิภาพนี้ ใช้คลังข้อมูลฝึกสอนขนาดประมาณ 20,000 คำที่จัดทำขึ้นเพื่อฝึกสอนให้โปรแกรมเรียนรู้ค่าสถิติสำหรับไว้ใช้คำนวณค่าความน่าจะเป็นในการตัดคำและการกำกับหมวดคำ และได้ทำการทดสอบกับคลังข้อมูลทดสอบขนาดประมาณ 5,000 คำ แล้วจึงประเมินผลประสิทธิภาพความถูกต้อง โดยเทียบผลลัพธ์ที่ได้จากโปรแกรมกับคำตอบการตัดคำ

และการกำกับหมวดคำของคลังข้อมูลทดสอบชุดเดียวกันที่ผู้วิจัยได้จัดทำไว้ล่วงหน้า ซึ่งจะถือว่ามีความถูกต้องในการตัดคำและกำกับหมวดคำเท่ากับ 100% ผลการทดลองปรากฏ ดังนี้

6.3.1 ผลการทดลองกำกับหมวดคำ

ผลการทดลองกำกับหมวดคำให้กับข้อความทดสอบ ดังตารางที่ 6-2 แสดงให้เห็นว่าโปรแกรมตัดคำและกำกับหมวดคำแบบเบ็ดเสร็จสามารถกำกับหมวดคำให้กับข้อความภาษาไทยได้อย่างมีประสิทธิภาพ โดยสามารถกำกับหมวดคำได้ถูกต้อง 89.590 % ซึ่งมีความถูกต้องสูงกว่าการกำกับหมวดคำแบบที่อาศัยค่าความน่าจะเป็นในการปรากฏของหมวดคำนั้นๆเพียงอย่างเดียวโดยไม่ได้นำหมวดคำข้างเคียงมาช่วย ซึ่งกำกับหมวดคำได้ถูกต้อง 84.110 %

	Precision	Recall	F-measure
POS Tagging	88.570 %	90.634 %	89.590 %
POS Tagging using simple probability	83.743 %	84.481 %	84.110 %

ตารางที่ 6-2 ประสิทธิภาพในการกำกับหมวดคำของโปรแกรมแบบเบ็ดเสร็จ

ผลการทดลองกำกับหมวดคำจากตารางที่ 6-2 ชี้ให้เห็นว่า แบบจำลองไตรแกรมซึ่งใช้ปริบทหมวดคำข้างเคียงในการคำนวณค่าความน่าจะเป็นของสายคำและสายหมวดคำเป็นกระบวนการที่เหมาะสมในการกำกับหมวดคำภาษาไทย ซึ่งตรงตามสมมติฐานที่ตั้งไว้ก่อนหน้านี้ ที่ว่า หมวดคำข้างเคียงเป็นปริบทที่เหมาะสมซึ่งสามารถนำมาใช้เพื่อช่วยแก้ปัญหาความกำกวมในการกำกับหมวดคำภาษาไทยได้

การที่แบบจำลองไตรแกรมซึ่งอาศัยปริบทหมวดคำข้างเคียงสามารถกำกับหมวดคำได้ถูกต้องสูง เนื่องจากการกำกับหมวดคำโดยใช้แบบจำลองไตรแกรมได้เลือกผลลัพธ์โดยพิจารณาให้สายคำและสายหมวดคำที่มีค่าความน่าจะเป็นสูงที่สุดเป็นผลลัพธ์ของการกำกับหมวดคำ และเนื่องจากการใช้ปริบทหมวดคำข้างเคียงในรูปไตรแกรมของหมวดคำนี้สามารถนำปริบทในการปรากฏของหมวดคำในขอบเขตที่จำกัด (ขอบเขตเท่ากับ 2 คำก่อนหน้า) มาช่วยในการคำนวณค่า

ความน่าจะเป็นได้ และยังทำให้สามารถเลือกสายคำและสายหมวดคำที่ถูกต้องซึ่งมีค่าความน่าจะเป็นสูงที่สุดได้เป็นจำนวนมาก ดังนั้น วิธีการนี้จึงมีประสิทธิภาพสูงในการกำกับหมวดคำ ในขณะที่หากไม่นำปริบทหมวดคำข้างเคียงมาช่วยคำนวณค่าความน่าจะเป็นในการปรากฏของหมวดคำจะทำให้เลือกหมวดคำที่เหมาะสมกับปริบทในการปรากฏได้ถูกต้องน้อยกว่า ตัวอย่างเช่น รูปคำว่า “ที่” เป็นคำหลายหน้าที่ที่สามารถเป็นได้มากกว่า 1 หมวดคำและแต่ละหมวดคำปรากฏในปริบทต่างกัน ตัวอย่างที่ 1 แสดงการกำกับหมวดคำที่ถูกต้องของคำว่า “ที่” ในปริบทต่างๆ

ตัวอย่างที่ 1:

ก. โฆษก/NCM-ยูเอ็น/NPP-ที่/PN-นคร/NCM-นิวยอร์ก/NPP

ข. หาก/PV-ต้องการ/VCV0-ที่/PCOMP-จะ/AV-อยู่รอด/V0

ค. เป็น/VN0-สายการบิน/NCM-ที่/PCOMP-ได้รับ/VN0-คำร้อง/NCM

จากตัวอย่าง จะเห็นได้ว่า คำว่า “ที่” ที่ปรากฏในปริบทต่างกัน จะมีหมวดคำที่ถูกต้องต่างกัน ทั้งนี้การใช้แบบจำลองไตรแกรมของหมวดคำจะสามารถเลือกกำกับหมวดคำที่เหมาะสมของคำว่า “ที่” ในแต่ละปริบทได้อย่างถูกต้องตามตัวอย่างที่ 1 เนื่องจากได้นำหมวดคำข้างเคียงมาพิจารณาร่วมด้วย ดังจะอธิบายได้ดังนี้

ในตัวอย่างที่ 1-ก แบบจำลองไตรแกรมที่ใช้หมวดคำข้างเคียงจะกำกับหมวดคำเป็น “ที่/PN” เพราะว่า เมื่อพิจารณาจากหมวดคำข้างเคียง การกำกับหมวดคำเป็น “PN” จะทำให้ลำดับหมวดคำที่เกี่ยวข้อง (ได้แก่ NCM-NPP-PN, NPP-PN-NCM, PN-NCM-NPP) มีค่าความน่าจะเป็นสูง และเมื่อคำนวณร่วมกับค่าความน่าจะเป็นในการปรากฏของรูปคำ “ที่” เมื่อกำหนดให้หมวดคำเป็น “PN” แล้ว จะทำให้สายหมวดคำดังกล่าวมีค่าความน่าจะเป็นสูงที่สุดในปริบทนี้ (สูงกว่าการกำกับหมวดคำเป็น “ที่/PCOMP”)

ในตัวอย่างที่ 1-ข แบบจำลองไตรแกรมที่ใช้หมวดคำข้างเคียงจะกำกับหมวดคำเป็น “ที่/PCOMP” เพราะว่า เมื่อพิจารณาจากหมวดคำข้างเคียง การกำกับหมวดคำเป็น “PCOMP” จะทำให้ลำดับหมวดคำที่เกี่ยวข้อง (ได้แก่ PV-VCV0-PCOMP, VCV0-PCOMP-AV, PCOMP-AV-V0) มีค่าความน่าจะเป็นสูง และเมื่อคำนวณร่วมกับค่าความน่าจะเป็นในการปรากฏของรูปคำ “ที่” เมื่อกำหนดให้หมวดคำเป็น “PCOMP” จะทำให้สายหมวดคำดังกล่าวมีค่าความน่าจะเป็นสูงที่สุด

ในปริบทนี้ (สูงกว่าการกำกับหมวดคำเป็น “ที่/PN”) ตัวอย่างที่ 1-ค ก็สามารถอธิบายได้ในลักษณะเดียวกัน

ส่วนการคำนวณค่าความน่าจะเป็นแบบที่ไม่นำปริบทหมวดคำข้างเคียงมาพิจารณาร่วมด้วยจะกำกับหมวดคำในทุกปริบทเป็นแบบเดียวเท่านั้น กล่าวคือ หมวดคำที่มีความถี่ในการปรากฏร่วมกับรูปคำดังกล่าวสูงที่สุดจะถูกเลือกให้เป็นผลลัพธ์ในการกำกับหมวดคำของรูปคำนั้นทุกครั้ง ไม่ว่ารูปคำดังกล่าวจะปรากฏในปริบทใด ซึ่งเมื่อพิจารณาคำว่า “ที่” แล้ว พบว่า “ที่/PCOMP” มีค่าความถี่ในการปรากฏสูงที่สุดในคลังข้อมูลฝึกสอน ดังนั้น หากทำการกำกับหมวดคำให้กับข้อความในตัวอย่างที่ 1 โดยไม่นำปริบทหมวดคำข้างเคียงมาพิจารณาร่วมด้วยแล้วจะได้ผลลัพธ์ดังตัวอย่างที่ 2

ตัวอย่างที่ 2:

ก. โฆษก/NCM-ยูเอ็น/NPP-ที่/PCOMP-นคร/NCM-นิวยอร์ก/NPP ✕

ข. หาก/PV-ต้องการ/VN0-ที่/PCOMP-จะ/AV-อยู่รอด/V0

ค. เป็น/VN0-สายการบิน/NCM-ที่/PCOMP-ได้รับ/VN0-คำร้อง/NCM

จากตัวอย่างที่ 2 จะเห็นได้ว่า เมื่อไม่นำปริบทหมวดคำข้างเคียงมาพิจารณาร่วมด้วย คำว่า “ที่” จะถูกกำกับหมวดคำเป็น “ที่/PCOMP” เหมือนกันหมดทุกครั้ง ซึ่งส่งผลให้ในปริบทที่ควรกำกับเป็น “ที่/PN” (ตัวอย่างที่ 2-ก) กลับได้ผลลัพธ์เป็น “ที่/PCOMP” ซึ่งเป็นหมวดคำที่ผิดในปริบทนั้น ดังนั้น จึงทำให้ค่าความถูกต้องในการกำกับหมวดคำด้วยวิธีการนี้ต่ำกว่าวิธีการที่ใช้แบบจำลองไตรแกรมที่ได้นำปริบทหมวดคำข้างเคียงมาช่วยคำนวณค่าความน่าจะเป็น

ด้วยเหตุผลดังที่กล่าวมา การกำกับหมวดคำโดยใช้แบบจำลองไตรแกรมซึ่งอาศัยปริบทหมวดคำของคำข้างเคียงในการคำนวณค่าความน่าจะเป็นของสายคำและสายหมวดคำจึงเป็นวิธีการที่เหมาะสมสำหรับใช้กำกับหมวดคำภาษาไทยวิธีการหนึ่งซึ่งมีความถูกต้องในการกำกับหมวดคำสูง

อย่างไรก็ตาม เมื่อพิจารณาจากตารางที่ 6-2 จะเห็นได้ว่า โปรแกรมแบบเบ็ดเสร็จมีข้อผิดพลาดในการกำกับหมวดคำอยู่ประมาณ 10.5 % ผู้วิจัยจึงได้นำข้อผิดพลาดเหล่านี้มาศึกษาเพื่อดูว่ามีการกำกับหมวดคำที่ผิดพลาดเกิดขึ้นกับหมวดคำใดบ้าง โดยเลือกเอาเฉพาะกรณี

ของคำที่ตัดคำถูกต้องแล้วแต่กำกับหมวดคำผิด และคำข้างเคียงภายในขอบเขต 2 คำก่อนหน้า และ 2 คำตามหลังของคำดังกล่าวมีการตัดคำและกำกับหมวดคำที่ถูกต้อง (ซึ่งเป็นกรณีที่เห็นได้ชัดเจนว่า การที่คำดังกล่าวกำกับหมวดคำผิดไม่ได้เป็นผลกระทบมาจากความผิดพลาดในการตัดคำหรือในการกำกับหมวดคำของคำข้างเคียง) ผู้วิจัยได้จัดผลจากการศึกษาจำแนกเป็นกลุ่มตามหมวดคำที่ถูกต้อง (คือ หมวดคำในไฟล์คำตอบที่ผู้วิจัยได้ทำการกำกับไว้) ซึ่งสามารถแสดงดังตารางที่ 6-3 ข้างล่างนี้ (ส่วนรายการของคำที่กำกับหมวดคำผิดพลาดทั้งหมดที่นำมาศึกษาจะแสดงไว้ในภาคผนวก ข)

หมวดคำที่ถูกต้อง	**จำนวนรูปคำทั้งหมดที่กำกับหมวดคำนี้ผิด**	**กรณีที่กำกับหมวดคำผิดและจำนวนรูปคำที่กำกับผิด**
V0	22	**VN0(11)**, AV(4), VPN0(3), VCV0(2), VV0(1), VNPN0(1)
VN0	18	**V0(7)**, VCV0(4), VAUX(3), VPN0(1), VNPN0(1), PV(1), PCOMP(1)
NCM	17	**NCSF(7)**, **Q(5)**, V0(2), NPRO(1), D(1), PN(1)
VCV0	9	VN0(4), V0(3), VPN0(1), VNCV0(1)
AV	8	V0(2), D(2), VPN0(1), VAUX(1), D(1), C(1), Q(1)
PN	8	PV(3), V0(1), VS0(1), NCM(1), PCOMP(1), C(1)
VPN0	7	V0(7), VCV0(1), VN0(1), VNPN0(1)
VNPN0	6	VN0(4), V0(1), VS0(1)
VS0	6	VN0(2), VCV0(1), VV0(1), PN(1), PCOMP(1)
NCSF	6	**NCM(6)**
PV	4	AV(2), PN(1), C(1)
Q	4	NCM(2), VN0(1), AV(1)
NPRO	3	NCM(2), NCSF(1)
VV0	3	V0(2), PV(1)
PCOMP	2	PN(2)
VADJ	2	V0(1), VN0(1)
VNCV0	2	VN0(2)

C	1	PV(1)
PT	1	AV(1)
VAUX	1	AV(1)
D	1	AV(1)
AVNEG	0	
NPP	0	
PFX	0	
PUNC	0	
VNN0	0	

ตารางที่ 6-3 ตารางแสดงหมวดคำและความถี่ที่กำกับผิด

เมื่อพิจารณาตามตารางที่ 6-3 จะเห็นว่า จากหมวดคำทั้งหมด 26 หมวดคำในวิทยานิพนธ์ฉบับนี้ มี 21 หมวดคำที่โปรแกรมแบบเบ็ดเสร็จกำกับหมวดคำผิด ซึ่งแสดงว่าความผิดพลาดในการกำกับหมวดคำมีการกระจายไปปรากฏกับหมวดคำต่างๆเกือบทุกหมวดคำ ผู้วิจัยจึงได้เลือกศึกษาวิเคราะห์โดยละเอียดเฉพาะกรณีของการกำกับหมวดคำผิดพลาดที่มีความถี่สูง เช่น กรณีที่หมวดคำ NCM ถูกกำกับผิดไปเป็น NCSF มีทั้งหมด 8 ครั้งซึ่งเกิดกับรูปคำทั้งหมด 7 รูปคำ เป็นต้น โดยผู้วิจัยเลือกเอาเฉพาะกรณีที่การกำกับหมวดคำผิดพลาดเกิดขึ้นกับรูปคำตั้งแต่ 5 รูปคำขึ้นไป ซึ่งพบว่ามีอยู่ 5 กรณี (ดังแสดงด้วยตัวอักษรหนาในตารางที่ 6-3) ได้แก่ V0 กำกับผิดไปเป็น VN0, VN0 กำกับผิดไปเป็น V0, NCM กำกับผิดไปเป็น NCSF, NCSF กำกับผิดไปเป็น NCM, และ NCM กำกับผิดไปเป็น Q แต่ผู้วิจัยจะไม่นำกรณีสุดท้าย(คือกรณีที่ NCM กำกับผิดไปเป็น Q) มาศึกษาเนื่องจากเห็นว่า รูปคำที่กำกับหมวดคำผิดพลาดของกรณีนี้ (ได้แก่ 7, 28, 30, 3 และ 31) น่าจะถือว่าเป็นชนิดเดียวกันได้ ส่วนกรณีที่เหลืออีก 4 กรณีจะได้นำมาศึกษาวิเคราะห์อย่างละเอียด ซึ่งสามารถจัดเป็นกรณีความผิดพลาดในการกำกับหมวดคำของคู่หมวดคำ 2 คู่ คือ ความผิดพลาดระหว่าง NCM-NCSF, และความผิดพลาดระหว่าง V0-VN0 ดังนี้

กรณีของการกำกับหมวดคำผิดพลาดระหว่างนามสามัญกับลักษณนาม (NCM-NCSF) เกิดจากลักษณะปัญหาความกำกวมของคำหลายหน้าที่ (polyseme) ในภาษาไทยที่โดยส่วนใหญ่จะเป็นความกำกวมระหว่างคำเนื้อหา (content word) เช่น นาม, กริยา กับคำไวยากรณ์ (grammatical word หรือ function word) เช่น ลักษณนาม, บุพบท, กริยาช่วย, สันธาน เป็นต้น

กรณีนี้ก็เป็นความกำกวมระหว่างคำนามซึ่งเป็นคำเนื้อหากับคำลักษณนามซึ่งเป็นคำไวยากรณ์ที่มีรูปคำเหมือนกัน เช่น “ไว้ใจคน/NCMที่มาจากรัฐบาล” กับ “ตัวประกัน 21 คน/NCSF” เป็นต้น ความผิดพลาดในการกำกับหมวดคำของโปรแกรมแบบเบ็ดเสร็จนี้มีทั้ง กรณีที่โปรแกรมกำกับเป็นหมวดคำลักษณนามในคำที่ควรจะกำกับเป็นนามสามัญ ตัวอย่างเช่น

คือ/C- /PUNC-ชนชั้น/NCSF-สูง/VADJ- /PUNC-หรือ/C-ผู้/NCM-ชนะ/VN0 ✕

ฝัง/VN0-ราก/NCM-ลึก/VADJ-ตั้งแต่/PN-ครั้ง/NCSF-ที่/PCOMP-นคร/NCM-โฮจิมินห์/NPP ✕

อันดับ/NCSF-แรก/D- /PUNC-คณะกรรมการ/NCM-สำนักงานความสัมพันธ์แรงงานแห่งชาติ/NPP ✕

ใน/PN-บ่าย/NCM-วัน/NCSF-เดียวกัน/D-นั้น/PT ✕

ถูก/VAUX-ประกบ/VN0-ข้าง/NCSF-ด้วย/PN-คน/NCM-แปลกหน้า/VADJ ✕

และ กรณีที่โปรแกรมกำกับเป็นหมวดคำนามสามัญในคำที่ควรจะกำกับเป็นลักษณนาม ตัวอย่างเช่น

มี/VN0-สุภาษิต/NCM-บท/NCM-หนึ่ง/D ✕

ทั้ง/Q- /PUNC-2/Q- /PUNC-ฝ่าย/NCM-ได้/VAUX-เปิด/VN0-การ/PFX-เจรจา/V0 ✕

เก็บ/VN0-ค่าธรรมเนียม/NCM- /PUNC-อาทิ/C- /PUNC-30/Q- /PUNC-เซนต์/NCSF-ต่อ/PN-เรื่อง/NCM ✕

ตั้งแต่/PN-ช่วง/NCM-หลาย/Q-ทศวรรษ/NCM-ที่/PCOMP-ผ่าน/V0-มา/AV ✕

หลังจาก/PV-เกิด/VN0-อุบัติเหตุ/NCM…ประมาณ/Q- /PUNC-5/Q- /PUNC-สัปดาห์/NCM ✕

ปัญหาความกำกวมระหว่างนามสามัญกับลักษณนามนี้ นอกจากจะเป็นปัญหาเรื่องคำหลายหน้าที่แล้ว ยังมีสาเหตุมาจากการที่คำนามสามัญและคำลักษณนามมักจะปรากฏได้ในตำแหน่งที่เหมือนกันด้วย เช่น สามารถปรากฏหน้ากริยา, หลังกริยา, หลังบุพบทได้เหมือนกัน และสามารถปรากฏหน้าตัวกำหนดได้เหมือนกันอีกด้วย ดังนั้นวิทยานิพนธ์นี้จึงจัดให้คำนามสามัญ

และคำลักษณนามเป็นหมวดคำย่อยที่อยู่ในหมวดคำหลักเดียวกัน คือ หมวดคำนาม และใช้ตำแหน่งการปรากฏเหล่านี้เพื่อเป็นเกณฑ์ในการจัดแบ่งหมวดคำนามแยกจากหมวดคำหลักอื่นๆ (ดังที่กล่าวไว้อย่างละเอียดในเรื่องการกำหนดชุดหมวดคำในบทที่ 4) ตัวอย่างของการปรากฏในตำแหน่งที่เหมือนกันแสดงได้ดังตัวอย่างที่ 3 (ในที่นี้พิจารณาจากการปรากฏร่วมของส่วนหลัก)

ตัวอย่างที่ 3:

ก	คน/NCM-ไม่/AVNEG-ชอบ/VN0-เจ้า/NCM-แมว/NCM-ทอม/NPP	ทุก/Q-เรื่อง/NCSF-ดึงดูด/VN0-อารมณ์ขัน/NCM
ข	เป็น/VN0-เรื่อง/NCM-ยากลำบาก/VADJ	มี/VN0-อีก/Q-ประเด็น/NCSF-เล็กๆ/VADJ
ค	อยู่/VPN0-ใน/PN-กำมือ/NCM	สถาบัน/NCM-การศึกษา/NCM-ใน/PN-บาง/Q-ประเทศ/NCSF
ง	พิจารณา/VN0-คดี/NCM-นี้/D	กรรมการผู้จัดการ/NCM-คน/NCSF-นั้น/D

ตัวอย่างที่ 3-ก แสดงตัวอย่างของคำนามสามัญและคำลักษณนามที่ปรากฏหน้ากริยา (เป็นประธานของกริยา) ตัวอย่างที่ 3-ข แสดงตัวอย่างของคำนามสามัญและคำลักษณนามที่ปรากฏหลังกริยา (เป็นส่วนเติมเต็มหรือกรรมของกริยา) ตัวอย่างที่ 3-ค แสดงตัวอย่างของคำนามสามัญและคำลักษณนามที่ปรากฏหลังคำบุพบท (เป็นส่วนประกอบหลักส่วนหนึ่งของบุพบทวลี) และตัวอย่างที่ 3-ง แสดงตัวอย่างของคำนามสามัญและคำลักษณนามที่ปรากฏหน้าตัวกำหนด (เป็นส่วนหลักของตัวกำหนด)

การที่คำนามสามัญและคำลักษณนามมักปรากฏในตำแหน่งที่เหมือนกันดังที่กล่าวมานี้ทำให้เกิดปัญหาความกำกวมในการตัดสินระหว่างหมวดคำทั้งสอง เนื่องจากโปรแกรมไม่สามารถเรียนรู้ปริบทหมวดคำข้างเคียงที่ชัดเจนที่จะช่วยตัดสินแยกหมวดคำทั้งสองจากกันได้ นอกจากนี้เมื่อผู้วิจัยได้ศึกษาวิเคราะห์ตัวอย่างของข้อผิดพลาดในการกำกับหมวดคำระหว่างคำนามสามัญกับคำลักษณนาม พร้อมกับได้พิจารณาถึงลักษณะของตัวอย่างการใช้ภาษาจริงจากคลังข้อมูลแล้ว ยังพบด้วยว่า ความกำกวมอาจเกิดขึ้นจากลักษณะของการใช้ภาษาจริงที่มักปรากฏโครงสร้างของภาษาที่ไม่ใช่โครงสร้างพื้นฐานอยู่เป็นจำนวนมาก อันได้แก่ การที่คำนามมากกว่าหนึ่งคำสามารถปรากฏร่วมกันในโครงสร้างนามวลีได้ และการที่วิทยานิพนธ์นี้อนุญาตให้คำนามสามัญสามารถปรากฏหลังตัวบอกจำนวนได้เช่นเดียวกับคำลักษณนาม ซึ่งสามารถอธิบายได้ดังตัวอย่างที่ 4 และ 5 ตามลำดับ

ตัวอย่างที่ 4:

ก. มี/VN0-สุภาษิต/NCM-บท/NCSF-หนึ่ง/D

ข. ใน/PN-บ่าย/NCM-วัน/NCM-เดียวกัน/D

ตัวอย่างที่ 4-ก เป็นตัวอย่างของโครงสร้างที่เป็นพื้นฐานในภาษาไทย ซึ่งสามารถปรากฏนามวลีที่มีโครงสร้าง [NCM-NCSF-D] ได้ เช่น “แมวตัวนี้” (แต่ *“แมวแมวนี้” ไม่ได้ ดังได้กล่าวอภิปรายไปในหัวข้อ 4.2.2.1.1) ส่วนตัวอย่างที่ 4-ข เป็นตัวอย่างที่พบในคลังข้อมูลที่แสดงให้เห็นว่า อันที่จริงแล้ว ในการใช้ภาษาจริง ในตำแหน่ง [NCM-__-D] ก็สามารถปรากฏเป็น NCM ได้เช่นกัน ดังนั้นจึงทำให้ตำแหน่งดังกล่าวเกิดความกำกวมและส่งผลให้โปรแกรมแบบเบ็ดเสร็จกำกับหมวดคำให้กับข้อความในตัวอย่างที่ 4-ก และ 4-ข ผิดไปเป็น “มี/VN0-สุภาษิต/NCM-บท/NCM-หนึ่ง/D”✕ และ “ใน/PN-บ่าย/NCM-วัน/NCSF-เดียวกัน/D”✕ ตามลำดับ

การที่ในภาษาไทยสามารถปรากฏโครงสร้างที่มีรูปแบบ [NCM-NCM-D] เกิดจากลักษณะโครงสร้างที่ซับซ้อนในภาษาไทยที่คำนามสามัญมากกว่าหนึ่งคำสามารถปรากฏร่วมกันเป็นนามวลีได้ เช่น นักศึกษา-มหาวิทยาลัย, หัวหน้า-ภาควิชา-ภาษาไทย เป็นต้น ซึ่งลักษณะของภาษาเช่นนี้ปรากฏเป็นจำนวนมากในคลังข้อมูล ตัวอย่างเช่น รูปแบบ/NCM-กรมธรรม์/NCM-นี้/D, สถาบัน/NCM-การศึกษา/NCM-ดังกล่าว/D, ปลาย/NCM-ปี/NCM-นี้/D, จำนวน/NCM-หุ้น/NCM-ทั้งหมด/D, เบื้องหลัง/NCM-เรื่อง/NCM-นี้/D เป็นต้น

นอกจากลักษณะดังตัวอย่างที่ 4 แล้ว ในคลังข้อมูลยังปรากฏความกำกวมระหว่างนามสามัญกับลักษณนามในโครงสร้างที่เกิดร่วมกับตัวบอกจำนวนด้วย ดังตัวอย่างที่ 5

ตัวอย่างที่ 5:

ก. ช่วง/NCM-หลาย/Q-ทศวรรษ/NCSF

ข. 78/Q- /PUNC-ส.ว./NCM

ตัวอย่างที่ 5-ก เป็นตัวอย่างของโครงสร้างที่ลักษณนามปรากฏร่วมกับตัวบอกจำนวนซึ่งเป็นรูปแบบที่เป็นพื้นฐานและปรากฏใช้อยู่ทั่วไปในภาษาไทย เนื่องจากลักษณนามมีหน้าที่หลักในการ

บอกจำนวนนับของคำนาม เช่น “แมว 2 ตัว” (แต่ *“แมว 2 แมว” ไม่ได้) อย่างไรก็ตาม เมื่อพิจารณาจากคลังข้อมูลที่เป็นตัวอย่างการใช้ภาษาจริงแล้ว พบว่า รูปแบบดังตัวอย่างที่ 5-ข คือ โครงสร้างที่คำนามสามัญปรากฏร่วมกับตัวบอกจำนวน ก็มีปรากฏเป็นจำนวนมากเช่นกัน ซึ่งมีทั้งโครงสร้าง [Q-NCM] และ [NCM-Q-NCM] ตัวอย่างเช่น ถก/V0-3/Q-ผู้บริหาร/NCM-การบิน/NCM, จำนวน/NCM- /PUNC-9/Q- /PUNC- สายการบิน/NCM, เก้าอี้/NCM-สาม/Q-ขา/NCM, บาง/Q-ธุรกิจ/NCM เป็นต้น ดังนั้น วิทยานิพนธ์นี้จึงอนุญาตให้คำนามสามัญสามารถปรากฏหลังตัวบอกจำนวนได้เช่นเดียวกับลักษณนาม ซึ่งก็ทำให้เกิดความกำกวมในตำแหน่งดังกล่าวและส่งผลให้โปรแกรมแบบเบ็ดเสร็จกำกับหมวดคำในตัวอย่างที่ 4-ก ผิดไปเป็น “ช่วง/NCM-หลาย/Q-ทศวรรษ/NCM”✕

ส่วนกรณีของการกำกับหมวดคำผิดพลาดระหว่างกริยาที่ปรากฏลำพังกับกริยาที่ต้องมีนามตามหลัง (V0-VN0) ก็เกิดจากการที่คำกริยาที่มีรูปคำเหมือนกันสามารถปรากฏในโครงสร้างกริยาวลีที่แตกต่างกันได้ ในกรณีนี้การที่คำกริยารูปหนึ่งๆมีทั้งที่ปรากฏโดยลำพังและมีทั้งที่ปรากฏร่วมกับคำนามที่ตามหลัง เช่น "สิ่ง/NCM-ที่/PCOMP-เปลี่ยน/V0-ไป/AV” กับ “วาย & อาร์/NPP-เปลี่ยน/VN0-เงื่อนไข/NCM-บาง/Q-ประการ/NCSF” ทำให้เกิดความกำกวมในการกำกับหมวดคำระหว่างคำกริยาสองประเภทนี้ ซึ่งข้อผิดพลาดในการทดลองนี้มีทั้งกรณีที่โปรแกรมได้กำกับหมวดคำ VN0 ในตำแหน่งที่ควรกำกับเป็น V0 ตัวอย่างเช่น

สถานี/NCM-ตำรวจ/NCM- /PUNC-ที่/PCOMP-ตั้ง/VN0-อยู่/AV-อีก/Q-ฟาก/NCSF ✕

สิ่ง/NCM-ที่/PCOMP-เปลี่ยน/VN0-ไป/AV ✕

หา/VN0-เงิน/NCM-อย่าง/PV-ดี/VADJ-ที่สุด/AV-เท่าที่/PV-จะ/AV-ทำ/VN0-ได้/VAUX ✕

แต่/C-พวกเขา/NPRO-คิด/VN0-ผิด/VADJ ✕

ผู้/NCM-ชนะ/VN0- /PUNC-จะ/AV-เป็น/VN0-ผู้/NCM-คว้า/VN0-รางวัล/NCM ✕

ใน/PN-บริษัท/NCM-ก่อสร้าง/VN0-แห่ง/NCSF-หนึ่ง/D ✕

ใน/PN-การ/PFX-เจรจา/VN0-ขั้น/NCM-สุดท้าย/D ✕

และ กรณีที่โปรแกรมได้กำกับหมวดคำ V0 ในตำแหน่งที่ควรกำกับเป็น VN0 ตัวอย่างเช่น

คุ ณ /NCM-ไม่ /AVNEG-มี /VN0-ท า ง /NCM-พู ด /V0-อ ะ ไ ร /NPRO-ไ ด้ /VAUX-หรอก/PT ✕

ยัง/AV-ไม่/AVNEG-ยุติ/V0-การ/PFX-ปิดล้อม/VN0-ฐานที่มั่น/NCM ✕

ดึงดูดใจ/VN0-นักเขียน/NCM-หนังสือ/NCM-การ์ตูน/NCM-เข้า/V0-เวบไซต์/NCM ✕

ละเมิด/V0-สิทธิ/NCM-ส่วนตัว/VADJ ✕

คือ/C-สิ่ง/NCM-ที่/PCOMP-พรรค/NCM-คอมมิวนิสต์/NPP-กำลัง/AV-เรียนรู้/V0 ✕

การ์ตูน/NCM-ชุด/NCM- /PUNC-"/PUNC-อาร์ซี/NPP-"/PUNC- /PUNC-ที่/PCOMP-เอา/VN0-มา/AV-ทำ/V0-ใหม่/AV ✕

นำ/VN0-ผลงาน/NCM-ของ/PN-ตน/NPRO-มา/AV-แสดง/V0-ไว้/AV ✕

เมื่อผู้วิจัยได้ศึกษาตัวอย่างของข้อผิดพลาดที่เกิดขึ้น ร่วมกับพิจารณาถึงลักษณะของตัวอย่างการใช้ภาษาจริงจากคลังข้อมูลแล้ว พบว่า แม้ว่ากริยาทั้งสองประเภทจะปรากฏในโครงสร้างกริยาวลีที่แตกต่างกัน คือ V0 ปรากฏโดยลำพังเป็นกริยาวลีได้ ส่วน VN0 ต้องปรากฏร่วมกับคำนามหนึ่งคำตามหลังจึงจะได้กริยาวลีที่สมบูรณ์ แต่โครงสร้างกริยาวลีที่พบในข้อมูลการใช้ภาษาจริงนั้นมักจะไม่ใช่โครงสร้างพื้นฐานที่เรียบง่าย แต่มักประกอบไปด้วยส่วนขยายและโครงสร้างที่ซับซ้อนขึ้น อันได้แก่ การที่กริยา VN0 ไม่จำเป็นต้องปรากฏในตำแหน่งที่ติดกับคำนามที่ทำหน้าที่เป็นกรรมเสมอไป และคำนามที่ปรากฏหลังกริยาก็อาจไม่ใช่ส่วนจำเป็นของกริยา นอกจากนี้ การที่ในโครงสร้างต่างๆอาจมีการละคำนามที่เป็นกรรมไปได้ โดยที่ในกรณีนี้วิทยานิพนธ์ฉบับนี้ก็ถือว่ามีสรรพนามไร้รูปในตำแหน่งดังกล่าว ก็อาจเป็นสาเหตุที่ทำให้เกิดความกำกวมได้ ซึ่งสามารถอธิบายได้ดังตัวอย่างที่ 6 และ 7 ดังนี้

ตัวอย่างที่ 6:

ก. ละเมิด/VN0-สิทธิ/NCM-ส่วนตัว/VADJ

ข. เข้า/V0-มา/V0-ซื้อ/VN0-ถือ/VN0-หุ้น/NCM

ค. ผงาด/V0-ค่าย/NCM-โฆษณา/NCM-อันดับ/NCM-1/NCM

ง. ใน/PN-บริษัท/NCM-ก่อสร้าง/V0-แห่ง/NCSF-หนึ่ง/D

ตัวอย่างที่ 6-ก เป็นตัวอย่างของโครงสร้างปกติในภาษาไทย ซึ่ง VN0 จะมีคำนามปรากฏตามหลังในตำแหน่งที่ติดกัน และคำนามในที่นี้ทำหน้าที่เป็นกรรมของกริยา ส่วนตัวอย่างที่ 6-ข เป็นตัวอย่างของกรณีที่คำนามที่ทำหน้าที่เป็นกรรมของกริยาไม่จำเป็นต้องปรากฏติดกับ VN0 โดยที่ระหว่างกริยากับคำนามนั้นอาจมีคำอื่นๆปรากฏแทรกอยู่ก็ได้ ส่วนตัวอย่างที่ 6-ค และ 6-ง แสดงให้เห็นว่าแม้ว่าคำกริยาจะมีคำนามปรากฏตามหลังแต่กริยาตัวดังกล่าวก็อาจไม่ใช่ VN0 ก็ได้ และคำนามที่ตามหลังอาจไม่ได้ทำหน้าที่เป็นกรรมของกริยาก็ได้ จากตัวอย่างจะเห็นได้ว่า คำกริยา V0 ก็สามารถปรากฏข้างหน้าคำนามได้เช่นเดียวกัน นอกจากนี้ เมื่อพิจารณาจากตัวอย่างที่ 6-ง แล้วจะเห็นได้ว่า คำกริยา “ก่อสร้าง” ไม่ใช่ภาคแสดงของคำนาม “บริษัท” แต่ทำหน้าที่เป็นส่วนขยายและนามวลี “แห่งหนึ่ง” ที่ปรากฏตามหลังก็ไม่ได้มีความสัมพันธ์ใดๆกับคำกริยาด้วย แต่นามวลี “แห่งหนึ่ง” ทำหน้าที่เป็นส่วนขยายของ “บริษัทก่อสร้าง” อีกทีหนึ่ง

ดังที่อธิบายมาข้างต้น แสดงให้เห็นว่า ในตำแหน่ง [__-NCM] นั้น กริยา V0 และ VN0 ต่างก็สามารถปรากฏได้เช่นเดียวกัน และกริยา VN0 ก็ไม่ได้มีคำนามปรากฏตามหลังในตำแหน่งที่ติดกันเสมอไป ดังนั้น จึงทำให้เกิดความกำกวมและส่งผลให้โปรแกรมกำกับหมวดคำในตัวอย่างที่ 6-ก และ 6-ง ผิดไปเป็น “ละเมิด/V0-สิทธิ/NCM-ส่วนตัว/VADJ” ✕ และ “ใน/PN-บริษัท/NCM-ก่อสร้าง/VN0-แห่ง/NCSF-หนึ่ง/D” ✕ ตามลำดับ

นอกจากนั้นแล้ว ในการใช้ภาษาจริงยังปรากฏว่าในภาษาไทยมีโครงสร้างที่มีการละคำนามหรือย้ายที่คำนามได้ ดังตัวอย่างโครงสร้างคุณานุประโยคขยายคำนามในตัวอย่างที่ 7

ตัวอย่างที่ 7:

ก. สิ่ง/NCM-ที่/PCOMP-เปลี่ยน/V0-ไป/AV
ข. ลูกจ้าง/NCM-ที่/PCOMP-ใช้/VN0-อี-เมล์/NCM
ค. คำ/NCM-ที่/PCOMP-คน/NCM-ใช้/VN0-กัน/AV
ง. ศาล/NCM-ที่/PCOMP-จัดตั้ง/VN0-ขึ้น/AV

ตัวอย่างที่ 7-ก แสดงตัวอย่างของคุณานุประโยคที่มีกริยา V0 เป็นกริยาหลักภายในคุณานุประโยค (ในตัวอย่างนี้ ประธานของกริยาในคุณานุประโยคถูกละไปเนื่องจากตำแหน่งที่เป็นประธานอ้างถึงสิ่งเดียวกับคำนามที่เป็นส่วนหลัก (คือ “สิ่ง”)) ส่วนตัวอย่างที่ 7-ข ถึง 7-ง แสดงตัวอย่างของคุณานุ

ประโยคที่มีกริยา VN0 เป็นกริยาหลักภายในคุณานุประโยค ซึ่งแต่ละตัวอย่างมีความแตกต่างกันตรงตำแหน่งของคำนามที่ละไป ในตัวอย่างที่ 7-ข มีการละประธานไป แต่ยังคงปรากฏคำนามที่เป็นกรรม (คือ “อี-เมล์”) ดังนั้น กรณีนี้จึงตรงตามเกณฑ์ในการปรากฏของ VN0 ที่ต้องมีคำนามตามหลัง ในตัวอย่างที่ 7-ค ปรากฏคำนามที่เป็นประธาน แต่ได้ละคำนามที่เป็นกรรมไปเพราะตำแหน่งที่เป็นกรรมอ้างถึงสิ่งเดียวกับคำนามที่เป็นส่วนหลัก (คือ “คำ”) วิทยานิพนธ์นี้ถือว่า ในตำแหน่งที่เป็นกรรมของกรณีนี้มีสรรพนามไร้รูปปรากฏอยู่ ดังนี้ “คำ/NCM-ที่/PCOMP-คน/NCM-ใช้/VN0-∅-กัน/AV” ดังนั้นกริยา “ใช้” จึงยังคงถูกจัดให้เป็นกริยา VN0 อยู่เหมือนเดิม ส่วนตัวอย่างที่ 7-ง ได้ละทั้งประธานและกรรมไป โดยที่ตำแหน่งที่เป็นกรรมอ้างถึงสิ่งเดียวกับคำนามที่เป็นส่วนหลัก (คือ “ศาล”) ด้วยเหตุผลเดียวกับตัวอย่างที่ 7-ค คือ ถือว่าตำแหน่งที่เป็นกรรมมีสรรพนามไร้รูปปรากฏอยู่ ดังนั้น กริยา “จัดตั้ง” จึงถูกจัดให้เป็น VN0

เมื่อพิจารณาเปรียบเทียบตัวอย่างที่ 7-ก กับตัวอย่างที่ 7-ง จะเห็นได้ว่า ในปริบท [NCM-PCOMP-__] นั้น ทั้งกริยา V0 และกริยา VN0 สามารถปรากฏได้ในตำแหน่งดังกล่าวทั้งคู่ ดังนั้นจึงทำให้เกิดความกำกวมและส่งผลให้โปรแกรมแบบเบ็ดเสร็จกำกับหมวดคำในตัวอย่างที่ 7-ก ผิดไปเป็น “สิ่ง/NCM-ที่/PCOMP-เปลี่ยน/VN0-ไป/AV” ✕

จากการศึกษาผลการกำกับหมวดคำดังที่กล่าวมาทั้งหมด พอจะสรุปได้ว่า แบบจำลองไตรแกรมซึ่งได้นำหมวดคำข้างเคียงมาช่วยในการคำนวณค่าความน่าจะเป็นของสายคำและสายหมวดคำสามารถกำกับหมวดคำภาษาไทยได้ถูกต้องสูง เนื่องจากได้นำปริบทหมวดคำข้างเคียงมาช่วยตัดสินเลือกหมวดคำที่เหมาะสมกับปริบทได้เป็นจำนวนมาก แต่อย่างไรก็ดี ก็ยังมีข้อผิดพลาดในการกำกับหมวดคำอยู่จำนวนหนึ่ง จากการศึกษาข้อผิดพลาดโดยที่วิทยานิพนธ์นี้ได้เลือกศึกษากรณีความผิดพลาดระหว่าง NCM-NCSF และ V0-VN0 ซึ่งเป็นกรณีที่ความผิดพลาดมีความถี่สูง สามารถชี้ให้เห็นได้ว่า สาเหตุหนึ่งที่ทำให้การกำกับหมวดคำของโปรแกรมผิดพลาดไปก็เนื่องมาจากปัญหาความกำกวมของคำหลายหน้าที่ในภาษาไทยมีลักษณะปัญหาที่ซับซ้อน การใช้ปริบทหมวดคำข้างเคียงเพียงอย่างเดียวไม่สามารถตัดสินเลือกหมวดคำที่ถูกต้องได้เสมอไป เนื่องจากในข้อเขียนภาษาไทยโครงสร้างภาษาที่ปรากฏมักไม่ใช่โครงสร้างพื้นฐานที่มีความเรียบง่าย แต่มักเป็นโครงสร้างที่มีความซับซ้อนขึ้น ตัวอย่างเช่น มีส่วนขยายต่างๆปรากฏอยู่, มีการย้ายที่คำนามหรือมีการละคำนามไป เป็นต้น ซึ่งทำให้ปริบทไม่สามารถช่วยบ่งชี้หมวดคำที่ถูกต้องได้อย่างชัดเจน นอกจากนี้ วิธีการที่วิทยานิพนธ์ฉบับนี้ใช้ตัดสินความกำกวมในการกำกับหมวดคำ

เช่น การอนุญาตให้คำนามสามัญปรากฏหลังตัวบอกจำนวนได้, การอนุญาตให้มีสรรพนามไร้รูปได้ อาจจะไม่ได้ช่วยให้ตัดสินเลือกหมวดคำได้ดีขึ้น ดังนั้น ประเด็นความกำกวมระหว่างหมวดคำนี้ยังน่าจะมีการศึกษาหาวิธีการที่สามารถแก้ปัญหาความกำกวมเหล่านี้ต่อไปได้ และการแก้ปัญหาความกำกวมในกรณีย่อยๆแต่ละกรณีอาจต้องมีส่วนงานเฉพาะที่ทำหน้าที่แก้ปัญหาความกำกวมเฉพาะกรณีนั้นๆมารองรับเพิ่มเติมหลังกระบวนการกำกับหมวดคำ

6.3.2 **ผลการทดลองตัดคำ**

ผลการทดลองตัดคำให้กับข้อความทดสอบ ดังตารางที่ 6-4 แสดงให้เห็นว่า โปรแกรมตัดคำและกำกับหมวดคำแบบเบ็ดเสร็จสามารถตัดคำให้กับข้อความภาษาไทยได้อย่างมีประสิทธิภาพ โดยสามารถตัดคำได้ถูกต้อง 96.087 %

	Precision	Recall	F-measure
Segmentation	94.993 %	97.207 %	96.087 %

ตารางที่ 6-4 ประสิทธิภาพการตัดคำของโปรแกรมแบบเบ็ดเสร็จ

อย่างไรก็ตาม เมื่อนำประสิทธิภาพการตัดคำของโปรแกรมแบบเบ็ดเสร็จซึ่งใช้แบบจำลองไตรแกรมที่นำหมวดคำข้างเคียงมาช่วยในการตัดคำ ไปเปรียบเทียบกับประสิทธิภาพการตัดคำของโปรแกรมแบบที่ใช้แบบจำลองไตรแกรมที่นำรูปคำข้างเคียงมาช่วยในการตัดคำโดยไม่ได้นำหมวดคำข้างเคียงมาช่วยแล้ว พบว่า การตัดคำของแบบจำลองที่ใช้รูปคำข้างเคียงมีประสิทธิภาพในการตัดคำสูงกว่า โดยสามารถตัดคำได้ถูกต้องสูงถึง 99.527 % ดังแสดงในตารางที่ 6-5

	Precision	Recall	F-measure
Segmentation using trigram of word form	99.600 %	99.454 %	99.527 %

ตารางที่ 6-5 ประสิทธิภาพการตัดคำของโปรแกรมแบบที่ใช้รูปคำข้างเคียงโดยไม่ใช้หมวดคำข้างเคียง

จากผลการวัดประสิทธิภาพและการเปรียบเทียบประสิทธิภาพของการตัดคำ ชี้ให้เห็นว่า แบบจำลองไตรแกรมซึ่งใช้ปริบทหมวดคำข้างเคียงช่วยในการคำนวณค่าความน่าจะเป็นของสายคำและสายหมวดคำ (ซึ่งเป็นแบบจำลองที่ใช้ในโปรแกรมแบบเบ็ดเสร็จ) สามารถตัดคำได้ถูกต้องสูง ถึงกระนั้นก็ตาม ก็ยังไม่ใช่กระบวนการที่ดีที่สุดในการตัดคำภาษาไทย ซึ่งแย้งกับสมมติฐานที่ตั้งไว้ก่อนหน้านี้ เนื่องจากการใช้รูปคำของคำข้างเคียงมาช่วยในการตัดคำเป็นปริบทที่เหมาะสมมากกว่า มีประสิทธิภาพสามารถตัดคำได้ถูกต้องสูงกว่า

จากการเปรียบเทียบประสิทธิภาพของการตัดคำที่ได้ซึ่งพบว่าขัดแย้งกับสมมติฐานที่ตั้งไว้ ทำให้ผู้วิจัยเกิดข้อสงสัยว่า การที่แบบจำลองไตรแกรมที่นำหมวดคำข้างเคียงเข้ามาช่วยในการตัดคำมีประสิทธิภาพในการตัดคำด้อยกว่าแบบจำลองไตรแกรมที่ใช้รูปคำข้างเคียงช่วยในการตัดคำ มีสาเหตุหลัก คือ ขนาดของคลังข้อมูลฝึกสอนมีขนาดเล็กเกินไป จริงหรือไม่ ดังนั้น ผู้วิจัยจึงได้นำโปรแกรมทั้งสองแบบไปทดสอบกับคลังข้อมูลที่มีขนาดใหญ่ขึ้น โดยได้นำคลังข้อมูลออร์คิด (Virach Sornlertlamvanich et al., 1997) ซึ่งแตกต่างจากคลังข้อมูลที่ผู้วิจัยจัดทำขึ้นทั้งในด้านขนาดของคลังข้อมูล, ชนิดของเนื้อหา และหมวดคำที่ใช้กำกับ มาใช้เป็นคลังข้อมูลฝึกสอนและเป็นข้อความทดสอบ โดยให้โปรแกรมเรียนรู้ค่าสถิติจากคลังข้อมูลฝึกสอนขนาดประมาณ 265,000 คำ และทำการทดสอบโดยใช้ข้อความทดสอบขนาดประมาณ 40,000 คำ จากนั้นประเมินผลประสิทธิภาพในการตัดคำและกำกับหมวดคำด้วยวิธีเดียวกัน คือ ทำการวัดค่า F-measure โดยเทียบผลลัพธ์จากโปรแกรมกับคำตอบที่ได้นำมาจากข้อความในคลังข้อมูลออร์คิดเองซึ่งมีการตัดคำและกำกับหมวดคำไว้แล้ว และจะถือว่ามีค่าความถูกต้อง 100 % ได้ผลการทดลองตัดคำและกำกับหมวดคำให้กับคลังข้อมูลออร์คิดดังแสดงในตารางที่ 6-6

	Precision	Recall	F-measure
POS Tagging	88.779 %	86.881 %	87.820 %
POS Tagging using simple probability	87.330 %	84.901 %	86.098 %
Segmentation	94.791 %	92.765 %	93.7673 %
Segmentation using trigram of word form	96.458 %	94.813 %	95.629 %

ตารางที่ 6-6 สรุปประสิทธิภาพการตัดคำและกำกับหมวดคำของโปรแกรมแบบเบ็ดเสร็จและประสิทธิภาพของการตัดคำของโปรแกรมแบบที่ไม่ใช้หมวดคำช่วยในการตัดคำเมื่อใช้กับคลังข้อมูลออร์คิด

จากผลการทดลองในตารางที่ 6-6 แสดงให้เห็นว่า เมื่อทดสอบกับคลังข้อมูลที่มีขนาดใหญ่ขึ้น ผลการเปรียบเทียบประสิทธิภาพในการตัดคำของแบบจำลองทั้ง 2 แบบนี้ยังออกมาในลักษณะเดิมอยู่ (แม้ว่าตัวเลขแสดงค่าความถูกต้องจะเปลี่ยนไป) กล่าวคือ แบบจำลองไตรแกรมที่ใช้หมวดคำข้างเคียงมาช่วยในการตัดคำยังคงตัดคำได้ถูกต้องน้อยกว่าแบบจำลองไตรแกรมที่ใช้รูปคำข้างเคียงมาช่วยในการตัดคำ ซึ่งแสดงว่า การที่คลังข้อมูลมีขนาดเล็กไม่ใช่ปัจจัยหลักที่ทำให้แบบจำลองที่ใช้หมวดคำข้างเคียงมีประสิทธิภาพด้อยกว่าแบบจำลองที่ใช้รูปคำข้างเคียง ผู้วิจัยคาดว่า ความแตกต่างของตัวแบบจำลองทั้งสองซึ่งใช้ปริบทแตกต่างกันในการแก้ปัญหาการตัดคำ (หมวดคำข้างเคียง กับ รูปคำข้างเคียง) ก็น่าจะเป็นปัจจัยที่มีผลต่อความแตกต่างของประสิทธิภาพการตัดคำระหว่างแบบจำลองทั้งสองด้วย อย่างไรก็ตาม ประเด็นนี้ยังไม่อาจสรุปได้อย่างแน่ชัดว่าสิ่งใดเป็นปัจจัยหลักที่ส่งผลให้ประสิทธิภาพแตกต่างกัน ดังนั้นจึงน่าจะมีการศึกษาในประเด็นนี้ต่อไป

จากการศึกษาผลการตัดคำของโปรแกรมทั้งสองแบบ ผู้วิจัยพบว่า ความผิดพลาดในการตัดคำของโปรแกรมทั้งสองแบบโดยส่วนใหญ่เป็นกรณีความผิดพลาดที่เกิดขึ้นกับคำประกอบ กล่าวคือ ในการแก้ปัญหาการตัดคำให้กับคำประกอบ จะเกิดความกำกวมในการตัดสินใจเลือกว่าจะตัดสายอักขระออกเป็นคำหนึ่งคำหรือตัดเป็นคำหลายคำเพราะมีความเป็นไปได้ในการตัดคำทั้งสองแบบ ซึ่งข้อผิดพลาดที่เกิดขึ้นมีทั้งที่เกิดจาก โปรแกรมตัดคำผิดเพราะได้ผลลัพธ์เป็นคำหลายคำในตำแหน่งที่ควรเป็นคำประกอบหนึ่งคำ ตัวอย่างเช่น ข้อความ “บริษัทอยู่ในฐานะหุ้นส่วน” ซึ่ง

คำตอบที่เหมาะสมกับปริบทควรเป็นคำประกอบหนึ่งคำ คือ “หุ้นส่วน” แต่โปรแกรมกลับตัดแยกคำเป็น “หุ้น” และ “ส่วน” (รายการคำที่ตัดคำผิดในลักษณะนี้ทั้งหมดดูได้ในตารางที่ ค-1 ในภาคผนวก ค) หรือในทางตรงกันข้าม โปรแกรมตัดคำผิดเพราะได้ผลลัพธ์เป็นคำประกอบหนึ่งคำในตำแหน่งที่ควรเป็นคำหลายคำ ตัวอย่างเช่น ข้อความ ”คนที่-ทำงานกับรัฐบาล” ซึ่งคำตอบที่เหมาะสมกับปริบทควรเป็น “ที่” และ “ทำงาน” แต่โปรแกรมกลับตัดเป็นคำประกอบหนึ่งคำ คือ “ที่ทำงาน” เป็นต้น (รายการคำที่ตัดคำผิดในลักษณะนี้ทั้งหมดดูได้ในตารางที่ ค-2 ในภาคผนวก ค) ดังนั้นจึงเห็นได้ว่า ลักษณะความกำกวมที่เกิดจากคำประกอบเป็นสาเหตุใหญ่ที่ทำให้เกิดข้อผิดพลาดในการตัดคำภาษาไทย

เมื่อพิจารณาจากความถี่ของข้อผิดพลาด (รายการคำที่ตัดคำผิดแสดงไว้ในภาคผนวก ค) พบว่า ข้อผิดพลาดที่เกิดจากการตัดแยกออกเป็นคำหลายคำในปริบทที่ควรเป็นคำหนึ่งคำมีจำนวนทั้งหมด 120 แห่ง(64 รูปคำ) ซึ่งมากกว่าข้อผิดพลาดที่เกิดจากการตัดเป็นคำหนึ่งคำในปริบทที่ควรเป็นคำหลายคำซึ่งมีจำนวนทั้งหมด 12 แห่ง(10 รูปคำ) ดังนั้น แสดงให้เห็นว่าการพยายามที่จะตัดแยกสายอักขระออกเป็นคำย่อยๆหลายคำจะก่อให้เกิดข้อผิดพลาดมากกว่าการตัดสินให้เป็นคำประกอบหนึ่งคำ ซึ่งอาจสะท้อนลักษณะของภาษาไทยได้ว่า ในภาษาไทยหากคำย่อยๆสามารถประกอบเข้าด้วยกันเป็นคำหนึ่งคำได้ ก็มักจะประกอบเข้าด้วยกันเป็นคำหนึ่งคำ นอกจากนี้ เมื่อพิจารณาถึงแนวโน้มของการตัดคำของโปรแกรมทั้ง 2 แบบจากข้อผิดพลาด พบว่าโปรแกรมแบบเบ็ดเสร็จที่ใช้แบบจำลองไตรแกรมที่ได้นำหมวดคำข้างเคียงมาใช้ในการคำนวณมักตัดคำประกอบแยกออกเป็นคำหลายคำ (ซึ่งสามารถดูได้จากตารางที่ ค-1 ว่าส่วนใหญ่เป็นข้อผิดพลาดของโปรแกรมแบบเบ็ดเสร็จ) ส่วนโปรแกรมที่ใช้แบบจำลองไตรแกรมที่ได้นำรูปคำข้างเคียงมาใช้ในการคำนวณมักตัดคำประกอบเป็นหนึ่งคำ (ซึ่งสามารถดูได้จากตารางที่ ค-2 ว่าส่วนใหญ่เป็นข้อผิดพลาดของโปรแกรมแบบที่ใช้รูปคำข้างเคียงมาช่วย)

โดยสรุปแล้ว ข้อผิดพลาดที่เกิดจากการตัดแยกออกเป็นคำหลายคำเป็นข้อผิดพลาดที่มีความถี่สูงกว่า และข้อผิดพลาดในลักษณะนี้ก็มักเป็นข้อผิดพลาดของโปรแกรมแบบเบ็ดเสร็จที่อาศัยหมวดคำข้างเคียงช่วยในการคำนวณ ดังนั้น จึงส่งผลให้โปรแกรมแบบเบ็ดเสร็จมีประสิทธิภาพความถูกต้องในการตัดคำด้อยกว่าโปรแกรมแบบที่ใช้รูปคำข้างเคียงช่วยในการคำนวณ สาเหตุที่โปรแกรมทั้งสองมักได้ผลลัพธ์การตัดคำที่แตกต่างกันเป็นเพียงเพราะลักษณะของตัวแบบจำลองที่ใช้นั้นมีความแตกต่างกันเท่านั้น ซึ่งสามารถแสดงให้เห็นได้ดังตัวอย่างที่ 8

ซึ่งเป็นตัวอย่างที่แบบจำลองไตรแกรมที่ใช้หมวดคำข้างเคียงตัดคำผิดแต่แบบจำลองไตรแกรมที่ใช้รูปคำข้างเคียงสามารถตัดคำได้ถูกต้อง ดังนี้

ตัวอย่างที่ 8:

ก. เอา/VN0-เงิน/NCM-เข้า/VN0-กระเป๋า/NCM-ตัวเอง/NPRO-ทุก/Q-ที/NCSF

ข. เอา/VN0-เงิน/NCM-เข้า/VN0-กระเป๋า/NCM-ตัว/NCSF-เอง/D-ทุก/Q-ที/NCSF ✕

ค. เอา-เงิน-เข้า-กระเป๋า-ตัวเอง-ทุก-ที

ตัวอย่างที่ 8-ก คือการตัดคำและกำกับหมวดคำในคำตอบซึ่งถือว่าถูกต้อง ตัวอย่างที่ 8-ข คือผลลัพธ์การตัดคำและกำกับหมวดคำของโปรแกรมแบบเบ็ดเสร็จซึ่งใช้แบบจำลองไตรแกรมที่นำหมวดคำข้างเคียงมาช่วยคำนวณความน่าจะเป็น ตัวอย่างที่ 8-ค คือผลลัพธ์การตัดคำของโปรแกรมแบบที่ใช้แบบจำลองไตรแกรมที่นำรูปคำของคำข้างเคียงมาช่วยโดยไม่ได้นำหมวดคำข้างเคียงมาช่วย

ในตัวอย่างที่ 8 โปรแกรมแบบเบ็ดเสร็จตัดคำได้เป็น “ตัว/NCSF-เอง/D” ซึ่งไม่ถูกต้อง สาเหตุที่ได้ผลลัพธ์การตัดคำเช่นนี้เป็นเพราะว่าค่าความน่าจะเป็นในการปรากฏของรูปคำของทั้ง “ตัว/NCSF” และ “เอง/D” มีค่าสูงกว่าค่าความน่าจะเป็นในการปรากฏของรูปคำ “ตัวเอง/NPRO” อยู่มาก และลำดับไตรแกรมของหมวดคำที่เกี่ยวข้องของการตัดคำเป็น “ตัว/NCSF-เอง/D” (ได้แก่ VN0-NCM-NCSF, NCM-NCSF-D, NCSF-D-Q และ D-Q-NCSF) ก็มีค่าความน่าจะเป็นสูงกว่าลำดับไตรแกรมของหมวดคำที่เกี่ยวข้องของการตัดคำเป็น “ตัวเอง/NPRO” (ได้แก่ VN0-NCM-NPRO, NCM-NPRO-Q, และ NPRO-Q-NCSF) ดังนั้น จึงทำให้โปรแกรมแบบเบ็ดเสร็จตัดคำได้เป็น “ตัว/NCSF-เอง/D” ซึ่งไม่ถูกต้อง ส่วนตัวอย่างที่ 8-ค แบบจำลองไตรแกรมที่อาศัยรูปคำข้างเคียงในการคำนวณโดยไม่ได้พิจารณาหมวดคำข้างเคียงสามารถเลือกการตัดคำที่ถูกต้องตรงกับคำตอบได้ เนื่องจากว่าการเลือกตัดคำเป็น “ตัวเอง” จะทำให้ลำดับไตรแกรมของรูปคำที่เกี่ยวข้อง (ได้แก่ เข้า-กระเป๋า-ตัวเอง, กระเป๋า-ตัวเอง-ทุก, ตัวเอง-ทุก-ที) มีค่าความน่าจะเป็นสูงกว่าลำดับไตรแกรมของรูปคำที่เกี่ยวข้องหากตัดคำเป็น “ตัว-เอง” (ได้แก่ เข้า-กระเป๋า-ตัว, กระเป๋า-ตัว-เอง, ตัว-เอง-ทุก, เอง-ทุก-ที) สาเหตุที่เป็นเช่นนี้เพราะว่า รูปคำ “ตัว” และ “เอง” ไม่เคยปรากฏต่อเนื่องกันเลยในคลังข้อมูลฝึกสอน (แม้ว่าแต่ละคำจะมีค่าความถี่ในการปรากฏที่สูงก็ตาม) จึงส่งผลให้ลำดับรูปคำ “ตัว-เอง” มีค่าความถี่ต่ำ ดังนั้น การที่ “ตัว” และ “เอง” มีความถี่ในการ

ปรากฏสูงกว่า “ตัวเอง” จึงไม่สามารถใช้เป็นปริบทที่จะทำให้การตัดคำเป็น “ตัว-เอง” มีค่าความน่าจะเป็นที่สูงขึ้นมาได้

จากคำอธิบายในตัวอย่างที่ 8 ข้างต้น จะเห็นได้ว่าความผิดพลาดในการตัดคำของโปรแกรมแบบเบ็ดเสร็จเกิดจากลักษณะการคำนวณค่าความน่าจะเป็นของแบบจำลองนี้และปริบทที่ใช้ในแบบจำลองนี้มีผลต่อค่าความน่าจะเป็นทำให้รูปแบบการตัดคำที่ถูกต้องมีค่าความน่าจะเป็นต่ำกว่ารูปแบบการตัดคำที่ผิด นอกจากนี้ การที่ “ตัวเอง/NPRO” มีค่าความถี่ต่ำมาก (ซึ่งก็เป็นปัจจัยหนึ่งที่ทำให้ค่าความถี่ของการตัดคำที่ถูกต้องต่ำ) ก็เพราะว่าคำนี้ปรากฏในคลังข้อมูลฝึกสอนน้อยครั้งหรือไม่เคยปรากฏเลย ซึ่งก็เป็นผลมาจากการที่คลังข้อมูลฝึกสอนมีขนาดเล็กจึงไม่สามารถครอบคลุมการปรากฏของรูปคำต่างๆได้หลากหลาย ส่วนแบบจำลองไตรแกรมที่ใช้รูปคำข้างเคียงได้ผลลัพธ์การตัดคำต่างไปเนื่องจากมีลักษณะการคำนวณค่าความน่าจะเป็นที่ต่างออกไปและได้เลือกใช้ปริบทที่ต่างออกไปเท่านั้น แบบจำลองไตรแกรมที่ใช้รูปคำข้างเคียงนี้ก็อาจตัดคำผิดในตัวอย่างอื่น ดังตัวอย่างที่ 9 ข้างล่างนี้

ตัวอย่างที่ 9:

ก. ชุด/NCM-เจรจา/V0-ของ/PN-ทางการ/NCM-ฟิลิปปินส์/NPP

ข. ชุด/NCM-เจรจา/VPN0-ของ/PN-ทางการ/NCM-ฟิลิปปินส์/NPP

ค. ชุด-เจรจา-ของ-ทาง-การ-ฟิลิปปินส์ ✕

ตัวอย่างที่ 9-ก คือคำตอบการตัดคำและกำกับหมวดคำในคำตอบซึ่งถือว่าถูกต้อง ตัวอย่างที่ 9-ข คือผลลัพธ์การตัดคำและกำกับหมวดคำของโปรแกรมแบบเบ็ดเสร็จ ซึ่งใช้แบบจำลองไตรแกรมที่นำหมวดคำข้างเคียงมาช่วยคำนวณความน่าจะเป็น ตัวอย่างที่ 9-ค คือผลลัพธ์การตัดคำของโปรแกรมแบบที่ใช้แบบจำลองไตรแกรมที่นำรูปคำของคำข้างเคียงมาช่วยโดยไม่ได้นำหมวดคำข้างเคียงมาช่วย

ในตัวอย่างที่ 9-ค แบบจำลองไตรแกรมที่นำรูปคำข้างเคียงมาช่วยตัดคำผิดไปเป็น “ทาง-การ” สาเหตุที่ได้ผลลัพธ์การตัดคำเป็นเช่นนี้เป็นเพราะว่าค่าความน่าจะเป็นในการปรากฏของรูปคำ “ทางการ” มีค่าต่ำกว่าค่าความน่าจะเป็นในการปรากฏของรูปคำของทั้ง “ทาง” และ “การ” อยู่มาก ประกอบกับมีลำดับรูปคำ “ของ-ทาง”, “ทาง-การ” ปรากฏต่อเนื่องกันในคลังข้อมูลฝึกสอน ซึ่งส่งผลให้ลำดับไตรแกรมของรูปคำที่เกี่ยวข้อง (ได้แก่ เจรจา-ของ-ทาง, ของ-ทาง-การ, ทาง-การ-ฟิลิปปินส์) มีค่าความน่าจะเป็นสูงกว่าลำดับไตรแกรมของรูปคำที่เกี่ยวข้องหากตัดคำเป็น

“ทางการ” (ได้แก่ เจรจา-ของ-ทางการ, ของ-ทางการ-ฟิลิปปินส์) ดังนั้น แบบจำลองที่อาศัยรูปคำข้างเคียงในการคำนวณค่าความน่าจะเป็นจึงตัดคำผิดในตัวอย่างนี้ ส่วนโปรแกรมแบบเบ็ดเสร็จในตัวอย่างที่ 9-ข ซึ่งมีลักษณะการคำนวณค่าความน่าจะเป็นและใช้ปริบทที่แตกต่างกัน คือ ใช้รูปคำข้างเคียงมาช่วยในการคำนวณ สามารถตัดคำในตัวอย่างนี้ได้ถูกต้อง ปัจจัยที่ทำให้ผลลัพธ์การตัดคำแตกต่างกันก็เหมือนกับที่ได้อธิบายไปในตัวอย่างที่ 8 คือ ลักษณะการคำนวณค่าความน่าจะเป็นและปริบทที่แบบจำลองแต่ละแบบเลือกใช้แตกต่างกัน

จากการศึกษาผลการตัดคำดังที่กล่าวมาทั้งหมด พอจะสรุปได้ว่า ข้อผิดพลาดในการตัดคำที่พบในวิทยานิพนธ์ฉบับนี้ส่วนใหญ่เกิดจากความกำกวมของคำประกอบ ซึ่งมีทั้งกรณีที่ตัดคำผิดเป็นคำหลายคำในปริบทที่ควรตัดเป็นคำประกอบหนึ่งคำ และกรณีที่ตัดคำผิดเป็นคำประกอบหนึ่งคำในปริบทที่ควรตัดเป็นคำหลายคำ โดยที่โปรแกรมแบบเบ็ดเสร็จที่ใช้หมวดคำข้างเคียงมาช่วยในการตัดคำมีแนวโน้มที่จะตัดคำแยกเป็นหลายคำมากกว่าที่จะตัดเป็นคำหนึ่งคำ ส่วนโปรแกรมที่ใช้รูปคำข้างเคียงมีแนวโน้มที่จะตัดคำเป็นหนึ่งคำมากกว่าที่จะตัดแยกเป็นหลายคำ ซึ่งเมื่อพิจารณาจากความถี่ของข้อผิดพลาด พบว่า ข้อผิดพลาดที่เกิดจากการตัดแยกคำเป็นหลายคำในปริบทที่ควรตัดเป็นคำหนึ่งคำเป็นข้อผิดพลาดที่ปรากฏมากกว่า ดังนั้น โปรแกรมแบบเบ็ดเสร็จซึ่งมักตัดแยกคำออกเป็นคำหลายคำจึงมีประสิทธิภาพด้อยกว่าโปรแกรมแบบที่ใช้รูปคำข้างเคียงซึ่งมักตัดคำเป็นคำประกอบคำเดียว นอกจากนี้ ข้อผิดพลาดดังกล่าวยังอาจสะท้อนลักษณะของภาษาไทยด้วยว่า หากคำย่อยๆหลายคำสามารถรวมกันเป็นคำประกอบหนึ่งคำได้ ก็มักจะรวมกันเป็นคำประกอบ ข้อผิดพลาดในการตัดคำของโปรแกรมแต่ละแบบแตกต่างกัน มีสาเหตุมาจากลักษณะในการคำนวณค่าความน่าจะเป็นของแบบจำลองที่ใช้และปริบทที่ใช้ในแบบจำลองแต่ละแบบแตกต่างกันเท่านั้น ดังนั้น หากมีการปรับแบบจำลองที่ใช้ในการตัดคำให้เหมาะสมก็อาจช่วยลดข้อผิดพลาดในการตัดคำลงได้

6.3.3 สรุปผลการทดลองตัดคำและกำกับหมวดคำ

ในหัวข้อที่ 6.3.1 และ 6.3.2 ได้นำเสนอและอภิปรายผลการกำกับหมวดคำและผลการตัดคำของโปรแกรมแบบเบ็ดเสร็จตามลำดับ โดยนำเสนอประสิทธิภาพของโปรแกรมเป็นค่าร้อยละ F-measure ของความถูกต้องในการกำกับหมวดคำและตัดคำ และได้อภิปรายถึงสาเหตุของข้อผิดพลาดในการตัดคำและกำกับหมวดคำของโปรแกรมแบบเบ็ดเสร็จ โดยยกตัวอย่างของข้อผิดพลาดที่พบในการทดสอบประกอบ

ซึ่งจากการศึกษา พบว่า โปรแกรมแบบเบ็ดเสร็จสามารถกำกับหมวดคำและตัดคำได้ถูกต้องเพราะได้นำปริบทหมวดคำข้างเคียงมาช่วยในการคำนวณค่าความน่าจะเป็นของสายคำและสายหมวดคำ ซึ่งสามารถช่วยตัดสินเลือกคำและหมวดคำที่เหมาะสมกับปริบทได้ดีพอสมควร แต่เมื่อพิจารณาจากข้อผิดพลาด พบว่า ในการกำกับหมวดคำข้อผิดพลาดส่วนหนึ่งเกิดจากลักษณะความกำกวมของคำหลายหน้าที่ในภาษาไทย ซึ่งการใช้ปริบทไม่สามารถช่วยแก้ปัญหาความกำกวมดังกล่าวได้อย่างถูกต้องเสมอไป เพราะโครงสร้างของภาษาไทยในการใช้ภาษาจริงมักจะมีลักษณะที่ซับซ้อน และส่วนหนึ่งเกิดจากวิธีการที่วิทยานิพนธ์นี้ใช้ตัดสินปัญหาความกำกวมในการกำกับหมวดคำด้วย ส่วนในการตัดคำ ข้อผิดพลาดเกิดจากลักษณะของตัวแบบจำลองเองซึ่งมีลักษณะการคำนวณค่าความน่าจะเป็นและการใช้ประโยชน์จากปริบทเพื่อแก้ปัญหาการตัดคำที่ยังไม่เหมาะสม

ตารางที่ 6-7 สรุปประสิทธิภาพของโปรแกรมแบบเบ็ดเสร็จในการทดลองตัดคำและกำกับหมวดคำให้กับข้อความทดสอบ ค่าความถูกต้องจากตารางแสดงให้เห็นว่า โปรแกรมแบบเบ็ดเสร็จสามารถตัดคำและกำกับหมวดคำได้อย่างมีประสิทธิภาพสูง และเมื่อพิจารณาเปรียบเทียบกัน จะเห็นได้ว่า การกำกับหมวดคำมีประสิทธิภาพต่ำกว่าการตัดคำอยู่พอสมควร ซึ่งเป็นเพราะว่าการกำกับหมวดคำที่ถูกต้องจะต้องประกอบไปด้วย การตัดคำที่ถูกต้องและการกำกับหมวดคำที่ถูกต้อง จึงมีหลายกรณีที่สามารถตัดคำถูกแต่กำกับหมวดคำผิด ในขณะที่ หากกรณีดังกล่าวกำกับหมวดคำถูกต้องก็แสดงว่าตัดคำได้ถูกต้องด้วย

	Precision	Recall	F-measure
POS Tagging	88.570 %	90.634 %	89.590 %
Segmentation	94.993 %	97.207 %	96.087 %

ตารางที่ 6-7 สรุปประสิทธิภาพการตัดคำและกำกับหมวดคำของโปรแกรมแบบเบ็ดเสร็จ

จากผลการทดลองที่ได้นำเสนอและอภิปรายมาทั้งหมด สามารถสรุปได้ว่า แบบจำลองไตรแกรมที่ใช้หมวดคำข้างเคียงในการคำนวณความน่าจะเป็นของสายคำและสายหมวดคำเป็นกระบวนการที่เหมาะสมในการกำกับหมวดคำภาษาไทยวิธีการหนึ่ง ซึ่งมีความถูกต้องในการกำกับหมวดคำสูงถึง 89.590 % แต่ก็ยังไม่ใช่กระบวนการที่ดีที่สุดสำหรับการตัดคำภาษาไทย แม้จะมี

ความถูกต้องในการตัดคำสูงถึง 96.087 % เนื่องจากแบบจำลองไตรแกรมที่ใช้รูปคำข้างเคียงในการคำนวณค่าความน่าจะเป็นของสายคำมีความถูกต้องในการตัดคำสูงกว่า ดังนั้นเพื่อให้ได้ประสิทธิภาพในการตัดคำและกำกับหมวดคำสูงที่สุด จึงอาจจะต้องแยกกระบวนการตัดคำและกำกับหมวดคำออกจากกัน โดยที่กระบวนการตัดคำเป็นกระบวนการขั้นต้น สมควรใช้แบบจำลองไตรแกรมที่นำรูปคำข้างเคียงมาช่วยในการตัดคำ หรือใช้ปริบททั้ง 2 อย่าง คือ รูปคำข้างเคียงและหมวดคำข้างเคียง ร่วมกันในการตัดคำ (ดังจะกล่าวในข้อเสนอแนะในบทที่ 7) จากนั้นจึงนำข้อความที่ตัดคำแล้วไปกำกับหมวดคำโดยใช้แบบจำลองไตรแกรมที่นำหมวดคำข้างเคียงมาช่วยในการกำกับหมวดคำ

บทที่ 7

สรุป และข้อเสนอแนะ

บทนี้จะได้กล่าวสรุปผลการทำงานของโปรแกรมแบบเบ็ดเสร็จตรวจสอบกับสมมติฐานที่ตั้งไว้ก่อนหน้าเพื่อดูว่า การศึกษาในวิทยานิพนธ์ฉบับนี้ได้ผลลัพธ์เป็นไปตามที่คาดไว้หรือไม่อย่างไร และจากการศึกษาครั้งนี้ สามารถเห็นแนวทางที่จะนำไปพัฒนาปรับปรุงอย่างไรบ้าง โดยในหัวข้อที่ 7.1 จะกล่าวถึงกระบวนการและผลการศึกษาทั้งหมดที่ได้นำเสนอไปในวิทยานิพนธ์ฉบับนี้โดยสรุป แล้วจึงเสนอแนะประเด็นต่างๆที่น่าจะสามารถนำไปพัฒนาต่อเพื่อเพิ่มประสิทธิภาพในการตัดคำและกำกับหมวดคำภาษาไทยหัวข้อที่ 7.2 ดังนี้

7.1 สรุปกระบวนการในการพัฒนาโปรแกรมและผลการทำงานของโปรแกรมแบบเบ็ดเสร็จ

วิทยานิพนธ์นี้ได้พัฒนาโปรแกรมสำหรับตัดคำและกำกับหมวดคำแบบเบ็ดเสร็จขึ้นมา โดยได้ประยุกต์แบบจำลองไตรแกรมมาใช้เพื่อแก้ปัญหาการตัดคำและการกำกับหมวดคำให้กับข้อความภาษาไทย แบบจำลองไตรแกรมที่ใช้สามารถนำปริบทหมวดคำของคำข้างเคียงมาช่วยคำนวณค่าความน่าจะเป็นของสายคำและสายหมวดคำเพื่อตัดสินเลือกสายคำและสายหมวดคำที่มีค่าความน่าจะเป็นสูงสุดเป็นผลลัพธ์สำหรับทั้งการตัดคำและการกำกับหมวดคำไปด้วยกัน โดยวิทยานิพนธ์นี้มองว่าปัญหาการตัดคำและการกำกับหมวดคำเป็นส่วนงานเดียวกันที่สามารถแก้ปัญหาไปพร้อมๆกันได้

เนื่องจากโปรแกรมแบบเบ็ดเสร็จซึ่งใช้แบบจำลองไตรแกรมต้องอาศัยค่าสถิติสำหรับคำนวณค่าความน่าจะเป็นเพื่อแก้ปัญหาการตัดคำและกำกับหมวดคำ โดยแบบจำลองไตรแกรมจะเรียนรู้ค่าความถี่จากคลังข้อมูลภาษาไทยที่มีการตัดคำและกำกับหมวดคำไว้แล้ว ดังนั้น ผู้วิจัยจึงได้จัดทำคลังข้อมูลภาษาไทยขึ้นมา โดยคัดเลือกและรวบรวมข้อมูลที่เป็นภาษาไทยที่ใช้กันทั่วไปในชีวิตประจำวันจากคลังข้อมูลของหนังสือพิมพ์กรุงเทพธุรกิจ และได้ทำการตัดคำและกำกับหมวดคำด้วยมือเพื่อใช้เป็นคลังข้อมูลฝึกสอน ผู้วิจัยมีกระบวนการขั้นตอนต่างๆในการจัดเตรียมคลังข้อมูลดังที่กล่าวไปในบทที่ 3 เพื่อให้คลังข้อมูลมีการตัดคำและกำกับหมวดคำที่มีความคงที่

และสม่ำเสมอตามเกณฑ์มากที่สุด นอกจากนี้ เพื่อให้สามารถทำการตัดคำและกำกับหมวดคำในคลังข้อมูลได้อย่างมีประสิทธิภาพ ผู้วิจัยจึงได้ศึกษาประเด็นเรื่องคำและได้นำเสนอเกณฑ์การตัดสินคำสำหรับนำมาใช้ตัดคำด้วยมือให้กับคลังข้อมูล ซึ่งอธิบายไว้โดยละเอียดในบทที่ 4 พร้อมกันนี้ ผู้วิจัยได้ศึกษาและคัดเลือกการจัดแบ่งหมวดคำภาษาไทยที่มีผู้เสนอไว้เพื่อนำมาทดสอบกับข้อมูลการใช้ภาษาจริง แล้วจึงได้พัฒนาและจัดแบ่งหมวดคำย่อยเพิ่มเติมสำหรับใช้เป็นชุดหมวดคำภาษาไทยของวิทยานิพนธ์ฉบับนี้ โดยวิทยานิพนธ์นี้ใช้เกณฑ์ทางวากยสัมพันธ์เป็นเกณฑ์หลักในการจัดแบ่งหมวดคำ ได้แก่ เกณฑ์การกระจายของคำ และ เกณฑ์การปรากฏร่วมของคำ และได้ผลลัพธ์เป็นชุดหมวดคำภาษาไทยที่ประกอบไปด้วย 9 หมวดคำหลัก ได้แก่ นาม, กริยา, ตัวกำหนด, ตัวบอกจำนวน, วิเศษณ์, คำหน้าหน่วยสร้างไร้ศูนย์, สันธาน, อนุภาค และเครื่องหมาย ซึ่งสามารถแบ่งหมวดคำย่อยให้เหมาะสมสำหรับการกำกับหมวดคำให้กับคลังข้อมูลและใช้เป็นป้ายหมวดคำสำหรับโปรแกรม ได้ผลลัพธ์เป็นป้ายหมวดคำภาษาไทยทั้งหมด 26 ป้ายหมวดคำ ดังที่ได้นำเสนอไว้โดยละเอียดในบทที่ 4 เช่นกัน ซึ่งหมวดคำที่ได้มานี้นับว่ามีความเหมาะสมในระดับหนึ่งเนื่องจากได้มาจากการวิเคราะห์ข้อมูลภาษาจริงและได้ทดลองใช้กำกับในคลังข้อมูลภาษา จากนั้น ผู้วิจัยจึงได้ทำการสร้างโปรแกรมสำหรับตัดคำและกำกับหมวดคำภาษาไทยแบบเบ็ดเสร็จตามแนวคิดของแบบจำลองไตรแกรมพร้อมทั้งนำขั้นตอนวิธีวิเทอร์บิเข้ามาใช้เพื่อช่วยเพิ่มประสิทธิภาพของโปรแกรม

ในการทดสอบประสิทธิภาพของโปรแกรมแบบเบ็ดเสร็จที่พัฒนาขึ้นมา ผู้วิจัยได้นำโปรแกรมแบบเบ็ดเสร็จไปทดลองทำการตัดคำและกำกับหมวดคำให้กับข้อมูลชุดทดสอบที่จัดเตรียมไว้ แล้วทำการประเมินผลประสิทธิภาพความถูกต้องของการตัดคำและการกำกับหมวดคำ ผลการทดลองปรากฏว่า โปรแกรมแบบเบ็ดเสร็จมีประสิทธิภาพสูงในการกำกับหมวดคำและตัดคำ แม้ว่าจะให้โปรแกรมเรียนรู้ค่าสถิติจากคลังข้อมูลที่มีขนาดเล็กก็ตาม โดยสามารถกำกับหมวดคำได้ถูกต้อง 89.590 % และสามารถตัดคำได้ถูกต้อง 96.087 % ดังที่ได้นำเสนอในบทที่ 6 แต่เมื่อเปรียบเทียบผลการตัดคำของโปรแกรมแบบเบ็ดเสร็จกับผลการตัดคำของโปรแกรมแบบที่ใช้แบบจำลองไตรแกรมที่นำรูปคำข้างเคียงมาช่วยคำนวณค่าความน่าจะเป็นของสายคำโดยไม่ได้นำหมวดคำข้างเคียงเข้ามาช่วย พบว่า โปรแกรมแบบเบ็ดเสร็จที่ใช้หมวดคำข้างเคียงมาช่วยในการตัดคำมีประสิทธิภาพในการตัดคำด้อยกว่าโปรแกรมที่ใช้รูปคำข้างเคียงมาช่วยในการตัดคำ ซึ่งแสดงให้เห็นว่า แบบจำลองไตรแกรมที่ใช้หมวดคำข้างเคียงในการคำนวณค่าความน่าจะเป็นของสายคำและสายหมวดคำยังไม่ใช่กระบวนการที่ดีที่สุดในการตัดคำภาษาไทย เนื่องจากแบบจำลอง

ไตรแกรมที่ใช้รูปคำข้างเคียงในการคำนวณความน่าจะเป็นของสายคำเป็นกระบวนการที่ตัดคำได้ถูกต้องสูงกว่า

ดังนั้น หากจะใช้แบบจำลองไตรแกรมเพื่อแก้ปัญหาการตัดคำและกำกับหมวดคำภาษาไทย การแยกกระบวนการตัดคำและกระบวนการกำกับหมวดคำออกเป็น 2 ส่วนงานจึงน่าจะเหมาะสมกว่า โดยกระบวนการตัดคำถือเป็นกระบวนการขั้นต้น ในกระบวนการตัดคำนี้สมควรใช้แบบจำลองไตรแกรมที่ใช้รูปคำข้างเคียงในการคำนวณค่าความน่าจะเป็นของสายคำแล้วจึงนำข้อความที่ตัดคำแล้วไปทำการกำกับหมวดคำ ในกระบวนการกำกับหมวดคำนี้สามารถใช้แบบจำลองไตรแกรมที่ใช้หมวดคำข้างเคียงในการคำนวณค่าความน่าจะเป็นของสายหมวดคำได้เนื่องจากเป็นกระบวนการที่เหมาะสมในการกำกับหมวดคำภาษาไทยกระบวนการหนึ่ง หรือหากจะทำการตัดคำและกำกับหมวดคำแบบเบ็ดเสร็จโดยใช้แบบจำลองไตรแกรมก็น่าจะมีการปรับวิธีการที่ใช้คำนวณค่าความน่าจะเป็นของแบบจำลองให้มีประสิทธิภาพที่สูงขึ้น (ดังจะได้กล่าวในข้อเสนอแนะ)

7.2 ข้อเสนอแนะในการศึกษาพัฒนาเพิ่มเติม

1. การนำโปรแกรมแบบเบ็ดเสร็จไปใช้ตัดคำและกำกับหมวดคำภาษาไทยให้กับข้อความอื่นๆให้ได้ประสิทธิภาพมากยิ่งขึ้น ควรใช้คลังข้อมูลฝึกสอนที่มีขนาดใหญ่ขึ้นเพื่อให้โปรแกรมได้เรียนรู้ค่าสถิติสำหรับใช้คำนวณค่าความน่าจะเป็นของสายคำและสายหมวดคำที่น่าเชื่อถือมากยิ่งขึ้น การเพิ่มขนาดของคลังข้อมูลฝึกสอนควรเพิ่มทั้งในเชิงปริมาณและในเชิงคุณภาพ กล่าวคือ คลังข้อมูลควรมีค่าความถี่ในการปรากฏของคำและค่าความถี่ของลำดับหมวดคำเพิ่มมากขึ้นตามสัดส่วน พร้อมทั้งคลังข้อมูลควรครอบคลุมถึงลักษณะความกำกวมแบบต่างๆในการตัดคำและกำกับหมวดคำได้มากยิ่งขึ้น อันได้แก่ มีรูปคำใหม่ๆและลำดับหมวดคำใหม่ๆปรากฏในคลังข้อมูลด้วย ซึ่งจะทำให้คลังข้อมูลสามารถสะท้อนลักษณะของความกำกวมในภาษาไทยได้ดียิ่งขึ้น

2. ผู้วิจัยเห็นว่า แบบจำลองไตรแกรมมีข้อจำกัดในการเรียนรู้ปริบทที่จะนำมาใช้แก้ปัญหาความกำกวม กล่าวคือ แบบจำลองไตรแกรมสามารถเรียนรู้เพียงลำดับหมวดคำที่ปรากฏเรียงกันอยู่ในขอบเขตจำกัดเท่านั้น ซึ่งแตกต่างจากมนุษย์ตรงที่แบบจำลองไตรแกรมไม่สามารถเรียนรู้ความสัมพันธ์ระหว่างคำต่างๆในเชิงโครงสร้างของภาษาได้ คือ ไม่สามารถเรียนรู้ว่าคำใด

ประกอบกับคำใดขึ้นเป็นหน่วยสร้างและหน่วยสร้างต่างๆมีลำดับในการเกิดก่อน-หลังอย่างไร ซึ่งส่งผลให้แบบจำลองไตรแกรมไม่สามารถแก้ปัญหาความกำกวมของโครงสร้างที่มีความซับซ้อนในภาษาไทยได้ เช่น โครงสร้างที่ส่วนหลักปรากฏอยู่ห่างจากกันเนื่องจากมีการเติมส่วนขยายแทรกอยู่ระหว่างส่วนหลักนั้น, ประโยคที่มีการละคำหรือมีการย้ายที่ส่วนประกอบภายในประโยคเพื่อเน้นความ ฯลฯ เพื่อลดข้อจำกัดดังกล่าว จึงน่าจะศึกษาหาวิธีนำปริบทในระดับวลีมาช่วยเสริมปริบทในระดับคำที่แบบจำลองไตรแกรมสามารถเรียนรู้ได้ หรือพัฒนาให้แบบจำลองสามารถเรียนรู้ข้อมูลในระดับวลีได้

3. จากการศึกษาผลการตัดคำเปรียบเทียบกัน เห็นได้ว่า สำหรับแบบจำลองไตรแกรม การนำรูปคำข้างเคียงมาใช้สามารถช่วยแก้ปัญหาความกำกวมในการตัดคำได้ถูกต้องมากกว่าการนำหมวดคำข้างเคียงมาใช้ จึงน่าจะหาวิธีใช้ประโยชน์จากรูปคำข้างเคียงเพื่อพัฒนาโปรแกรมที่ทำการแก้ปัญหาแบบเบ็ดเสร็จต่อไปได้ ดังนั้น จากแบบจำลองไตรแกรมอันเดิมซึ่งทำการหาสายคำและสายหมวดคำไปพร้อมๆกัน อาจจะเปลี่ยนมุมมองในการคำนวณค่าความน่าจะเป็นได้ คือ จาก PROB($W_{1,n}$, $T_{1,n}$) แทนที่จะแก้ปัญหาโดยแปลงเป็นแบบจำลองที่ทำการคำนวณ PROB($T_{1,n}$) * PROB($W_{1,n}$ | $T_{1,n}$) เหมือนที่ได้นำเสนอไปในวิทยานิพนธ์ฉบับนี้ อาจจะมองกลับกันได้โดยแปลงเป็นแบบจำลองใหม่ที่ทำการคำนวณ PROB($W_{1,n}$) * PROB($T_{1,n}$ | $W_{1,n}$) แทน เพื่อที่จะได้ใช้ประโยชน์จากไตรแกรมของรูปคำมาทำการคำนวณ PROB(W_i | W_{i-1} W_{i-2}) ได้ อย่างไรก็ดี ควรมีการประมาณค่า PROB($T_{1,n}$| $W_{1,n}$) ให้เหมาะสมเพื่อไม่ให้มีการให้น้ำหนักแก่หมวดคำที่เกิดบ่อยที่สุดสำหรับรูปคำหนึ่งๆมากเกินไป เพราะหากคำนวณโดยเพียงประมาณค่าว่าเท่ากับผลคูณรวมของ PROB(T_i | W_i) น้ำหนักการตัดสินเลือกผลลัพธ์การตัดคำและกำกับหมวดคำโดยส่วนใหญ่ก็น่าจะขึ้นอยู่กับว่าแต่ละรูปคำเกิดเป็นหมวดคำใดบ่อยที่สุด นอกจากนี้ ผู้วิจัยเห็นว่าอาจสามารถนำปริบททั้งสองแบบ คือ ปริบทหมวดคำข้างเคียง และปริบทรูปคำข้างเคียง มาใช้ร่วมกันเพื่อเพิ่มประสิทธิภาพของโปรแกรมที่ทำการแก้ปัญหาแบบเบ็ดเสร็จให้สูงขึ้นไปได้ เนื่องจากผลการทดลองตัดคำแสดงให้เห็นว่า ปริบททั้งสองแบบสามารถใช้แก้ไขข้อบกพร่องซึ่งกันและกันได้ โดยที่น่าจะสามารถคำนวณค่าความน่าจะเป็นจากไตรแกรมของหมวดคำและไตรแกรมของรูปคำไปด้วยกันได้

4. พัฒนาคลังข้อมูลภาษาไทยให้มีความถูกต้องในการตัดคำและกำกับหมวดคำมากยิ่งขึ้นเพื่อใช้เป็นฐานความรู้ที่มีประสิทธิภาพสำหรับแบบจำลองไตรแกรม ดังนั้น จึงน่าจะมีการศึกษามโนทัศน์เรื่องคำเพิ่มเติมเพื่อค้นหาเกณฑ์การตัดคำที่เหมาะสมสำหรับภาษาไทยที่จะนำมาใช้

ในการตัดคำให้กับคลังข้อมูลได้ โดยเฉพาะอย่างยิ่งประเด็นเรื่องคำประสมซึ่งเป็นประเด็นที่ทำให้เกิดปัญหาต่อการตัดคำเป็นอย่างมาก นอกจากนี้ ควรมีการศึกษาและพัฒนาชุดหมวดคำภาษาไทยที่จะนำมาใช้ในงานประมวลผลภาษาธรรมชาติต่อไปในแง่มุมใหม่ๆ หรือพัฒนาและแก้ปัญหาของชุดหมวดคำที่มีอยู่เดิมให้ดียิ่งขึ้น ผู้วิจัยขอเสนอว่า น่าจะมีการจัดทำชุดหมวดคำภาษาไทยแบบอัตโนมัติโดยใช้หลักของการ clustering โดยจับรวมคำต่างๆที่มีการกระจาย (distribution) คล้ายคลึงกันเข้าไว้เป็นหมวดคำเดียวกันโดยอาศัยค่าสถิติ ซึ่งสามารถศึกษาแนวทางได้จากงานวิจัยที่มีผู้ทำไว้ในภาษาอังกฤษ เช่น Schutze (1995) ผลการจัดแบ่งหมวดคำโดยอาศัยค่าสถิตินี้สามารถนำมาใช้ประกอบการวิเคราะห์เรื่องหมวดคำ เพื่อสนับสนุนการจัดหมวดคำโดยมนุษย์ได้

รายการอ้างอิง

ภาษาไทย

กำชัย ทองหล่อ. 2515. หลักภาษาไทย. กรุงเทพมหานคร: รวมสาส์น.

จรัสดาว อินทรทัศน์. 2539. การที่คำกริยากลายเป็นคำบุพบทในภาษาไทย. วิทยานิพนธ์ปริญญาเอก สาขาวิชาภาษาศาสตร์. จุฬาลงกรณ์มหาวิทยาลัย. อ้างถึงใน อมรา ประสิทธิ์รัฐสินธุ์. 2543. ชนิดของคำในภาษาไทย: การวิเคราะห์ทางวากยสัมพันธ์โดยอาศัยฐานข้อมูลภาษาไทยปัจจุบันสองล้านคำ. รายงานการวิจัย เสนอ สำนักงานกองทุนสนับสนุนการวิจัย (ทุนวิจัยองค์ความรู้ใหม่ที่เป็นพื้นฐานต่อการพัฒนา). (เอกสารไม่ตีพิมพ์)

นววรรณ พันธุเมธา. 2527. ไวยากรณ์ไทย. กรุงเทพมหานคร: รุ่งเรืองสาส์น.

บรรจบ พันธุเมธา. 2514. ลักษณะภาษาไทย. กรุงเทพมหานคร: มหาวิทยาลัยรามคำแหง.

บุญเสริม กิจศิริกุล. 2541. การกำกับหมวดคำสำหรับข้อความภาษาไทย. กรุงเทพมหานคร: สถาบันวิจัยและพัฒนาคณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย.

ปราณี กุลละวณิชย์, กัลยา ติงศภัทิย์, สุดาพร ลักษณียนาวิน และอมรา ประสิทธิ์รัฐสินธุ์. 2535. ภาษาทัศนา. พิมพ์ครั้งที่ 2. กรุงเทพมหานคร: วัฒนชัยการพิมพ์.

พิสิทธิ์ พรมจันทร์. 2540. การวิเคราะห์แนวทางการเปรียบเทียบสมรรถนะของโปรแกรมแยกคำภาษาไทย. วิทยานิพนธ์ปริญญามหาบัณฑิต ภาควิชาวิศวกรรมคอมพิวเตอร์ บัณฑิตวิทยาลัย จุฬาลงกรณ์มหาวิทยาลัย.

ไพศาล เจริญพรสวัสดิ์. 2541. การตัดคำภาษาไทยโดยใช้คุณลักษณะ. วิทยานิพนธ์ปริญญามหาบัณฑิต ภาควิชาวิศวกรรมคอมพิวเตอร์ บัณฑิตวิทยาลัย จุฬาลงกรณ์มหาวิทยาลัย.

รัตติกร วรากูลศิริพันธ์, จงกล งามวิวิทย์, สมศักดิ์ จันวัน, สุธาทิพย์ จิวัธยากูล และศักดิ์ชัย ทิพย์จักษุรัตน์. 2538ก. การตัดคำจากประโยคภาษาไทยด้วยวิธีการเทียบคำที่ยาวที่สุด. Papers on Natuaral Language Processing, Compiled by Virach Sornlertlamvanich.

รัตติกร วรากูลศิริพันธ์, วราภรณ์ สุชัยชิต, สมศักดิ์ จันวัน และศักดิ์ชัย ทิพย์จักษุรัตน์. 2538ข. การวิเคราะห์เลือกประโยคที่ถูกต้องจากความถี่ของการใช้คำ. Papers on Natuaral Language Processing, Compiled by Virach Sornlertlamvanich.

ราชบัณฑิตยสถาน. 2525. พจนานุกรมฉบับราชบัณฑิตยสถาน พ.ศ. 2525. กรุงเทพมหานคร: อักษรเจริญทัศน์.

ราชบัณฑิตยสถาน. 2530. หลักเกณฑ์การใช้เครื่องหมายวรรคตอน และเครื่องหมายอื่นๆ. เอกสารเผยแพร่ชุดที่ 4 เนื่องในวันสถาปนาราชบัณฑิตยสถาน.

เรืองเดช ปันเขื่อนขัติย์. 2541. ภาษาศาสตร์ภาษาไทย. นครปฐม: มหาวิทยาลัยมหิดล.

วิจินตน์ ภาณุพงศ์. 2532. โครงสร้างของภาษาไทย: ระบบไวยากรณ์. พิมพ์ครั้งที่ 10. กรุงเทพมหานคร: มหาวิทยาลัยรามคำแหง.

วิรัช ศรเลิศล้ำวาณิช. 2536. การตัดคำในระบบแปลภาษา (Word Segmentation for Thai in Machine Translation System). การแปลภาษาด้วยคอมพิวเตอร์. NECTEC. หน้า 50-55.

สมชาย ลำดวน. 2526. ไวยากรณ์ไทย. กรุงเทพมหานคร: โอเดียนสโตร์.

สมปรารถนา รัทยานนท์. 2535. โครงสร้างข้อมูลสำหรับพจนานุกรมอิเล็กทรอนิกส์ภาษาไทย. วิทยานิพนธ์ปริญญามหาบัณฑิต ภาควิชาวิศวกรรมคอมพิวเตอร์ บัณฑิตวิทยาลัย จุฬาลงกรณ์มหาวิทยาลัย.

สุโขทัยธรรมาธิราช. 2533. ภาษาไทย 3. พิมพ์ครั้งที่ 4. กรุงเทพมหานคร: มหาวิทยาลัยสุโขทัยธรรมาธิราช.

อมรา ประสิทธิ์รัฐสินธุ์. 2543. ชนิดของคำในภาษาไทย: การวิเคราะห์ทางวากยสัมพันธ์โดยอาศัยฐานข้อมูลภาษาไทยปัจจุบันสองล้านคำ. รายงานการวิจัย เสนอ สำนักงานกองทุนสนับสนุนการวิจัย(ทุนวิจัยองค์ความรู้ใหม่ที่เป็นพื้นฐานต่อการพัฒนา). (เอกสารไม่ตีพิมพ์)

อมรา ประสิทธิ์รัฐสินธุ์, ยุพาพรรณ หุ่นจำลอง และสรัญญา เศวตมาลย์. 2544. ทฤษฎีไวยากรณ์. กรุงเทพมหานคร: สำนักพิมพ์จุฬาลงกรณ์มหาวิทยาลัย.

อุดม วโรตม์สิกขดิตถ์. 2535. ความรู้เบื้องต้นเกี่ยวกับภาษา. พิมพ์ครั้งที่ 3. กรุงเทพมหานคร: มหาวิทยาลัยรามคำแหง.

อุดม วโรตม์สิกขดิตถ์. 2538. ไวยากรณ์ไทยในภาษาศาสตร์. กรุงเทพมหานคร: ชวนพิมพ์ อ้างถึงใน อมรา ประสิทธิ์รัฐสินธุ์. 2543. ชนิดของคำในภาษาไทย: การวิเคราะห์ทางวากยสัมพันธ์โดยอาศัยฐานข้อมูลภาษาไทยปัจจุบันสองล้านคำ. รายงานการวิจัย เสนอ สำนักงานกองทุนสนับสนุนการวิจัย(ทุนวิจัยองค์ความรู้ใหม่ที่เป็นพื้นฐานต่อการพัฒนา). (เอกสารไม่ตีพิมพ์)

อุปกิตศิลปสาร, พระยา. 2514. หลักภาษาไทย. กรุงเทพมหานคร: ไทยวัฒนาพานิช.

ภาษาอังกฤษ

Allen, J. 1995. Natural language understanding. 2nd ed. California: Benjamin/Cummings.

Asanee Kawtrakul et. al. 1995. A lexicon model for writing production assistance system, in Proceedings of the Symposium on Natural Language Processing on Thailand '95. cited in Surapant Meknavin and Boonserm Kijsirikul. 2000. Thai grapheme-to-phoneme conversion, In D.Burnham, S.Luksaneeyanawin, C.Davis and M.Lafourcade (eds.) Interdisciplinary approaches to language processing. Bangkok: NECTEC: pp. 214-223.

Bloomfield, Leonard. 1933. Language. New York: Henry Holt and Company.

Brill, Eric. 1993. Automatic grammar induction and parsing free text: a transformation-based approach. Proceedings of ACL-93.

Brown, E.K. and J.E. Miller. 1980. Syntax: A linguistic introduction to sentence structure. London: Hutchison.

Chaniak, A., C. Hendrickson, N. Jacobson and M. Perkowitz. 1993. Equations for part-of-speech tagging. Proceedings of the 11th National Conference on Artificial Intelligence. อ้างถึงใน บุญเสริม กิจศิริกุล. 2541. การกำกับหมวดคำสำหรับข้อความภาษาไทย. กรุงเทพมหานคร: สถาบันวิจัยและพัฒนาคณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย.

Chao, Yuen Ren. 1970. A grammar of spoken Chinese. Berkeley: Univerity of California Press. อ้างถึงใน อมรา ประสิทธิ์รัฐสินธุ์. 2543. ชนิดของคำในภาษาไทย: การวิเคราะห์ทางวากยสัมพันธ์โดยอาศัยฐานข้อมูลภาษาไทยปัจจุบันสองล้านคำ. รายงานการวิจัยเสนอ สำนักงานกองทุนสนับสนุนการวิจัย(ทุนวิจัยองค์ความรู้ใหม่ที่เป็นพื้นฐานต่อการพัฒนา). (เอกสารไม่ตีพิมพ์)

Church, W. 1988. A Stochastic Parts Program and Noun Phrase Parser for Unrestricted Text. Second Conference on Applied Natural Language Processing. ACL: pp. 136-143. อ้างถึงใน บุญเสริม กิจศิริกุล. 2541. การกำกับหมวดคำสำหรับข้อความภาษาไทย. กรุงเทพมหานคร: สถาบันวิจัยและพัฒนาคณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย.

Crystal, David. 1971. Linguistics. Penguin Books: Harmondsworth.

Golding, Andrew R. 1995. A bayesian hybrid method for context-sensitive spelling correction. Proceedings of the Third Workshop on Very Large Corpora. อ้างถึงใน บุญเสริม กิจศิริกุล. 2541. การกำกับหมวดคำสำหรับข้อความภาษาไทย. กรุงเทพมหานคร: สถาบันวิจัยและพัฒนาคณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย.

Kramsky, Jiri. 1969. The word as a linguistic unit. Paris: Mouton.

Lehmann, Wilfred P. 1983. Language: An introduction. New York: Random House.

Manning, Christopher D. and H. Schutze. 1999. Foundations of statistical natural language processing. Cambridge: MIT Press.

Matthews, P.H. 1981. Syntax. Cambridge: Cambridge University Press.

Miller, George A. 1991. The science of word. New York: Scientific American Library.

Nida, Eugene A. 1949. Morphology: The descriptive analysis of words. 2nd ed. Ann Arbor: University of Michigan Press.

Pike, Kenneth L. 1967. Language in relation to the unified theory of the human behavior. Paris: Mouton.

Pike, Kenneth L. and E.G. Pike. 1977. Grammatical analysis. Dallas: The Summer Institute of Linguistics and the University of Texas at Arlington.

Robins, R.H. 1964. General linguistics: An introductory survey. London: Longman.

Ross, John R. 1973. Nouniness, In Osama Fujimura (ed.), Three dimensions in Linguistic theory. Tokyo & TEC Corporation. pp. 137-258. อ้างถึงใน สุโขทัยธรรมาธิราช. 2533. ภาษาไทย 3. พิมพ์ครั้งที่ 4. กรุงเทพมหานคร: มหาวิทยาลัยสุโขทัยธรรมาธิราช. หน้า 39-40.

Schutze, Hinrich. 1995. Distributional Part-of-Speech Tagging[Online]. Available from: ftp://csli.stanford.edu/pub/prosit/DisPosTag.ps.

Starosta, Stanley.1988. The case for lexicase. London: Pinter. อ้างถึงใน อมรา ประสิทธิ์รัฐสินธุ์. 2543. ชนิดของคำในภาษาไทย: การวิเคราะห์ทางวากยสัมพันธ์โดยอาศัยฐานข้อมูลภาษาไทยปัจจุบันสองล้านคำ. รายงานการวิจัย เสนอ สำนักงานกองทุนสนับสนุนการวิจัย (ทุนวิจัยองค์ความรู้ใหม่ที่เป็นพื้นฐานต่อการพัฒนา). (เอกสารไม่ตีพิมพ์)

Surapant Meknavin and Boonserm Kijsirikul. 2000. Thai grapheme-to-phoneme conversion, In D.Burnham, S.Luksaneeyanawin, C.Davis and M.Lafourcade

(eds.) Interdisciplinary approaches to language processing. Bangkok: NECTEC: pp. 214-223.

Surapant Meknavin. 1995. Towards 99.99% accuracy of Thai word segmentation. Oral Presentation at the Symposium on Natural Language Processing in Thailand '95. cited in Surapant Meknavin and Boonserm Kijsirikul. 2000. Thai grapheme-to-phoneme conversion, In D.Burnham, S.Luksaneeyanawin, C.Davis and M.Lafourcade (eds.) Interdisciplinary approaches to language processing. Bangkok: NECTEC: pp. 214-223.

Surapant Meknavin, Paisarn Charoenpornsawat, Boonserm Kijsirikul. 1997. Feature-based Thai word segmentation. Proceedings of the Natuaral Language Processing Pacific Rim Symposium (NLPRS) 1997.

Van Rijsbergen, C.J. 1979. Information retrieval. London: Butterworths. cited in Manning, Christopher D. and Hinrich Schutze. 1999. Foundations of statistical natural language processing. Cambridge: MIT Press.

Virach Sornlertlamvanich, T. Charoenporn, and H. Isahara. 1997. ORCHID: Thai part-of-speech tagged corpus. In Technical report Orchid corpus. Bangkok: NECTEC: pp. 5-19.

Wirote Aroonmanakun. 1999. Extending focusing for zero pronoun resolution in Thai. Doctoral dissertation, Graduate School of Arts and Science, Georgetown University.

Witten, Ian H. and Timothy C. Bell. 1991. The zero-frequency problem: Estimating the probabilities of novel events in adaptive text compression. IEEE Transactions on Information Theory 37: pp. 1085-1094. cited in Jurafsky, Daniel and J.H. Martin. 2000. Speech and language processing: An introduction to natural language processing, speech recognition, and computational linguistics. New Jersey: Prentice-Hall.

ภาคผนวก

ภาคผนวก ก

รายการคำศัพท์ในไฟล์พจนานุกรมจัดเรียงตามหมวดคำ

AV

กร้าว, กลับ, กลางอากาศ, กว่า, กัน, กำลัง, ก็, ก็ดี, ก็ตาม, ก็มี, ก็ได้, ก่อน, ก่อนหน้า, ก้มหน้าก้มตา, ขนานใหญ่, ขึ้น, ขึ้นๆ, คง, คลุ้ง, คว้าง, ค่อนข้าง, ค่อยๆ, จริง, จริงๆ, จะ, จะๆ, จึง, จู่ๆ, ชัก, ซ้ำ, ดังนี้, ดู, ดูเหมือน, ดูเหมือนว่า, ด้วย, ตลอด, ตลอดไป, ตาม, ตามใจชอบ, ตามๆกัน, ต่อ, ต่อมา, ต่อไป, ต่อไปนี้, ต่าง, ต่างหาก, ถึงกับ, ถึงกับว่า, ถ้วนหน้า, ทันที, ทั่วไป, ทั้ง, ทั้งนั้น, ทั้งสิ้น, ทั้งหมด, ทำไม, ทีเดียว, ที่สุด, ที่หนึ่ง, ทุกเมื่อ, บางที, บ้าง, ประจำ, พอสมควร, มัก, มั่ง, มา, มีแต่, ยัง, ยังคง, ยิ่ง, ย่อม, ร่วม, ร่อมร่อ, ลง, ลับหลัง, ล่วงหน้า, ล่าสุด, ล้วน, วันดีคืนดี, วันๆ, สุด, สุดท้าย, สุดๆ, หน่อย, หน้าสลอน, หมด, หมาดๆ, อยู่, อย่า, อย่างยิ่ง, อย่างไร, ออก, อาจ, อีก, ฮึ่มๆ, เกิน, เกือบ, เข้า, เฉยๆ, เช่นกัน, เช่นเดียวกัน, เช่นไร, เดี๋ยวนี้, เต็มทน, เท่านั้น, เท่าใด, เท่าไหร่, เป็นกระตั๊ก, เป็นครั้งคราว, เป็นต้นมา, เป็นต้นไป, เป็นระยะๆ, เป็นลำดับ, เป็นหลัก, เป็นแถว, เพิ่ง, เพียง, เพียงแต่, เพียงไร, เรื่อย, เรื่อยๆ, เลย, เสมอๆ, เหมือนกัน, เอง, เอา, เอาไว้, เอ้เต้, แก่ใจ, แค่, แค่ไหน, แต่, แต่อย่างใด, แต่เพียง, แต่แรก, แทน, แทบ, แท้จริง, แน่, แน่นอน, แม้, แม้แต่, แล้ว, โดยรวม, โดยเฉพาะ, โดยเฉพาะอย่างยิ่ง, โดยแท้, ในที่สุด, ใหม่, ให้, ได้แต่, ไป, ไปหมด, ไม่นานนี้, ไว้

AVNEG

มิ, เปล่า, ไม่

C

กระนั้น, กล่าวคือ, กับ, ก็, ก็คือ, คือ, ค่อย, จน, จนกระทั่ง, จนถึง, ซ้ำร้าย, ดังนั้น, ดังเช่น, ตรงกันข้าม, ถึงอย่างไร, ทั้ง, ทั้งนี้, นอกจากนั้น, นอกจากนี้, มิฉะนั้น, มิฉะนั้นแล้ว, ยิ่งกว่านั้น, ยิ่งไปกว่านั้น, รวมทั้ง, ส่วน, หรือ, หรือว่า, หรือไม่ก็, อนึ่ง, อย่างเช่น, อย่างไรก็ดี, อย่างไรก็ตาม, อาทิ, อีกทั้ง, เช่น, เดี๋ยว, แต่, แต่ถึงอย่างไรเสีย, แต่ทว่า, แต่ว่า, แต่แล้ว, แถม, และ, แล้ว, โดย, ได้แก่, ไปจนถึง, ไม่ว่า

D

กว่า, ก่อน, ข้างหน้า, ครึ่ง, ฉะนี้, ดังกล่าว, ด้วยกัน, ต่อมา, ต่างๆ, ถัดมา, ทั่วไป, ทั้งหมด, ทั้งหลาย, ที่ว่า, ที่แล้ว, นั่น, นั้น, นั้นๆ, นี้, ล่าสุด, สุดท้าย, หนึ่ง, หน้า, หลัง, หลังๆ, อย่างใดอย่างหนึ่ง, อะไร, อื่น, อื่นๆ, เดิม, เดิมๆ, เดียว, เดียวกัน, เศษ, เหล่านั้น, เหล่านี้, เอง, แรก, โน้น, ใด, ใดๆ, ไหน

NCM

(1), (2), (3), (4), (5), (6), (ก), (ข), 1, 1., 1.1, 1.2, 1.3, 10, 2, 2., 2.1, 2.2, 24, 2503, 2513, 2519, 2528, 2533, 2536, 2538, 2540, 2541, 2542, 2543, 2547, 261, 27, 28, 29, 3, 3., 30, 31, 4, 4., 5, 7, Call Warrant, Code Sharing, Consultant, D.W., Dental Surgeon, Dentist, Derivative warrant, GP, General practitioner, MLR, Medical doctor, Newspaper, PO, PP, SET 50 Index, SET Index, Surgeon, back to back, identical option, ก.ศ., กฎ, กฎระเบียบ, กฎหมาย, กฎเหล็ก, กบฏ, กรณี, กรณีศึกษา, กรม, กรมธรรม์, กรรมการ, กรรมการบริหาร, กรรมการผู้จัดการ, กรรมการผู้จัดการใหญ่, กรรมการผู้อำนวยการใหญ่, กรรมาธิการ, กรอบ, กระทรวง, กระท่อม, กระบวนการ, กระบวนทัศน์, กระเป๋า, กระแส, กรุง, กลยุทธ, กลาง, กลิ่น, กลุ่ม, กลุ่มทุน, กล้อง, กษัตริย์, กองทัพ, กองทุน, การ, การกระทำ, การค้า, การธนาคาร, การบรรยาย, การบิน, การประชุม, การประชุมใหญ่, การศึกษา, การสัมมนา, การเกษตร, การเงิน, การเมือง, การเรียนการสอน, การเลือกตั้ง, การเลือกตั้งซ่อม, การ์ตูน, กาแฟ, กำมือ, กำลัง, กำหนดการ, กำแพง, กำไร, กิจกรรม, กิจการ, ก๊าซ, ก๊าซธรรมชาติ, ก๊าซฯ, ขณะ, ขนาด, ขบวนการ, ขวา, ขอบ, ขอบเขต, ขั้น, ขั้นตอน, ขา, ขาประจำ, ขาใหญ่, ขีดจำกัด, ขุนทหาร, ข่าว, ข่าวลือ, ข่าวสาร, ข้อกล่าวหา, ข้อกำหนดสิทธิ, ข้อความ, ข้อจำกัด, ข้อตกลง, ข้อติง, ข้อบังคับ, ข้อผิดพลาด, ข้อมูล, ข้อสงสัย, ข้อสรุป, ข้อสังเกต, ข้อเท็จจริง, ข้อเรียกร้อง, ข้อเสนอ, ข้าง, ข้างต้น, ข้างมาก, ข้างๆ, ข้าราชการ, คณะ, คณะกรรมการ, คณะกรรมการการเลือกตั้ง, คณะกรรมการบริหาร, คณะกรรมาธิการ, คณะทำงาน, คณะรัฐมนตรี, คดี, คดีความ, คน, คนดู, คนในเครื่องแบบ, คนไข้, ครม., ครอบครัว, ครั้ง, ครู, คลินิก, ความคิด, ความคิดเห็น, ความจริง, ความจุ, ความถี่, ความผิด, ความยาว, ความรู้, ความรู้สึก, ความสนใจ, ความสามารถ, ความหนา, ความหมาย, ความเป็นจริง, ความเป็นอยู่, ความเมือง, ความเร็ว, ความเห็น, คอมพิวเตอร์, คอมพิวเตอร์ส่วนบุคคล, คอมพิวเตอร์แบบตั้งโต๊ะ, คอรัปชั่น, คอลัมน์, คะแนน, คะแนนจัดตั้ง, คะแนนนิยม, คะแนนสงสาร, คาถา, คำ, คำกล่าว, คำขวัญ, คำขอ, คำขออนุญาต, คำขาด, คำถาม, คำบอกเล่า, คำปรึกษา, คำปรึกษาแนะนำ, คำพิพากษา, คำย่อ, คำร้อง, คำร้องเรียน, คำสัมภาษณ์, คำสั่ง, คำสำทับ, คิว, คืน, คุณ, คุณชาย, คุณภาพ, คุณสมบัติ, คุณูปการ, คู่, คู่ควง, คู่แข่ง, ค่า, ค่าจ้าง, ค่าที่ปรึกษา, ค่าธรรมเนียม, ค่าผ่านท่อ, ค่าย, ค่ารอยัลตี้, ค่าใช้จ่าย, ค่าไถ่, งบ, งบประมาณ, งาน, งานประจำ, งานเขียน, งานเลี้ยง, จดหมาย, จักรยานยนต์, จังหวัด, จำกัด, จำนวน, จิตวิญญาณ, จิตใจ, จุด, จุดบิน, จุดประสงค์, จุดอ่อน, จุดแข็ง, ชนชั้น, ชนวน, ชนิด, ชมรม, ชัย, ชัยชนะ, ชั้น, ชั้นศาล, ชาติ, ชาย, ชายฝั่ง, ชาวต่างชาติ, ชาวบ้าน, ชาวประมง, ชาวหมาหอน, ชิพ, ชีวภาพ, ชีวิต, ชื่อ, ชื่อเสียง, ชุด, ชุดวิชา, ชุมชน, ช่วง, ช่อง, ช่องทาง, ช่องโหว่, ชุ้ม, ญาติพี่น้อง, ฐาน, ฐานคะแนน, ฐานที่มั่น, ฐานะ, ดร., ดอก, ดอกเบี้ย, ดอกเบี้ยจ่าย, ดอกเบี้ยรับ, ดอกเบี้ยอ้างอิง, ดัชนี, ดัชนีอ้างอิง, ดิน, ดินแดน, ดึก, ดุลพินิจ, ด้าน, ตราประทับ, ตลาด, ตลาดสด, ตลาดโลก, ตอน, ตังค์, ตัว, ตัวการ์ตูน, ตัวตลก, ตัวนำ, ตัวบุคคล, ตัวประกัน, ตัวละคร, ตัวอย่าง, ตัวแทน, ตัวแปร, ตั๋ว, ตา, ตาราง, ตำรวจ, ตำแหน่ง, ตึน, ต่างชาติ, ต่างประเทศ, ต้น, ต้นคิด, ต้นทุน, ต้นเหตุ, ต้นไม้, ต้นๆ, ถนน, ถังขยะ, ถุง, ถ้อยคำ, ทรัพยากร, ทรัพย์, ทรัพย์สิน, ทวีป, ทศวรรษ, ทหาร, ทอง, ทักษะ, ทันตแพทย์, ทัศนะ, ทาง, ทางการ, ทางออก, ทางอ้อม, ทางเลือก, ทำนอง, ทิศทาง, ทีท่า, ทีม, ทีมงาน, ที่ 2, ที่ 4, ที่ 7, ที่ 8, ที่คุมขัง, ที่ทำการ, ที่ทำงาน, ที่นั่ง, ที่ปรึกษา, ที่พักพิง, ที่มา, ที่สอง, ที่สาม, ที่อยู่อาศัย, ทุกข์สุข, ทุกวันนี้, ทุน, ทุนการศึกษา, ทุนจดทะเบียน, ทุนนิยม, ทุนสนับสนุน, ทุนเงินออม, ทูต, ท่อส่งก๊าซฯ, ท่า, ท่าที, ท่าน, ท่า

อากาศยาน, ท้องถิ่น, ธงชาติ, ธนาคาร, ธนาคารชาติ, ธนาคารพาณิชย์, ธุรกิจ, ธุระปะปัง, นก, นกหวีด, นคร, นศ., นักการเมือง, นักข่าว, นักค้าน้ำลาย, นักธุรกิจ, นักบริหาร, นักบิน, นักรบ, นักลงทุน, นักวิจัย, นักวิเคราะห์, นักศึกษา, นักออกแบบ, นักเขียน, นักเดินทาง, นักเรียน, นักเอาใจชุมชนชาวหมาหอน, นาง, นางสาว, นานาชาติ, นาม, นาย, นายกรัฐมนตรี, นายกฯ, นายหน้า, นิตยสาร, นิทรรศการ, นิสัย, นโยบาย, น็อต, น้ำตา, น้ำหนัก, น้ำเสียง, บ., บง., บท, บทบรรณาธิการ, บทบาท, บทภาพยนตร์, บทลงโทษ, บมจ., บรรณาธิการ, บรรยากาศ, บรรษัท, บริการ, บริษัท, บริษัทจดทะเบียน, บริหาร, บริเวณ, บอร์ด, บัญชี, บัญชีรายชื่อ, บัญชีสำรองส่วนเกิน, บัดนี้, บัตร, บันทึก, บันเทิง, บาดแผล, บาป, บาร์-โค้ด, บาร์เตอร์, บุคคล, บุคลากร, บุรุษ, บ่าย, บ้าน, บ้านเกิดเมืองนอน, บ้านเมือง, บ้านใกล้เรือนเคียง, ปฏิรูป, ประกัน, ประกันชีวิต, ประกันภัย, ประการ, ประชาชน, ประชาชาติ, ประชาพิจารณ์, ประชาสัมพันธ์, ประชุม, ประธาน, ประธานาธิบดี, ประมง, ประมุข, ประวัติ, ประวัติการณ์, ประสบการณ์, ประสา, ประสิทธิภาพ, ประเด็น, ประเทศ, ประเทศชาติ, ประเภท, ประโยชน์, ปรากฏการณ์, ปริญญา, ปริญญาตรี, ปริญญาเอก, ปริมณฑล, ปริมาณ, ปลัด, ปลา, ปลาย, ปลายเหตุ, ปัจจัย, ปัจจุบัน, ปัญญา, ปัญหา, ปาก, ปี, ปีกลาย, ปีการศึกษา, ปีงบประมาณ, ปุ๋ย, ปูนซีเมนต์, ปูมหลัง, ป่า, ผล, ผลกระทบ, ผลงาน, ผลบังคับใช้, ผลประกอบการ, ผลประโยชน์, ผลผลิต, ผลพวง, ผลสำรวจ, ผลสำเร็จ, ผลเสีย, ผู้, ผู้กำกับ, ผู้ก่อการร้าย, ผู้จัดการ, ผู้จัดงาน, ผู้ชม, ผู้ชาย, ผู้ถือหุ้น, ผู้นำ, ผู้บริสุทธิ์, ผู้บริหาร, ผู้บริโภค, ผู้ปกครอง, ผู้ประกอบการ, ผู้ผลิต, ผู้ผลิตรายการ, ผู้พิพากษา, ผู้ร่วมทุน, ผู้ลงทุน, ผู้ว่า, ผู้ว่าการ, ผู้ว่างงาน, ผู้ว่าจ้าง, ผู้สมัคร, ผู้สังเกตการณ์, ผู้สื่อข่าว, ผู้หญิง, ผู้อำนวยการ, ผู้อ่าน, ผู้เชี่ยวชาญ, ผู้เรียน, ผู้เล่น, ผู้แทน, ผู้แทนฯ, ผู้โดยสาร, ผู้ใหญ่, ฝัน, ฝ่าย, พ.ศ., พนักงาน, พนักงานต้อนรับ, พรรค, พรรคการเมือง, พระบาทสมเด็จพระเจ้าอยู่หัว, พระราชินี, พระราชโอรส, พระองค์, พระเกียรติ, พระเจ้า, พฤติกรรม, พลัง, พลังความร้อน, พลังงาน, พลังน้ำ, พลาสติก, พวก, พันธบัตร, พันธมิตร, พาณิชย์, พาดหัวข่าว, พิซซ่า, พีโอ, พี่, พื้นฐาน, พื้นที่, พื้นที่เคลื่อนไหว, พื้นบ้าน, พ่อครัว, พ่อพระ, พ่อแม่, ฟาร์ม, ฟิวส์, ภรรยา, ภริยา, ภาค, ภาคพื้น, ภาพ, ภาพพจน์, ภาพยนตร์, ภาพรวม, ภาพลักษณ์, ภายนอก, ภายหลัง, ภายใน, ภารกิจ, ภาระ, ภาวะ, ภาษา, ภาษี, ภาษีมูลค่าเพิ่ม, ภูมิภาค, ม., มงคลสมัย, มติ, มนุษยชาติ, มนุษยธรรม, มนุษยสัมพันธ์, มนุษย์, มหาชน, มหาวิทยาลัย, มาตรการ, มาตรฐาน, มิติ, มือ, มุข, มุมมอง, มุมมืด, มูลค่า, ม้า, ยอด, ยอดขาย, ยักษ์ใหญ่, ยาม, ยิ้ม, ยุค, ยุติธรรม, ยุทธวิธี, ย่าน, รถยนต์, รถยนต์นั่ง, รถแท็กซี่, รมช., รมต., รศ., รสชาติ, รอง, รอบ, รอบตัว, รอย, รอยยิ้ม, ระดับ, ระบบ, ระบบคุณภาพมาตรฐาน, ระบอบ, ระยะ, ระยะเวลา, ระเบียบ, รัฐ, รัฐธรรมนูญ, รัฐบาล, รัฐมนตรี, รัฐมนตรีว่าการ, รัฐสภา, ราก, รากเหง้า, ราคา, รางวัล, ราชการ, ราชกิจจานุเบกษา, ราชสำนัก, ราย, รายการ, รายงาน, รายจ่าย, รายรับ, รายละเอียด, รายได้, ราษฎร, รุ่น, รู, รูป, รูปแบบ, ร่างกฎหมาย, ร่างกาย, ร้าน, ร้านค้า, ร้านรวง, ร้านแดกด่วน, ลงทุน, ละคร, ลักษณะ, ลาภลอย, ลำดับ, ลูก, ลูกค้า, ลูกจ้าง, ลูกตา, ลูกทีม, ลูกน้อง, ลูกน้อย, ลูกมือ, ลูกหนี้, ลูกหลาน, ล่าง, วง, วงการ, วงจร, วงเงิน, วงเงินคุ้มครอง, วงใน, วัง, วัฒนธรรม, วัตถุดิบ, วัตถุประสงค์, วัน, วันทำการ, วันที่, วันนี้, วันรุ่งขึ้น, วันหยุด, วัย, วาจา, วานนี้, วารสาร, วาระ, วาระแห่งชาติ, วิกฤติ, วิกาล, วิจัย, วิชาการ, วิดีโอ เกม, วิถีชีวิต, วิถีทาง, วิทยาเขต, วิธี, วิธีการ, วิธีทำ, วินัย, วิศวกร, วิศวกรรม, วุฒิสภา, วุฒิสมาชิก, ว่าที่, ศักดิ์ศรี, ศักยภาพ, ศัตรู, ศัพท์, ศาล, ศาลยุติธรรม, ศาลล้มละลาย, ศาลอาญา, ศาสนา, ศิลป์, ศูนย์, ศูนย์บริการการศึกษา, ศูนย์เทคโนโลยี, ศูนย์เทคโนโลยีทางการศึกษา, ส.ว., ส.ส., สงคราม, สถานการณ์, สถานที่, สถานที่ทำงาน, สถานภาพ, สถานศึกษา,

สถานี, สถาบัน, สถาบันการจัดอันดับความน่าเชื่อถือ, สถาบันจัดอันดับ, สถาบันจัดอันดับความน่าเชื่อถือ, สถิติ, สนาม, สนามบิน, สปิริต, สภา, สภาผู้แทนราษฎร, สภาพ, สภาพคล่อง, สภาพแวดล้อม, สภาวะ, สภาวะการณ์, สภาสูง, สมอง, สมัย, สมาคม, สมาชิก, สรรพกำลัง, สหภาพ, สหาย, สอง, สังคม, สังคมนิยม, สัญชาติ, สัญญา, สัญญาณ, สัดส่วน, สัตยาบัน, สัตว์, สันดอน, สันดาน, สันติภาพ, สัปดาห์, สัมพันธไมตรี, สากล, สาขาวิชา, สาม, สามัญสำนึก, สามี, สาย, สายการบิน, สายตา, สายส่ง, สารสนเทศ, สาระ, สาว, สาวๆ, สาเหตุ, สำนัก, สำนักงาน, สำนักงานปลัด, สำนักงานใหญ่, สำนักผู้ว่าการ, สำนักพิมพ์, สิทธิ, สินค้า, สินค้าทุนนำเข้า, สินค้าฟุ่มเฟือย, สินค้าอุปโภคบริโภค, สินเชื่อ, สินเชื่อไม่ก่อให้เกิดรายได้, สิ่ง, สิ่งทอ, สี, สี่, สื่อ, สื่อการศึกษา, สื่อการสอน, สื่อมวลชน, สุขภาพ, สุดที่รัก, สุดสัปดาห์, สุนทรพจน์, สุนัข, สุภาพสตรีหมายเลขหนึ่ง, สุภาษิต, สูตร, สเปรด, ส่วน, ส่วนต่าง, ส่วนรวม, ส่วนร่วม, ส่วนแบ่ง, หนัง, หนังสือ, หนังสือพิมพ์, หนังสือแสดงเจตจำนงฯ, หนี้, หนี้สิน, หนี้ไม่ก่อให้เกิดรายได้, หนึ่ง, หนึ่งเดียว, หนู, หน่วยงาน, หน้า, หน้าตา, หน้าต่าง, หน้าที่, หน้าที่การงาน, หน้าแข้ง, หมอ, หมายเหตุ, หมี, หมู่, หลวง, หลัก, หลักการ, หลักฐาน, หลักทรัพย์, หลักประกัน, หลักสูตร, หลักเกณฑ์, หลังมือ, หลุม, หอศิลป์, หัว, หัวหน้า, หัวเรื่อง, หัวใจ, หุ่น, หุ้น, หุ้นส่วน, หุ้นอ้างอิง, หุ้นเพิ่มทุน, ห้องน้ำ, ห้างสรรพสินค้า, องค์กร, องค์ประกอบ, อดีต, อธิการบดี, อธิบดี, อนาคต, อนามัย, อนุพันธ์, อนุสัญญา, อรรถรส, อัตรา, อัตราส่วน, อัธยาศัย, อันดับ, อันตราย, อาการ, อาจารย์ผู้สอน, อาชญากร, อาชญากรรม, อาณานิคม, อานิเมชั่น, อายุ, อารมณ์, อารมณ์ขัน, อาวุโส, อาหาร, อาหารการกิน, อาหารว่าง, อาหารเย็น, อำนาจ, อินเทอร์เน็ต, อิเล็กทรอนิกส์, อี-เมล์, อุดมศึกษา, อุตสาหกรรม, อุบัติเหตุ, อุปกรณ์, อุปสรรค, เกณฑ์, เกม, เกร็ดข่าว, เกษตร, เกาะ, เกียรติยศ, เก้าอี้, เขตเดียวเบอร์เดียว, เครดิตฟองซิเอร์, เครือข่าย, เครื่อง, เครื่องกีดขวาง, เครื่องจักร, เครื่องบิน, เครื่องปั่นผลไม้, เครื่องมือ, เครื่องยนต์, เครื่องใช้, เคเบิลทีวี, เค้า, เค้าลาง, เงิน, เงินกองทุนสภาพคล่องสุทธิ, เงินกู้, เงินช่วยเหลือ, เงินตรา, เงินต้น, เงินทุน, เงินประกัน, เงินฝาก, เงินลงทุน, เงินสด, เงินสภาพคล่องสุทธิ, เงินออม, เงินเดือน, เงินใต้โต๊ะ, เงินๆทองๆ, เงื่อนไข, เจตนา, เจ้า, เจ้าของ, เจ้าตัว, เจ้าบ้าน, เจ้าพนักงานพิทักษ์ทรัพย์, เจ้าภาพ, เจ้าหนี้, เจ้าหน้าที่, เชิง, เชื้อชาติ, เชื้อสาย, เชื้อเพลิง, เช้า, เดินสาย, เดิม, เดือน, เด็ก, เทคนิค, เทคนิคัล, เทคโนโลยี, เทศกาล, เที่ยวบิน, เท้า, เนื้อ, เนื้อหา, เบี้ยประกัน, เบี้ยประกันภัย, เบื้องต้น, เบื้องหน้า, เบื้องหลัง, เป้า, เป้าหมาย, เพลง, เพศ, เพื่อน, เพื่อนบ้าน, เพื่อนฝูง, เพื่อนร่วมงาน, เมาส์, เมือง, เมืองกรุง, เมืองหลวง, เมื่อก่อน, เมื่อวานนี้, เม็ดเงิน, เยาวชน, เรือ, เรื่อง, เรื่องราว, เร็วๆนี้, เลขาธิการ, เลือกตั้ง, เลือกตั้งซ่อม, เลือด, เวบ, เวบไซต์, เวลา, เศรษฐกิจ, เศรษฐมิติ, เศรษฐศาสตร์, เสถียรภาพ, เสรีภาพ, เสียง, เสียงโวย, เสี่ย, เสื้อ, เสื้อยืด, เส้นทาง, เส้นสาย, เหตุ, เหตุผล, เหยื่อ, เห็ด, เอกชน, เอกราช, เอกลักษณ์, เอกสาร, เอกอัครราชทูต, เอฟที, เอสพีแอล, เอเยนซี, เอ็นซีอาร์, เอ็นพีแอล, แกนนำ, แถบ, แถว, แท็กซี่, แนว, แนวคิด, แนวทาง, แนวร่วม, แนวหน้า, แนวโน้ม, แบงก์, แบบ, แบบฟอร์ม, แบบสอบถาม, แผงวงจร, แผน, แผนงาน, แผนฟื้นฟู, แผ่น, แผ่นดินใหญ่, แผ่นฟิล์ม, แพทย์, แพทย์ผ่าตัด, แฟนๆ, แมว, แม็กซ์, แม่ข่าย, แม่บ้าน, แรง, แรงกดดัน, แรงงาน, แรงงานสัมพันธ์, แสน, แหล่ง, แหล่งข่าว, แหล่งอ้างอิง, โครงการ, โครงสร้าง, โฆษก, โฆษณา, โต๊ะ, โต๊ะบัคคาร่า, โท, โทรทัศน์, โทรศัพท์แบบหมุน, โทษทัณฑ์, โบราณ, โปร, โปรแกรมเมอร์, โภชนาการ, โรค, โรคท้องร่วง, โรง, โรงงาน, โรงพยาบาล, โรงเรียน, โรงแรม, โรงไฟฟ้า, โลก, โอกาส, ใต้, ใบสมัคร, ใบสำคัญ, ใบหน้า, ใบอนุญาต, ไข้โป้ง, ไซต์, ไตรมาส, ไปรษณีโทรเลข, ไฟ, ไฟฟ้า, ไฟฟ้าผันแปร, ไฟล์, ไฟแนนซ์, ไม้กวาด, ไอ้

NCSF

กรณี, ก้อน, ข้าง, คน, ครั้ง, คราว, คำ, คำรบ, คืน, คู่, งวด, งวดๆ, จังหวัด, จำนวน, ฉบับ, ชนชั้น, ชนิด, ชั่วโมง, ชุด, ชุดวิชา, ดอลลาร์, ด้าน, ตัว, ทศวรรษ, ทาง, ที, นาย, บท, บริษัท, บาท, บีทียู, ประการ, ประเด็น, ประเทศ, ประเภท, ปี, ฝีก้าว, ฝ่าย, พรรค, ฟาก, รอบ, ราย, รายการ, ริ้วๆ, ลักษณะ, วัน, วิชา, สัปดาห์, สาขา, ส่วน, หน, หน้ากระดาษ, หุ้น, องค์, อย่าง, อัน, อันดับ, เครื่องยนต์, เซนต์, เดือน, เมกะวัตต์, เรื่อง, เอสดีอาร์, แนวทาง, แผ่น, แห่ง, โครงการ, โรง, ไตรมาส, ไมครอน

NPP

14 สิงหาคม 2541, Fantasticon.com, HSCB, KFW, Krungthep Turakij, Shockwave.com, Thrust Reverse, Toonscape.com, joecartoon.com, ก.ค., ก.ย., ก.ล.ต., กกต., กทม., กฟผ., กรุงเทพ, กรุงเทพฯ, กรุงไทย, กฤษณพงศ์, กศน., กษมา, กสิกรไทย, กอคเวเซอร์, กองทุนการเงินระหว่างประเทศ, กองทุนฟื้นฟูฯ, กองทุนเพื่อการฟื้นฟูและพัฒนาระบบสถาบันการเงิน, กันยายน, กัมพูชา, การคลัง, การบินไทย, การประกันภัย, การประดิษฐ์การ์ตูนแนวใหม่, การปิโตรเลียมแห่งประเทศไทย, การศึกษานอกโรงเรียน, การไฟฟ้าฝ่ายผลิตแห่งประเทศไทย, กำแพงเพชร, กีรติกร, กุมภาพันธ์, ขวัญใจธัญญา, คณะกรรมการข้าราชการครู, คณะอนุกรรมการมาตรฐานสินเชื่อที่อยู่อาศัย, คมนาคม, คริส, คริสต์มาส, คลัง, คลินตัน, ความหวังใหม่, คอมมิคส์ รีเทลเลอร์, คอมมิวนิสต์, คำรณ, คินน็อค, ง็อก, จองชัย, จอมล้วงลูก, จอมล้วงลูกไร้รอย, จอร์แดน, จอห์น, จัตุมงคล, จันทบุรี, จันทร์, จัสมิน อินเตอร์เนชั่นแนล, จีน, จึงสงวนสิทธิ์, จูเนียร์, ชวน, ชัช, ชับบ์, ชับบ์ ประกันภัย, ชาญ, ชาติไทย, ชาวกัมพูชา, ชาวพรรคประชาธิปัตย์, ชาวฟิลิปปินส์, ชาวมุสลิม, ชาวอังกฤษ, ชาวอเมริกัน, ชาวเวียดนาม, ชิลด์ส, ชุมพร, ชโรเดอร์, ช็อคเวฟ, ซังแตร์, ซับไมครอน, ซัวเถา, ซานดิเอโก, ซาอุดีอาระเบีย, ซิติ กรุ๊ป, ซีเมนส์, ซูซานนา, ฌากส์, ญี่ปุ่น, ดอง ฟาม, ดอนเมือง, ดับเบิลยูพีพี, ดับเบิลยูพีพี กรุ๊ป, ดัลลัส, ดีน, ดูปองต์, ต.ค., ตรัง, ตลาดรองสินเชื่อเพื่อที่อยู่อาศัย, ตลาดหลักทรัพย์, ตังทัตสวัสดิ์, ตันศิริ, ตูนสเคป, ทนง, ทบวงมหาวิทยาลัย, ทบวงฯ, ทองอินทร์, ทอม, ทอมกับเจอร์รี่, ทอมแอนด์เจอร์รี่, ทักษิณ, ทากาฮาชิ, ทารา์, ทำความเข้าใจกับการ์ตูน, ทำเนียบ, ทำเนียบขาว, ทิม, ทิศทางของศูนย์การศึกษาในอนาคต, ทีเอ็นที, ทีโออาร์, ทูซอน, ธงชัย, ธนชาติ, ธนาคารพัฒนาเอเชีย, ธนาคารแห่งประเทศไทย, ธนาคารโลก, ธปท., ธรรมนูญ, ธอส., ธารินทร์, ธี, ธีรภัทร, นครหลวงไทย, นรวัฒน์, นิวซีแลนด์, นิวยอร์ก, นิวยอร์กออบเซอร์เวอร์, นีล, นูร์, น้ำเทิน 2, บรัสเซลส์, บอมเบย์, บอริส, บังกลาเทศ, บางปะกง, บาสิลัน, บิล, บีทีเอส, บุญเลิศ, ปตท., ประกันชีวิต, ประกันภัย, ประชาธิปัตย์, ปราจีนบุรี, ปริศนานันทกุล, ปองพล, ปอยเปต, ปัตตานี, ปากีสถาน, ปิแอร์, ปูติน, ผลิตไฟฟ้า, ฝรั่งเศส, พ.ค., พนมหัตถ์, พนมเปญ, พม่า, พรรคแม่ธรณีบีบมวยผม, พฤษภาคม, พะเยา, พับบลิกซิส, พับลิกซิส, พิทยะ, พุธ, ฟรอก เบลเดอร์, ฟรังค์ฟูร์เทอร์อัลเกไมเนอไซทุง, ฟลอริดา, ฟอร์ด มอเตอร์, ฟาม, ฟาร์อีสเทิร์น, ฟาร์อีสเทิร์น อีโคโนมิครีวิว, ฟิตช์ อิบคา ดีซีอาร์, ฟิลิปปินส์, ฟูจิโมริ, ม.ค., มกราคม, มสธ., มหานคร, มอร์ลีย์ ฟันด์ แมเนจเมนท์, มอร์แกน, มอสโก, มัทราส, มาร์ควิซ, มาเก๊า, มาเลเซีย, มาแสง, มิชิแกน, มิซูอารี, มิถุนายน, มินห์, มิยาซาวา, มิลเลอร์, มี.ค., มีนาคม, มุสลิม, มูดี้ส์, ยัง & รูบิแคม, ยัง&รูบิแคม, ยาดานา, ยุโรป, ยูเอ็น, ยูไนเต็ด, ยูไนเต็ด แอร์ไลน์, ย่างกุ้ง, ระบบขนส่งมวลชนกรุงเทพ, รัตนาภรณ์, รัตโนภาส, ราชบุรี, ราฟิค, รีพับบ

ลิก, ลอนดอน, ลอว์เรนซ์, ลักเซมเบิร์ก, ลาว, ลาสเวกัส, ลินช์, ลิสบอน, ลี, ลุฟท์ ฮันซ่า, ลุฟท์ฮันซ่า, ลเอสเปรสโซ่, วงศ์โสธร, วรวรรณ ณ อยุธยา, วลาดิมีร์, วอชิงตัน, วัน, วันชัย, วัลด์รอน, วาย & อาร์, วิชัย, วิทช์เบลด, วีระ, ศธ., ศรีนคร, ศรีลังกา, ศิริชนะ, ศึกษาธิการ, ศึกษาฯ, ศุกร์, สกอตต์, สก็อตต์, สงครามเวียดนาม, สตาร์ อัลลายแอนซ์, สมศักดิ์, สมอ., สมเกียรติ, สมเชาว์, สศช., สหประชาชาติ, สหภาพยุโรป, สหรัฐ, สหรัฐอเมริกา, สอง, สันติวงษ์, สำนักงานคณะกรรมการกำกับหลักทรัพย์และตลาดหลักทรัพย์, สำนักงานคณะกรรมการพัฒนาการเศรษฐกิจและสังคมแห่งชาติ, สำนักงานคณะกรรมาธิการ, สำนักงานคณะกรรมาธิการสหภาพยุโรป, สำนักงานความสัมพันธ์แรงงานแห่งชาติ, สำนักงานมาตรฐานผลิตภัณฑ์อุตสาหกรรม, สิงคโปร์, สิงคโปร์ แอร์ไลน์ส, สิงหาคม, สิงห์, สิทธิพร, สิปาดัน, สุรพล, สุราษฎร์ธานี, สุวรรณ, สุวัฒน์, สุโขทัยธรรมาธิราช, สแกนดิเนเวีย, สแตน, ส่งเสริมอุตสาหกรรม, หลีกภัย, หวั่งหลี, องค์การณ์, อดิศร, อดิเรกสาร, อภิสิทธิ์, อริโซนา, อลาสก้า แอร์ไลน์, ออลนิปปอน แอร์เวย์, ออสเตรเลีย, อะแมชิง สไปเดอร์แมน, อังกฤษ, อับดุลเลาะห์, อับดุลเลาะห์ที่ 1, อัลเบอร์โต, อัศวโชค, อาคารสงเคราะห์, อาทิตย์, อาบู เซยาฟ, อาร์ชี, อาร์ลิน, อาเซียน, อิตาลี, อินธนูไชย, อินเดีย, อินโดจีน, อินโดนีเซีย, อินไซด์มอสโก, อิสราเอล, อิสลาม, อุตสาหกรรม, อุตฯ, อเมริกัน, อเมริกา, อ่อนวิมล, อ่าวไทย, ฮวง, ฮานอย, ฮิกุชิ, ฮิตชิงส์, ฮิตชิงห์, ฮิลลารี, ฮุน, ฮุสเซน, เกษประทุม, เกอร์บิล อิน อะ ไมโครเวฟ, เกอร์ฮาร์ด, เขมร, เขมรแดง, เคน, เครซอง, เคอร์รี, เคียร์นีย์, เงินฉ่ำ, เงินทุนอุตสาหกรรมแห่งประเทศไทย, เจ., เจ.วอลเทอร์ ธอมพ์สัน, เจดดะห์, เจอร์รี่, เชียงใหม่, เซน, เซาธ์ พาร์ค, เซาธ์เวสต์, เซาธ์เวสต์ แอร์ไลน์, เซียะเหมิน, เซเว่น-อีเลฟเว่น, เดวิด, เดอ ลา ครูซ, เดอบอสซิเออ, เดอร์ สปิเคล, เดอะซันเดย์ไทม์, เดอะรีพิวเทชั่น อินสติติว, เดอะเฮีย, เตาปูน, เต่า, เท็กซัส, เนเธอร์แลนด์, เบลล์, เบอร์ตัน, เปรู, เพียงเกษ, เม.ย., เมษายน, เมอร์ริล ลินช์ ภัทร, เยตะกุน, เยลต์ซิน, เยอรมนี, เยอรมัน, เลอ มงด์, เวชชาชีวะ, เวียดนาม, เวียดนามคอมมิวนิสต์, เวียดนามเหนือ, เวียดนามใต้, เสาร์, เหงียน, เอ ดี บี, เอดิธ, เอที&ที, เอบีเอ็น แอมโร, เอส แอนด์ พี มูดีส์ อินเวสเตอร์ เซอร์วิส, เอสเอเอส, เอสเอ็มซี, เอเชีย, เอเชียอาคเนย์, เอเอฟพี, เอ็กโก, เอ็นแอลอาร์บี, เอ็มดี-80, แกรนด์ แรปิดส์, แคนาดา, แคลิฟอร์เนีย, แนนซี่, แบงก์ชาติ, แบงก์ออฟอเมริกา, แบร์รี, แบร์รี่, แบลร์, แพรทท์ แอนด์ วิทนีย์, แพร่, แพ็ตเทน, แมคคลาวด์, แมคลาร์เลน, แมน, แมร์, แม็คคาร์ทนีย์, แม็คคาร์ธี, แลนด์ แอนด์ เฮ้าส์, แวร์ริก, แอนดรู, แอนเซส ออสเตรเลีย, แอร์แคนาดา, แฮร์ริส อินเตอร์แอคทีฟ, โกลดา, โจ, โจรสลัดแห่งเกาะตะลุเตา, โจโล, โทนี, โธมัส, โปรดี, โปรตุเกส, โมโตโรล่า, โรมาโน, โสภณ, โอกิลวี่ &เมเธอร์, โอไฮโอ, โฮจิมินห์, โฮจิมินห์ ซิตี้, ไข, ไคโอ, ไซ่ง่อน, ไทม์คีปปิง ซิสเต็ม, ไทย, ไทยธนาคาร, ไทยพาณิชย์, ไทยรักไทย, ไทยรับประกันภัยต่อ, ไบรอัน, ไมร์เนอร์ แอนด์ เวอร์ส, ไมเคิล, ไมโครซอฟท์, ไรนา, ไลน์เวเบอร์, ไอเอฟซีที, ไอเอสโอ, ไอเอสโอ 9000, ไอเอ็มเอฟ

NPRO

กัน, คุณ, ตน, ตนเอง, ตรงนี้, ตรงไหน, ตัวเอง, ทั้งคู่, ทั้งหมด, ที่นั่น, ที่นี่, ที่ไหน, นั่น, นั้น, นาง, นี่, นี้, น้อง, ผม, ผู้, พระองค์, พวกข้า, พวกนั้น, พวกนี้, พวกเขา, พวกเรา, พวกเอ็ง, มัน, อะไร, อะไรๆ, อัน, เขา, เช่นนั้น, เช่นนี้, เท่าไร, เท่าไหร่, เธอ, เพื่อก, เรา, เหล่านี้, เอ็ง, เฮา, ใคร, ใครๆ

PCOMP

ซึ่ง, ที่, ว่า, ให้

PFX

การ, การที่, ความ

PN

กลาง, กว่า, กับ, ก่อน, ก่อนหน้า, ของ, ข้าม, จนถึง, จาก, ณ, ดัง, ด้วย, ตรง, ตลอด, ตั้งแต่, ตาม, ต่อ, ถึง, ทั่ว, ทาง, ที่, ท่ามกลาง, นอก, นอกจาก, นับจาก, นับตั้งแต่, นับแต่, บน, ประจำ, ผ่าน, พร้อม, พร้อมกับ, ภายใต้, ภายใน, ยัง, ระหว่าง, ราวกับ, ละ, สำหรับ, สู่, หน้า, หลัง, หลังจาก, อย่าง, เกี่ยวกับ, เช่น, เช่นเดียวกับ, เดียวกับ, เนื่องเพราะ, เนื่องใน, เพราะ, เพื่อ, เมื่อ, เหนือ, แก่, แต่, แห่ง, โดย, ใน, ในระหว่าง, ให้, ให้กับ, ให้แก่, ไปถึง

PT

ก็แล้วกัน, ครับ, จ๊ะ, ซี, นัก, นั่นเอง, นั่นแหละ, นั้น, นา, นี่, นี่แหละ, นี้, ฟะ, มั้ย, ยังไงล่ะ, ละ, ละก็, ละก้อ, ล่ะ, สิ, หรอก, หรือยัง, หรืออย่างไร, หรือไง, หรือไม่, หา, อิ, อ้อ, ฮ่า, เข้านั่น, เซ่, เป็นต้น, เฟ้ย, เลย, เสีย, เสียเลย, เสียแล้ว, เอ๊ย, แล, แหละ, โถ, โว้ย, โห, ไง

PUNC

! " % ' ( ) , - . : <slash> ?

PV

กว่า, ก่อน, ก่อนที่, ก่อนหน้าที่, ขณะ, ขณะที่, จน, จนกว่า, จาก, จากที่, ดัง, ด้วย, ตราบเท่าที่, ตั้งแต่, ตาม, ถึง, ถึงแม้, ถึงแม้ว่า, ถ้า, ถ้าหาก, ทั้งๆที่, นอกจาก, นับตั้งแต่, พร้อม, พร้อมกับ, พอ, ระหว่าง, สำหรับ, หลัง, หลังจาก, หลังจากที่, หาก, หากว่า, อย่าง, เช่น, เท่าที่, เนื่องจาก, เนื่องจากว่า, เป็นเพราะ, เพราะ, เพราะว่า, เพื่อ, เพื่อสำหรับ, เมื่อ, เวลา, เหมือนกับว่า, แทนที่, แม้, แม้ว่า, โดย, โดยที่, ในขณะที่, ในระหว่าง

Q

0.01, 0.18, 0.19, 0.35, 1, 1,000, 1,500, 1,655, 1.28, 1.3, 1.5, 1.8, 1.95, 10, 13, 13,470, 14, 149, 15, 16, 2, 2,000, 2,400, 2,500, 2.2, 2.8, 20, 20,867, 200, 21, 24, 25, 25,000, 258, 27, 275, 28, 29, 3, 3,000, 3,103, 3,450, 3,53, 3.5, 3.63, 30, 300, 31, 320, 34, 340, 35, 35,000, 4, 4.5, 40, 400, 407, 49, 49.5, 5, 50,000, 50,322, 500, 6, 60, 600, 61, 62, 625, 65, 7, 75, 78, 79.6, 8, 8.25, 800, 82, 825, 850, 9, 90, 93, กว่า, กี่, กึ่ง, ตั้ง, ถึง, ทั่ว, ทั้ง, ทุก, บรรดา, บาง, ประมาณ, พัน, ยี่สิบ, ราว, ร่วม, ล., ล้าน, สอง, สองสาม, สัก, สาม, สิบ, สิ้น, สี่, หมื่น, หลาย, หลายๆ, ห้า, อีก, เจ็ด, เฉพาะ, เพียง, แค่, แต่, แต่ละ, แต่เพียง, แสน, ไม่กี่

V0

กดดัน, กระจาย, กระชิบ, กระทำ, กระโดด, กลับ, กลับคืน, กลาย, กลุ้ม, กังวลใจ, กำไร, กินน้ำใต้ศอก, กู้, ก่อสร้าง, ก้าวหน้า, ขยับ, ขยายตัว, ขอกันกินมากกว่านี้, ขออนุญาต, ขออภัย, ขัดแย้ง, ขาด, ขาดทุน, ขาย, ขึ้น, ขึ้นศาล, ข่มขู่, ข้องใจ, คงค้าง, คงเหลือ, ครบ, ครบกำหนด, คร่อมหลัง, คล้อง, ควบคุม, คอย, คอรัปชั่น, คับแค้นใจ, คิด, คิดอ่าน, คืนชีพ, คืบ, คืบหน้า, คุ้มครอง, ค้ำประกัน, ง่วง, จดทะเบียน, จบ, จบการศึกษา, จัดการ, จัดซื้อจัดจ้าง, จ่าย, ฉลุย, ฉิบหาย, ฉ้อราษฎร์บังหลวง, ชนะ, ชำรุด, ชุมนุม, ชูโรง, ช่วยเหลือ, ซุกซ่อน, ซุบซิบ, ซ้ำซาก, ดาหน้า, ดำเนิน, ดำเนินการ, ดำเนินงาน, ดื่ม, ดูหมิ่น, ดูแล, ดูแลเอาใจใส่, ตก, ตกต่ำ, ตกลง, ตบท้าย, ตรงกันข้าม, ตรงต่อเวลา, ตรงเวลา, ตรวจสอบ, ตะขิดตะขวงใจ, ตัดสิน, ตัดสินใจ, ตั้ง, ตั้งรับ, ตั้งใจ, ตาม, ติดตาม, ต่อต้าน, ต่อรอง, ต่อสู้, ต่อเนื่อง, ต่าง, ถก, ถล่ม, ถอนตัว, ถอยร่น, ถังแตก, ทราบ, ทะเลาะ, ทะเลาะเบาะแว้ง, ทัน, ทับซ้อน, ทั่ว, ทำ, ทำงาน, ทุจริต, ท้อแท้, นอน, นั่ง, นำ, นิยม, บริการ, บริหาร, บริโภค, บิน, บุก, บูดบึ้ง, ปกครอง, ปฏิบัติ, ปนเป, ประกันภัย, ประกาศ, ประจาน, ประจำ, ประชุม, ประมูล, ปรับ, ปรับตัว, ปรับทุกข์, ปรากฏ, ปรากฏตัว, ปลอม, ปล่อยกู้, ปะปน, ปั่น, ปิดตัว, ผงาด, ผลัด, ผลิต, ผวา, ผันผวน, ผ่าน, ผ่านพ้น, ฝึกงาน, พบปะ, พยักเพยิด, พยายาม, พร้อมใจ, พลิกล็อค, พอ, พอใจ, พัฒนา, พัวพัน, พาดหัว, พิจารณา, พิทักษ์ทรัพย์เด็ดขาด, พุงกาง, พูด, พูดคุย, พ้อง, ฟื้น, ฟื้นตัว, ฟื้นฟู, ฟ้อง, มอง, มั่นใจ, มา, มี, มีกิน, ยกเกรด, ยั่วยุ, ยิ้มแย้มแจ่มใส, ยืดอก, ยืดเยื้อ, ยืน, ยืนยัน, ยุติ, รณรงค์, รบ, รวม, รวมความ, รอ, รอท่า, ระเหี่ยใจ, รับผิดชอบ, ริเริ่ม, รุก, รุดหน้า, รุม, รู้, ร่วม, ร่วมทุน, ร่วมมือ, ร่ำเรียน, ร้อนวิชา, ลง, ลงขัน, ลงคะแนน, ลงจอด, ลงตัว, ลงทุน, ลงโทษ, ลด, ลอยนวล, ละเมิด, ล็อกถล่ม, ล่ม, ล่มจม, ล่วงละเมิด, ล้มละลาย, ล้มเหลว, ล้วงลูก, ล้อเลียน, วนเวียน, วางจำหน่าย, วางแผน, วิจัย, วิวาท, วิ่ง, วิ่งวิบาก, ว่า, ว่างงาน, ศึกษา, สนับสนุน, สบอารมณ์, สมควร, สมัคร, สรุป, สร้างงาน, สร้างสรรค์, สวิง, สอน, สอบ, สะสม, สับสน, สัมภาษณ์, สำเร็จ, สำเร็จการศึกษา, สิ้นพระชนม์, สิ้นสุด, ส่งมอบ, ส้มหล่น, หงายหลัง, หน้าแตก, หมด, หมดสภาพ, หมดเขตรับสมัคร, หมุน, หยอกล้อ, หยุด, หลบหนี, หลอกลวง, หลับตา, หลุด, หวั่นไหว, หัวเราะ, หัวเสีย, หางานทำ, หาย, หารือ, หาเสียง, อนุญาต, อยู่, อยู่รอด, ออก, ออกลูก, ออมทรัพย์, อัด, อิจฉา, อุปถัมภ์, อุปโภคบริโภค, อ้อยเข้าปากช้าง, เกิด, เกี่ยวข้อง, เข็ด, เข้า, เข้าทีม, เข้าป้าย, เข้าเรียน, เข้าใจ, เคลื่อนไหว, เงียบ, เจตนา, เจรจา, เจริญพรชนมายุ, เจ๊ง, เฉลี่ย, เซ็ง, เซ็นสัญญา, เดิน, เดินขบวนพาเหรด, เดินทาง, เดินพาเหรด, เดินหน้า, เดินเครื่อง, เดินเหิน, เดือดร้อน, เตรียมการ, เติบโต, เถียง, เบี่ยงเบน, เบื่อ, เปลี่ยน, เปลี่ยนแปลง, เปิด, เปิดปาก, เปิดสอน, เป็น, เป็นงง, เป็นไป, เผยแพร่, เพิ่ม, เพียงพอ, เฟื่องฟู, เรียกขาน, เรียน, เรียนรู้, เรืองอำนาจ, เรื้อรัง, เลือน, เลื่อนชั้นเลื่อนตำแหน่ง, เล่น, เล่าเรียน, เว้นว่าง, เสด็จ, เสนอขาย, เสร็จ, เสร็จสิ้น, เสีย, เสียหาย, เสี่ยง, เหนือ, เหลือ, เห็นชอบ, เห็นดีเห็นงาม, เห็นด้วย, เอวัง, เอาจริงเอาจัง, เอ่ยปาก, แก้ไข, แข่งขัน, แค้น, แตกต่าง, แตกแยก, แต่ง, แถลง, แพร่หลาย, แพ้, แยก, แล้ว, แล้วเสร็จ, แล้วไปใหญ่, แวะเยี่ยม, แวะเวียน, แสดง, แหก, โกง, โฆษณา, โปรด, โรดโชว์, โล่งใจ, โวยวาย, โหมโรง, โอน, ใกล้ชิด, ใกล้เคียง, ใจจดใจจ่อ, ใช้การ, ใช้งาน, ใช้จ่าย, ใช้ประโยชน์, ให้, ให้สัมภาษณ์, ให้อภัย, ได้, ได้ผล, ได้เสีย, ไป, ไปไหนมาไหน, ไล่เสือเข้าป่า ปิ้งปลาประชดแมว, ไหล

VADJ

กว้าง, กว้างๆ, กะทัดรัด, กะทันหัน, คึกคัก, คุ้ม, ง่าย, ง่ายๆ, จริง, จริงจัง, จริงๆจังๆ, จำกัด, จำเป็น, จิ๋ว, จุกจิก, ฉลาด, ชอบธรรม, ชัดเจน, ชั่ว, ชื่อดัง, ช้า, ช้าๆ, ซับเซา, ซื่อ, ซ่อนเร้น, ดัง, ดี, ดีงาม, ดีๆ, ดึกดื่น, ดึกๆ, ดูดี, ดูไม่จืด, ด่วน, ตรง, ตลก, ติดดิน, ตื่นตัว, ต่ำ, ต้องห้าม, ถนัด, ถี่, ถูก, ทรงคุณวุฒิ, ทรงอิทธิพล, ทันสมัย, ทุจริต, ธรรมดา, นาน, น่านิยมยกย่อง, น่าประทับใจ, น่าประหลาดใจ, น่าพิศมัย, น่าสนใจ, น่าอึดอัด, น่าเจ็บใจ, น่าเชื่อ, น่าเชื่อถือ, น่าแปลกใจ, น้อย, บริบูรณ์, บังเอิญ, บ่อยครั้ง, บ่อยๆ, ปกติ, ปราดเปรื่อง, ปลอดภัย, ผิด, ผิดกฎหมาย, ผิดพลาด, ผ่อนปรน, พร้อม, พิสดาร, พิเศษ, ฟรี, มหาศาล, มั่นคง, มาก, มากน้อย, มากมาย, มิดชิด, มีค่า, มีหน้ามีตา, ยั่งยืน, ยาก, ยากจน, ยากลำบาก, ยาว, ยาวนาน, ยิ่งหย่อน, ยิ่งใหญ่, ยืดหยุ่น, ยุ่งยาก, ย่อย, ย่ำแย่, รวดเร็ว, รวย, รอง, รอบคอบ, รุนแรง, ร่ำรวย, ร้าย, ร้ายแรง, ลบ, ละเอียด, ละเอียดอ่อน, ลำบาก, ลึก, ลึกๆ, ล่าช้า, ล้าสมัย, ล้าหลัง, วิเศษ, วุ่น, ว่าง, สกปรก, สงบ, สด, สดใส, สนุก, สนุกสนาน, สบาย, สบายๆ, สมัยใหม่, สวย, สว่าง, สะดวก, สะดวกสบาย, สะเทินน้ำสะเทินบก, สั้น, สามัญ, สำคัญ, สำคัญๆ, สำเร็จ, สุดยอด, สูง, ส่วนตัว, ส่วนบุคคล, หนัก, หนักหนา, หนักแน่น, หนุ่ม, หยาบ, หยาบคาย, หลัก, หลักๆ, หลากหลาย, หวือหวา, หอมหวาน, หัวกะทิ, หัวหมอ, หัวแข็ง, หัวใส, หิน, ออกนอกหน้า, อัตโนมัติ, อันตราย, อาวุโส, อิดโรย, อิสระ, อิสรเสรี, อ่อนแอ, เกินงาม, เกินจริง, เก่ง, เก่งๆ, เก่า, เก่าก่อน, เก่าแก่, เก่าๆ, เก๋, เขียว, เข้ม, เข้มข้น, เข้มงวด, เข้มแข็ง, เข้าท่า, เคร่งครัด, เงินหนา, เงียบเหงา, เจ้าเล่ห์, เจ๋ง, เฉพาะกิจ, เชื่องช้า, เดียวดาย, เต็มจำนวน, เต็มที่, เต็มรูปแบบ, เต็มสูบ, เถื่อน, เบา, เบาบาง, เบาๆ, เบ็ดเสร็จ, เปิดเผย, เป็นดี, เป็นทางการ, เป็นที่แน่นอน, เป็นธรรม, เป็นธรรมดา, เป็นปกติ, เป็นประจำ, เป็นพิเศษ, เป็นล่ำเป็นสัน, เป็นส่วนตัว, เป็นส่วนใหญ่, เป็นอันหนึ่งอันเดียว, เป็นเอกฉันท์, เป็นใหญ่เป็นโต, เป็นไร, เยอะ, เยี่ยม, เย้ายวน, เรียบ, เรียบร้อย, เร็ว, เลว, เละ, เละเทะ, เล็ก, เล็กน้อย, เล็กๆ, เล็กๆน้อยๆ, เว่อร์, เสมือนจริง, เสรี, เสียเส้น, เสี่ยง, เสื่อมถอย, เหงา, เหนียม, เหมาะสม, เหม็น, เหลือเฟือ, แข็งแกร่ง, แข็งแรง, แจ่มใส, แดง, แน่ชัด, แน่น, แน่นอน, แปลกหน้า, แปลกใหม่, แพง, แยบยล, แรง, แสนกล, แสบ, โดดเด่น, โต, โปร่งใส, โอบอ้อมอารี, ใกล้ชิด, ใจกว้าง, ใจดี, ใต้ดิน, ใหญ่, ใหญ่ๆ, ใหม่, ใหม่ๆ, ไกลๆ, ได้เปรียบ

VAUX

ควร, คอย, ค่อย, จำเป็นต้อง, ดู, ต้อง, ถูก, ทรง, ทำท่า, น่า, พอ, พึง, สามารถ, ออก, อาจ, เคย, เป็น, เห็น, โดน, ใช่, ได้, ได้รับ

VCV0

กล่าว, กำหนด, ขอ, ขู่, คาด, คาดการณ์, คิด, ค่อนแคะ, จัด, จำ, ชักชวน, ชี้, ชี้แจง, ช่วย, ดีใจ, ดึงดูดใจ, ดู, ตกลง, ตรวจสอบ, ตอบ, ตัดสินใจ, ตีความ, ต้องการ, ถือ, ทราบ, ทึกทัก, นับ, นิยม, นึก, บอก, บอกเล่า, บังคับ, ปฏิเสธ, ประกาศ, ประสงค์, ประเมิน, ปรับตัว, ปรากฏ, ปรารถนา, ปล่อย, ผ่อนปรน, พบ, พิจารณา, พูด, มอง, มั่นใจ, มุ่งมั่น, ยอม, ยอมรับ, ยึดมั่น, ยืนยัน, รณรงค์, รอ, ระบุ, รับทราบ, รับรอง, รายงาน, รู้, รู้สึก, ลืม, วิเคราะห์, สนใจ, สะท้อน, สังเกต, สั่งการ, ส่งผล, ส่งเสริม, หมายความ, หวัง, หวังดี, อนุญาต, อนุมัติ, อยาก,

อ้าง, เกรง, เข้าใจ, เชื่อ, เปิดช่อง, เปิดทาง, เปิดเผย, เปิดเผยตัว, เปิดโอกาส, เป็นผล, เผย, เรียก, เรียกร้อง, เลือก, เล่า, เสนอ, เสมือน, เสริม, เห็น, แจ้ง, แถลง, โน้มน้าว, โปรโมท, ไว้วางใจ

VN0

กรอก, กระจาย, กระตุ้น, กระทำ, กลับ, กลั่นแกล้ง, กล่าว, กล่าวหา, กล้ำกลืน, กวาด, กวาดล้าง, กอง, กำ, กำกับ, กำหนด, กีดกัน, กู้, ก่อ, ก่อตั้ง, ก่อสร้าง, ก่อเกิด, ขน, ขบคิด, ขยับ, ขยาย, ขอ, ขออนุมัติ, ขัด, ขับ, ขับเคลื่อน, ขาด, ขาดแคลน, ขานรับ, ขาย, ขี่, ขุด, ข้าม, คง, ครบ, ครอง, ครอบครอง, ครอบคลุม, คลิ๊ก, คลุม, คล้าย, ควบคุม, ควัก, คว่ำ, คว้า, คัด, คัดเลือก, คิด, คิดถึง, คิดเป็น, คืน, คุม, คุ้มครอง, งัด, งาบ, จด, จัด, จัดการ, จัดซื้อ, จัดตั้ง, จัดทำ, จัดวาง, จัดเตรียม, จับ, จับกุม, จับมือ, จำกัด, จำหน่าย, จำแนก, จ่อ, จ่าย, ฉลอง, ฉาย, ชดเชย, ชน, ชนะ, ชม, ชอบ, ชอบหน้า, ชะลอ, ชำระ, ชิง, ชี้, ชี้แจง, ชี้แนะ, ชื่นชม, ชื่นชอบ, ชู, ช่วย, ชบ, ซื้อ, ดับ, ดำรง, ดำเนิน, ดึง, ดึงดูด, ดึงดูดใจ, ดู, ดูด, ดูหมิ่นดูแคลน, ดูแล, ด่า, ตก, ตบ, ตรวจจับ, ตรวจตรา, ตรวจสอบ, ตอก, ตอบ, ตอบแทน, ตัด, ตั้ง, ตาม, ตามติด, ติด, ติดตั้ง, ติดตาม, ติดต่อ, ติดๆ, ตีกลับ, ต่อ, ต่อต้าน, ต้องการ, ต้าน, ถก, ถาม, ถึง, ถือ, ถือเป็น, ทดสอบ, ทดแทน, ทราบ, ทำ, ทิ้ง, ทุ่ม, ท้วงติง, ท้าทาย, นับเป็น, นั่ง, นำ, นำเข้า, นำเสนอ, นิยม, นึกถึง, บรรจุ, บรรลุ, บรรเทา, บริหาร, บวก, บอก, บังเกิด, บัญญัติ, บำรุงรักษา, ปกครอง, ปกป้อง, ปฏิรูป, ปฏิวัติ, ปฏิเสธ, ประกบ, ประกอบ, ประกันภัย, ประกาศ, ประจาน, ประชด, ประทับใจ, ประมวล, ประมาณ, ประยุกต์, ประยุกต์ใช้, ประสบ, ประเมิน, ปรับ, ปรับปรุง, ปรับเปลี่ยน, ปราบปราม, ปราม, ปลด, ปลอม, ปล่อย, ปิด, ปิดฉาก, ปิดล้อม, ป้อน, ผนวก, ผนึก, ผลิต, ผสมผสาน, ผ่อนคลาย, ผ่าน, ฝัง, ฝาก, ฝืน, พก, พบ, พบเห็น, พอใจ, พัก, พักงาน, พัฒนา, พิจารณา, พิทักษ์ทรัพย์เด็ดขาด, พิมพ์, พึ่งพิง, พูด, พ้น, ฟัง, ฟื้น, ฟื้นฟู, มอง, มองข้าม, มองดู, มองหา, มอบ, มี, ยก, ยกย่อง, ยกเครื่อง, ยกเลิก, ยกเว้น, ยอมรับ, ยิง, ยึดถือ, ยึด, ยืนยัน, ยื่น, ยุติ, ย่ำ, ย้อน, ย้าย, รณรงค์, รวบรวม, รวม, รอ, รองรับ, ระดม, ระบุ, ระมัดระวัง, รัก, รักษา, รับ, รับตรวจ, รับผิดชอบ, รับฟัง, รับรอง, รับสมัคร, รายงาน, รุมล้อม, รู้, รู้จัก, ร่วม, ร้องเรียน, ลง, ลงทุน, ลงโทษ, ลด, ลดทอน, ลอกเลียน, ลอง, ละทิ้ง, ละเมิด, ลักพา, ล้มละลาย, ล้อเลียน, วาง, วิจัย, วิเคราะห์, วิ่งไล่จับ, ว่าจ้าง, ว่าด้วย, ศึกษา, สกัด, สงวน, สงสาร, สนับสนุน, สนใจ, สรุป, สร้าง, สร้างสรรค์, สะกดรอยตาม, สะท้อน, สะสาง, สัมภาษณ์, สั่งงาน, สำรวจ, สำรอง, สิ้น, สืบแทน, สูญเสีย, ส่ง, ส่งมอบ, ส่งออก, ส่งเสริม, ส่งเสีย, ส่อ, ส่าย, หนี, หมด, หมั่นไส้, หมายถึง, หยุด, หลอก, หลอม, หลีกเลี่ยง, หวาดผวา, หัก, หัน, หา, ห้าม, อนุญาต, อนุมัติ, อภัยให้, อยู่, ออก, ออกแบบ, อาย, อาศัย, อุ้ม, อ่าน, อ้าง, ฮุบ, เกรงใจ, เกลียด, เกิด, เกิน, เกี่ยวข้อง, เก็บ, เก็บเกี่ยว, เขมือบ, เขียน, เข้า, เข้าข้าง, เข้าซื้อ, เข้าถึง, เข้าร่วม, เข้าใจ, เจรจา, เจาะข่าว, เฉพาะ, เฉลิม, เฉลิมฉลอง, เฉลี่ย, เชิญ, เชียร์, เชี่ยวชาญ, เชื่อ, เด้ง, เตรียม, เติมต่อ, เท, เทียบ, เท่า, เน้น, เบิก, เปรียบเทียบ, เปรียบเหมือน, เปลี่ยน, เปลือง, เปิด, เปื้อน, เป็น, เป็นหนี้, เป่า, เผชิญ, เผยแพร่, เพิกถอน, เพิ่ม, เยี่ยม, เยือน, เย็บ, เริ่ม, เรียก, เรียกประชุม, เรียกร้อง, เรียนรู้, เร่ง, เลย, เลี้ยงดู, เลี้ยงโต๊ะจีน, เลือก, เลือกตั้ง, เลื่อน, เล่น, เล่า, เว้น, เสนอ, เสนอขาย, เสนอแนะ, เสมือน, เสริมสร้าง, เสิร์ฟ, เสีย, เสียบ, เหมาจ่าย, เหมือน, เห็น, เอา, เอาชนะ, เอารัดเอาเปรียบ, เอ่ย, แกล้ง, แก้, แก้ไข, แขวน, แจก, แจกแจง, แจ้ง, แตะ, แต่งตั้ง, แถลง, แทน, แทนที่, แนะนำ, แปลง, แลก, แลกซื้อ, แล่น, แวะ, แสดง, แสวงหา, โกงกิน, โกย, โค่น

อำนาจ, โจมตี, โดดเดี่ยว, โบก, โปรด, โมโห, โอน, ใช่, ใช้, ใส่, ให้, ให้กำเนิด, ได้, ได้รับ, ไต่สวน, ไถ, ไป, ไร้, ไล่, ไล่ออก, ไว้หน้า, ไว้ใจ

VNCV0

กล่าวถึง, กล่าวหา, กาหัว, ชี้แนะ, ตราหน้า, ติดตาม, ถาม, ทำ, บังคับ, ปรับ, พัฒนา, พิจารณา, ยกย่อง, สร้าง, สอน, สั่ง, เจาะ, เตะ, เตือน, เปรียบ, เปรียบเทียบ, เสนอ, แก้, แก้ไข, แจ้ง

VNN0

ชำระ, ใช้, ให้

VNPN0

ขอ, ขัง, ขาย, คืน, จ่าย, ชำระ, ซื้อ, ถก, ทุ่มเท, นำ, ปลดแอก, ป้อน, ผนวก, ร่วม, สร้าง, ส่ง, ออก, เก็บ, เข้าซื้อ, เทียบ, เปรียบ, เสนอ, เสนอขาย, ให้, ได้, ได้รับ, ไว้

VPN0

กระทบ, กลับ, ขัดแย้ง, ขึ้น, คล้าย, คำนึง, จินตนาการ, ตก, ตกลง, ตรง, ตรงข้าม, ตาม, ติด, ติดต่อสื่อสาร, ต่าง, ถอนตัว, ถัด, ทรมาน, ทราบ, ทารุณ, นอกเหนือ, บวก, บีบคั้น, บุ้ยใบ้, ปฏิบัติ, ประกอบ, ปรากฏตัว, ปลดแอก, ผนวก, ผละ, พบ, พบปะ, พอฟัดพอเหวี่ยง, พิจารณา, พ่ายแพ้, ฟ้องร้อง, มั่นใจ, มา, มุ่ง, ยึดติด, ยืนยัน, ย้ำ, รวม, รอดพ้น, รับทราบ, รับผิดชอบ, รับมือ, ร่วม, ลงนาม, ละเมิด, วิตกกังวล, ศึกษา, สนับสนุน, สนใจ, สม, สอดคล้อง, สะท้อน, สูง, สู้, ส่งเสริม, หารือ, อยู่, อุทธรณ์, อ่อนไหว, เกิด, เกี่ยว, เกี่ยวข้อง, เข้มงวดกวดขัน, เข้า, เข้าทำงาน, เข้าพบ, เข้าร่วม, เข้าศึกษา, เข้าใจ, เจรจา, เต็ม, เทียบ, เนื่อง, เบื่อหน่าย, เปิดเผย, เป็นกันเอง, เพียงพอ, เริ่ม, เริ่มต้น, เล่า, เสียเปรียบ, เสี่ยง, เหมาะ, เหมือน, เห็น, แข่ง, แข่งขัน, แจ้ง, แตะ, แวะ, ไป

VS0

ขู่, คาด, ชี้, ทำ, ทำนาย, ทำเอา, ทำให้, ฟัง, มั่นใจ, มี, ยกเว้น, รู้, ว่า, สั่ง, หวัง, หวั่น, ห้าม, เกรง, เจอ, เฉลิม ฉลอง, เท่ากับ, เปิดทาง, เผย, เหมือน, เห็น, แจง, ให้, ได้ยิน

VV0

กล้า, ขอ, ขออนุญาต, คิด, จงใจ, ชอบ, ชิง, ช่วย, ซ้อม, ดำเนินการ, ตกลง, ตากหน้า, ต้องการ, ทน, ทยอย, ทำเป็น, นัด, นิยม, ประสงค์, ผลักดัน, พยายาม, พร้อม, พากัน, พิจารณา, มัว, ยอม, ยื่น, รอ, รับ, รีบ, รู้สึก, ร่วม, ลอง, สนใจ, สมควร, สมัคร, สั่ง, หลง, หัน, ห้าม, อด, อนุญาต, อยาก, เจรจา, เตรียม, เน้น, เบี้ยว, เป็นอัน, เผอิญ, เริ่ม, เร่ง, เร่งรัด, เลิก, เลือก, เล็ง, เสนอ, เสนอตัว, แก้, แวะ, แห่, ใกล้

ภาคผนวก ข

รายการคำที่กำกับหมวดคำผิด

หมวดคำที่ถูกต้อง	หมวดคำกำกับผิด	รูปคำที่กำกับหมวดคำผิด	จำนวนรูปคำ	จำนวนครั้ง
NCM	NCSF	ชนชั้น, ครั้ง, ประเทศ, อันดับ, วัน, ชนิด, ข้าง	7	8
	NPRO	ผู้	1	2
	D	หนึ่ง	1	1
	Q	7, 28, 30, 3, 31	5	5
	PN	หน้า	1	1
	V0	โฆษณา, เจตนา	2	9
NCSF	NCM	บท, ฝ่าย, ส่วน, เรื่อง, ทศวรรษ, สัปดาห์	6	6
NPRO	NCM	คุณ, ผู้	2	12
	NCSF	อัน	1	1
V0	AV	มา, อยู่, ไป, ขึ้น	4	9
	VPN0	ปฏิบัติ, รวม, เกิด	3	4
	VCV0	ตกลง, ปรากฏ	2	3
	VN0	ตั้ง, ทำ, ชนะ, ก่อสร้าง, คิด, เปลี่ยน, เจรจา, ควบคุม, นำ	9	11
	VV0	ตกลง	1	1
	VNPN0	นำ	1	1
VPN0	V0	มั่นใจ, ยืนยัน, เกี่ยวข้อง, รวม	4	5
	VCV0	ตกลง	1	1
	VN0	พบ	1	1
	VNPN0	ปลดแอก	1	1
VCV0	V0	พูด, อนุญาต, ตัดสินใจ	3	3
	VPN0	เปิดเผย	1	2
	VN0	กล่าว, รายงาน, กำหนด,เล่า	4	4
	VNCV0	พิจารณา	1	1

VN0	V0	พูด, เรียนรู้, ยุติ, เข้า, ทำ, แสดง, ละเมิด	7	8
	VPN0	เกิด	1	1
	VCV0	กล่าว, ยอมรับ, ชี้แจง, เรียกร้อง	4	7
	VNPN0	ผนวก	1	2
	VAUX	เห็น, ใช่, ได้	3	3
	PV	แทนที่	1	1
	PCOMP	ให้	1	1
VNPN0	V0	เข้า	1	1
	VN0	ได้รับ, ร่วม, สร้าง, ให้	4	9
	VS0	ให้	1	1
VNCV0	VN0	ถาม, กล่าวหา	2	2
VV0	V0	สมัคร, ร่วม	2	2
	PV	พร้อม	1	1
VS0	VCV0	ขู่	1	1
	VN0	ให้, ห้าม	2	2
	VV0	ห้าม	1	1
	PN	ให้	1	1
	PCOMP	ให้	1	1
VADJ	V0	สำเร็จ	1	2
	VN0	เยี่ยม	1	1
VAUX	AV	ดู	1	2
D	AV	เอง	1	1
Q	VN0	ตั้ง	1	1
	NCM	1, 1.3	2	2
	AV	อีก	1	1
AV	V0	ไป, ออก	2	4
	VPN0	อยู่	1	2
	VAUX	อาจ	1	4
	D	ต่อมา, เอง	2	2
	C	ก็	1	1
	Q	แค่	1	1

PN	V0	ผ่าน	1	1
	VS0	ให้	1	1
	NCM	ทาง	1	1
	PV	อย่าง, เพื่อ, หลัง	3	3
	PCOMP	ที่	1	3
	C	จนถึง	1	1
PV	AV	ก่อ น, กว่า	2	4
	PN	เพราะ	1	1
	C	โดย	1	6
PCOMP	PN	ที่, ให้	2	4
C	PV	โดย	1	2
PT	AV	เลย	1	3

ภาคผนวก ค

รายการคำที่ตัดคำผิด

ตารางที่ ค-1 รายการคำที่ตัดคำผิดเนื่องจากตัดคำหนึ่งคำแยกเป็นหลายคำ

การตัดคำที่ถูกต้องของผู้วิจัย	**การตัดคำที่ผิดของ** POS trigram	**การตัดคำที่ผิดของ** WFORM trigram	**จำนวนครั้ง**
13/Q	1/Q-3/Q		1
21/Q	2/Q-1/Q		2
34/Q	3/Q-4/Q		1
ก่อนหน้าที่/PV	ก่อนหน้า/PN-ที่/PCOMP		1
ข้างมาก/NCM	ข้าง/NCSF-มาก/VADJ		1
ขึ้นศาล/V0	ขึ้น/AV-ศาล/NCM		1
เข้าข้าง/VN0	เข้า/V0-ข้าง/NCSF		1
เข้าถึง/VN0	เข้า/VPN0-ถึง/PN		1
ความเป็นอยู่/NCM	ความ/PFX-เป็น/VN0-อยู่/AV		1
ความผิด/NCM	ความ/PFX-ผิด/VADJ		5
ความยาว/NCM	ความ/PFX-ยาว/VADJ		1
คำร้องเรียน/NCM	คำร้อง/NCM-เรียน/V0		1
คิดถึง/VN0	คิด/V0-ถึง/PN		1
เครื่องใช้/NCM	เครื่อง/NCM-ใช้/VN0		1
งานเขียน/NCM	งาน/NCM-เขียน/VN0		1
เจ้าของ/NCM	เจ้า/NCM-ของ/PN		4
ชื่อดัง/VADJ	ชื่อ/NCM-ดัง/PV		1
ดูเหมือนว่า/AV	ดูเหมือน/AV-ว่า/PCOMP		2
ได้แต่/AV		ได้-แต่	1
ได้แต่/AV	ได้/VAUX-แต่/C		1
ตรงเวลา/V0	ตรง/PN-เวลา/NCM		1
ต่อมา/AV	ต่อ/AV-มา/AV		1
ต่อรอง/V0	ต่อ/PN-รอง/NCM		6
ตัวการ์ตูน/NCM	ตัว/NCM-การ์ตูน/NCM		2
ตัวแทน/NCM	ตัว/NCM-แทน/VN0		1

ตัวประกัน/NCM	ตัว/NCM-ประกัน/NCM		14
ตัวเอง/NPRO	ตัว/NCSF-เอง/D		1
แต่เพียง/Q	แต่/Q-เพียง/Q		1
ทางการ/NCM		ทาง-การ	2
ทำให้/VS0	ทำ/V0-ให้/PCOMP		1
ที่ทำงาน/NCM	ที่/PCOMP-ทำงาน/V0		5
ที่นั่ง/NCM	ที่/PCOMP-นั่ง/V0		1
ที่นี่/NPRO	ที่/PN-นี่/NPRO		1
ที่แล้ว/D		ที่-แล้ว	1
ที่ไหน/NPRO	ที่/PCOMP-ไหน/D		2
ทุกเมื่อ/AV	ทุก/Q-เมื่อ/PN		1
นักเขียน/NCM	นัก/PT-เขียน/VN0		2
นางสาว/NCM	นาง/NCM-สาว/NCM		1
แนวหน้า/NCM	แนว/NCM-หน้า/NCM		1
ปรับเปลี่ยน/VN0	ปรับ/VN0-เปลี่ยน/VN0		1
เปิดปาก/V0	เปิด/VN0-ปาก/NCM		1
ไปจนถึง/C	ไป/VPN0-จนถึง/PN		1
ผู้กำกับ/NCM	ผู้/NCM-กำกับ/VN0		1
ผู้ชม/NCM	ผู้/NCM-ชม/VN0		1
ผู้อ่าน/NCM	ผู้/NCM-อ่าน/VN0		2
พวกนั้น/NPRO	พวก/NCM-นั้น/D		1
พวกเรา/NPRO	พวก/NCM-เรา/NPRO		1
พักงาน/VN0	พัก/VN0-งาน/NCM		1
มองดู/VN0	มอง/VN0-ดู/VN0		1
เมื่อวานนี้/NCM	เมื่อ/PN-วานนี้/NCM		2
แม้ว่า/PV	แม้/AV-ว่า/PCOMP		1
แม้ว่า/PV	แม้/PV-ว่า/V0		1
แม้ว่า/PV	แม้/PV-ว่า/VS0		1
แรงงาน/NCM	แรง/NCM-งาน/NCM		6
ลูกน้อย/NCM	ลูก/NCM-น้อย/VADJ		1
ไล่ออก/VN0	ไล่/VN0-ออก/AV		1
วันหยุด/NCM	วัน/NCM-หยุด/V0		2
วิธีการ/NCM	วิธี/NCM-การ/PFX		1

ศาลยุติธรรม/NCM	ศาล/NCM-ยุติธรรม/NCM		1
สถานที่ทำงาน/NCM	สถานที่/NCM-ทำงาน/V0		1
ส่วนบุคคล/VADJ	ส่วน/C-บุคคล/NCM		3
สังคมนิยม/NCM	สังคม/NCM-นิยม/V0		3
หน้าต่าง/NCM	หน้า/NCM-ต่าง/AV		1
หัวหน้า/NCM	หัว/NCM-หน้า/D		3
หัวหน้า/NCM	หัว/NCM-หน้า/NCM		2
หัวหน้า/NCM	หัว/NCM-หน้า/PN		2
หุ้นส่วน/NCM	หุ้น/NCM-ส่วน/NCM		1
ออกแบบ/VN0	ออก/VN0-แบบ/NCM		1
เอ่ยปาก/V0	เอ่ย/VN0-ปาก/NCM		1

ตารางที่ ค-2 รายการคำที่ตัดคำผิดเนื่องจากตัดคำหลายคำรวมเป็นคำหนึ่งคำ

การตัดคำที่ถูกต้องของผู้วิจัย	**การตัดคำที่ผิดของ** POS trigram	**การตัดคำที่ผิดของ** WFORM trigram	**จำนวนครั้ง**
ก็/AV-มี/VN0		ก็มี	1
การ/PFX-ศึกษา/V0	การศึกษา/NCM		2
การ/PFX-ศึกษา/V0	การศึกษา/NCM	การศึกษา	1
เกี่ยว/VPN0-กับ/PN	เกี่ยวกับ/PN	เกี่ยวกับ	1
ตัวอย่าง/NCM-เช่น/C	ตัว/NCSF-อย่างเช่น/C		1
ที่/PCOMP-ทำงาน/V0		ที่ทำงาน	1
ที่/PCOMP-นั่ง/V0		ที่นั่ง	1
เท่า/VN0-นั้น/NPRO	เท่านั้น/AV	เท่านั้น	1
ไป/AV-ถึง/Q		ไปถึง	1
ยัง/AV-คง/VN0	ยังคง/AV	ยังคง	1
หุ้น/NCM-ส่วน/NCSF		หุ้นส่วน	1

ประวัติผู้เขียนวิทยานิพนธ์

นายนัฐวุฒิ ไชยเจริญ เกิดวันที่ 17 มกราคม พ.ศ. 2521 กรุงเทพมหานคร สำเร็จการศึกษาระดับปริญญาตรี อักษรศาสตรบัณฑิต สาขาภาษาญี่ปุ่น ภาควิชาภาษาตะวันออก จุฬาลงกรณ์มหาวิทยาลัย ในปีการศึกษา 2540 และเข้าศึกษาต่อในหลักสูตรอักษรศาสตรมหาบัณฑิต ภาควิชาภาษาศาสตร์ ที่จุฬาลงกรณ์มหาวิทยาลัย เมื่อปี พ.ศ. 2541