

การศึกษาริบทบ่งบอกชื่อเฉพาะในภาษาไทยตามแนวภาษาศาสตร์คอมพิวเตอร์

นางสาว อมรทิพย์ กวินปณิธาน

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาอักษรศาสตรมหาบัณฑิต

สาขาวิชาภาษาศาสตร์ ภาควิชาภาษาศาสตร์

คณะอักษรศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย

ปีการศึกษา 2546

ISBN 974-17-4494-3

ลิขสิทธิ์ของจุฬาลงกรณ์มหาวิทยาลัย

A COMPUTATIONAL LINGUISTIC STUDY
OF CONTEXT CLUES OF PROPER NAMES IN THAI

Miss Amornthip Kawinpanithan

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 2003

ISBN 974-17-4494-3

อมรทิพย์ กวินปณิธาน : การศึกษาบริบทบ่งบอกชื่อเฉพาะในภาษาไทยตามแนวภาษาศาสตร์
คอมพิวเตอร์. (A COMPUTATIONAL LINGUISTIC STUDY OF CONTEXT CLUES OF
PROPER NAMES IN THAI) อ. ที่ปรึกษา : ผู้ช่วยศาสตราจารย์ ดร. วิโรจน์ อรุณมานะกุล,
102 หน้า. ISBN 974-17-4494-3.

งานวิจัยนี้มีวัตถุประสงค์เพื่อศึกษาบริบทที่สามารถเป็นบริบทบ่งบอกการปรากฏของชื่อเฉพาะภาษาไทย
สำหรับนำไปใช้ในการพัฒนางานด้านภาษาศาสตร์คอมพิวเตอร์ โดยศึกษาจากคลังข้อมูลภาษาที่มีการกำกับ
ชื่อเฉพาะ ผู้วิจัยได้ทำการจัดสร้างคลังข้อมูลภาษาขึ้น โดยรวบรวมข้อความภาษาไทยจากหนังสือพิมพ์ธุรกิจ
และนิตยสารสกุลไทยแล้วทำการกำกับชื่อเฉพาะและบริบทบ่งบอกตามแนวทางการกำกับชื่อเฉพาะของ TEI

ผลการศึกษาชื่อเฉพาะทั้งหมด 2,547 ชื่อ พบว่าชื่อเฉพาะที่ปรากฏบริบทบ่งบอกคิดเป็น 63% ชื่อเฉพาะที่ไม่
ปรากฏบริบทบ่งบอกคิดเป็น 37% บริบทที่สามารถบ่งบอกชื่อเฉพาะภาษาไทยได้แก่ (1) คำที่ปรากฏติดกับชื่อเฉพาะ
หรือคำบ่งบอก (2) การเว้นวรรค และ (3) ตัวบ่งบอกระดับปริจเฉท คำบ่งบอกและการเว้นวรรคเป็นตัวบ่งบอกที่
ปรากฏกับชื่อเฉพาะโดยส่วนใหญ่ คำบ่งบอกมักจะปรากฏติดอยู่หน้าชื่อเฉพาะ ในการศึกษาที่พบคำบ่งบอก 161 คำ
ซึ่งสามารถจำแนกเป็นกลุ่มตามอรรถลักษณะร่วมในกลุ่มนั้นได้ทั้งสิ้น 19 กลุ่ม คำบ่งบอกยังสามารถบ่งชี้ประเภทของ
ชื่อเฉพาะโดยชื่อเฉพาะประเภทหนึ่งจะปรากฏร่วมกับคำบ่งบอกกลุ่มเฉพาะกลุ่มหนึ่ง ประเภทของชื่อเฉพาะที่พบใน
การศึกษานี้มี 6 ประเภท ได้แก่ ชื่อคน ชื่อองค์กร ชื่อสถานที่ ชื่อกลุ่มคนร่วมเชื้อสาย ชื่อกฎหมาย และชื่ออื่นๆ
นอกจากนี้ยังพบว่าคำบ่งบอกมีขีดความสามารถในการบ่งบอกชื่อเฉพาะแตกต่างกันแม้ว่าจะเป็นคำบ่งบอกในกลุ่ม
เดียวกัน ส่วนการเว้นวรรคปรากฏในตำแหน่งส่วนหน้าหรือส่วนหลังของชื่อเฉพาะและเป็นบริบทบ่งบอกที่มี
ความสัมพันธ์กับคำบ่งบอก การวิจัยพบว่าชื่อเฉพาะจำนวน 71% มีการเว้นวรรคร่วมด้วย และ 80% ของชื่อเฉพาะที่มี
คำบ่งบอกจะมีการเว้นวรรคด้วย หากพิจารณาการเว้นวรรคร่วมกับคำบ่งบอกจะทำให้คำบ่งบอกมีขีดความสามารถใน
การบ่งบอกชื่อเฉพาะสูงขึ้น และตัวบ่งบอกระดับปริจเฉท แม้จะพบเป็นจำนวนน้อยแต่มีรูปแบบที่ชัดเจน สามารถ
บ่งบอกการปรากฏชื่อเฉพาะที่ไม่ปรากฏคำบ่งบอกได้ โดยส่วนใหญ่พบว่าเป็นลักษณะอ้างตามชื่อที่ปรากฏก่อน
ในกรณีชื่อเฉพาะ 931 ชื่อที่ไม่ปรากฏบริบทบ่งบอกพบว่า เป็นชื่อเฉพาะที่มีโครงสร้างของรูปภาษาในลักษณะทับศัพท์
ภาษาต่างประเทศ หรือเป็นชื่อเฉพาะที่ผู้ใช้ภาษามีภูมิความรู้ร่วมกัน จึงเอื้อต่อการนำชื่อเฉพาะนั้นไปใช้โดยไม่ต้องมี
ตัวบ่งบอกในการระบุว่ารูปลภาษานั้นเป็นชื่อเฉพาะ ในการศึกษาครั้งนี้ยังพบว่า นอกจากชื่อเฉพาะจะปรากฏเป็น
หน่วยชื่อเฉพาะที่มีคุณสมบัติเฉพาะประจำรูปแล้ว ยังสามารถปรากฏเป็นส่วนประกอบหนึ่งของรูปภาษาอื่นที่ไม่ใช่
ชื่อเฉพาะอีกด้วย

| | | |
|------------|------------|---------------------------------|
| ภาควิชา | ภาษาศาสตร์ | ลายมือชื่อนิสิต..... |
| สาขาวิชา | ภาษาศาสตร์ | ลายมือชื่ออาจารย์ที่ปรึกษา..... |
| ปีการศึกษา | 2546 | |

4380199822 : MAJOR LINGUISTICS

KEY WORD: PROPER NAME / NAME / CLUE / CONTEXT / CORPUS / THAI

AMORNTHIP KAWINPANITHAN : A COMPUTATIONAL LINGUISTIC STUDY OF
CONTEXT CLUES OF PROPER NAMES IN THAI. THESIS ADVISOR : ASST. PROF.
WIROTE AROONMANAKUN, Ph.D., 102 pp. ISBN 974-17-4494-3.

This study is a computational linguistic study of context clues. It aims to analyse context clues that could be used to identify proper names in Thai. A corpus is collected from Bangkok Business newspaper and Sakulthai magazine. Proper names and context clues are manually tagged by using the Text Encoding Initiative (TEI)'s guideline.

Based on the studying of 2,547 proper names, 63 percents of proper names occur with context clues, while 37 percents do not have any context clues. The context clues consist of (1) clue words, (2) spaces and (3) discourse context. Clue words are normally in front of proper names. 161 clue words are found in this study. They can be grouped in according to their semantic features into 19 groups. Each group of clue words can be used to indicate types of proper names. Six types of proper names are identified in this study, namely person, organization, place, affinity, law, and miscellaneous. Furthermore, different clue words tend to have different predictive power in identifying proper names, even when they are in the same group. A space usually occurs together with a clue word. It can appear either immediately before or after the proper name. It is found that 71 percents of proper names occur with a space, and 80 percents of proper names that have clue words do occur with a space also. Most clue words have more predictive power when spacing is taking into consideration. When proper names do not occur with clue words or spaces, discourse context can be used to identify proper names. Discourse context here refers to anaphoric relation or cataphoric relation, but it is mostly anaphoric one. These proper names can be used without any clue words or spaces because it refers to the same discourse entity mentioned before. However, there are 931 proper names that cannot be identified by any context clues. These proper names are either foreign words or well-known names. Therefore, these proper names can be used without context clues. In addition, we also found that proper names can be used as a part of linguistic units that are not proper names.

| | | |
|----------------|-------------|--------------------------|
| Department | Linguistics | Student's signature..... |
| Field of study | Linguistics | Advisor's signature..... |
| Academic year | 2003 | |

กิตติกรรมประกาศ

ผู้วิจัยขอขอบพระคุณ ผู้ช่วยศาสตราจารย์ ดร. วิโรจน์ อรุณมานะกุล อาจารย์ที่ปรึกษาวิทยานิพนธ์ เป็นอย่างสูงที่ได้ช่วยเหลือและให้คำปรึกษาแนะนำแนวทางในทุกขั้นตอนตลอดจนแก้ไขและขัดเกลาวิทยานิพนธ์ฉบับนี้จนสำเร็จลุล่วงลงได้ และผู้วิจัยขอขอบพระคุณ ผู้ช่วยศาสตราจารย์ ดร. สุดาพร ลักษณะนิยานาวิน และ ศาสตราจารย์ ดร.ปราณี กุลละวณิชย์ กรรมการสอบวิทยานิพนธ์ที่ได้ให้คำปรึกษาและเสียสละเวลาเพื่อตรวจแก้ไขวิทยานิพนธ์ฉบับนี้

นอกจากนี้ ผู้วิจัยขอขอบพระคุณคณาจารย์ภาควิชาภาษาศาสตร์ทุกท่านที่ได้มอบโอกาสให้ผู้วิจัยได้เข้ามาเรียนรู้แขนงวิชาภาษาศาสตร์รวมทั้งได้ประสิทธิ์ประสาทความรู้ให้แก่ผู้วิจัย และขอขอบคุณเจ้าหน้าที่ภาควิชาภาษาศาสตร์ทุกคนที่ช่วยอำนวยความสะดวก รวมทั้งขอขอบคุณเพื่อนๆ ภาควิชาภาษาศาสตร์ทุกคนที่ได้ช่วยเหลือและให้กำลังใจในระหว่างการทำวิทยานิพนธ์ฉบับนี้

สุดท้ายนี้ ผู้วิจัยขอขอบคุณกัลยาณมิตรของผู้วิจัยทุกคนที่ได้ช่วยเหลือผู้วิจัยในด้านการงานและในด้านการสนับสนุนให้กำลังใจรวมทั้งคอยกระตุ้นเตือนผู้วิจัยตลอดเวลาของการทำวิทยานิพนธ์ฉบับนี้ และขอขอบพระคุณครอบครัวของผู้วิจัยสำหรับการสนับสนุนค่าใช้จ่ายระหว่างการศึกษา

สารบัญ

| | หน้า |
|--|------|
| บทคัดย่อภาษาไทย | ง |
| บทคัดย่อภาษาอังกฤษ | จ |
| กิตติกรรมประกาศ | ฉ |
| สารบัญ | ช |
| สารบัญตาราง..... | ญ |
| สารบัญภาพ | ฎ |
| บทที่ | |
| 1 บทนำ..... | 1 |
| 1.1 ความเป็นมาและความสำคัญของปัญหา | 1 |
| 1.2 วัตถุประสงค์ | 6 |
| 1.3 สมมุติฐาน | 6 |
| 1.4 ขอบเขตการวิจัย | 6 |
| 1.5 เครื่องมือที่ใช้ในการวิจัย | 7 |
| 1.6 วิธีดำเนินการวิจัย | 7 |
| 1.7 ประโยชน์ที่คาดว่าจะได้รับจากการวิจัย..... | 7 |
| 1.8 โครงร่างของบทต่างๆในวิทยานิพนธ์ | 8 |
| 2 ทบทวนวรรณกรรม..... | 9 |
| 2.1 ลักษณะสำคัญของชื่อเฉพาะภาษาไทย..... | 9 |
| 2.2 ปัญหาการประมวลผลชื่อเฉพาะในงานประมวลผลภาษาธรรมชาติ | 10 |
| 2.3 การศึกษาชื่อเฉพาะภาษาไทยที่ผ่านมา..... | 16 |
| 3 ชื่อเฉพาะและการกำกับชื่อเฉพาะ | 21 |
| 3.1 ที่มาของเกณฑ์นิยาม | 21 |
| 3.2 แนวทางการกำกับชื่อเฉพาะของ TEI..... | 30 |
| 3.2.1 การกำกับข้อมูลในคลังข้อมูลภาษา..... | 30 |
| 3.2.2 แนวทางการกำกับชื่อเฉพาะของ TEI | 31 |

สารบัญ (ต่อ)

| บทที่ | หน้า |
|---|------|
| 4 การจัดสร้างคลังข้อมูลและการวิเคราะห์บริบทบ่งบอก | 39 |
| 4.1 การจัดสร้างคลังข้อมูลภาษา | 39 |
| 4.1.1 ออกแบบโครงสร้างของข้อมูล | 41 |
| 4.1.2 เก็บรวบรวมข้อมูล..... | 42 |
| 4.1.3 กำกับข้อมูล..... | 42 |
| 4.1.3.1 การกำกับชื่อเฉพาะ | 43 |
| 4.1.3.1.1 การวิเคราะห์หารูปภาษาที่เป็นชื่อเฉพาะ | 43 |
| 4.1.3.1.2 แท็กสำหรับการกำกับชื่อเฉพาะ | 44 |
| 4.1.3.2 การกำกับบริบทบ่งบอก | 45 |
| 4.2 การวิเคราะห์บริบทบ่งบอก | 48 |
| 5 ชื่อเฉพาะและบริบทบ่งบอกชื่อเฉพาะ | 50 |
| 5.1 ลักษณะโครงสร้างของรูปภาษาที่เป็นชื่อเฉพาะ | 50 |
| 5.2 คำบ่งบอกชื่อเฉพาะ..... | 52 |
| 5.2.1 คำที่ปรากฏเป็นตัวบ่งบอก (คำบ่งบอก) | 53 |
| 5.2.2 ลักษณะการปรากฏของคำบ่งบอก | 54 |
| 5.2.3 รูปและความหมายของคำบ่งบอก | 55 |
| 5.2.4 การจัดกลุ่มคำบ่งบอกและประเภทชื่อเฉพาะ | 56 |
| 5.2.4.1 การจัดกลุ่มคำบ่งบอก | 56 |
| 5.2.4.2 การจำแนกประเภทของชื่อเฉพาะโดยใช้คำบ่งบอก | 59 |
| 5.2.4.3 ความคาบเกี่ยวของกลุ่มคำบ่งบอกและประเภทชื่อเฉพาะ | 63 |
| 5.2.5 ขีดความสามารถในการบ่งบอกถึงชื่อเฉพาะของคำบ่งบอก..... | 64 |
| 5.2.5.1 แนวทางการวิเคราะห์ขีดความสามารถในการบ่งบอก | 65 |
| 5.2.5.2 ขีดความสามารถของคำบ่งบอก | 66 |
| 5.3 การเว้นวรรค (Space) | 69 |
| 5.4 คำบ่งบอกปรากฏร่วมกับการเว้นวรรค | 70 |
| 5.5 ตัวบ่งบอกระดับปริจเฉท | 74 |
| 5.6 ชื่อเฉพาะที่ไม่ปรากฏบริบทบ่งบอก..... | 76 |
| 5.7 ชื่อเฉพาะปรากฏเป็นองค์ประกอบของรูปภาษาอื่น | 79 |

สารบัญ (ต่อ)

| บทที่ | หน้า |
|---|------|
| 6 สรุปและข้อเสนอแนะ | 82 |
| 6.1 สรุปกระบวนการศึกษาและผลการศึกษหาบริบทบ่งบอกชื่อเฉพาะภาษาไทย | 82 |
| 6.2 ข้อเสนอแนะ | 86 |
| รายการอ้างอิง | 89 |
| ภาคผนวก ก | 94 |
| ภาคผนวก ข | 95 |
| ประวัติผู้เขียนวิทยานิพนธ์ | 102 |

สารบัญตาราง

| ตาราง | หน้า |
|---|------|
| ตารางที่ 5-1 คำบ่งบอกกลุ่มต่างๆ | 58 |
| ตารางที่ 5-2 คำบ่งบอกที่มีสองความหมาย..... | 64 |
| ตารางที่ 5-3 ขีดความสามารถในการบ่งบอกของคำบ่งบอกแต่ละคำ | 67 |
| ตารางที่ 5-4 ค่าพิสัย ค่าเฉลี่ย และค่าเบี่ยงเบนมาตรฐานของขีดความสามารถในการบ่งบอกของ แต่ละกลุ่มคำบ่งบอก..... | 68 |
| ตารางที่ 5-5 จำนวนชื่อเฉพาะที่ปรากฏการเว้นวรรคในลักษณะต่างๆ | 70 |
| ตารางที่ 5-6 จำนวนชื่อเฉพาะที่แจกแจงตามการปรากฏหรือไม่ปรากฏของคำบ่งบอกและการ เว้นวรรค..... | 71 |
| ตารางที่ 5-7 ขีดความสามารถในการบ่งบอกถึงชื่อเฉพาะเมื่อพิจารณาบริบทเว้นวรรคร่วมกับ คำบ่งบอก | 72 |
| ตารางที่ 5-8 จำนวนชื่อเฉพาะแจกแจงตามประเภท..... | 76 |

สารบัญภาพ

| ภาพประกอบ | หน้า |
|--|------|
| รูปที่ 4-1 กระบวนการขั้นตอนจัดสร้างคลังข้อมูลและวิเคราะห์บริบทบ่งบอกถึงชื่อเฉพาะ | 40 |
| รูปที่ 4-2 ตัวอย่างการกำกับคุณสมบัติเพิ่มเติมเพื่อการศึกษาบริบทบ่งบอก | 46 |
| รูปที่ 4-3 ภาพรวมของการกำกับข้อมูลในคลังข้อมูลภาษาของวิทยานิพนธ์..... | 47 |
| รูปที่ 4-4 รายการข้อมูลที่แสดงชื่อเฉพาะพร้อมบริบทของโปรแกรม CU Thai Concordance | 49 |
| รูปที่ 5-1 สัดส่วนของชื่อเฉพาะที่ปรากฏคำบ่งบอก | 53 |
| รูปที่ 5-2 รายการคำที่เป็นตัวบ่งบอกถึงชื่อเฉพาะ | 53 |
| รูปที่ 5-3 ตัวอย่างชื่อเฉพาะจากคลังข้อมูลที่ผ่านการดัดแปลง (ตัดคำบ่งบอก)..... | 59 |
| รูปที่ 5-4 ตัวอย่างชื่อเฉพาะจากคลังข้อมูลที่ปรากฏเหมือนต้นฉบับ..... | 59 |
| รูปที่ 5-5 ผลการจำแนกชื่อเฉพาะออกเป็นประเภทต่างๆ | 62 |
| รูปที่ 5-6 ชีตความสามารถของคำบ่งบอกที่ปรากฏร่วมกับเว้นวรรคในการบ่งบอกถึงชื่อเฉพาะ .. | 73 |
| รูปที่ 5-7 ตัวอย่างตัวบ่งบอกระดับปริจเฉทในข้อความต่อเนื่อง | 75 |

บทที่ 1

บทนำ

1.1 ความเป็นมาและความสำคัญของปัญหา

เทคโนโลยีสารสนเทศโดยเฉพาะคอมพิวเตอร์มีความเกี่ยวข้องกับภาษาอย่างใกล้ชิด ความเกี่ยวข้องประการแรกคือ การใช้คอมพิวเตอร์เป็นเครื่องมือในการประมวลผลและจัดเก็บข้อมูล เช่น ใช้ในการประมวลผลคำ ใช้ในการจัดเก็บข้อมูลปริมาณมากอย่างมีระบบเพื่อความสะดวกรวดเร็วในการนำข้อมูลนั้นมาใช้งาน นอกจากนี้ในยุคปัจจุบันข้อมูลข่าวสารต่างๆได้มีบทบาทสำคัญต่อชีวิตประจำวันของมนุษย์ ซึ่งทำให้มนุษย์มีความต้องการในการบริโภคข้อมูลสูง และต้องการความสะดวกรวดเร็วในการเข้าถึงข้อมูล ทำให้คอมพิวเตอร์กลายเป็นเครื่องมือสำคัญในการตอบสนองความต้องการนี้ ตัวอย่างที่เห็นได้ชัดเจนคือ การเจริญเติบโตอย่างรวดเร็วของระบบข้อมูลข่าวสารบนอินเทอร์เน็ตที่มีข้อมูลปริมาณมากมายจากแหล่งต่างๆทั่วโลก สิ่งที่เกิดขึ้นตามมาคือ มนุษย์ต้องคิดวิธีให้คอมพิวเตอร์จัดการกับปริมาณข้อมูลมหาศาลนี้เพื่อตอบสนองความต้องการในการเข้าใช้ข้อมูลได้อย่างมีประสิทธิภาพ เช่น ความต้องการค้นหาข้อมูล การใช้ข้อมูลติดต่อสื่อสารกันทั่วโลก เป็นต้น ข้อมูลที่กล่าวถึงนี้มนุษย์ได้ถ่ายทอดผ่าน "ภาษา" ซึ่งเป็นเครื่องมือหลักในการติดต่อสื่อสารของมนุษย์ จึงอาจกล่าวได้ว่าข้อมูลก็คือตัวภาษานั้นเอง ความเกี่ยวข้องอีกประการหนึ่งคือภาษาเครื่อง ซึ่งเป็นภาษาที่มนุษย์กำหนดขึ้นเพื่อใช้สื่อสารกับคอมพิวเตอร์ ต้องมีการเรียนรู้ทำความเข้าใจภาษาเครื่องโดยเฉพาะ เมื่อเป็นเช่นนี้จึงเป็นธรรมดาที่มนุษย์พยายามหาวิธีการที่จะทำให้คอมพิวเตอร์เข้าใจภาษาที่มนุษย์ใช้สื่อสารกันอยู่ ความเกี่ยวข้องทั้งสองประการได้แสดงให้เห็นว่า ภาษามนุษย์มีความเกี่ยวข้องกับคอมพิวเตอร์อย่างเห็นได้ชัดเจน จนเกิดความพยายามที่จะศึกษาหาวิธีการทำให้คอมพิวเตอร์สามารถเข้าใจภาษามนุษย์และประมวลผลได้ถูกต้องเหมือนมนุษย์

การศึกษาวิธีการทำให้คอมพิวเตอร์เข้าใจภาษามนุษย์หรือภาษาธรรมชาติคือศาสตร์ที่เรียกว่า การประมวลผลภาษาธรรมชาติ (Natural Language Processing) ซึ่งการทำให้คอมพิวเตอร์เข้าใจภาษามนุษย์นั้นเป็นเรื่องยากมาก ส่วนหนึ่งเพราะภาษามนุษย์เป็นภาษาที่มีความหลากหลายและประกอบด้วยองค์ความรู้หลายส่วนที่มีความซับซ้อนรวมทั้งทุกภาษามีความกำกวมอยู่ไม่มากนักน้อย การศึกษาวิเคราะห์เพื่อทำความเข้าใจคุณสมบัติของรูปภาษา ขบวนการขั้นตอน และกลไกในการประมวลผลภาษามนุษย์ จึงเป็นพื้นฐานที่สำคัญในการนำไปจำลองให้กับ

เครื่องคอมพิวเตอร์ เช่น การทำให้คอมพิวเตอร์สามารถแยกประโยคที่ใช้ในภาษามนุษย์ออกเป็นคำๆ และพิจารณาหน้าที่ของคำเหล่านี้เพื่อทำความเข้าใจความหมายของประโยค จึงทำให้การศึกษาด้านการประมวลผลภาษาธรรมชาตินี้มีการค้นคว้าวิจัยทางด้านภาษาศาสตร์ในรูปแบบใหม่ที่เรียกว่า ภาษาศาสตร์คอมพิวเตอร์ (Computational Linguistics) ผลลัพธ์ที่ได้สามารถนำไปพัฒนางานด้านต่างๆ ที่เกี่ยวข้องกับภาษา เช่น การแปลภาษามนุษย์จากภาษาหนึ่งไปเป็นอีกภาษาหนึ่ง (Machine Translation), การค้นคืนสารสนเทศ (Information Retrieval), การสกัดสารสนเทศ (Information Extraction)

สำหรับภาษาไทย ได้มีความพยายามในการที่จะพัฒนาให้เครื่องคอมพิวเตอร์เข้าใจและสามารถประมวลผลภาษาไทยมานานเป็นทศวรรษ แต่ยังมีปัญหามากมายที่ต้องทำการศึกษาวิจัย เพียงเรื่องของการประมวลผลระดับคำยังพบปัญหามากมาย เนื่องจากภาษาไทยมีลักษณะเป็นภาษาคำโดด คือ คำแต่ละคำจะมีเพียงรูปเดียว ไม่มีการเปลี่ยนแปลงรูปคำหรือเติมวิภัติปัจจัย ลักษณะของคำมีทั้งคำเดี่ยวและคำที่เกิดจากวิธีสร้างคำแบบคำประสมสร้างประโยคโดยนำคำมาเรียงต่อกันไม่มีตัวบ่งชี้ขอบเขตของคำหรือประโยคที่ชัดเจนเหมือนภาษาอังกฤษ อาศัยลำดับการเรียงคำ (word order) ที่ต่างกันประโยคเป็นตัวอย่างที่และความหมายของแต่ละคำ ลักษณะเช่นนี้ทำให้เกิดปัญหาในเรื่องของการตัดแบ่งคำและการระบุแยกแยะประเภทของคำ (word identification) ซึ่งเป็นสิ่งสำคัญในการประมวลผลระดับคำ หน่วยภาษาระดับคำเป็นหน่วยภาษาในระดับพื้นฐานสำหรับการประมวลผลหน่วยภาษาระดับที่ซับซ้อนมากขึ้น เช่น การแจงส่วนประโยค (Sentence Parsing) การตรวจสอบไวยากรณ์ (Grammar Check) เป็นต้น ดังนั้นถ้าเครื่องคอมพิวเตอร์ไม่สามารถระบุขอบเขตคำและแยกแยะคำได้อย่างถูกต้องจะมีผลต่อการประมวลผลภาษาระดับที่ซับซ้อน เช่น มีผลทำให้ความเข้าใจข้อมูลในการแจงส่วนประโยคหรือกระบวนการค้นคืนข้อมูลมีประสิทธิภาพลดลง ปัญหาการประมวลผลระดับคำในภาษาไทยจึงนับเป็นปัญหาที่สำคัญและยังคงต้องการการศึกษาค้นคว้าวิจัยอีกมาก ประเด็นปัญหาหนึ่งซึ่งเป็นประเด็นที่น่าสนใจและยังคงเป็นปัญหาในการวิจัยพัฒนาเกี่ยวกับการตัดแบ่งคำและการระบุแยกแยะประเภทของคำ (word identification) ในภาษาไทยคือการประมวลผลชื่อเฉพาะ (Proper Name)

ชื่อเฉพาะ (Proper Name) เป็นหน่วยภาษาที่จัดว่าเป็นหน่วยหลักหน่วยหนึ่งที่ทำให้การสื่อสารของมนุษย์มีประสิทธิภาพ โดยธรรมชาติของชื่อเฉพาะสามารถเกิดขึ้นได้เสมอและเกิดได้กับทุกสิ่งทุกอย่างขึ้นอยู่กับมนุษย์ว่าให้ความสนใจสิ่งใดเป็นพิเศษและต้องการระบุแยกแยะสิ่งนั้นเพียงสิ่งเดียวออกมาจากสิ่งอื่นๆ อย่างชัดเจน ซึ่งการกำหนดโดยระบุรูปภาษาในลักษณะของชื่อเฉพาะจะง่ายกว่าการระบุโดยบรรยายคุณสมบัติของสิ่งนั้นเพื่อแสดงความเจาะจง เช่น ต้องการ

จะกล่าวถึง "มากาเร็ต แทตเชอร์" การใช้รูปภาษานี้จะทำให้การสื่อสารเป็นไปโดยสะดวกและยังคงมีความชัดเจน เป็นผลให้การสื่อสารมีประสิทธิภาพมากกว่าการกล่าวถึงโดยบรรยายว่า "นายกรัฐมนตรีของประเทศอังกฤษที่ดำรงตำแหน่งในปีคริสต์ศักราช 1989" (McArthur, 1992) ดังนั้นในการใช้ภาษาจริงจึงมีการสร้างชื่อเฉพาะใหม่ตลอดเวลา นอกจากนี้ชื่อเฉพาะในภาษาไทยยังเป็นรูปภาษาที่มีโครงสร้างหลากหลายรูปแบบ ตัวอย่างเช่น

(จังหวัด)ตรัง (พรรค)ประชาธิปัตย์ - คำเดี่ยว

(อำเภอ)กำแพงแสน (พรรค)เสรีธรรม - คำประสม

(กรุง)กัวลาลัมเปอร์ นิวยอร์ก - ทับศัพท์ภาษาต่างประเทศ

(พรรค)ไทยรักไทย (พรรค)ชาติพัฒนา - มีโครงสร้างคล้ายประโยค เช่น ไทย(ประธาน)-รัก(กริยา)-ไทย(กรรม)

โดยรูปแบบต่างๆเหล่านี้พบในคำนามทั่วไปเช่นกัน รวมทั้งในภาษาไทยไม่มีสิ่งที่บ่งชี้ได้ว่ารูปภาษานั้นเป็นชื่อเฉพาะดังเช่นบางภาษา เช่น ภาษาอังกฤษที่ใช้การขึ้นต้นด้วยตัวอักษรพิมพ์ใหญ่เป็นตัวบ่งชี้

การที่ชื่อเฉพาะในภาษาไทยมีลักษณะของรูปภาษาเช่นนี้ ทำให้การประมวลผลชื่อเฉพาะในภาษาไทยกลายเป็นปัญหาหลักอันหนึ่งในการวิจัยพัฒนาระบบประมวลผลภาษาธรรมชาติ ปัญหาลักษณะหนึ่งในการประมวลผลชื่อเฉพาะคือ ปัญหาของคำที่คอมพิวเตอร์ไม่รู้จัก (unknown word) โดยปกติแล้วในการประมวลผลระดับคำมักจะใช้วิธีให้คอมพิวเตอร์รู้จักคำต่างๆผ่านทางคลังคำศัพท์หรือพจนานุกรม ซึ่งไม่สามารถจะครอบคลุมคำที่เกิดขึ้นใหม่ได้อย่างครบถ้วน ดังนั้นชื่อเฉพาะคำใหม่ๆจึงกลายเป็นคำที่คอมพิวเตอร์ไม่รู้จักด้วย ดังเช่นในบทความวิจัยเรื่อง "Automatic Thai Unknown Word Recognition" (Kawtrakul et al., 1997) ได้แสดงให้เห็นว่าประเภทคำที่พบมากที่สุดในกลุ่มคำที่คอมพิวเตอร์ไม่รู้จักในงานวิจัยนี้คือ ชื่อเฉพาะ (proper noun) โดยพบประมาณ 43 เปอร์เซ็นต์ นอกจากนี้ในการวิจัยยังพบว่าคอมพิวเตอร์มีความสับสนเกี่ยวกับการประมวลผลคำนามทั่วไป (common noun) ที่คอมพิวเตอร์รู้จักซึ่งไปประกอบเป็นส่วนหนึ่งของชื่อเฉพาะที่มีโครงสร้างเป็นแบบคำประสม คณะผู้วิจัยยังชี้ให้เห็นถึงความสำคัญของการแก้ปัญหาคำที่คอมพิวเตอร์ไม่รู้จักว่า ผลการศึกษาคำที่คอมพิวเตอร์ไม่รู้จักนี้สามารถนำไปใช้พัฒนาระบบการกำกับข้อมูลแบบอัตโนมัติ (automatic tagger system) ที่มีความเกี่ยวข้องโดยตรงต่อประสิทธิภาพการประมวลผลภาษาธรรมชาติ เช่น กรณีที่ระบบกำกับข้อมูลอัตโนมัติไม่สามารถจัดการกับคำที่คอมพิวเตอร์ไม่รู้จักได้ จะส่งผลให้ในการแจงส่วนประโยคที่ประกอบด้วยคำที่คอมพิวเตอร์ไม่รู้จักมีความผิดพลาด ทำให้ประสิทธิภาพของการทำความเข้าใจข้อความหรือกระบวนการค้นคืนข้อมูลด้อยลง

ตัวอย่างที่เห็นได้ชัดเจนและพบได้บ่อยซึ่งแสดงให้เห็นถึงปัญหาในการประมวลผลชื่อเฉพาะคือ การใช้บริการค้นหาข้อมูล (search engine) ที่เป็นรูปแบบหนึ่งของงานค้นคืนสารสนเทศ (Information Retrieval) ตามความต้องการของผู้ใช้ โดยผู้ใช้งานจะระบุข้อมูลที่ต้องการให้ค้นคืนในลักษณะต่างๆ เช่น ระบุเป็นคำ ระบุเป็นวลี จากนั้นระบบบริการค้นหาข้อมูลจะมองหาเอกสารที่มีคำหรือวลีนั้นโดยใช้การเทียบดูรูปภาษาแล้วแสดงเอกสารที่มีรูปภาษานั้น ข้อมูลที่เป็นผลลัพธ์โดยส่วนใหญ่จะมีข้อมูลส่วนที่ไม่ต้องการปรากฏอยู่ไม่มากก็น้อย ซึ่งนอกจากจะไม่ได้ข้อมูลที่ถูกต้องตามความต้องการของผู้ใช้แล้วยังมีผลให้ต้องเสียเวลาในการค้นคืนข้อมูลนาน ดังเช่นตัวอย่างข้อมูลต่อไปนี้ซึ่งเป็นผลลัพธ์ที่ได้จากการใช้บริการค้นหาข้อมูลในอินเทอร์เน็ต (ตัวอย่างจาก <http://www.siamguru.com>)

ก. เมื่อค้นคำว่า "แก้ว" พบว่าได้ข้อมูลที่ไม่ตรงตามความต้องการประมาณ 56 เปอร์เซนต์ (ข้อมูลที่ไม่ตรงตามความต้องการ / ข้อมูลที่ค้นคืนได้ทั้งหมด เท่ากับ 34 / 61) ซึ่งข้อมูลที่ไม่ตรงตามความต้องการนี้ส่วนใหญ่เป็นชื่อเฉพาะที่มีคำว่า "แก้ว" เป็นส่วนประกอบ เช่น

| | |
|----------------|---------------------------------|
| หลวงพ่อกแก้ว | ทรายแก้ววิลล่า (ซีรีสอร์ต) |
| มูลนิธิดวงแก้ว | دنۇفل گەۋگاجۇجنى |
| จังหวัดสระแก้ว | โรงพยาบาลตรวิศด้ากาดสวนแก้ว |
| วัดพระแก้วมรกต | บริษัทอุตสาหกรรมแก้วคงถาวรจำกัด |

ข. เมื่อค้นคำว่า "กรุงเทพ" ได้ข้อมูลไม่ตรงตามความต้องการประมาณ 26 เปอร์เซนต์ (ข้อมูลที่ไม่ตรงตามความต้องการ / ข้อมูลที่ค้นคืนได้ทั้งหมด เท่ากับ 38 / 147) ข้อมูลส่วนนี้เป็นชื่อเฉพาะอื่นๆที่มีคำว่า "กรุงเทพ" เป็นส่วนประกอบ เช่น

| | |
|--------------------------|---------------------------------|
| บริษัทกรุงเทพเหล็กกล้า | กรุงเทพไอทีมาร์ท (ซีอีอาน) |
| บริษัทการบินกรุงเทพ | องค์การขนส่งมวลชนกรุงเทพ(ขสมก.) |
| โรงเรียนกรุงเทพคริสเตียน | |

ค. ค้นเกี่ยวกับ "(จังหวัด) น่าน" พบข้อมูลที่ไม่ตรงตามความต้องการประมาณ 32 เปอร์เซนต์ (ข้อมูลที่ไม่ตรงตามความต้องการ / ข้อมูลที่ค้นคืนได้ทั้งหมด เท่ากับ 37 / 117) โดยข้อมูลที่ไม่ตรงตามความต้องการเป็นคำนามทั่วไป เช่น น่านฟ้า น่านน้ำ

ในงานด้านการสกัดสารสนเทศ (Information Extraction) สำหรับภาษาไทย องค์ประกอบที่สำคัญซึ่งเป็นพื้นฐานของการพัฒนางานด้านนี้คือ ส่วนของการทำให้คอมพิวเตอร์รู้จักองค์ประกอบต่างๆของภาษาเพื่อนำไปเชื่อมโยงกับกระบวนการส่วนอื่นๆของระบบการประมวลผลภาษา (natural language processing) โดยการศึกษาวิจัยหลายๆขององค์ประกอบพื้นฐานส่วนนี้คือ เรื่องของกระบวนการตัดคำ (word segmentation) ปัญหาที่ยุ่งยากอันหนึ่งที่พบ

ในระหว่างการตัดคำก็คือ ข้อมูลชื่อเฉพาะ ที่เกิดปัญหาว่าระบบไม่เคยเห็นชื่อเฉพาะนี้มาก่อนในพจนานุกรมหรือในระหว่างการฝึกสอนคลังข้อมูลและไม่สามารถแยกแยะออกจากคำศัพท์ทั่วไปงานในส่วนของการระบุแยกแยะชื่อเฉพาะนับว่าเป็นสิ่งที่สำคัญสำหรับการสกัดสารสนเทศ เพราะชื่อเฉพาะเป็นหน่วยภาษาที่ใช้บอกถึงข้อมูลสำคัญสำหรับการประมวลผลให้ได้ผลลัพธ์ที่มีประสิทธิภาพ (Sukhahuta and Smith, 2001)

ส่วนการศึกษาค้นคว้าด้านการแปลภาษาด้วยเครื่องคอมพิวเตอร์ (Machine Translation) นั้น แน่นนอนว่าภาษาในระดับคำของภาษาไทยยังคงเป็นปัญหาพื้นฐานที่สำคัญเช่นกัน ในการศึกษาต้องมีการวิเคราะห์ภาษาในเรื่องของ การตัดแบ่งคำ การระบุแยกแยะประเภทของคำ ความกำกวมในเรื่องความหมายของคำ เป็นต้น ดังนั้นในเรื่องของการตัดแบ่งคำและการระบุแยกแยะประเภทของคำย่อมจะพบปัญหาที่เกี่ยวกับการระบุชื่่ว่าเป็นชื่อเฉพาะ เมื่อเครื่องคอมพิวเตอร์ไม่รู้จำก็ย่อมมีผลต่อการแปลทั้งในด้านวากยสัมพันธ์และความหมายที่ถูกต้อง ตัวอย่างเช่นเมื่อพบข้อมูล "พรรคไทยรักไทย" โปรแกรมแปลภาษาไม่ควรพยายามที่จะแปลคำว่า "ไทย" หรือ "รัก" แต่ควรแปลเพียงคำว่า "พรรค" ไปเป็นภาษาเป้าหมายโดยเลือกคำที่เหมาะสมที่สามารถบ่งชี้ได้ว่าเป็นชื่อเฉพาะในภาษาเป้าหมายนั้นและคงส่วนที่เหลือไว้อย่างเดิมโดยไม่ต้องมีการแปลความหมายใดๆ

จะเห็นได้ว่าชื่อเฉพาะ (Proper Name) ในภาษาไทย เป็นประเด็นที่ยังเป็นปัญหาและมีความสำคัญต่อการประมวลผลภาษาในการพัฒนางานด้านต่างๆเกี่ยวกับภาษาที่ปรากฏในคอมพิวเตอร์ โดยเฉพาะงานด้านค้นคืนสารสนเทศ (Information Retrieval), การสกัดสารสนเทศ (Information Extraction) หรือ การแปลภาษาด้วยเครื่องคอมพิวเตอร์ (Machine Translation) ซึ่งเข้ามามีบทบาทในการสนองความต้องการบริโภคข้อมูลข่าวสารของมนุษย์มากขึ้นเรื่อยๆ จึงต้องมีการพัฒนาระบบประมวลผลให้มีประสิทธิภาพทั้งในด้านความถูกต้องด้านภาษาและด้านความแม่นยำรวดเร็ว ปัญหาที่กล่าวถึงนั้นเนื่องมาจากโดยธรรมชาติของชื่อเฉพาะแล้วมีแนวโน้มที่จะเกิดขึ้นใหม่อยู่ตลอดเวลาในการใช้ภาษา การใช้วิธีการสอนให้คอมพิวเตอร์รู้จักชื่อเฉพาะโดยการใช้คลังคำศัพท์หรือพจนานุกรมจึงไม่สามารถครอบคลุมชื่อเฉพาะที่เกิดขึ้นใหม่ ปัญหาที่เกี่ยวข้องอีกประการหนึ่งคือปัญหาจากลักษณะโครงสร้างของภาษาไทยและลักษณะโครงสร้างของชื่อเฉพาะเองที่มีรูปแบบในลักษณะเดียวกับค่านามทั่วไป เป็นผลให้ระบบคอมพิวเตอร์ไม่สามารถแยกแยะชื่อเฉพาะออกจากค่านามทั่วไปได้ ผลจากการที่คอมพิวเตอร์ไม่รู้จำและไม่สามารถแยกแยะชื่อเฉพาะได้ ทำให้ความสามารถในการประมวลผลภาษาโดยรวมของเครื่องคอมพิวเตอร์ด้อยประสิทธิภาพลง การศึกษาหาวิธีการที่จะทำให้คอมพิวเตอร์สามารถแยกแยะและบ่งชี้ได้ว่าหน่วยภาษานั้นเป็นชื่อเฉพาะได้จึงเป็นเรื่องสำคัญเรื่องหนึ่งในการประมวลผลภาษาไทย โดยปกติ

แล้วการประมวลผลภาษาของมนุษย์ทั้งในเรื่องของการวิเคราะห์แยกแยะและทำความเข้าใจหน่วยภาษาต่างๆจะต้องอาศัยเรื่องของบริบท (Context) ในลักษณะต่างๆเข้ามาเป็นตัวช่วยที่สำคัญที่ทำให้สามารถประมวลผลภาษาลักษณะต่างๆที่เกิดขึ้นในการใช้ภาษาจริงได้อย่างไม่มีที่สิ้นสุด ดังนั้นในการที่จะทำให้คอมพิวเตอร์สามารถแยกแยะและบ่งชี้ชื่อเฉพาะได้อย่างครอบคลุม ผู้วิจัยจึงเห็นว่าเรื่องของบริบท (Context) น่าจะใช้เป็นตัวบ่งชี้ไปที่ชื่อเฉพาะได้ โดยใช้แนวทางภาษาศาสตร์เข้ามาวิเคราะห์ตัวภาษาหาบริบทบ่งบอกการปรากฏของชื่อเฉพาะในลักษณะที่สอดคล้องกับการนำไปใช้เป็นพื้นฐานทางภาษาของการพัฒนางานด้านต่างๆทางภาษาศาสตร์คอมพิวเตอร์

1.2 วัตถุประสงค์

1. วิเคราะห์หารูปภาษาที่สามารถใช้เป็นตัวบ่งบอกชื่อเฉพาะในภาษาไทย
2. จำแนกประเภทตัวบ่งบอกชื่อเฉพาะตามเกณฑ์ทางความหมาย
3. วิเคราะห์บริบทอื่นๆที่สามารถใช้เป็นตัวบ่งบอกชื่อเฉพาะนอกเหนือไปจากตัวบ่งบอกที่อยู่ติดกับชื่อเฉพาะ
4. ประเมินขีดความสามารถของตัวบ่งบอกและบริบทในการบ่งบอกชื่อเฉพาะ

1.3 สมมุติฐาน

1. ตัวบ่งบอกชื่อเฉพาะส่วนใหญ่เป็นรูปภาษาที่ปรากฏติดกับชื่อเฉพาะ
2. ประเภทของตัวบ่งบอกชื่อเฉพาะสัมพันธ์กับประเภทของชื่อเฉพาะ
3. บริบทในระดับปริจเฉทสามารถใช้บ่งบอกชื่อเฉพาะได้
4. ตัวบ่งบอกชื่อเฉพาะแต่ละประเภทมีขีดความสามารถในการบ่งบอกไม่เท่ากัน

1.4 ขอบเขตการวิจัย

1. ศึกษาบริบทบ่งบอกชื่อเฉพาะจากข้อความภาษาไทยที่อยู่ในรูปตัวเขียน
2. ศึกษาบริบทบ่งบอกกับชื่อเฉพาะที่อยู่ในกรอบของเกณฑ์นิยาม (definitional criteria) ที่กำหนด (ดูหัวข้อที่ 3.1)
3. ศึกษาบริบทบ่งบอกจากคลังข้อมูลภาษาที่พบชื่อเฉพาะจำนวนอย่างน้อย 2,000 ชื่อ

1.5 เครื่องมือที่ใช้ในการวิจัย

1. โปรแกรม WinHTTrack Website Copier
พัฒนาโดย Xavier Roche and Other Contributors, Free Software Foundation, Inc.
2. โปรแกรม Dehtml และ CU Thai Concordance
พัฒนาโดย Wirote Aroonmanakun: Department of Linguistics, Chulalongkorn University
3. ภาษากำกับชื่อเฉพาะ โดย Text Encoding Initiative (TEI)
4. คลังข้อมูลภาษา จัดสร้างโดยผู้วิจัย
5. โปรแกรมภาษา Perl ของบริษัท Active Perl

1.6 วิธีดำเนินการวิจัย

1. ศึกษาโน้ตค้นทางภาษาศาสตร์เกี่ยวกับชื่อเฉพาะและงานวิจัยที่ศึกษาเกี่ยวกับชื่อเฉพาะ
2. กำหนดเกณฑ์นิยาม (definitional criteria) สำหรับการวิเคราะห์ว่ารूपภาษาส่วนใดในคลังข้อมูลภาษาเป็นชื่อเฉพาะที่จะศึกษาหาบริบทบ่งบอก
3. ศึกษาวรรณกรรมเกี่ยวกับภาษากำกับข้อมูล (markup language) ที่ใช้สำหรับกำกับชื่อเฉพาะเพื่อประยุกต์ใช้หน่วยกำกับ (Tags) กำกับชื่อเฉพาะในคลังข้อมูลที่จัดสร้างขึ้นและประยุกต์ใช้หน่วยกำกับ (Tags) เป็นเครื่องมือตัวหนึ่งในการศึกษาวิเคราะห์หาบริบทบ่งบอก
4. จัดสร้างคลังข้อมูล ตามขั้นตอนต่อไปนี้
 - 4.1. ออกแบบโครงสร้างข้อมูล
 - 4.2. เก็บรวบรวมข้อมูลตามโครงสร้างที่ออกแบบไว้
 - 4.3. จัดเตรียมข้อมูลให้อยู่ในรูปแบบที่มีข้อความเพียงอย่างเดียว (plain text)
 - 4.4. ใช้เกณฑ์นิยามวิเคราะห์หารูปภาษาที่เป็นชื่อเฉพาะและกำกับรูปภาษานั้นด้วยหน่วยกำกับ (Tags) ที่ประยุกต์จากมาตรฐานในการกำกับข้อมูลของ TEI
5. ศึกษาวิเคราะห์หาบริบทบ่งบอกจากรายการผลลัพธ์ของโปรแกรม CU Thai Concordance และจากข้อความในคลังข้อมูลที่จัดสร้างขึ้น

1.7 ประโยชน์ที่คาดว่าจะได้รับการวิจัย

1. ใช้เป็นพื้นฐานทางภาษาสำหรับการพัฒนางานทางด้านภาษาศาสตร์คอมพิวเตอร์
2. เป็นแนวทางในการศึกษาภาษาศาสตร์โดยใช้คลังข้อมูลภาษา

1.8 โครงร่างของบทต่างๆในวิทยานิพนธ์

ในบทที่ 2 ผู้วิจัยได้ศึกษาทบทวนว่าชื่อเฉพาะภาษาไทยมีลักษณะเช่นใดและมีผลทำให้การประมวลผลชื่อเฉพาะภาษาไทยเกิดปัญหาอย่างไรในงานประมวลผลภาษาธรรมชาติ (หัวข้อที่ 2.1 และหัวข้อที่ 2.2) รวมถึงทบทวนแนวคิดวิธีการศึกษาชื่อเฉพาะภาษาไทยที่ผ่านมา (หัวข้อที่ 2.3) ในบทที่ 3 ผู้วิจัยศึกษาทบทวนเกี่ยวกับมโนทัศน์ของชื่อเฉพาะเพื่อนำไปใช้ในขั้นตอนของการจัดสร้างคลังข้อมูล ซึ่งมโนทัศน์ชื่อเฉพาะดังกล่าวเป็นที่มาของเกณฑ์นิยามชื่อเฉพาะของวิทยานิพนธ์ฉบับนี้ (หัวข้อที่ 3.1) รวมทั้งทบทวนเกี่ยวกับแนวทางการกำกับข้อมูลที่จะนำมาใช้สำหรับขั้นตอนการจัดสร้างคลังข้อมูลเช่นกัน โดยแนวทางการกำกับชื่อเฉพาะที่ใช้ในวิทยานิพนธ์นี้เป็นแนวทางการกำกับชื่อเฉพาะของ TEI (Text Encoding Initiative) (หัวข้อที่ 3.2) บทที่ 4 ในหัวข้อที่ 4.1 จะนำเสนอขั้นตอนวิธีในการจัดสร้างคลังข้อมูลภาษาของวิทยานิพนธ์นี้และหัวข้อที่ 4.2 นำเสนอขอบวนวิธีวิเคราะห์บริบทบ่งบอกชื่อเฉพาะจากคลังข้อมูลภาษาที่สร้างขึ้น

บทที่ 5 เป็นการนำเสนอและอภิปรายผลการศึกษาที่พบซึ่งจะประกอบด้วยประเด็นเกี่ยวกับโครงสร้างของรูปภาษาที่เป็นชื่อเฉพาะ โดยกล่าวไว้ในหัวข้อที่ 5.1 ต่อมาเป็นประเด็นเกี่ยวกับบริบทบ่งบอกที่สามารถบ่งบอกถึงชื่อเฉพาะภาษาไทยโดยกล่าวไว้ในหัวข้อที่ 5.2 หัวข้อที่ 5.3 หัวข้อที่ 5.4 และหัวข้อที่ 5.5 นอกจากนี้จะนำเสนอเกี่ยวกับชื่อเฉพาะซึ่งไม่พบบริบทบ่งบอกในหัวข้อที่ 5.6 รวมทั้งชื่อเฉพาะอีกส่วนหนึ่งที่พบว่าปรากฏเป็นส่วนประกอบหนึ่งของรูปภาษาอื่นในหัวข้อที่ 5.7 และบทที่ 6 เป็นบทสุดท้าย กล่าวสรุปกระบวนการศึกษาและผลการศึกษหาบริบทบ่งบอกชื่อเฉพาะภาษาไทย (หัวข้อที่ 6.1) และนำเสนอแนวทางการการนำบริบทบ่งบอกที่พบในการศึกษาครั้งนี้ไปประยุกต์ใช้ในการประมวลผลชื่อเฉพาะของงานทางด้านภาษาศาสตร์คอมพิวเตอร์รวมถึงแนวทางการศึกษาพัฒนาเพิ่มเติมต่อไป (หัวข้อที่ 6.2)

บทที่ 2

ทบทวนวรรณกรรม

วรรณกรรมต่างๆที่เกี่ยวข้องซึ่งผู้วิจัยได้ทำการทบทวนไว้จะประกอบด้วยสองส่วนหลักคือ ส่วนแรกจะเกี่ยวข้องกับแนวคิดประเด็นต่างๆในการศึกษาวิจัยชื่อเฉพาะและส่วนที่สองเกี่ยวข้องกับการนำไปใช้ในการจัดสร้างคลังข้อมูลภาษาของวิทยานิพนธ์ฉบับนี้ ซึ่งจะกล่าวถึงรายละเอียดของแต่ละส่วนไว้ในสองบท โดยในบทนี้จะเป็นรายละเอียดต่างๆเกี่ยวกับแนวคิดประเด็นต่างๆในการศึกษาวิจัยชื่อเฉพาะที่ประกอบด้วยหัวข้อที่ 2.1 ประเด็นเกี่ยวกับลักษณะสำคัญของชื่อเฉพาะภาษาไทย หัวข้อที่ 2.2 ประเด็นปัญหาของการประมวลผลชื่อเฉพาะในงานประมวลผลภาษาธรรมชาติ และหัวข้อที่ 2.3 การศึกษาเกี่ยวกับชื่อเฉพาะภาษาไทยที่ผ่านมา ส่วนรายละเอียดเนื้อหาในประเด็นที่เกี่ยวข้องกับการนำไปใช้ในการสร้างคลังข้อมูลภาษาจะนำไปกล่าวไว้ในบทที่ 3

2.1 ลักษณะสำคัญของชื่อเฉพาะภาษาไทย

ชื่อเฉพาะเป็นหน่วยภาษาที่จัดว่าเป็นหน่วยหลักหน่วยหนึ่งที่ทำให้การสื่อสารของมนุษย์มีประสิทธิภาพและชื่อเฉพาะเป็นหน่วยภาษาที่มีคุณสมบัติเฉพาะตัวที่มีความสำคัญต่อการประมวลผลภาษาธรรมชาติ ซึ่งคุณสมบัติอันหนึ่งที่เป็นคุณสมบัติร่วมในทุกภาษาคือ ชื่อเฉพาะเกิดขึ้นได้เสมอและเกิดได้กับทุกสิ่งทุกอย่าง ส่วนคุณสมบัติสำคัญที่พบในชื่อเฉพาะภาษาไทยเป็นคุณสมบัติเกี่ยวกับลักษณะโครงสร้างของรูปภาษาที่เป็นชื่อเฉพาะที่มีความซับซ้อนสำหรับการประมวลผลด้วยเครื่องคอมพิวเตอร์ โดยลักษณะสำคัญดังกล่าวของชื่อเฉพาะภาษาไทยโดยรวมแล้วสรุปได้ดังนี้คือ

- (1) โดยธรรมชาติของชื่อเฉพาะสามารถเกิดขึ้นได้เสมอและเกิดได้กับทุกสิ่งทุกอย่างขึ้นอยู่กับมนุษย์ว่าให้ความสนใจสิ่งใดเป็นพิเศษและต้องการระบุแยกแยะสิ่งนั้นเพียงสิ่งเดียวออกมาจากสิ่งอื่นๆอย่างชัดเจน ซึ่งการกำหนดโดยระบุภาษาในลักษณะของชื่อเฉพาะจะง่ายกว่าการระบุโดยบรรยายคุณสมบัติของสิ่งนั้นเพื่อแสดงความเจาะจง เช่น ต้องการจะกล่าวถึง "มากาเร็ต แทตเชอร์" การใช้รูปภาษานี้จะทำให้การสื่อสารเป็นไปโดยสะดวกและยังคงมีความชัดเจน เป็นผลให้การสื่อสารมีประสิทธิภาพมากกว่าการกล่าวถึงโดยบรรยายว่า "นายกรัฐมนตรีของประเทศอังกฤษที่ดำรงตำแหน่งในปีคริสตศักราช 1989" (McArthur, 1992) ดังนั้นในการใช้ภาษาจริงจึงมีการสร้างชื่อเฉพาะใหม่ตลอดเวลา

- (2) ไม่มีตัวบ่งชี้สำหรับระบุแสดงว่ารูปภาษานั้นเป็นชื่อเฉพาะ เหมือนที่มีการกำหนดไว้ในบางภาษา เช่น ภาษาอังกฤษ ที่ใช้การขึ้นต้นด้วยตัวอักษรพิมพ์ใหญ่เป็นตัวบ่งชี้
- (3) เป็นรูปภาษาที่มีโครงสร้างหลากหลายรูปแบบ เช่น คำเดี่ยว, คำประสม, ทับศัพท์ ภาษาต่างประเทศ, โครงสร้างคล้ายประโยค ซึ่งรูปแบบต่างๆเหล่านี้พบในคำนามทั่วไปเช่นกัน ดังตัวอย่างต่อไปนี้
- (จังหวัด)ตรัง (พรรค)ประชาธิปไตย - คำเดี่ยว
 - (อำเภอ)กำแพงแสน (พรรค)เสรีธรรม - คำประสม
 - (กรุง)กัวลาลัมเปอร์ นิวยอร์ก - ทับศัพท์ภาษาต่างประเทศ
 - (พรรค)ไทยรักไทย (พรรค)ชาติพัฒนา - มีโครงสร้างคล้ายประโยค เช่น ไทย(ประธาน)-รัก(กริยา)-ไทย(กรรม)
- (4) ชื่อเฉพาะภาษาไทยสามารถสร้างขึ้นจากคำที่มีใช้อยู่แล้วในภาษา เช่น "สอน" นำไปสร้างเป็นชื่อเฉพาะ "นายสอน" (พระยาอุปกิตศิลปสาร, 2544: 71), "ต้นพะยอม" นำไปสร้างเป็นชื่อเฉพาะ "นางสาวพะยอม" (กำชัย ทองหล่อ, 2525: 269), "เอกภาพ" (หรือ เอก+ภาพ) นำไปสร้างเป็นชื่อเฉพาะ "พรรคเอกภาพ" (เปรมจิต ชนะวงศ์, 2538: 193) เป็นต้น ซึ่งการสร้างชื่อเฉพาะขึ้นมาจะใช้คำจำนวนเท่าใดก็ได้ไม่มีการกำหนดกฎเกณฑ์ไว้

โดยลักษณะต่างๆของชื่อเฉพาะดังกล่าวมีผลทำให้การประมวลผลชื่อเฉพาะกลายเป็นปัญหาที่ยู่ยากในการประมวลผลภาษาไทย ซึ่งจะชี้ให้เห็นถึงปัญหาที่เกิดจากชื่อเฉพาะในการพัฒนาวิจัยการประมวลผลภาษาไทยที่ผ่านมาในหัวข้อถัดไป

2.2 ปัญหาของการประมวลผลชื่อเฉพาะในงานประมวลผลภาษาธรรมชาติ

การพัฒนางานด้านต่างๆทางคอมพิวเตอร์ที่เกี่ยวข้องกับภาษา ดังเช่นงานด้านการค้นคืนสารสนเทศ (Information Retrieval), การสกัดสารสนเทศ (Information Extraction) หรือการแปลภาษาด้วยเครื่อง (Machine Translation) ซึ่งเป็นงานที่กำลังได้รับความสนใจในปัจจุบันนั้น ล้วนแต่เป็นงานที่ต้องอาศัยความรู้ทางภาษาเป็นองค์ประกอบจำเป็นสำหรับการพัฒนาระบบประมวลผล โดยเฉพาะอย่างยิ่งสำหรับภาษาที่ไม่มีตัวบ่งชี้ขอบเขตของคำหรือประโยคที่ชัดเจนอย่างเช่นภาษาไทย

การค้นคืนสารสนเทศเป็นงานที่เกี่ยวข้องกับการค้นหาข้อมูลหรือเอกสารที่มีคำวลี หรือสายอักขระใด ๆ ที่ต้องการ ภายในระบบของงานด้านนี้จะมียังมีองค์ประกอบของฐานข้อมูลที่ประกอบด้วยเอกสารข้อมูลต่างๆ ซึ่งการพัฒนาให้ระบบค้นคืนสามารถให้ผลลัพธ์ที่มีประสิทธิภาพ

มากกว่าการเทียบเพียงสายอักขระให้ตรงกับคำที่ต้องการค้นจะต้องอาศัยการประมวลผลภาษาธรรมชาติ โดยการพัฒนาองค์ประกอบแรกสำหรับภาษาไทยคือ ทำการตัดคำในข้อความต่อเนื่องของแต่ละเอกสารในฐานข้อมูล เพราะการมีความสามารถในการรู้จำคำจะทำให้สามารถจัดการกับข้อความได้อย่างถูกต้องและสะดวก ผลลัพธ์จากการประมวลผลภาษาจะเป็นฐานความรู้สำหรับการสร้างองค์ประกอบส่วนที่เป็นขบวนการวิธีค้นคืน เช่น นำคำที่เป็นคำสำคัญ (keyword) ของแต่ละเอกสารไปสร้างรายการดัชนี (index) โดยคำที่เป็นดัชนีแต่ละคำจะชี้ไปยังรายการเอกสารที่เกี่ยวข้องกับคำนั้นทำให้ระบบสามารถค้นหาข้อมูลที่ต้องการได้

ส่วนการสกัดสารสนเทศและการแปลภาษาด้วยเครื่องยังเป็นงานที่ต้องอาศัยการประมวลผลภาษาที่ซับซ้อนมากขึ้น ต้องเกี่ยวข้องกับการประมวลผลภาษาระดับที่สูงกว่าคำกล่าวคือการสกัดสารสนเทศเป็นงานที่เกี่ยวกับการดึงใจความสำคัญจากตัวบทโดยการมองผ่านเข้าไป (skim) ในแหล่งรวมข้อมูลขนาดใหญ่แล้วจับใจความสำคัญที่สอดคล้องกับขอบเขตข้อมูลที่ระบุไว้ออกมาในรูปแบบวลี ประโยค หรือย่อหน้า การทำงานกับข้อมูลภาษาไทยโดยอาศัยเพียงกลไกที่อยู่บนพื้นฐานของการเทียบสายอักขระหรือกฎนั้น ไม่เพียงพอและไม่สามารถนำไปใช้ได้โดยตรงเนื่องจากภาษาไทยเป็นภาษาที่มีความกำกวมของโครงสร้างภาษามาก จึงต้องรวมขบวนการประมวลผลภาษาธรรมชาติเข้ามาเป็นองค์ประกอบหนึ่งของระบบเพื่อจัดการกับปัญหาความกำกวมของภาษา ซึ่งจะทำให้ขบวนการวิธีในการสกัดสารสนเทศมีประสิทธิภาพดีขึ้นมากคือ สามารถประมวลได้ข้อมูลที่มีความหมายเป็นจำนวนมากขึ้น ดังนั้นการพัฒนาวิจัยระบบสกัดสารสนเทศส่วนใหญ่จึงมุ่งไปที่การศึกษาเรื่องของการประมวลผลภาษาธรรมชาติ เช่น การตัดคำ การกำกับหมวดคำ การระบุถึงคำที่คอมพิวเตอร์ไม่รู้จัก (unknown word) รวมถึงการประมวลผลภาษาระดับที่ซับซ้อนกว่าคำ ก่อนที่จะเข้าไปทำในส่วนของการสกัดสารสนเทศ (extraction engine) ซึ่งการวิเคราะห์ระดับคำได้รับการพิจารณาว่าเป็นส่วนที่สำคัญสำหรับงานค้นคืนสารสนเทศและงานสกัดสารสนเทศในภาษาไทยรวมทั้งภาษาในเอเชียที่ไม่มีตัวแสดงขอบเขตของคำ เพราะประสิทธิภาพของการค้นหาคำศัพท์ (term) จะขึ้นอยู่กับขบวนการตัดคำอย่างมาก

ส่วนการแปลภาษาด้วยเครื่องมีลักษณะเป็นการพยายามเลียนแบบการแปลภาษาของมนุษย์ จึงเป็นงานที่เกี่ยวข้องกับการประมวลผลภาษาในระดับสูงคือ ระบบแปลภาษาต้องมีความสามารถในการประมวลผลตั้งแต่ภาษาระดับคำทั้งด้านวากยสัมพันธ์และความหมายเพื่อนำไปสู่การประมวลผลภาษาระดับที่ซับซ้อนขึ้นไปเป็นลำดับ จนสามารถแปลจากภาษาต้นฉบับไปเป็นภาษาเป้าหมายได้ โดยพื้นฐานที่สำคัญอันดับแรกของการประมวลผลระดับคำในภาษาไทยคือการแยกแยะคำและการรู้จำถึงบทบาทหน้าที่ของคำ เราสามารถกล่าวได้ว่า

การศึกษาพัฒนาก้าวแรกสำหรับการนำเข้าสู่ระบบแปลภาษาด้วยเครื่องก็คือ การตัดคำและการกำกับหมวดคำ

จะเห็นได้ว่าองค์ประกอบพื้นฐานที่จำเป็นสำหรับการประมวลผลภาษาธรรมชาติในการพัฒนางานด้านต่างๆที่เกี่ยวข้องกับภาษาไทยคือ เรื่องของการตัดคำและระบุแยกแยะประเภทของคำ (word identification) ซึ่งวิธีการตัดคำและกำกับหมวดคำภาษาไทยได้มีการศึกษาพัฒนาอย่างต่อเนื่อง โดยส่วนใหญ่ของวิธีการตัดคำมักจะอิงกับวิธีการใช้พจนานุกรมหรือคลังคำศัพท์ทั้งโดยตรงและโดยอ้อม วิธีการตัดคำที่เกี่ยวข้องโดยตรงกับการใช้พจนานุกรมจะทำการเก็บคำภาษาไทยไว้ในพจนานุกรม แล้วนำข้อความที่ป้อนเข้า (input) ไปค้นหาและเทียบสายอักขระกับคำในพจนานุกรมเพื่อหาว่าข้อความดังกล่าวควรตัดคำในบริเวณใดและประกอบด้วยคำใดบ้าง หรือใช้พจนานุกรมช่วยรู้จำคำภาษาไทย ปัญหาที่พบในการตัดคำด้วยวิธีนี้คือ เป็นไปไม่ได้ที่จะเก็บคำทุกคำในภาษาไทยลงในพจนานุกรมได้ โดยเฉพาะคำวิสามานยนาม เช่น ชื่อคน ชื่อสถานที่, ตัวเลข หรือคำที่เกิดขึ้นใหม่ (วิรัช ศรีเลิศล้ำวานิช, 2536 อ้างถึงใน นัฐวุฒิ ไชยเจริญ, 2544: 21) คำที่ไม่ได้ปรากฏในพจนานุกรมจึงกลายเป็นคำที่คอมพิวเตอร์ไม่รู้จำ (unknown word) ไม่สามารถประมวลผลได้อย่างถูกต้อง และที่สำคัญว่านั้นในกรณีที่วิสามานยนามหรือชื่อเฉพาะประกอบขึ้นจากคำที่มีอยู่แล้วในพจนานุกรม เช่น "สมชาย" - สม ชาย, "กนกพร" - กนก พร เมื่อเทียบกับคำในพจนานุกรมโปรแกรมตัดคำจะเกิดความสับสนในการประมวลผล สามารถทำให้เกิดการตัดคำที่ไม่ถูกต้องเพราะทำการตัดแยกสายอักขระออกเป็นคำหลายคำในกรณีที่สายอักขระนั้นควรจะพิจารณาเป็นชื่อเฉพาะหนึ่งรูปคำ ต่อมาวิธีการตัดคำได้พัฒนาโดยอาศัยแนวคิดเรื่องความน่าจะเป็นและคลังข้อมูล ซึ่งยังคงมีความเกี่ยวข้องอยู่กับการใช้คลังคำศัพท์เป็นฐานความรู้สำหรับขบวนการตัดคำ โดยคลังคำศัพท์มักจะได้จากการเรียนรู้คลังข้อมูลฝึกสอน (training corpus) จึงอาจจะไม่ครอบคลุมคำศัพท์ที่ปรากฏในข้อความต่อเนื่องใดๆที่จะทำการตัดคำและเกิดคำที่ไม่ปรากฏในคลังคำศัพท์หรือคำที่คอมพิวเตอร์ไม่รู้จำ โดยเฉพาะอย่างยิ่งชื่อเฉพาะที่มีธรรมชาติของการเกิดใหม่อยู่ตลอดเวลา มีแนวโน้มสูงที่จะกลายเป็นคำที่ไม่รู้จักสำหรับคอมพิวเตอร์เมื่อคลังคำศัพท์นั้นสร้างจากคลังข้อมูลฝึกสอนที่มีเนื้อหาต่างกันหรือคลังข้อมูลนั้นเก็บรวบรวมในช่วงเวลาที่ต่างกัน แม้ต่อมาในระยะหลังๆที่วิธีการตัดคำจะมีลักษณะเป็นวิธีการแบบผสมผสาน (hybrid approach) เช่น วิธีการตัดคำโดยอาศัยคุณลักษณะของคำ (feature-based word segmentation) (Meknavin et al., 1997 อ้างถึงใน นัฐวุฒิ, 2544: 29) ซึ่งยังคงต้องให้ความสำคัญกับคำที่คอมพิวเตอร์ไม่รู้จำโดยจำกัดให้เกิดคำที่ไม่มีในพจนานุกรมน้อยที่สุดก่อนด้วยวิธีการตัดคำที่เรียกว่า maximal matching (เป็นวิธีการตัดคำวิธีหนึ่งของหลักการตัดคำที่อิง

การใช้พจนานุกรม*) แล้วจึงจะอาศัยเรื่องของคุณลักษณะ (feature) เช่น การเกิดร่วมกันของคำ (collocation) เป็นต้น ที่จะทำให้เกิดการขยายขอบเขตของบริบทเพื่อใช้เป็นตัวช่วยแก้ปัญหาความกำกวมในการตัดคำ สำหรับเรื่องของการกำกับหมวดคำพบปัญหาความกำกวมจากชื่อเฉพาะ เช่นเดียวกัน Somlertlamvanich, Charoenporn, and Isahara (1997) ได้กล่าวถึงปัญหาที่พบในการกำกับหมวดคำในการจัดสร้างคลังข้อมูลออร์คิด (ORCHID) ซึ่งปัญหาสำคัญประเด็นหนึ่งคือการกำกับหมวดคำให้กับคำนามทั่วไป (common noun) และชื่อเฉพาะ (proper noun) มีความกำกวมในการแยกแยะหมวดคำทั้งสอง เนื่องจากในภาษาไทยรูปแบบการเขียนของคำทั้งสองชนิดนี้ไม่มีความแตกต่างกัน

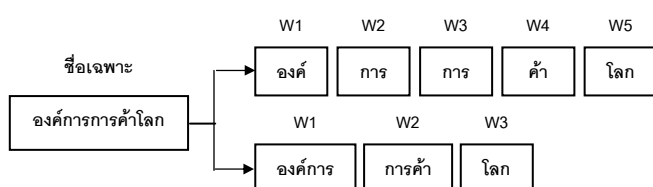
ในภาษาไทยการเกิดคำที่คอมพิวเตอร์ไม่รู้จักรัก (unknown word) นั้น ชื่อเฉพาะ (proper name or proper noun) เป็นประเภทคำที่พบบ่อยที่สุด Kawtrakul et al. (1997) ศึกษาวิจัยเรื่อง "Automatic Thai Unknown Word Recognition" ได้ทำการวิเคราะห์ประเภทของคำที่เกิดเป็นคำที่คอมพิวเตอร์ไม่รู้จักรักโดยการวิเคราะห์จากคลังข้อมูล และแสดงผลให้เห็นว่าประเภทคำที่พบบ่อยที่สุดในกลุ่มคำที่คอมพิวเตอร์ไม่รู้จักรักคือชื่อเฉพาะ ที่พบประมาณ 43 เปอร์เซ็นต์ เมื่อพิจารณาแจกแจงตามประเภทของเอกสารพบว่า หนังสือพิมพ์และวรรณกรรมที่ไม่ใช่นวนิยายปรากฏคำไม่รู้จักรักประเภทชื่อเฉพาะด้วยจำนวนที่มากที่สุด และ Kawtrakul et al. (2002: 2-3) ได้เสนอบทความเรื่อง "A State of the Art of Thai Language Resources and Thai Language Behavior Analysis and Modeling" ได้กล่าวถึงประเด็นการศึกษาเกี่ยวกับการขยายตัวของคลังคำศัพท์ (Lexicon Growth) ในช่วงเวลาต่างๆ โดยผลการสำรวจครั้งล่าสุดแสดงให้เห็นว่า 63.75 เปอร์เซ็นต์ของกลุ่มคำที่ไม่รู้จักรักเป็นคำประเภทชื่อเฉพาะ และเอกสารประเภทหนังสือพิมพ์กับนิตยสารจะพบคำไม่รู้จักรักประเภทชื่อเฉพาะเป็นจำนวนมากที่สุดถึง 91.74, 81.39 เปอร์เซ็นต์ตามลำดับ นอกจากนี้ชื่อเฉพาะในภาษาอื่น เช่น ภาษาจีน ก็เป็นชนิดคำที่พบโดยส่วนใหญ่ของกลุ่มคำที่คอมพิวเตอร์ไม่รู้จักรักที่ปรากฏในข้อความต่อเนื่องของการใช้ภาษาจริง (Chen et al., 1998: 233) ประเด็นปัญหาเกี่ยวกับชื่อเฉพาะที่กลายเป็นคำที่คอมพิวเตอร์ไม่รู้จักรักในการตัดคำและกำกับหมวดคำ จึงยังเป็นปัญหาหลักที่สำคัญในการประมวลผลระดับคำซึ่งเป็นพื้นฐานจำเป็นสำหรับการประมวลผลภาษาที่ถูกรวมเป็นองค์ประกอบสำคัญในการพัฒนาระบบงานทางด้านภาษาศาสตร์คอมพิวเตอร์ดังที่ได้กล่าวมาแล้ว

นอกจากชื่อเฉพาะจะกลายเป็นรูปภาษาที่คอมพิวเตอร์ไม่เคยเห็นมาก่อนในพจนานุกรมหรือในการเรียนรู้จากคลังข้อมูลฝึกสอนแล้ว โดยลักษณะโครงสร้างของชื่อเฉพาะ

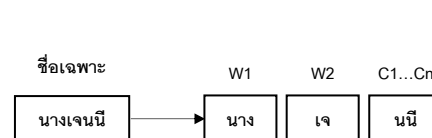
* วิชา ศรเลิศล้ำาณิษ, 2536

ภาษาไทย (ดังที่กล่าวไว้ในหัวข้อที่ 2.1) ยังทำให้ปัญหาในการประมวลผลระดับคำในภาษาไทยมีความซับซ้อนมากขึ้นไปอีก กล่าวคือ การระบุซึ่งถึงชื่อเฉพาะหรือระบุขอบเขตที่แท้จริงของชื่อเฉพาะจะมีความกำกวมสูงกว่าการระบุขอบเขตของคำทั่วไป เพราะชื่อเฉพาะภาษาไทยสามารถประกอบด้วยจำนวนคำที่แตกต่างกัน ไม่สามารถระบุได้แน่นอน และชนิดของคำที่นำมาประกอบกันรวมทั้งลักษณะรูปแบบของการประกอบกันมีความหลากหลาย นอกจากนี้คำที่เป็นส่วนประกอบของชื่อเฉพาะอาจเป็นคำที่มีการกำหนดใช้อยู่แล้วในภาษา (ปรากฏอยู่ในพจนานุกรมหรือคลังคำศัพท์) ทำให้เกิดความสับสนในการตัดคำ ได้ผลการตัดคำที่ไม่ถูกต้อง เช่น ชื่อเฉพาะที่ควรจะต้องแบ่งเป็นหนึ่งรูปคำกลับถูกตัดแบ่งเป็นหลายคำตามตัวอย่างที่ 1 หรือ ตัดคำในสายอักขระของชื่อเฉพาะแล้วเกิดสายอักขระที่อาจจะไม่สามารถเป็นคำที่มีความหมายได้ดังตัวอย่างที่ 2 *

ตัวอย่างที่ 1



ตัวอย่างที่ 2



โดยที่ W แทนคำหนึ่งคำ
C₁...C_n แทนสายอักขระใดๆ

การที่ระบบประมวลผลภาษาธรรมชาติไม่สามารถรู้จำชื่อเฉพาะได้ นอกจากจะส่งผลให้ชื่อเฉพาะนั้นไม่ถูกรู้จำอย่างถูกต้องแล้ว ยังมีผลกระทบไปถึงการรู้จำคำรอบข้าง ทำให้เกิดความผิดพลาดในการเข้าใจและประมวลผลภาษาระดับที่ซับซ้อนขึ้นไปและมีผลต่อเนื่องไปยังระบบประมวลผลโดยรวม ดังเช่นกรณีต่อไปนี้

ในการพัฒนาระบบกำกับข้อมูลแบบอัตโนมัติ (automatic tagger system) สำหรับการนำไปใช้ในงานประมวลผลภาษาธรรมชาติด้านต่างๆ ถ้านำไปใช้ในการแจ่งส่วนประโยคแล้วระบบกำกับข้อมูลอัตโนมัติไม่สามารถรู้จำชื่อเฉพาะได้อย่างถูกต้อง ย่อมส่งผลให้การกำกับหมวดคำเกิดความผิดพลาดและมีผลทำให้ประสิทธิภาพของการทำความเข้าใจเนื้อหาของประโยคที่ประกอบด้วยชื่อเฉพาะนั้นด้อยลงมาก

* อ้างอิงจาก Charoenpomsawat, Kijirikul and Meknavin (1998)

ในการสกัดสารสนเทศซึ่งเกี่ยวข้องกับการดึงใจความสำคัญออกจากข้อมูล การประมวลผลในส่วนของกระบวนการแยกแยะและรู้จำชื่อเฉพาะ นับว่าเป็นสิ่งที่สำคัญเนื่องจากชื่อเฉพาะเป็นหน่วยภาษาที่ทำหน้าที่ในการเก็บข้อมูลที่ใช้สื่อถึงสิ่ง* (entity) พื้นฐาน เช่น ใคร (ชื่อคน ชื่อองค์กร), ที่ไหน (ชื่อสถานที่) ซึ่งข้อมูลที่บอกว่าใคร ที่ไหน เป็นใจความส่วนสำคัญของการใช้ภาษาให้สามารถสื่อสารได้อย่างมีประสิทธิภาพ ถ้าระบบสกัดสารสนเทศไม่สามารถจัดการกับปัญหาการประมวลผลชื่อเฉพาะ จะได้ผลลัพธ์ที่ไม่สื่อความหมายหรือสื่อความหมายไม่ตรงกับสิ่งที่ต้องการ เช่น ต้องการให้ดึงใจความสำคัญที่เกี่ยวข้องกับเรื่องของ "เหล็ก" ที่เป็นวัสดุสิ่งของ ถ้าระบบไม่สามารถแยกแยะได้ว่า "ทำซีเหล็ก" เป็นชื่อเฉพาะประเภทสถานที่และไม่เกี่ยวข้องกับเรื่อง "เหล็ก" ผลลัพธ์ที่ได้จะเป็นผลลัพธ์ที่ประกอบด้วยข้อมูลที่ไม่ตรงกับความต้องการ

ในปัจจุบันการพัฒนาระบบค้นคืนสารสนเทศได้มีการพัฒนาไปถึงเรื่องของการค้นคืนสารสนเทศข้ามภาษา (Cross-Language Information Retrieval) ซึ่งประเด็นที่มีความสำคัญอันดับแรกของการศึกษาวิจัยงานด้านนี้ จะเป็นเรื่องของความกำกวมในการแปลคำร้องขอสำหรับการสืบค้น (query) จากภาษาหนึ่งไปเป็นอีกภาษาหนึ่งซึ่งคำร้องขอสำหรับการสืบค้นอาจเป็นรูปภาษาระดับคำหรือเป็นวลี เช่น งานวิจัยของ Lin et al. (1999) ศึกษาเกี่ยวกับการค้นคืนสารสนเทศระหว่างภาษาจีนกับภาษาอังกฤษ โดยระบุคำร้องขอเป็นภาษาจีนแล้วทำการค้นคืนเอกสารหรือข้อมูลภาษาอังกฤษที่เกี่ยวข้องกับคำร้องขอภาษาจีนนั้น จึงต้องมีการทำในส่วนของงานแปลคำสำคัญสำหรับการค้นคืนที่ระบุเป็นภาษาจีนไปเป็นภาษาอังกฤษ ซึ่งการพัฒนาสำหรับภาษาจีนที่มีลักษณะโครงสร้างภาษาคัดลอกภาษาไทยนั้น จะต้องเกี่ยวข้องกับพื้นฐานของการประมวลผลระดับคำคือ การตัดคำ ที่ยังคงพบปัญหาในการตัดคำที่ไม่ถูกต้องจากการตัดแบ่งชื่อเฉพาะออกเป็นคำหลายคำ ทำให้การแปลคำสำคัญสำหรับสืบค้นผิดพลาด คณะผู้วิจัยได้กล่าวไว้ว่า การใช้วิธีการทางสถิติเป็นวิธีการที่มีประสิทธิภาพ สามารถลดความผิดพลาดในการแปลได้แต่ในขณะเดียวกันผลการวิจัยก็แสดงให้เห็นว่าการประมวลผลชื่อเฉพาะยังเป็นประเด็นปัญหาที่มีความกำกวมอยู่มากจึงมีความสำคัญต่อการพัฒนางานค้นคืนสารสนเทศข้ามภาษาระหว่างภาษาจีนกับภาษาอังกฤษ (Chinese-English Cross Language Retrieval)

กล่าวโดยสรุป ชื่อเฉพาะเป็นปัญหาสำคัญในการประมวลผลภาษา เพราะชื่อเฉพาะนั้นมีได้ไม่จำกัด จึงไม่สามารถจัดเก็บในพจนานุกรมได้ครบถ้วน และชื่อเฉพาะอาจประกอบด้วยคำทั่วไปทำให้เป็นปัญหาในการระบุขอบเขตของชื่อเฉพาะ การไม่สามารถหาชื่อ

* ศัพท์บัญญัติตามพจนานุกรมศัพท์ปรัชญาอังกฤษ-ไทย ฉบับราชบัณฑิตยสถาน พ.ศ. 2532 หน้า 37 อ้างถึงใน สุริยา รัตนกุล (2544:70)

เฉพาะได้ถูกต้องนี้ จะส่งผลให้เกิดข้อผิดพลาดในการประมวลผลเรื่องอื่นต่อไป เช่น การตัดคำ การแปลภาษา การสกัดสารสนเทศ

2.3 การศึกษาชื่อเฉพาะภาษาไทยที่ผ่านมา

การศึกษาเพื่อหาวิธีการแก้ปัญหาความกำกวมในการประมวลผลชื่อเฉพาะภาษาไทยที่ผ่านมา นั้น จะมีทั้งการศึกษาเจาะจงถึงชื่อเฉพาะและศึกษาคำที่คอมพิวเตอร์ไม่รู้จัก (unknown word) ทุกประเภท วิธีการที่นำเสนอในงานวิจัยต่างๆจะมีลักษณะเป็นวิธีการแบบผสมผสานที่อาศัยเทคนิควิธีการประมวลผลภาษาธรรมชาติหลายวิธีการ เช่น วิธีการทางสถิติ (Statistical approach), วิธีการอาศัยคุณลักษณะ (Feature-based approach), วิธีการที่อิงพจนานุกรม (Dictionary-based approach) เป็นต้น และมีแนวทางการศึกษาวิเคราะห์บางประเด็นเป็นไปในแนวทางเดียวกัน ซึ่งวิธีการและประเด็นการศึกษาวิเคราะห์เกี่ยวกับชื่อเฉพาะในงานวิจัยต่างๆ สามารถสรุปได้ดังนี้

(1) ในงานวิจัยเกี่ยวกับชื่อเฉพาะ ดังเช่นงานวิจัยของ Charoenpornasawat, Kijirikul and Meknavin (1998) และ Kawtrakul et al. (1997) พิจารณาลักษณะของชื่อเฉพาะโดยวิเคราะห์ว่าประกอบด้วยคำหรือสายอักขระ (word or character string) ที่มีในพจนานุกรมหรือไม่ในพจนานุกรม เมื่อพิจารณาโดยรวมแล้วสามารถสรุปได้ 3 ลักษณะคือ

(1.1) คำหรือสายอักขระทุกตัวไม่ปรากฏในพจนานุกรม

ตัวอย่างเช่น แสนเว่น แฮรี่ สุนีย์ นครเรศ

(1.2) ประกอบด้วยคำหรือสายอักขระที่ปรากฏในพจนานุกรม

และไม่ปรากฏในพจนานุกรมอยู่รวมกัน

ตัวอย่างเช่น สุมานี คธาพงศ์ ไมโครซอฟต์

(1.3) คำหรือสายอักขระทุกตัวปรากฏอยู่ในพจนานุกรม

ตัวอย่างเช่น สมชาย กนกพร รุ่งศักดิ์

Charoenpornasawat et al. (1998) จัดให้ลักษณะ (1.1) และ (1.2) เป็นประเภท Proper name type I คือเป็นชื่อเฉพาะประเภทที่ประกอบด้วยคำที่ไม่มีในพจนานุกรม ส่วน (1.3) เป็นประเภท Proper name type II คือเป็นชื่อเฉพาะประเภทที่ไม่พบคำที่ไม่มีในพจนานุกรม ซึ่งการจัดแบ่งดังกล่าวนี้เป็นการจัดแบ่งโดยพิจารณาจากคำหรือสายอักขระที่ไม่มีในพจนานุกรมเป็นหลัก ในขณะที่ Kawtrakul et al. (1997) จะพิจารณาจากคำที่มีในพจนานุกรมเป็นหลัก

โดยจัดให้ชื่อเฉพาะลักษณะ (1.1) เป็นกลุ่มที่เรียกว่า Explicit ส่วน (1.2) และ (1.3) จัดอยู่ในกลุ่มที่เรียกว่า Hidden เพราะมีบางส่วนสามารถเป็นคำในพจนานุกรม

- (2) การระบุนขอบเขตของชื่อเฉพาะเป็นประเด็นปัญหาสำคัญที่เป็นพื้นฐานของการแก้ปัญหาการประมวลผลชื่อเฉพาะ ซึ่งการระบุนขอบเขตของชื่อเฉพาะเกี่ยวข้องกับขั้นตอนที่เกิดขึ้นในการตัดคำ โดยวิธีการที่ใช้ในการระบุนขอบเขตของชื่อเฉพาะจะอยู่บนพื้นฐานของการอาศัยค่าความน่าจะเป็นและการอิงพจนานุกรมได้แก่ การใช้แบบจำลองไตรแกรม (trigram model) (Charoenpornasawat et al., 1998; Kawtrakul et al., 1997) วิธีการหารายการคำให้ได้จำนวนคำที่มีในพจนานุกรมน้อยที่สุดและมีจำนวนคำที่คอมพิวเตอร์ไม่รู้จำก้น้อยที่สุด (Kanlayanawat and Prasitjutrakul, 1997)
- (3) ในการพัฒนาระบบประมวลผลชื่อเฉพาะที่ประกอบด้วยขั้นตอนการตัดคำเป็นพื้นฐานแล้ว ยังต้องประกอบด้วยขั้นตอนของการรู้จำชื่อเฉพาะและแก้ปัญหาความกำกวมจากชื่อเฉพาะ ซึ่งวิธีการที่มีการนำเสนอไว้ได้แก่วิธีการต่อไปนี้

(3.1) วิธีการพิจารณาคุณลักษณะ (Feature-based approach)

(Charoenpornasawat et al., 1998)

วิธีการนี้ใช้แนวทางการพิจารณาคุณลักษณะที่ได้จากการเรียนรู้คลังข้อมูลภาษาในการระบุชื่อเฉพาะภาษาไทยที่ผ่านการตัดคำและกำกับหมวดคำแล้ว โดยใช้โมเดลการเรียนรู้แบบวินโนว (Winnow Algorithm) เรียนรู้คุณลักษณะจากคลังข้อมูลและใช้ในการระบุชื่อเฉพาะ คุณลักษณะที่กล่าวถึงนี้ได้แก่ รูปคำที่ปรากฏข้างเคียงชื่อเฉพาะ (context word) และการปรากฏร่วม (collocation) ซึ่งในขั้นตอนการเรียนรู้การพิจารณารูปคำข้างเคียงจะพิจารณาในขอบเขต +/-10 คำ กล่าวคือวิเคราะห์คำที่อยู่ถัดจากชื่อเฉพาะไปทางขวา 10 คำ และถัดไปทางซ้าย 10 คำ ส่วนการปรากฏร่วมจะพิจารณาการเกิดร่วมกันของหมวดคำในขอบเขต +/-2 คำ แล้วแปรความสัมพันธ์ของคุณลักษณะดังกล่าวออกมาในรูปคะแนน นอกจากนี้จะให้เรียนรู้คุณลักษณะที่ปรากฏกับชื่อเฉพาะแล้ว วินโนวยังเรียนรู้คุณลักษณะของคำต่างๆที่ไม่ใช่ชื่อเฉพาะด้วยกลวิธีเดียวกัน ฐานความรู้ที่ได้นี้วินโนวจะนำไปใช้สำหรับการระบุชื่อเฉพาะ

เมื่อประโยคป้อนเข้าที่ทำการตัดคำและกำกับหมวดคำแล้วปรากฏคำที่คอมพิวเตอร์ไม่รู้จำ คำๆนั้นจะถูกรวมเข้ากับรูปคำที่อยู่ติดกันทั้งทางซ้ายและขวา ผลลัพธ์ที่ได้คือจะมีสายอักขระใหม่เกิดขึ้นจำนวนหนึ่ง ซึ่งจะเป็นตัวเลือก (candidates) สำหรับการวิเคราะห์ว่าตัวเลือกใดเป็นชื่อเฉพาะ ดังเช่นตัวอย่าง ประโยคป้อนเข้า คือ "นางเจนนี่ไปเดินตากลมเล่น" ถ้าประโยคทำการตัดคำและกำกับหมวดคำแล้วมีลักษณะดังนี้

นาง/NTTL เจ/NCMN นนี่/NPRP ไป/VACT เดิน/VACT ตาก/PPRS ลม/NCMN เล่น/ADVN

ในประโยคที่ได้นี้ "นนี่" เป็นคำที่คอมพิวเตอร์ไม่รู้จัก ถ้ากำหนดว่าให้รวมกับรูปคำข้างเคียง ถัดไป 2 คำ ผลลัพธ์ของทางเลือกที่ได้ทั้งหมดคือ นนี่, นนี่ไป, นนี่ไปเดิน, เจนนี่, เจนนี่ไป, เจนนี่ไปเดิน, นางเจนนี่, นางเจนนี่ไป, นางเจนนี่ไปเดิน ซึ่งตัวเลือกเหล่านี้จะถูกมองเสมือนเป็นชื่อเฉพาะ แล้วนำตัวเลือกแต่ละตัวไปรวมกับคำที่เหลือของประโยคเดิมแล้วกำกับหมวดคำใหม่ จากนั้นใช้วินโนววิเคราะห์ชื่อเฉพาะเสมือน (proper name candidates) ว่าตัวเลือกใดเป็นชื่อเฉพาะจากประโยคที่สร้างใหม่แต่ละประโยค โดยวินโนวจะอาศัยฐานความรู้ที่เรียนรู้จากคลังข้อมูลในการระบุชื่อเฉพาะ

ผลการวิจัยที่ใช้วิธีพิจารณาคูณสมบัตินี้ แสดงผลไว้ว่ามีความถูกต้องประมาณ 92.17 เปอร์เซ็นต์ แต่อย่างไรก็ตาม ผู้วิจัยคิดว่ามีประเด็นที่น่าสนใจเกี่ยวกับกลวิธีของวิธีการนี้ คือ ประการแรก ในการพิจารณาบริบทข้างเคียงโดยใช้กลวิธีรวมคำที่เป็นเป้าหมายเข้ากับรูปคำที่อยู่ติดกันในขอบเขตที่กำหนด เพื่อให้ระบบเรียนรู้ความเป็นไปได้ในการปรากฏร่วมกันของคำต่างๆ หรือเพื่อสร้างตัวเลือกที่มีโอกาสเป็นชื่อเฉพาะ หรือเพื่อใช้ประโยชน์ในการระบุว่าตัวเลือกใดเป็นชื่อเฉพาะนั้น ควรจะพิจารณาในขอบเขตจำนวนที่คำจึงจะเหมาะสมและมีความครอบคลุมพอที่จะสกัดหาชื่อเฉพาะภาษาไทย การตั้งคำนึงถึงประเด็นนี้ก็เนื่องจากในภาษาไทยนั้น ชื่อเฉพาะสามารถประกอบด้วยจำนวนคำที่แตกต่างกัน ไม่มีกฎเกณฑ์ในการสร้างที่แน่นอนและโครงสร้างภายในของชื่อเฉพาะอาจมีการเรียงลำดับของคำที่มาประกอบกันไม่แตกต่างจากโครงสร้างของรูปภาษาในระดับที่สูงกว่าคำ เช่น "...กองทุนเพื่อการฟื้นฟูและพัฒนาระบบสถาบันการเงิน (ชื่อเฉพาะ) เป็นกองทุนเพื่อช่วยเหลือสถาบันการเงิน (วลี) ที่ประสบปัญหาจากภาวะเศรษฐกิจ..." ถ้าคำเป้าหมายเป็นคำที่อยู่ส่วนต้นหรือส่วนท้ายของสายอักขระที่ประกอบกันเป็นชื่อเฉพาะแล้วมีการกำหนดให้สร้างตัวเลือกที่เสมือนเป็นชื่อเฉพาะโดยใช้ขอบเขต +/- 2 คำ โอกาสที่ชื่อเฉพาะรูปนี้จะถูกรู้จำอย่างไม่ถูกต้องนั้นมีความเป็นไปได้สูงทั้งสาเหตุจากการไม่เกิดตัวเลือกที่ถูกต้องหรือเกิดความสับสนในเรื่องลำดับคำที่ระบบอาจมองว่าเป็นคำที่ประกอบกันเป็นหน่วยสร้างวลีไม่ใช่ชื่อเฉพาะ หรือถ้ากำหนดให้ขอบเขตขยายมากขึ้น อาจเกิดตัวเลือกที่ถูกต้องแต่ในขณะเดียวกันจะเกิดตัวเลือกจำนวนมาก และต้องทำการตรวจสอบตัวเลือกโดยไม่จำเป็น จึงมีความเป็นไปได้หรือไม่ที่จะทำการวิเคราะห์จากบริบทข้างเคียงโดยไม่จำเป็นต้องกำหนดขอบเขตของคำ

ประการต่อมา ในการศึกษาเกี่ยวกับชื่อเฉพาะที่ใช้ฐานความรู้จากการเรียนรู้ในคลังข้อมูลฝึกสอน ซึ่งผู้วิจัยคิดว่าอาจประสบกับปัญหาว่าชื่อเฉพาะที่วิเคราะห์พบในข้อความ เป็นชื่อเฉพาะที่ไม่ปรากฏอยู่ในคลังข้อมูลฝึกสอน หมายความว่าชื่อเฉพาะนั้นก็จะเป็น

คำที่ระบบไม่สามารถประมวลผลได้ เกิดเป็นปัญหาของคำที่คอมพิวเตอร์ไม่รู้จักอีก นอกจากนี้ ในการสร้างคลังข้อมูลที่มีการตัดคำและกำกับหมวดคำด้วยมือเพื่อให้ระบบรู้จำภาษา จะต้องใช้ทักษะและเวลาในการทำมาก จะมีทางเลือกใดที่จะช่วยลดปัญหาการปรากฏของชื่อเฉพาะที่กระจายอยู่ในข้อมูลภาษาแตกต่างกันเมื่อข้อมูลภาษานั้นมาจากองค์ประกอบแวดล้อมที่แตกต่างกัน รวมทั้งสามารถลดภาระในการศึกษาวิจัยที่มีความเกี่ยวข้องกับชื่อเฉพาะแต่ละครั้งซึ่งมีการสร้างคลังข้อมูลในลักษณะคล้ายๆกัน

(3.2) วิธีการอื่นนอกเหนือจากวิธีพิจารณาคูณลักษณะที่มีการนำเสนอไว้ จะเป็นวิธีการที่ใช้เทคนิคเฉพาะร่วมกับวิธีการทางสถิติ เพื่อดูค่าความน่าจะเป็นของคำและกลุ่มความหมายที่คำมาประกอบกัน (Kawtrakul et al., 1997) หรือเป็นวิธีการที่ประกอบด้วยวิธีอาศัยพจนานุกรมร่วมกับเทคนิคเฉพาะที่ใช้วิธีตั้งน้ำหนักของการเชื่อมต่อกันของสายอักขระชุดต่างๆให้แตกต่างกัน (Kanlayanawat and Prasitjutrakul, 1997) ซึ่งวิธีการทั้งสองมีการเกี่ยวข้องกับเรื่องของการอาศัยพจนานุกรมหรือคลังข้อมูลฝึกสอน ข้อจำกัดอันเกิดจากปัญหาการกระจายของชื่อเฉพาะในข้อมูลย่อมเกิดขึ้น รวมทั้งเทคนิคเฉพาะที่ใช้ก็มีแนวคิดในการแก้ปัญหาการประมวลผลชื่อเฉพาะที่เกี่ยวข้องกับการสร้างตัวเลือกหรือทางเลือกหลายๆรูปแบบ จึงอาจเกิดประเด็นปัญหาในการพัฒนาวิธีการคล้ายๆกับปัญหาที่พบจากวิธีการพิจารณาคูณลักษณะ

ดังนั้นในการแยกแยะและบ่งชี้ชื่อเฉพาะภาษาไทยในงานประมวลผลภาษาธรรมชาติ ให้สามารถครอบคลุมชื่อเฉพาะลักษณะต่างๆและสามารถรู้จำชื่อเฉพาะที่เกิดขึ้นใหม่ ผู้วิจัยเห็นว่าควรใช้แนวทางที่อาศัยความรู้หรือข้อสรุปทางภาษาที่มีลักษณะสอดคล้องกับแนวทางการประมวลผลของคอมพิวเตอร์ เป็นฐานความรู้สำคัญสำหรับการรู้จำชื่อเฉพาะ โดยในการวิเคราะห์ความรู้ทางภาษาเกี่ยวกับชื่อเฉพาะ น่าจะเป็นเรื่องของการวิเคราะห์บริบท (Context) ที่บ่งบอกการปรากฏของชื่อเฉพาะ เพราะโดยปกติของการประมวลผลภาษาของมนุษย์ทั้งในเรื่องของการวิเคราะห์แยกแยะและทำความเข้าใจหน่วยภาษาต่างๆจะต้องอาศัยเรื่องของบริบท (Context) ในลักษณะต่างๆเข้ามาเป็นตัวช่วยที่สำคัญที่สุด ที่ทำให้สามารถประมวลผลภาษา ลักษณะต่างๆที่เกิดขึ้นในการใช้ภาษาจริงได้อย่างไม่มีสิ้นสุด การใช้บริบทที่บ่งบอกการปรากฏของชื่อเฉพาะจึงน่าจะประมวลผลชื่อเฉพาะที่ปรากฏในการใช้ภาษาจริงลักษณะต่างๆได้โดยไม่มีข้อจำกัดจากปัจจัยต่างๆ นอกจากนี้การที่เราสามารถหาบริบทบ่งบอกการปรากฏชื่อเฉพาะภาษาไทยโดยวิเคราะห์ตัวภาษาในคลังข้อมูลภาษาให้เป็นข้อสรุปทางภาษาของชื่อเฉพาะที่

สามารถนำไปใช้ได้ทั่วไปกับงานประมวลผลภาษาธรรมชาติ นอกจากจะเป็นประโยชน์ต่อการเพิ่มประสิทธิภาพในการวิจัยพัฒนางานด้านต่างๆแล้ว ยังเป็นการลดภาระในด้านการสร้างคลังข้อมูลที่มีลักษณะคล้ายกันและเพื่อนำไปใช้ประโยชน์ในจุดประสงค์เดียวกันของการวิจัยแต่ละครั้ง

บทที่ 3

ชื่อเฉพาะและการกำกับชื่อเฉพาะ

บทนี้จะได้กล่าวถึงวรรณกรรมอีกส่วนหนึ่งที่ผู้วิจัยได้ทำการศึกษาไว้ โดยเป็นประเด็นที่เกี่ยวข้องกับการนำไปใช้ในการจัดสร้างคลังข้อมูลภาษาของวิทยานิพนธ์ ซึ่งเป็นคลังข้อมูลประเภทที่มีการกำกับข้อมูล (annotated corpus) ในการจัดสร้างคลังข้อมูลนี้มีขั้นตอนสำคัญที่จะทำให้คลังข้อมูลสามารถเป็นเครื่องมือที่มีประโยชน์เต็มที่สำหรับการศึกษาบริบทบ่งบอกชื่อเฉพาะ โดยขั้นตอนดังกล่าวคือ การกำกับชื่อเฉพาะไว้ในคลังข้อมูล ซึ่งการดำเนินการในขั้นตอนนี้มีประเด็นที่ต้องพิจารณาอยู่ด้วยกันสองประการได้แก่ เกณฑ์สำหรับวิเคราะห์หาชื่อเฉพาะและสัญลักษณ์สำหรับการกำกับชื่อเฉพาะ ในประการแรกนั้นมโนทัศน์ของชื่อเฉพาะจะเป็นความรู้พื้นฐานที่เข้ามามีบทบาทในการพิจารณาและเป็นที่มาของเกณฑ์นิยาม (definitional criteria) ที่จะใช้ในวิทยานิพนธ์เล่มนี้ สำหรับประการหลังจะเป็นเรื่องของภาษาสำหรับกำกับข้อมูล (markup language) ซึ่งในวิทยานิพนธ์ฉบับนี้ได้ใช้แนวทางการกำกับชื่อเฉพาะของ Text Encoding Initiative (TEI) ทั้งนี้จะกล่าวถึงหัวข้อที่ 3.1 ที่มาของเกณฑ์นิยามและหัวข้อที่ 3.2 แนวทางการกำกับชื่อเฉพาะของ TEI ในส่วนถัดไป

3.1 ที่มาของเกณฑ์นิยาม

ในหัวข้อนี้จะกล่าวถึงมโนทัศน์ของชื่อเฉพาะในวรรณกรรมต่างๆ โดยประกอบด้วยมโนทัศน์ชื่อเฉพาะในภาษาไทยและภาษาอังกฤษ สาเหตุที่ศึกษามโนทัศน์ชื่อเฉพาะภาษาอังกฤษด้วยนั้น เนื่องมาจากวรรณกรรมภาษาไทยที่ได้กล่าวถึงมโนทัศน์ของชื่อเฉพาะมีจำนวนน้อยและอธิบายเกี่ยวกับชื่อเฉพาะเพียงคร่าวๆ นอกจากนี้แล้วจะกล่าวถึงประเด็นที่เป็นปัญหาของมโนทัศน์เหล่านี้ซึ่งเป็นปัจจัยที่ทำให้มีการกำหนดเกณฑ์นิยาม (definitional criteria) สำหรับการศึกษาคำนี้

หากกล่าวถึงชื่อเฉพาะในภาษาอังกฤษแล้ว ลักษณะเด่นที่จะมีการกล่าวถึงคือการขึ้นต้นรูปภาษาที่เป็นชื่อเฉพาะด้วยตัวอักษรพิมพ์ใหญ่ ซึ่งลักษณะนี้เป็นการกำหนดให้มีตัวบ่งชี้ (marker) เพื่อใช้แสดงว่ารูปภาษานั้นเป็นชื่อเฉพาะ แต่สำหรับชื่อเฉพาะในภาษาไทยยังไม่เคยมีการกำหนดตัวบ่งชี้ในลักษณะนี้รวมทั้งไม่มีการศึกษาว่าลักษณะโครงสร้างทางไวยากรณ์ใดของรูปภาษาที่จะบ่งบอกได้ ดังนั้นในการวิเคราะห์หาชื่อเฉพาะจึงต้องวิเคราะห์จากคุณสมบัติเชิงความหมาย กล่าวคือ พิจารณาจากคำนิยามหรือคำอธิบายของชื่อเฉพาะ ซึ่งพบว่าในวรรณกรรม

ต่างๆจะมีรายละเอียดบางส่วนแตกต่างกันและขอบเขตของรูปภาพที่จัดว่าเป็นชื่อเฉพาะก็มีความแตกต่างกัน รวมทั้งผู้วิจัยได้ทดสอบนำคำนิยามหรือคำอธิบายชื่อเฉพาะในวรรณกรรมต่างๆไปวิเคราะห์ห้ข้อมูลในคลังข้อมูลที่ประกอบด้วยข้อความภาษาไทย (plain text) พบว่าไม่สามารถตัดสินใจได้อย่างชัดเจนว่ารูปภาพที่วิเคราะห์เป็นชื่อเฉพาะหรือไม่เป็นชื่อเฉพาะ ผู้วิจัยจึงจะกล่าวถึงคำนิยามหรือคำอธิบายในวรรณกรรมดังกล่าวและยกตัวอย่างให้เห็นถึงประเด็นที่เป็นปัญหาซึ่งพบในการทดสอบวิเคราะห์ข้อมูล ดังนี้คือ

พระยาอุปกิตศิลปสาร (2541: 71-72) ได้ให้นิยามและคำอธิบายเกี่ยวกับชื่อเฉพาะไว้ในเนื้อหาส่วนที่เกี่ยวข้องกับคำนามและใช้คำว่า "วิสามานยนาม" ในการกล่าวถึงชื่อเฉพาะ สรุปได้ดังนี้

วิสามานยนาม คำนามที่เป็นชื่อเฉพาะที่สมมุติตั้งขึ้นสำหรับเรียกคน สัตว์ และสิ่งของบางอย่างเพื่อให้รู้ชัดว่าคนนี้ สัตว์ตัวนี้ ของสิ่งนี้ ฯลฯ

| | | |
|----------|---|---------------------------------|
| ตัวอย่าง | ชื่อคน - สอน ลิน | ชื่อเมือง - นนทบุรี ราชบุรี |
| | ชื่อสกุล - มาลากุล ณ สงขลา | ชื่อสัตว์ - มงคล (ช้าง) พังเป้น |
| | ชื่อบรรดาศักดิ์ - ญาณภิรมย์ อุดมจินดา วัน - อาทิตย์ จันทร | |
| | ชื่อสิ่งของ - เสือทะเลยานชน (เรือ) เสือคำรณสินธุ์ (เรือ) | |

วิสามานยนามนี้ต้องเป็นคำที่ตั้งขึ้นใช้เรียกคนคนเดียว สัตว์ตัวเดียว และของสิ่งเดียว ถึงจะเป็นชื่อของหมู่คณะก็ต้องเป็นหมู่เดียว คณะเดียว เช่น 'ชาติ-ไทย' หมายความว่าไทยชาติเดียว และ 'สโมสรร-สามัคยาจารย์สมาคม' ก็หมายความว่าถึงสโมสรรนั้นแห่งเดียว ถึงแม้จะเผชิญมีชื่อซ้ำกันบ้างก็หมายความว่าเฉพาะคน เฉพาะสิ่งไม่ทั่วถึงกันได้ ถ้าอยู่ใกล้กันก็ต้องเติมสร้อยหรือบอกเครื่องหมายท้ายชื่อให้สังเกตได้ว่า คนนี้ สิ่งนี้ เช่น นาย-แดง (เล็ก) นาย-แดง (ใหญ่) วัดสามจีนเหนือ วัดสามจีนใต้ เป็นต้น

ตัวอย่างที่เป็นปัญหาในการวิเคราะห์ตามนิยามของพระยาอุปกิตศิลปสาร ได้แก่

(ก) "(อำเภอ)เมือง" มีการใช้ซ้ำกันคือใช้อ้างถึงสถานที่ตั้งซึ่งเป็นศูนย์กลางของจังหวัดต่างๆโดยส่วนใหญ่ในประเทศไทย แต่เป็นในลักษณะหมายความว่าเฉพาะสิ่งไม่ได้ใช้อ้างถึงแบบทั่วถึงกันกล่าวคือใช้อ้างถึงเฉพาะสถานที่ตั้งที่เป็นศูนย์กลางของจังหวัดและสามารถรู้ได้ว่าอ้างอิงถึงบริเวณใดโดยมีชื่อจังหวัดเป็นสิ่งบ่งชี้ ถ้าพิจารณาตามนี้รูปภาพนี้จึงตัดสินใจได้ว่าเป็นวิสามานยนาม แม้จะไม่มีเติมสร้อยหรือบอกเครื่องหมายท้ายชื่อดังที่พระยาอุปกิตศิลปสารได้อธิบายไว้ แต่ก็สามารถที่จะรู้ได้อย่างชัดเจนว่าหมายถึงบริเวณใดโดยพิจารณาจากบริบทคือชื่อจังหวัดซึ่งอาจปรากฏติดกับ "(อำเภอ)เมือง" หรือ มีการกล่าวถึงชื่อจังหวัดมาก่อน แต่ถ้าวิเคราะห์ตามคำนิยามอื่นรูปภาพนี้ไม่สามารถตัดสินใจได้ว่าเป็นชื่อเฉพาะ ซึ่งจะกล่าวถึงต่อไป

- (ข) คำที่ใช้อ้างอิงถึงหน่วยเงินจะมีการใช้ในสองลักษณะคือ เป็นคำนามและเป็นคำลักษณะนาม
- กรณีเป็นคำนาม เช่น "...เนื่องจากดอลลาร์ได้รับแรงกดดัน..." ในกรณีคำนามนี้เมื่อวิเคราะห์ตามคำอธิบายของพระยาอุปกิตศิลปสารจะเป็นวิสามานยนาม เพราะใช้กล่าวอ้างถึงสกุลเงินที่ใช้ในประเทศต่างๆแต่ละประเทศซึ่งแต่ละประเทศจะมีสกุลเงินเพียงอันเดียวและไม่ซ้ำกับประเทศอื่น
- กรณีเป็นคำลักษณะนาม เช่น "...เมื่อค่าเงินเปโซได้อ่อนค่าลงสู่ระดับ 49.00-50.50 เปโซ/ดอลลาร์..." จากนิยามจะกล่าวถึงว่าวิสามานยนามสามารถใช้อ้างถึงหมู่เดียวคณะเดียวได้นั้นหมายถึงต้องมีการรวมกันเป็นหน่วยเดียว (Unit) หรือไม่ ซึ่งถ้าพิจารณาถึงตัวอย่างที่พระยาอุปกิตศิลปสารได้กล่าวไว้ ควรจะหมายถึงว่าวิสามานยนามจะใช้อ้างถึงกลุ่มของสิ่งใดๆที่มีการรวมกันเป็นหน่วยเดียว ส่วนในกรณีนี้คำที่ใช้อ้างอิงถึงหน่วยเงินในสภาพที่เป็นคำลักษณะนาม เป็นการกล่าวอ้างถึงกลุ่มของสิ่งของที่ไม่ได้มีการรวมกันเป็นหน่วยเดียว (Unit) แต่เป็นการกล่าวอ้างถึงกลุ่มของสิ่งของที่มีลักษณะเหมือนกันเท่านั้น ซึ่งในกรณีเช่นนี้เป็นปัญหาว่าจะพิจารณาให้เป็นชื่อเฉพาะหรือไม่เป็นชื่อเฉพาะ
- (ค) คำที่ใช้อ้างอิงถึงเครื่องหมายการค้าหรือยี่ห้อหรือตราสินค้า (brand name) เป็นรูปภาษาที่มีลักษณะของวิสามานยนามตามคำอธิบายนี้คือ เครื่องหมายการค้าจะมีเรื่องของการจดลิขสิทธิ์ซึ่งทำให้เกิดปรากฏการณ์ในลักษณะที่ไม่สามารถนำรูปภาษานี้ไปใช้อ้างอิงถึงสินค้าใดได้อีก จึงเป็นชื่อที่ใช้เรียกเฉพาะสินค้านั้นอย่างเดียวนำไปใช้เรียกสินค้าอื่นไม่ได้ แต่ในการอ้างอิงนั้นไม่ได้อ้างอิงถึงสิ่งที่เฉพาะเจาะจงเป็นขั้นๆ จะเป็นการอ้างอิงถึงกลุ่มของสิ่งของหรือสิ่งของจำนวนมากที่มีลักษณะเหมือนกันซึ่งไม่ได้มีการรวมกันเป็นหน่วยเดียว (Unit) จึงมีความคลุมเครือในลักษณะเดียวกับกรณีของหน่วยเงินที่เป็นคำลักษณะนาม นอกจากนี้ยังมีลักษณะของการขยายไปใช้อ้างอิงถึงสิ่งอื่นๆ โดยการนำรูปภาษาของเครื่องหมายการค้าไปใช้เรียกสิ่งอื่นที่มีลักษณะสำคัญหรือคุณสมบัติสำคัญเหมือนกับสิ่งที่เครื่องหมายการค้าอ้างถึงและมีความถี่ในการเกิดลักษณะนี้สูงจนเสมือนว่ารูปภาษานั้นไม่ใช่ชื่อเฉพาะที่ตั้งขึ้นเพื่อใช้อ้างอิงถึงสิ่งใดสิ่งเดียวโดยเฉพาะ กล่าวคือ มีพฤติกรรมเหมือนคำนามทั่วไป เช่น ตัวอย่างในภาษาไทย "มาม่า" เป็นเครื่องหมายการค้าของบะหมี่กึ่งสำเร็จรูปที่ผลิตโดยบริษัทหนึ่ง ซึ่งยี่ห้อนี้เป็นยี่ห้อแรกๆที่อยู่ในท้องตลาด ดังนั้นเมื่อมีการกล่าวอ้างถึงบะหมี่กึ่งสำเร็จรูปจะใช้รูปภาษานี้ในการระบุถึงและมีการใช้ต่อมาเรื่อยๆ แม้ว่าต่อมาจะมีบริษัทอื่นๆผลิตออกมาโดยใช้เครื่องหมายการค้าอื่นที่ไม่ใช่ "มาม่า" ก็ตาม ดังนั้นรูปภาษาที่ใช้อ้างอิงถึงเครื่องหมายการค้าจึงมีลักษณะที่คลุมเครือ มีลักษณะบางส่วนของวิสามานยนามตามคำอธิบายนี้ บางส่วนขัดแย้งกับคำอธิบายตามแนวคิดนี้

- (ง) คำที่ใช้ระบุหรือกล่าวอ้างถึงชื่อนิติยสาร ชื่อหนังสือพิมพ์ ซึ่งมีพฤติกรรมคล้ายกับรูปภาษาที่ใช้กล่าวอ้างถึงเครื่องหมายการค้า คือ เป็นรูปภาษาที่ไม่สามารถนำไปใช้ซ้ำได้ แต่ก็ถูกนำไปใช้อ้างอิงถึงกลุ่มของสิ่งของจำนวนมาก จึงพบปัญหาในการวิเคราะห์เช่นกัน

นวรรธณ พันธุมเมธา (2527: 4-5) มีคำอธิบายและคำนิยามชื่อเฉพาะในเนื้อหาเกี่ยวกับค่านาม โดยใช้คำศัพท์ในการกล่าวถึงชื่อเฉพาะว่า "ค่านามวิสามัญ"

ค่านามวิสามัญ คือ คำที่หมายถึงสิ่งใดสิ่งหนึ่งโดยเฉพาะ เช่น บังอร กาญจนบุรี รักเร่ ตูบ ค่านามวิสามัญนั้นหมายถึงสิ่งใดสิ่งหนึ่งโดยเฉพาะอยู่แล้ว จึงมักไม่ต้องมีคำขยายช่วยจำกัดความหมาย นอกจากนี้ค่านามวิสามัญคำใดใช้เรียกชื่อมากกว่า 2 สิ่ง จึงจำเป็นต้องหาคำอื่นมาจำกัดความหมายอีกทีหนึ่ง

- ตัวอย่าง
- 1) *แต่้ม* รักเจ้าของ
 - 2) *แต่้ม* ตัวนั้นรักเจ้าของ

ประโยคที่2) จะมีที่ใช้น้อยกว่าประโยคที่1) มาก เพราะคำว่า *แต่้ม* มีความหมายเฉพาะอยู่แล้วไม่ควรจะต้องมีคำ 'ตัวนั้น' ช่วยจำกัดความหมายลงไปอีก ประโยคที่2) ทำให้เข้าใจว่ามีแมวที่ชื่อแต่้มหลายตัว ผู้พูดจึงต้องระบุว่าเป็นตัวไหน

เปรมจิต ชนะวงส์ (2538: 193) กล่าวถึงชื่อเฉพาะโดยใช้คำศัพท์ว่า "วิสามานยนาม" ดังนี้

วิสามานยนาม หมายถึง ค่านามที่เป็นชื่อเฉพาะ เป็นการระบุลงไปยังอย่างชัดเจนว่าเป็นคนไหน ตัวไหน สิ่งไหน ได้แก่ ชื่อคน ชื่อสัตว์ ชื่อพืช ชื่อสิ่งของ หรือ ชื่ออื่น ๆ

วิสามานยนามจะหมายถึงสิ่งใดสิ่งหนึ่งโดยเฉพาะ จึงไม่ต้องใช้คำขยายมาช่วยจำกัดความหมาย นอกจากนี้วิสามานยนามนั้นใช้เรียกชื่อมากกว่าสองสิ่ง จึงจำเป็นต้องหาคำอื่นมาจำกัดความหมาย

- ตัวอย่าง
- 1) นายอุทัย พิมพ์ใจชน เป็นหัวหน้าพรรคเอกภาพ
 - 2) เรากำลังไปทัศนศึกษาที่ศูนย์ประชุมแห่งชาติสิริกิติ์
 - 3) อภัสราเป็นนางงามจักรวาลคนแรกของไทย
 - 4) ศิลปวัฒนธรรมเป็นวารสารเกี่ยวกับประวัติศาสตร์ไทย
 - 5) เจ้าแต่้มของเขาตายเสียแล้ว

คำอธิบายเกี่ยวกับวิสามานยนามตามแนวคิดของ นวรรธณ (2527) และ เปรมจิต (2538) จะไปในแนวทางเดียวกันและตัวอย่างที่เป็นปัญหาคล้ายกับพระยาอุปกิตศิลปสาร เช่น กรณีของรูปภาษา "(อำเภอ)เมือง" นอกจากนี้ในการให้คำนิยาม ให้ไว้ค่อนข้างกว้างและไม่มีการ

กล่าวถึงอย่างชัดเจนว่าวิสามานยนามสามารถอ้างถึงกลุ่มของสิ่งใดที่เป็นกลุ่มเดียวได้ด้วยหรือไม่ จึงอาจวิเคราะห์ได้ว่าชื่อตระกูล ชื่อหน่วยเงิน ชื่อเครื่องหมายการค้า ชื่อนิตยสาร ชื่อหนังสือพิมพ์ ไม่เป็นวิสามานยนาม ในขณะที่เดียวกันอาจวิเคราะห์ในทางตรงกันข้ามได้ด้วยคือ คำอธิบายหมายรวมถึงกลุ่มของสิ่งใดกลุ่มเดียวด้วยเนื่องจากในตัวอย่างของเปรมจิต (2538) มีตัวอย่างของชื่อวารสารซึ่งมีลักษณะเดียวกันกับชื่อนิตยสาร ชื่อหนังสือพิมพ์ คือ *ศิลปวัฒนธรรม* ในตัวอย่างที่ 4

กำชัย ทองหล่อ (2525: 263) ใช้คำศัพท์ว่า "วิสามานยนาม"

วิสามานยนามคือ คำนามที่เป็นชื่อเฉพาะของคน สัตว์ สถานที่ และสิ่งของ เป็นชื่อที่บัญญัติขึ้นสำหรับใช้เรียกร่องซึ่งเฉพาะเจาะจงลงไปว่า เป็นใครหรืออะไร

ตัวอย่าง สมศรี เป็นหญิงสาวที่เด่นในสังคม
 นายแดงและนายดำเป็นพี่น้องกัน
 เอรಾವถน์ เป็นช่างอยู่ในสวรรค์ชั้นดาวดึงส์
 เขาไปสมัครงานที่กระทรวงศึกษา
 สามก๊กเป็นหนังสือพงศาวดารจีน

นิตยตามแนวคิดนี้ให้ไว้ในลักษณะกว้างๆ การชี้เฉพาะเจาะจงลงไปว่าเป็นใครหรืออะไร ไม่มีขอบเขตที่ชัดเจน เมื่อนำไปใช้ในการวิเคราะห์ตัวอย่างที่เป็นปัญหานั้นที่ได้กล่าวมาแล้ว เช่น "(อำเภอ) เมือง" ชื่อหน่วยเงิน ชื่อเครื่องหมายการค้า ชื่อนิตยสาร-หนังสือพิมพ์ อาจวิเคราะห์ได้ว่า เป็นวิสามานยนาม ในขณะที่เดียวกันมีคำที่ถือว่ามี การชี้เฉพาะเจาะจงเช่นกันแต่คำคำนี้โดยปกติของการใช้ภาษาทั่วไปแล้วจะเป็นสามานยนาม เช่น "พริกขี้หนู" เป็นคำที่มีการชี้เฉพาะเจาะจงว่าเป็นพริกพันธุ์ใด ซึ่งถ้าวิเคราะห์โดยใช้คำอธิบายของพระยาอุปกิตศิลปสาร "พริกขี้หนู" ไม่สามารถตัดสินได้ว่าเป็นวิสามานยนาม เพราะไม่ได้อ้างถึงสิ่งใดสิ่งเดียว แต่อ้างถึงสิ่งที่มีอยู่โดยทั่วไป

Lehrer (1994: 3372) นิยามชื่อเฉพาะ (Proper Name) ไว้ว่า คือ หน่วยในภาษาที่ใช้สำหรับแทนแต่ละคน แต่ละตัว แต่ละอัน หรือ ใช้แทนแต่ละกลุ่ม ที่มีลักษณะเฉพาะเจาะจง โดยชื่อคนเป็นตัวอย่างที่ชัดเจนที่สุดเพราะในทุกๆสังคมแต่ละคนในสังคมต้องมีชื่ออย่างน้อย 1 ชื่อ รวมทั้งในสังคมเมืองขนาดใหญ่ย่อมต้องมีชื่อตระกูล (family name) ด้วย

(" Proper Names are terms for denoting specific individuals or individual groups, with personal names being the most typical. In every society each person has at least one personal name, and in large technological societies, a family name as well. ")

นิยามนี้อธิบายถึงชื่อเฉพาะ (Proper Name) ในลักษณะที่ขอบเขตในการใช้อ้างอิงถึงจะมีขอบเขตกว้างคือ อ้างอิงถึงแต่ละสิ่ง หรือ แต่ละกลุ่มที่มีลักษณะเฉพาะเจาะจง ดังนั้น "(อำเภอ) เมือง" อาจวิเคราะห์ให้เป็นชื่อเฉพาะได้เพราะใช้อ้างอิงซึ่งสถานที่ตั้งเพียงที่เดียวที่มีความเฉพาะเจาะจงตรงที่อยู่คนละจังหวัด ส่วนรูปภาพที่ใช้อ้างอิงถึงเครื่องหมายการค้ายังคงมีความกำกวมในลักษณะที่มีการขยายนำไปใช้อ้างอิงถึงสิ่งอื่นอีกดังที่ได้อภิปรายไปแล้ว

Vaida (1996 : 99) อธิบายว่า ชื่อเฉพาะ (Proper Name) หนึ่งชื่อใช้อ้างอิงถึงสิ่งใดสิ่งเดียว ซึ่งเป็นสิ่งที่มีหนึ่งเดียวในโลก ("...a proper name has a single unique referent...")

เมื่อพิจารณาตามนิยามนี้ รูปภาพ "(อำเภอ)เมือง" หรือ รูปภาพที่ใช้อ้างอิงถึงเครื่องหมายการค้า ชื่อตระกูล ชื่อนิตยสาร-หนังสือพิมพ์ ไม่จัดว่าเป็นชื่อเฉพาะ

Fromkin and Rodman (1998: 169) อธิบายว่า ชื่อเฉพาะ (Proper Name) อ้างอิงถึงวัตถุหรือสิ่งใดที่มีอยู่อันเดียวในโลกโดยอยู่ภายในบริบท รวมทั้งชื่อเฉพาะจะชี้เฉพาะแน่นอน (definite) ซึ่งหมายถึงว่าชื่อเฉพาะจะอ้างอิงถึงในลักษณะที่ผู้พูดและผู้ฟังมีส่วนร่วมต่อกัน เช่น ถ้าพูดว่า "Mary Smith is coming to dinner" ผู้ฟังจะเข้าใจว่า "Mary Smith" ใช้อ้างอิงถึงเพื่อนของผู้พูดและผู้ฟังไม่ใช่อ้างอิงถึง "Mary Smiths" ที่มีจำนวนมากในสมุดโทรศัพท์

("...proper names refer to unique (within context) objects or entities...Proper name are definite, which means they refer to a unique object insofar as the speaker and listener are concerned. If I say

Mary Smith is coming to dinner

my spouse understands Mary Smith to refer to our friend Mary Smith, and not to one of the dozens of Mary Smiths in the phone book....")

เมื่อวิเคราะห์ตามนิยามนี้ รูปภาพ "(อำเภอ)เมือง" ตัดสินได้ว่าเป็นชื่อเฉพาะ เนื่องจากเมื่ออยู่ในบริบทจะอ้างอิงถึงสถานที่ตั้งเพียงบริเวณเดียว เช่น "อำเภอเมือง จ. เพชรบูรณ์" จะหมายถึงสถานที่ตั้งซึ่งเป็นศูนย์กลางของ จ.เพชรบูรณ์ ซึ่งมีได้เพียงพื้นที่บริเวณเดียว แต่อย่างไรก็ตามยังไม่ครอบคลุมถึงบางกรณี เช่น รูปภาพเครื่องหมายการค้า ชื่อหนังสือพิมพ์-นิตยสาร ซึ่งมีลักษณะที่ไม่สามารถนำรูปภาษานั้นไปใช้อ้างอิงถึงสิ่งอื่นสิ่งใดได้อีก ในขณะที่เดียวกันก็อ้างอิงถึงกลุ่มของสิ่งของจำนวนมากไม่ว่าจะปรากฏในบริบทหรือไม่ก็ตาม

Bowman (1998) นิยามชื่อเฉพาะ (Proper Name) ว่าเป็นหน่วยในภาษาที่มีหน้าที่ทางความหมายในการระบุแยกแยะบุคคลใดบุคคลหนึ่ง โดยใช้อ้างอิงถึงได้อย่างชัดเจนและเป็นอิสระโดยไม่ต้องอาศัยบริบท (Context) ใดๆในการทำความเข้าใจว่าชื่อเฉพาะอ้างอิงถึงสิ่งใด นั่นคือ ไม่ว่าชื่อเฉพาะจะปรากฏในประโยค หรือ ปรากฏเป็นคำก็เป็นการอ้างอิงถึงสิ่งเดิม

...proper name function, regardless of context, as self-contained designations for the persons to whom they refer. For example, the noun " Paul " in any context - a sentence, a list or a name badge - has as its semantic function to identify a person... (Bowman 1998: online)

คำอธิบายของ Bowman จะแย้งกับ Fromkin and Rodman ในประเด็นเกี่ยวกับการปรากฏของชื่อเฉพาะ ซึ่ง Bowman มีแนวคิดที่ชื่อเฉพาะใช้สำหรับการสื่อความโดยไม่ต้องอาศัยบริบท ในขณะที่ Fromkin and Rodman มีแนวคิดที่ชื่อเฉพาะจะสื่อความได้อย่างชัดเจนต้องอาศัยบริบทร่วมด้วย

จากนิยามและคำอธิบายต่างๆที่ได้กล่าวมารวมทั้งตัวอย่างข้อมูลที่เป็นปัญหาในการวิเคราะห์ แสดงให้เห็นว่านิยามหรือคำอธิบายต่างๆเกี่ยวกับชื่อเฉพาะยังเป็นแนวคิดที่แตกต่างกันไม่มากนัก แนวคิดหลักในเรื่องของชื่อเฉพาะที่ทุกคนมีร่วมกันคือ เป็นคำที่ใช้อ้างอิงถึงสิ่งใดสิ่งหนึ่งโดยเฉพาะเจาะจง ไม่ว่าจะเป็น คน สัตว์ หรือสิ่งของ และเป็นสิ่งที่มีอยู่หนึ่งเดียวแต่จะมีแนวคิดแตกต่างกันในประเด็นปลีกย่อยคือ การอ้างอิงถึงกลุ่มของสิ่งที่มีจำนวนมากว่าหนึ่งจำเป็นที่จะต้องอ้างอิงถึงกลุ่มของที่แต่ละหน่วยแตกต่างกันและมีลักษณะของการรวมกันเป็นหน่วยเดียว (Unit)หรือไม่ เช่น คนที่มีนามสกุลร่วมกัน หรือสามารถอ้างอิงถึงสิ่งที่มีลักษณะเหมือนกันและไม่จำเป็นต้องมีการรวมกันเป็นหน่วยเดียว เช่น นิตยสารสารคดี รถโตโยตาไซลูน่า หรือกรณีชื่อที่ใช้นั้นใช้เหมือนค่านามทั่วไป เช่น มาม่า หมายถึงบะหมี่กึ่งสำเร็จรูป แพ็บ หมายถึงผงซักฟอก ทั้งนี้อาจเนื่องมาจากโดยลักษณะธรรมชาติของภาษาที่มักจะมีความยืดหยุ่นในการนำไปใช้เพื่อให้ได้ประโยชน์สูงสุด (ดังที่ Fromkin and Rodman, 1998 ได้กล่าวไว้ในหน้า 171 ว่า "...language is nothing if not flexible...") จึงเกิดรูปภาษาที่มีลักษณะกำกวมระหว่างลักษณะของชื่อเฉพาะและค่านามทั่วไป ชื่อเฉพาะจึงมีลักษณะที่มีความต่อเนื่องมากกว่าจะเป็นการแยกว่าใช่หรือไม่ใช่ โดยที่ปลายหัวหนึ่งแทนคำที่เป็นชื่อเฉพาะต้นแบบ (Prototypical Proper Name) เช่น ชื่อคน และอีกปลายหัวหนึ่งเป็นค่านามทั่วไปที่อ้างอิงถึงสิ่งที่มีลักษณะร่วมกัน เช่น คน และคำที่อยู่ระหว่างกลางๆที่

เป็นชื่อที่ตั้งขึ้นมาเพื่อเรียกสิ่งของบางอย่างให้มีความเฉพาะเจาะจง แต่ก็ใช้อ้างถึงกลุ่มของจำนวนมากกว่าหนึ่ง เช่น ชื่อรถโตโยตา

ประเด็นต่างๆที่ได้กล่าวมาแล้วนั้น สะท้อนให้เห็นว่าในการใช้ภาษาที่เกิดขึ้นจริง ชื่อเฉพาะเป็นหน่วยภาษาที่มีขอบเขตกว้างมาก ดังที่ McArthur (1992) ได้กล่าวไว้ในหัวข้อ Names and Naming ว่า ลักษณะธรรมชาติของชื่อเฉพาะคือ สามารถจะเกิดขึ้นใหม่ได้เสมอเพื่อใช้ในการอ้างอิงถึงสิ่งใดสิ่งหนึ่งอย่างต้องการระบุแยกแยะออกมาจากสิ่งอื่นๆ โดยมีการยกตัวอย่างให้เห็นว่าชื่อเฉพาะสามารถเกิดขึ้นได้จากการตั้งชื่อของสิ่งใดๆ ตัวอย่างเช่น กำหนดชื่อให้สัตว์เลี้ยง หมูดาว นาฬิกา บ้านอาคาร ยานพาหนะสาธารณะ รูปภาพ รูปปั้นแกะสลัก บทประพันธ์ตัวละคร นวนิยาย เหตุการณ์สำคัญ ต้นไม้และก้อนหิน ฯลฯ นอกจากนี้ McDonald (1993) ซึ่งทำการศึกษาเรื่อง Internal and External Evidence in the Identification and Semantic Categorization of Proper Names ได้กล่าวสรุปเกี่ยวกับชื่อเฉพาะไว้ประเด็นหนึ่งว่า เมื่อมีการขยายขอบเขตการวิจัยให้ครอบคลุมชื่อเฉพาะมากขึ้น พบว่ามีสิ่งต่างๆมากมายที่สามารถมีชื่อเฉพาะได้ เช่น ภาพยนตร์ หนังสือ กฎหมาย สวนสาธารณะ ปรากฏการณ์ทางดาราศาสตร์ และอื่นๆ ดังนั้นจึงมีความเป็นไปได้ค่อนข้างจำกัดที่จะศึกษาชื่อเฉพาะให้ครอบคลุมทั้งหมดภายใน การวิจัยของวิทยานิพนธ์นี้ วิทยานิพนธ์ฉบับนี้จึงกำหนดขอบเขตที่จะทำการศึกษาชื่อเฉพาะ ต้นแบบซึ่งมีคุณสมบัติทางความหมายตามเกณฑ์นิยาม (definitional criteria) ที่กำหนดขึ้น โดยเกณฑ์นิยามในข้อ (1) และ (2) กำหนดขึ้นจากคำนิยามหรือคำอธิบายของวรรณกรรมต่างๆที่กล่าวผ่านมาโดยอนุมานจากแนวคิดหลักที่พบร่วมกันในวรรณกรรมต่างๆ ส่วนเกณฑ์นิยามในข้อ (3) กำหนดขึ้นเพื่อแยกแยะชื่อเฉพาะออกจากวลีทั่วไป ซึ่งเกณฑ์นิยามของชื่อเฉพาะในการศึกษานี้ มีรายละเอียดดังนี้

- (1) เป็นรูปภาษาที่ใช้อ้างอิงสิ่งที่มีอยู่หนึ่งเดียว กล่าวคือ ชื่อเฉพาะกับสิ่งที่อ้างอิงนั้นมีความสัมพันธ์ในลักษณะหนึ่งต่อหนึ่ง เช่น สมจิตร เป็นชื่อที่ใช้อ้างถึงบุคคลคนหนึ่ง
- (2) สิ่งที่รูปภาษาอ้างอิงอาจเป็นสิ่งเดียวหรือกลุ่มของสิ่งใดๆที่มีลักษณะของการรวมกันเป็นหน่วยเดียว (Unit) เช่น คณะกรรมการการศึกษาแห่งชาติ เป็นชื่อที่ใช้อ้างถึงกลุ่มคนที่รวมกันเป็นหน่วยเดียวที่มีหน้าที่เฉพาะบางอย่าง
- (3) ความหมายของชื่อเฉพาะเป็นไปเพื่อการอ้างอิง กล่าวคือ แม้ว่าชื่อเฉพาะนั้นอาจประกอบด้วยคำต่างๆที่มีความหมาย ความหมายของชื่อเฉพาะนั้นไม่ได้มาจากการรวมความหมายย่อยๆของส่วนต่างๆ เช่น ชื่อพรรคไทยรักไทย แม้ว่าจะประกอบด้วย

คำที่มีความหมาย แต่ชื่อนี้จะถูกใช้ในความหมายของการอ้างถึงโดยไม่สนใจต่อองค์ประกอบความหมายย่อยที่มีในชื่อนั้น

เมื่อพิจารณาตัวอย่างต่อไปนี้ จากเกณฑ์นิยามที่กำหนดใช้ในการศึกษานี้ จะได้ผลดังนี้

"สำนักงานปฏิรูปที่ดินเพื่อเกษตรกรรม" รูปภาษานี้ใช้สำหรับอ้างถึงหน่วยราชการหน่วยหนึ่ง รูปภาษานี้หนึ่งรูปภาษากำหนดขึ้นโดยนำไปใช้เชื่อมโยงกับสิ่งใดเพียงหนึ่งสิ่ง จึงมีความสัมพันธ์ต่อกันในลักษณะหนึ่งต่อหนึ่ง เป็นไปตามเกณฑ์นิยามข้อ (1) และสิ่งที่รูปภาษานี้อ้างถึงมีลักษณะเป็นกลุ่มคนที่มาประกอบกันเป็นหน่วยเดียวเพื่อวัตถุประสงค์ใดวัตถุประสงค์หนึ่ง จึงเป็นตามเกณฑ์นิยามข้อ (2) สุดท้ายรูปภาษานี้ไม่ได้นำไปใช้ในเชิงความหมายว่าทำการปฏิรูปที่ดินสำหรับด้านการเกษตรซึ่งจะเป็นลักษณะที่มีการอาศัยความหมายขององค์ประกอบย่อยมารวมกัน จึงเป็นไปตามเกณฑ์นิยามข้อ (3) รูปภาษานี้จึงเป็นชื่อเฉพาะที่นำมาศึกษาหาบริบทบ่งบอก

แต่ "สำนักงานจัดหางาน" รูปภาษานี้สามารถใช้อ้างถึงองค์กรเอกชนที่ทำธุรกิจเกี่ยวกับการจัดหางานใดๆก็ได้ ซึ่งองค์กรเอกชนลักษณะนี้มีเป็นจำนวนมาก รูปภาษานี้หนึ่งรูปภาษาถูกใช้กับองค์กรเอกชนมากกว่าหนึ่งองค์กร จึงไม่ใช่ความสัมพันธ์แบบหนึ่งต่อหนึ่ง (เกณฑ์นิยามข้อ 1) รูปภาษานี้จึงไม่ใช่ชื่อเฉพาะ

"หนังสือพิมพ์ดิ เอเชียน วอลล์สตรีท" รูปภาษานี้นำมาใช้สำหรับอ้างถึงเครื่องหมายการค้า ซึ่งเครื่องหมายการค้าจะมีลักษณะของการจดลิขสิทธิ์ จึงเป็นชื่อใช้เรียกเฉพาะสินค้านั้นอย่างเดียวนำไปใช้เรียกสินค้าอื่นได้ รูปภาษานี้จึงมีความสัมพันธ์กับสิ่งที่อ้างถึงในลักษณะหนึ่งต่อหนึ่ง เป็นไปตามเกณฑ์นิยามข้อ (1) แต่เมื่อพิจารณาในมุมมองของสิ่งที่รูปภาษานี้อ้างถึง จะพบว่าเป็นสิ่งของจำนวนมากที่มีลักษณะเหมือนกันแต่ไม่มีการรวมกันเป็นหนึ่งหน่วย จึงไม่เป็นไปตามเกณฑ์นิยามข้อ (2) แม้ว่าการใช้ภาษาโดยทั่วไปอาจมีผู้ใช้ภาษาตัดสินว่ารูปภาษานี้เป็นชื่อเฉพาะ แต่สำหรับวิทยานิพนธ์ฉบับนี้จะไม่นำรูปภาษานี้มาศึกษา เนื่องจากไม่เป็นไปตามเกณฑ์นิยามข้อ (2) นอกจากตัวอย่างนี้แล้ว ชื่อรถยนต์ เช่น โตโยตา ฮอนดา ก็มีลักษณะนี้เช่นกัน ไม่ได้ถูกใช้อ้างถึงสิ่งที่มีหนึ่งเดียวหรือเป็นหน่วยเดียว จึงไม่จัดเป็นชื่อเฉพาะในการศึกษานี้

3.2 แนวทางการกำกับชื่อเฉพาะของ TEI

ในหัวข้อนี้จะเริ่มต้นจากการกล่าวแนะนำเกี่ยวกับการกำกับข้อมูลภาษาในคลังข้อมูลภาษาโดยย่อเป็นหัวข้อที่ 3.2.1 หลังจากนั้นจะกล่าวถึงแนวทางการกำกับชื่อเฉพาะตามมาตรฐานของ TEI (Text Encoding Initiative) เป็นหัวข้อที่ 3.2.2 ดังนี้คือ

3.2.1 การกำกับข้อมูลในคลังข้อมูลภาษา

โดยพื้นฐานของการสร้างคลังข้อมูลภาษานั้น จะได้คลังข้อมูลในลักษณะที่ประกอบด้วยข้อความเพียงอย่างเดียว (plain text) ซึ่งแตกต่างกับข้อความที่ปรากฏในแฟ้มข้อมูลจากโปรแกรมประมวลผลคำ (word processor) ที่จะมีการกำกับรหัสพิเศษเพื่อใช้ในการจัดรูปแบบ (format) แทรกอยู่ด้วย (วิโรจน์ อรุณมานะกุล, 2545 : 143) เช่น การขีดเส้นใต้ให้กับข้อความที่ต้องการเน้น สามารถเกิดขึ้นได้โดยอาศัยการกำกับด้วยรหัสพิเศษนี้ ดังนั้นในการที่จะนำคลังข้อมูลภาษาที่ประกอบด้วยข้อความเพียงอย่างเดียวมาใช้ประโยชน์ในการศึกษาภาษาเฉพาะเรื่องดังเช่นการศึกษาภาษาในวิทยานิพนธ์ฉบับนี้ จึงต้องมีการกำกับข้อมูลภาษาส่วนที่จะทำการศึกษา ซึ่งเปรียบเสมือนการกำกับรหัสพิเศษเข้าไปในข้อความที่มีใช้เพื่อให้เกิดผลในการจัดรูปแบบแต่เพื่อใช้สำหรับการวิเคราะห์ภาษาตามวัตถุประสงค์ที่จะทำการศึกษา

การกำกับข้อมูลในที่นี้ จึงหมายถึงการใส่ข้อมูลอื่นเข้าไปในข้อความที่ปรากฏในคลังข้อมูลเพื่อใช้ประโยชน์ในการวิเคราะห์ภาษาตามวัตถุประสงค์ โดยตัวข้อมูลที่ใส่เพิ่มเข้าไปนี้จะมีลักษณะเฉพาะตัวทั้งเรื่องรูปแบบและความหมายที่แตกต่างจากตัวข้อความในคลังข้อมูลและมีคุณสมบัติของการเป็นภาษาหนึ่งเรียกว่าเป็นภาษาสำหรับกำกับข้อมูล (markup language) ซึ่งในปัจจุบันมีการวางแนวทางการกำกับข้อมูลที่เป็นมาตรฐานไว้คือ แนวทางการกำกับข้อมูลภาษาของ TEI โดยองค์ประกอบสำคัญที่ใช้สำหรับกำกับข้อมูลคือ หน่วยกำกับข้อมูลที่ประกอบด้วยแท็ก (Tag) ที่ใช้เครื่องหมาย <...> และมีรูปแบบเป็นแท็กเปิดและแท็กปิดกำกับขอบเขตของข้อมูลภาษาที่ต้องการกำกับตามนี้คือ

<ชื่อแท็ก>ข้อมูลภาษาหน่วยใดๆ</ชื่อแท็ก>

นอกจากนี้ภายในแท็กเปิดอาจประกอบด้วยคุณสมบัติ (Attribute) ที่ใช้สำหรับแสดงให้เห็นคุณสมบัติของข้อมูลภาษาหน่วยนั้น ดังจะเห็นตัวอย่างของรูปแบบการกำกับข้อมูลด้วยแท็กและการกำหนดคุณสมบัติในส่วนถัดไปที่จะได้กล่าวถึงแนวทางการกำกับชื่อเฉพาะของ TEI

3.2.2 แนวทางการกำกับชื่อเฉพาะของ TEI

รายการแท็กที่เกี่ยวข้องกับการใช้กำกับชื่อเฉพาะตามแนวทางของ TEI ได้มีการกล่าวไว้ในรายงานของ TEI อย่างละเอียด ซึ่งสรุปได้ 2 แนวทางหลักคือ วิธีการกำกับข้อมูลในลักษณะที่ไม่ต้องการรายละเอียด และวิธีการกำกับข้อมูลในลักษณะที่ต้องการวิเคราะห์อย่างละเอียดถึงส่วนประกอบต่างๆหรือโครงสร้างภายในของชื่อเฉพาะ (บทที่ 6 หัวข้อที่ 6.4.1 Referring Strings และบทที่ 20 Names and Dates)

วิธีการกำกับข้อมูลในลักษณะที่ไม่ต้องการรายละเอียด เป็นการกำกับข้อมูลอย่างง่ายโดยประกอบด้วยหน่วยกำกับข้อมูล (element) หลัก 2 หน่วยคือ <rs> และ <name> เพื่อให้ระบุว่ารูปลักษณ์นั้นเป็นรูปลักษณ์เพื่อการอ้างถึง (referring strings) โดยรูปลักษณ์เพื่อการอ้างถึงนี้หมายถึง วลีที่อ้างถึงคน (person) สถานที่ (place) วัตถุ (object) และอื่นๆ ซึ่งอาจประกอบด้วยชื่อเฉพาะเพียงอย่างเดียว เช่น เช่น "Mr. Bennet" หรือประกอบด้วยคำนามทั่วไปและชื่อเฉพาะ เช่น "His Excellency the Life President, Ngwazi Dr H. Kamuzu Banda" หรือประกอบด้วยรูปภาษาชนิดอื่นๆ เช่น "his lady" หน่วยกำกับ<rs>สามารถใช้กำกับรูปลักษณ์เพื่อการอ้างถึงในลักษณะใดๆก็ได้ ในขณะที่หน่วยกำกับ<name>จะใช้กำกับกรณียูปลักษณ์เพื่อการอ้างถึงที่เป็นชื่อเฉพาะเพียงอย่างเดียว โดยชื่อเฉพาะนั้นอาจอยู่ในลักษณะคำนามหรือคำนามวลี (proper noun or noun phrase) ก็ได้ ดังเช่นตัวอย่าง 1 และตัวอย่าง 2

ตัวอย่าง 1 My dear <rs type=person>Mr. Bennet</rs>,
said <rs type=person>his lady</rs> to him one day,
have you heard that <rs type=place>Netherfield Park</rs>
is let at last?

ตัวอย่าง 2 My dear <name type=person>Mr. Bennet</name>,
said <rs type=person>his lady</rs> to him one day,
have you heard that <name type=place>Netherfield Park</name>
Is let at last?

นอกจากหน่วยกำกับ<name>อาจถูกใช้ในลักษณะที่มีความหมายเหมือนหน่วยกำกับ<rs>แล้วยังสามารถใช้ซ่อนอยู่ในหน่วยกำกับ<rs> ถ้ารูปลักษณ์เพื่อการอ้างถึงนั้นประกอบด้วยคำนามทั่วไป (common noun) และชื่อเฉพาะ (proper name) ผสมกันตามตัวอย่างต่อไปนี้

```
<rs>His Excellency the Life President,
<name>Ngwazi Dr H. Kamuzu Banda</name></rs>
```

และภายในหน่วยกำกับข้อมูลทั้งสองสามารถที่จะกำหนดคุณสมบัติ (Attributes) ให้กับหน่วยกำกับด้วยคุณสมบัติสามคุณสมบัติต่อไปนี้คือ

type ใช้บ่งชี้ชนิดของสิ่งที่ถูกอ้างถึงให้มีความเฉพาะมากขึ้น ค่าของคุณสมบัติ (value) อาจเป็น "person" "place" "ship" และอื่นๆ

ในตัวอย่าง 1 และตัวอย่าง 2 คุณสมบัตินี้ของหน่วยกำกับ<rs>หรือ<name>คือ *type=person* และ *type=place*

key เป็นตัวบ่งชี้ที่ต่างกันสำหรับสิ่งที่ชี้ที่นั่นอ้างถึง เช่น database record key คุณสมบัติ key อาจมีประโยชน์ในแง่ของการรวบรวมรูปภาพอ้างอิงที่อ้างถึงสิ่งที่มีลักษณะเหมือนให้มาอยู่ด้วยกันหรือในแง่การกระจายของตำแหน่งที่ปรากฏของรูปภาพนั้นตลอดเอกสาร ดังตัวอย่างต่อไปนี้คือ

```
My dear <rs type="person" key="BENM1">Mr. Bennet</rs>,
said <rs type="person" key="BENM2">his lady</rs> to him one day,
have you heard that <rs type="place" key="NETP1">Netherfield Park</rs> is let
at last?
```

reg แสดงรูปแบบปกติ (normalized or regularized form) ของชื่อ คุณสมบัติ reg (regularization) กำหนดขึ้นเพื่อเป็นวิธีในการกำกับรูปมาตรฐาน (standard form) ของรูปภาพเพื่อการอ้างถึง (referring strings) ดังที่แสดงไว้ข้างล่างนี้

```
<name type="person" key="WADLM1" reg="de la Mare, Walter">
Walter de la Mare</name> was born at
<name key="Ch1" type="place">Charlton</name>, in
<name key="KT1" type="county">Kent</name>, in 1873.
```

สำหรับวิธีการกำกับข้อมูลอีกแนวทางหนึ่งเป็นวิธีการกำกับข้อมูลในลักษณะที่ซับซ้อน ซึ่งเหมาะสำหรับการกำกับข้อมูลที่ต้องการวิเคราะห์อย่างละเอียดถึงส่วนประกอบต่างๆ

หรือโครงสร้างภายในของชื่อเฉพาะ แนวทางการกำกับอย่างละเอียดนี้กล่าวไว้ในรายงาน TEI บทที่ 20 โดยกล่าวถึงหน่วยกำกับข้อมูล (element) อย่างละเอียดที่กำหนดขึ้นสำหรับกำกับชื่อเฉพาะแต่ละประเภทแบ่งแยกเป็นแต่ละชุดอย่างชัดเจน ซึ่งในแต่ละชุดหน่วยกำกับจะมีวิธีการกำกับเป็นไปในแนวทางเดียวกัน ผู้วิจัยจึงนำแนวทางวิธีการกำกับชื่อเฉพาะมากล่าวไว้ในวิทยานิพนธ์เพียงสองประเภทคือ ชื่อคนและชื่อสถานที่ เพื่อเป็นการแสดงให้เห็นภาพตัวอย่างการกำกับชื่อเฉพาะไปถึงระดับองค์ประกอบหรือโครงสร้างภายในชื่อเฉพาะ

(1) การกำกับชื่อคน (personal names)

หน่วยกำกับสำหรับชื่อคนประกอบด้วยชุดแท็กต่างๆดังนี้

- <persName> สำหรับกำกับชื่อเฉพาะทั้งที่เป็นคำนามและวลีที่อ้างถึงคน ซึ่งอาจประกอบด้วยส่วนประกอบต่างๆคือ ชื่อ (forename) นามสกุล (surname) คำนำหน้าชื่อ (honorific) ส่วนประกอบเพิ่มเติมของชื่อ (added name) และอื่นๆ
- <surname> สำหรับกำกับส่วนประกอบของชื่อคนที่แสดงถึงนามสกุล (family name or inherited name)
- <foreName> สำหรับกำกับส่วนประกอบของชื่อคนที่แสดงถึงชื่อ (forename, given or baptismal name)
- <roleName> สำหรับกำกับส่วนประกอบของชื่อที่บ่งชี้ว่าผู้ที่ชื่ออ้างถึงมีบทบาท (role) หรือตำแหน่ง (position) ที่เฉพาะ เช่น official titles, rank
- <addName> สำหรับกำกับส่วนประกอบเพิ่มเติมของชื่อ เช่น ชื่อเล่น ฉายา ชื่อปลอม หรือวลีใดๆที่บรรยายอยู่ในชื่อคน
- <nameLink> สำหรับกำกับตัวเชื่อมโยง (connecting phrase or link) ที่ใช้อยู่ภายในชื่อ แต่ไม่ถือว่าเป็นส่วนหนึ่งของชื่อ เช่น "van der" "of"
- <genName> สำหรับกำกับส่วนประกอบของชื่อที่ใช้ในการบ่งชี้ถึงข้อมูลเกี่ยวกับชั้นอายุ (generation) เช่น "Junior" หรือตัวเลขที่ใช้ในชื่อของกษัตริย์

สำหรับคุณสมบัตินี้ (Attributes) ของหน่วยต่างๆในชุดนี้ TEI ได้กำหนดคุณสมบัตินี้ร่วมกัน (share attributes) สำหรับใช้กับทุกหน่วยกำกับ แต่อาจมีรายละเอียดปลีกย่อยบางส่วนของที่แตกต่างไปในบางหน่วย โดยกล่าวสรุปได้ดังนี้

- <persName> ประกอบด้วยคุณสมบัติต่อไปนี้
- type อธิบายเกี่ยวกับชื่อคนให้สมบูรณ์มากขึ้นโดยใช้รายการคำหรือวลีแบบ
 ปลายเปิด ซึ่งจะช่วยให้การนำรูปภาษาไปใช้ประโยชน์ (function) ได้
 เช่น "married name" "maiden name" "pen name" "religious name"
 เป็นต้น
 - key เป็นตัวบ่งชี้ที่ต่างกันสำหรับสิ่งที่ชื่อนั้นอ้างถึง เช่น database record key
 คุณสมบัตินี้ key อาจมีประโยชน์ในแง่ของการรวบรวมรูปภาษาอ้างอิงที่
 อ้างถึงสิ่งที่มีลักษณะเหมือนให้มาอยู่ด้วยกันหรือในแง่การกระจายของ
 ตำแหน่งที่ปรากฏของรูปภาษานั้นตลอดเอกสาร
 - reg แสดงรูปแบบปกติ (normalized or regularized form) ของชื่อ
- <surname>, <foreName>, <roleName>, <addName>, <nameLink> และ
 <genName> ประกอบด้วยคุณสมบัติซึ่งเป็นคุณสมบัติรวมของทุกหน่วยดังนี้คือ
- type ให้ข้อมูลที่เฉพาะเจาะจงเพื่อใช้จัดหมวดหมู่ส่วนประกอบ
 - key เป็นตัวบ่งชี้ที่ต่างกันสำหรับสิ่งที่ชื่อนั้นอ้างถึง เช่น database record key
 คุณสมบัตินี้ key อาจมีประโยชน์ในแง่ของการรวบรวมรูปภาษาอ้างอิงที่
 อ้างถึงสิ่งที่มีลักษณะเหมือนให้มาอยู่ด้วยกันหรือในแง่การกระจายของ
 ตำแหน่งที่ปรากฏของรูปภาษานั้นตลอดเอกสาร
 - reg แสดงรูปแบบปกติ (normalized or regularized form) ของชื่อ
 - full บ่งชี้การปรากฏ
 - ค่ามาตรฐานของคุณสมบัติ (legal value) คือ
 - yes ส่วนประกอบของชื่อเขียนรูปเต็ม
 - abb ส่วนประกอบของชื่อ ปรากฏในรูปตัวย่อ
 - init ส่วนประกอบของชื่อ ปรากฏเพียงส่วนเริ่มต้น
 - sort ระบุลำดับขององค์ประกอบของชื่อโดยเรียงลำดับตามความสัมพันธ์กับ
 องค์ประกอบอื่นในชื่อคนนั้น

ในส่วนถัดไปจะยกตัวอย่างบางส่วนจากรายงานของ TEI มาแสดงให้เห็นลักษณะการกำกับชื่อ
 เฉพาะที่วิเคราะห์กำกับส่วนประกอบต่างๆภายในชื่อเฉพาะอย่างละเอียด

```

<persName key="DUDO1">
  <roleName type="honorific" full="abb">Mme</roleName>
  <nameLink>de la</nameLink>
  <surname>Rochefoucault</surname>
</persName>

<persName key="EGBR1">
  <roleName type="office">Governor</roleName>
  <foreName sort="2">Edmund</foreName>
  <foreName reg="Gerald" full="init" sort="3">G</foreName>.
  <addName type="nick">Jerry</addName>
  <addName type="epithet">Moonbeam</addName>
  <surname sort="1">Brown</surname>
  <genName full="abb">Jr</genName>.
</persName>

```

(2) ชื่อสถานที่ (place names)

ในกรณีของชื่อสถานที่ TEI ได้พิจารณาจำแนกชื่อสถานที่ออกเป็น 3 ลักษณะและกำหนดชุดแท็กสำหรับการกำกับรายละเอียดต่างๆของชื่อสถานที่แต่ละลักษณะไว้ โดยชื่อสถานที่ทั้ง 3 ลักษณะได้แก่

- (2.1) Geo-political Place Names คือชื่อสถานที่ที่เกี่ยวข้องกับเขตการปกครอง มักมีการเรียงลำดับตามขนาดหรือความสำคัญทางการปกครองโดยเรียงแบบน้อยไปหามาก (ascending)
- (2.2) Geographic Names คือชื่อสถานที่ที่ถูกตั้งขึ้นโดยใช้ศัพท์ที่แสดงคุณลักษณะทางภูมิศาสตร์ (geographic features) เช่น ภูเขา ทะเลสาบ แม่น้ำ
- (2.3) Relative Place Names คือชื่อสถานที่ที่ระบุโดยการอ้างอิงถึงอีกสถานที่หนึ่ง ตัวอย่างเช่น "10 miles northeast of Paris" หรือ "near the top of Mount Sinai"

สำหรับชุดแท็กที่ใช้กำกับชื่อสถานที่และองค์ประกอบย่อยภายในโครงสร้างของชื่อสถานที่จะกล่าวสรุปดังนี้คือ

| | | |
|---|-------------|---|
| หน่วยกำกับรวม ของชื่อสถานที่ ทั้ง 3 กลุ่ม | <placeName> | สำหรับกำกับชื่อสถานที่ที่ใช้แทนสถานที่เพียงที่เดียว (absolute place name) หรือชื่อสถานที่ที่อ้างถึงโดยอิงกับ ที่อีกที่หนึ่ง (relative place name) |
|---|-------------|---|

| | |
|------------------------------|--|
| Geo-political Place Names | <p><settlement> สำหรับกำกับชื่อสถานที่ที่มีการตั้งถิ่นฐาน เช่น หมู่บ้าน (village) เมือง (town) เมืองสำคัญ (city)</p> <p><region> สำหรับกำกับหน่วยบริหารการปกครอง (administrative unit) ซึ่งเป็นเขตปกครองที่มีขนาดใหญ่กว่าหรือสำคัญกว่าชื่อสถานที่ที่มีลักษณะการตั้งถิ่นฐาน (settlement) แต่มีขนาดเล็กกว่าหรือมีความสำคัญน้อยกว่าชื่อสถานที่ที่หมายถึงประเทศ (country) ตัวอย่างของหน่วยนี้ดังเช่น จังหวัดหรือมณฑล (county) รัฐ (state)</p> <p><country> สำหรับกำกับสถานที่ที่ถือเป็นเขตการปกครองของชาติ</p> <p><bloc> สำหรับกำกับกลุ่มของสถานที่ที่มีลักษณะตามสิ่งที่ถูกกำกับด้วยหน่วยกำกับ <country></p> |
|------------------------------|--|

ตัวอย่าง 3

```
<placeName key="RNY1">
  <settlement type="city">Rochester</settlement>,
  <region type="state">New York</region>
</placeName>

<placeName key="LSEA1">
  <country type="nation">Laos</country>,
  <bloc type="sub-continent">Southeast Asia</bloc>
</placeName>
```

| | |
|---------------------|--|
| Geographic Names | <p><geogName> ชื่อที่มีคุณลักษณะทางภูมิศาสตร์ (geographic feature) เช่น หุบเขา ใน "Windrush Valley" ภูเขา ใน "Mount Sinai"</p> <p><geog> สำหรับกำกับสามัญนามที่ปรากฏอยู่ ซึ่งเป็นตัวชี้แสดงถึงคุณลักษณะทางภูมิศาสตร์บางอย่าง เช่น หุบเขา ภูเขา</p> |
|---------------------|--|

ตัวอย่าง 4

```
<geogName key="MIRI1" type="river">
  <name>Mississippi</name>
  <geog>River</geog>
</geogName>
```

| | |
|-------------------------|---|
| Relative Place Names | <p>แตกต่างกันสำหรับกำกับชื่อสถานที่ที่ได้กล่าวมาแล้วนั้น เป็นแท็กที่ใช้กำกับชื่อสถานที่ที่ใช้อ้างถึงสถานที่เพียงที่เดียว (absolute place name) ซึ่งการระบุถึง</p> |
|-------------------------|---|

สถานที่อาจระบุในรูปของการอ้างอิงสถานที่นั้นกับอีกสถานที่หนึ่ง (relative place name) ตัวอย่างเช่น “10 miles northeast of Paris” หรือ “near the top of Mount Sinai” ซึ่งการระบุสถานที่ในลักษณะเชื่อมโยงนี้จะประกอบด้วย

ชื่อสถานที่ (place name) ซึ่งทำหน้าที่เป็นตัวให้อ้าง (referent) เช่น "Paris" "Mount Sinai"

คำหรือวลีที่แสดงให้เห็นถึงตำแหน่งของสถานที่นั้นโดยอ้างอิงกับตัวให้อ้าง เช่น "the top of" "north of"

ระยะทาง (distance) อาจมีการระบุถึงระยะทางระหว่างสถานที่ให้อ้างกับสถานที่ที่ต้องการกล่าวถึง โดยสามารถระบุระยะทางโดยประมาณได้ เช่น "10 miles" "near"

การกำกับองค์ประกอบของชื่อสถานที่ในลักษณะนี้ สามารถใช้หน่วยต่อไปนี้

<distance> ใช้ระบุถึงระยะทาง

<offset> ใช้แสดงถึงทิศทาง

ตัวอย่าง 5 <placeName key="NEPA1">
 <distance>10 miles</distance>
 <offset>north of</offset>
 <settlement type="city">Paris</settlement>
 </placeName>

จากตัวอย่าง 3-5 จะเห็นได้ว่าการกำกับชื่อสถานที่จะมีการกำหนดคุณสมบัติเช่นกัน ซึ่งเป็นคุณสมบัติรวมของทุกหน่วยอันได้แก่

type ให้ข้อมูลที่เฉพาะเจาะจงเพื่อใช้จัดหมวดหมู่ของส่วนประกอบ

key เป็นตัวบ่งชี้ที่ต่างกันสำหรับสิ่งที่ชื่อนั้นอ้างอิง เช่น database record key

คุณสมบัติ key อาจมีประโยชน์ในแง่ของการรวบรวมรูปภาพอ้างอิงที่อ้างอิงถึงสิ่งที่มีลักษณะเหมือนให้มาอยู่ด้วยกันหรือในแง่การกระจายของตำแหน่งที่ปรากฏของรูปภาพนั้นตลอดเอกสาร

reg แสดงรูปแบบปกติ (normalized or regularized form) ของชื่อ

full ป่งชี้การปรากฏ

ค่ามาตรฐานของคุณสมบัติ (legal value) คือ

yes ส่วนประกอบของชื่อเขียนรูปเต็ม

abb ส่วนประกอบของชื่อ ปรากฏในรูปตัวย่อ

init ส่วนประกอบของชื่อ ปรากฏเพียงส่วนเริ่มต้น

เมื่อพิจารณาโดยรวมตามแนวทางการกำกับชื่อเฉพาะของ TEI ถ้าไม่คำนึงถึงการกำกับส่วนประกอบภายในของชื่อเฉพาะ ตัวแท็ก <persName> มีความหมายเหมือนกับ <name type="person"> หรือ <placeName> มีความหมายเหมือนกับ <name type="place"> ยกเว้นแต่ว่าคุณสมบัติ type ของ <persName> และ <placeName> จะมีความสามารถในการแบ่งออกเป็นประเภทย่อยไปได้อีก ตัวอย่างเช่น <persName> คุณสมบัติ type จะระบุชื่อคนนั้นว่าเป็นประเภท "married" "maiden" "pen" "pseudo" หรือ "religious"

โดยสรุปแล้ว ในวิทยานิพนธ์ฉบับนี้จะกำกับชื่อเฉพาะที่มีคุณสมบัติตามเกณฑ์นิยามที่กำหนดและนำแนวทางการกำกับชื่อเฉพาะของ TEI มาใช้ โดยการใส่แท็ก <rs> และ <name> (ดูหัวข้อ 4.1.3.1.2) ส่วนแท็กสำหรับกำกับบริบทบ่งบอกนั้นได้นำเกณฑ์การกำกับข้อมูลของ TEI มาประยุกต์ใช้ (ดูหัวข้อ 4.1.3.2) ซึ่งนอกจากการใช้แท็กสำหรับกำกับชื่อเฉพาะและข้อมูลที่เกี่ยวข้องกับการศึกษาบริบทบ่งบอกแล้ว ยังมีการใช้แท็กสำหรับกำกับข้อมูลที่เกี่ยวข้องกับโครงสร้างของคลังข้อมูล เช่น แหล่งที่มาของข้อมูลภาษาที่รวบรวมมาสร้างเป็นคลังข้อมูล ซึ่งผู้วิจัยประยุกต์มาจากแนวทางของ TEI เช่นกัน โดยรายละเอียดของแท็กที่ใช้กำกับข้อมูลจะกล่าวไว้ในบทถัดไปคือบทที่ 4

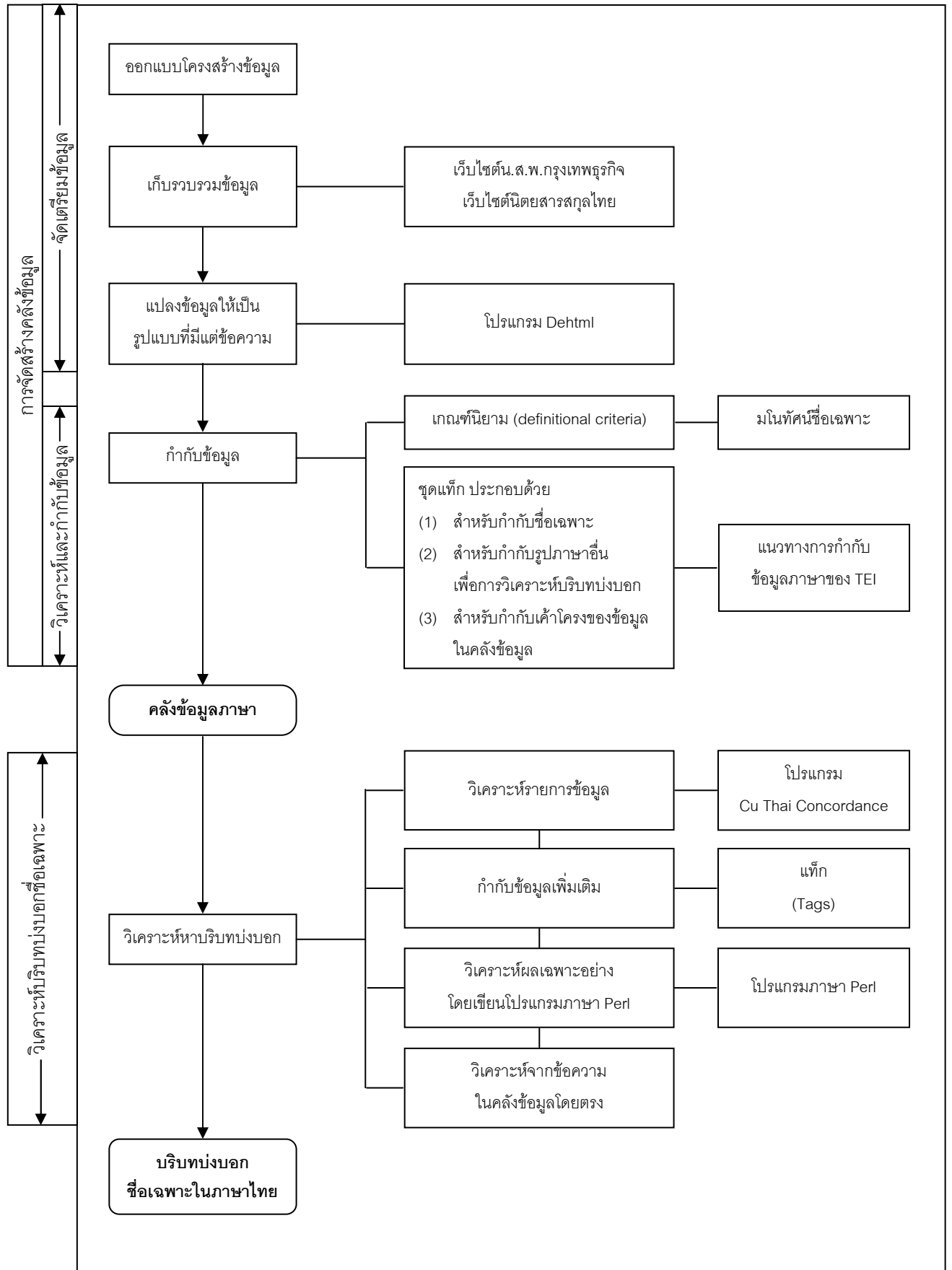
บทที่ 4

การจัดสร้างคลังข้อมูลและการวิเคราะห์บริบทบ่งบอก

ในบทนี้เป็นเนื้อหาเกี่ยวกับการดำเนินการวิจัยของขั้นตอนหลักสองขั้นตอน คือ การจัดสร้างคลังข้อมูลภาษาและการวิเคราะห์หาบริบทบ่งบอกที่เฉพาะจากคลังข้อมูล โดยจะกล่าวถึงรายละเอียดเกี่ยวกับการจัดสร้างคลังข้อมูลไว้ในหัวข้อที่ 4.1 และรายละเอียดเกี่ยวกับการวิเคราะห์บริบทบ่งบอกในหัวข้อที่ 4.2 ซึ่งลำดับขั้นตอนต่างๆของวิธีดำเนินการวิจัยในการจัดสร้างคลังข้อมูล และการวิเคราะห์หาบริบทบ่งบอก สามารถสรุปเป็นภาพรวมดังแผนภาพที่ปรากฏในรูปที่ 4-1

4.1 การจัดสร้างคลังข้อมูลภาษา

เนื่องจากการศึกษาบริบทบ่งบอกที่เฉพาะในภาษาไทยของวิทยานิพนธ์นี้ มีเป้าหมายจะศึกษาหาข้อสรุปทางภาษาที่สามารถใช้เป็นความรู้ทางภาษาสำหรับการพัฒนางานด้านภาษาศาสตร์คอมพิวเตอร์หรือการประมวลผลภาษาธรรมชาติ ซึ่งในยุคปัจจุบันให้ความสำคัญกับการประมวลผลข้อมูลภาษาที่เกิดขึ้นจริงในการใช้ภาษา ดังนั้นวิทยานิพนธ์ฉบับนี้จึงศึกษาปรากฏการณ์ทางภาษาของข้อเฉพาะที่สามารถเป็นบริบทบ่งบอกจากข้อมูลภาษาที่เกิดขึ้นจริง ซึ่งก็คือ คลังข้อมูลภาษา โดยคลังข้อมูลภาษาที่ใช้ในการศึกษาครั้งนี้เป็นคลังข้อมูลภาษาที่ผู้วิจัยจัดสร้างขึ้นเพื่อการศึกษาบริบทบ่งบอกโดยเฉพาะ คลังข้อมูลที่จัดสร้างขึ้นนี้เป็นคลังข้อมูลประเภทที่มีการกำกับข้อมูล (annotated corpus) กล่าวคือ เป็นคลังข้อมูลที่ประกอบด้วยข้อความภาษาไทยและมีการกำกับข้อมูลส่วนที่เป็นข้อเฉพาะรวมทั้งมีการกำกับรูปภาษาอื่นนอกเหนือจากข้อเฉพาะเพื่อสนับสนุนการวิเคราะห์เกี่ยวกับบริบทบ่งบอก



รูปที่ 4-1 กระบวนการขั้นตอนจัดสร้างคลังข้อมูลและวิเคราะห์บริบทของข้อเฉพาะถึงข้อเฉพาะ

สำหรับวิธีดำเนินการจัดสร้างคลังข้อมูลจะประกอบด้วยกระบวนการย่อย 2 กระบวนการคือ การจัดเตรียมข้อมูลและการวิเคราะห์กำกับข้อมูล (ดูรูปที่ 4-1) ซึ่งกระบวนการจัดเตรียมข้อมูลจะมีลำดับขั้นตอนดังนี้คือ ขั้นตอนแรกเป็นการออกแบบโครงสร้างของข้อมูล (หัวข้อที่ 4.1.1) ขั้นตอนถัดมาเป็นการเก็บรวบรวมข้อมูลตามโครงสร้างที่ได้ออกแบบไว้แล้วทำการแปลงข้อมูลให้มีสภาพเป็นข้อความเพียงอย่างเดียว (หัวข้อที่ 4.1.2) สำหรับกระบวนการวิเคราะห์และกำกับข้อมูล (หัวข้อที่ 4.1.3) จะประกอบด้วยการกำกับข้อมูลสามส่วนคือ ส่วนที่หนึ่งเป็นการกำกับชื่อเฉพาะ (หัวข้อที่ 4.1.3.1) ในขั้นตอนของการกำกับชื่อเฉพาะนี้ จะกล่าวถึงวิธีการวิเคราะห์หารูปภาษาที่เป็นชื่อเฉพาะตามเกณฑ์นิยาม (หัวข้อย่อยที่ 4.1.3.1.1) และกล่าวถึงแท็กสำหรับการกำกับชื่อเฉพาะ (หัวข้อย่อยที่ 4.1.3.1.2) การกำกับข้อมูลส่วนที่สองเป็นการกำกับบริบทบ่งบอก (หัวข้อที่ 4.1.3.2) และการกำกับข้อมูลส่วนสุดท้ายเป็นการกำกับเค้าโครงของข้อมูลในคลังข้อมูล (หัวข้อที่ 4.1.3.2 เช่นกัน)

4.1.1 ออกแบบโครงสร้างของข้อมูล

ในความเป็นจริงนั้น แหล่งข้อมูลทางภาษามีความหลากหลายและปรากฏเป็นจำนวนมากมาย การจัดสร้างคลังข้อมูลภาษาจึงต้องมีการกำหนดขอบเขตของข้อมูลภาษาที่จะรวบรวมเป็นคลังข้อมูลเพื่อให้ตรงกับจุดประสงค์ของการนำคลังข้อมูลไปใช้งาน โดยขอบเขตของข้อมูลภาษาจะพิจารณาจากจุดประสงค์เป็นหลัก รวมทั้งควรคำนึงถึงปัจจัยประการอื่นที่มีความเกี่ยวข้องหรืออาจจะเป็นข้อจำกัด เช่น ขนาดของคลังข้อมูลหรือปริมาณของหน่วยภาษาเป้าหมายที่ปรากฏในคลังข้อมูลซึ่งปัจจัยนี้มีผลต่อความน่าเชื่อถือในผลการศึกษา ความสะดวกในการเข้าถึงข้อมูลหรือการเก็บรวบรวมข้อมูล เป็นต้น ผู้วิจัยจึงได้ออกแบบโครงสร้างข้อมูลให้มีขอบเขตของข้อมูลดังนี้คือ

- (1) ข้อมูลมีรูปแบบเป็นภาษาเขียนที่ปรากฏบนเอกสารอิเล็กทรอนิกส์
- (2) เป็นเอกสารอิเล็กทรอนิกส์ในสื่อประเภทหนังสือพิมพ์และนิตยสารทั่วไป เนื่องจากภาษาที่ปรากฏในสื่อทั้งสองประเภทเป็นการใช้ภาษาที่เกิดขึ้นจริงที่พบเห็นได้ทั่วไป
- (3) แหล่งข้อมูลอิเล็กทรอนิกส์ที่เป็นสื่อประเภทหนังสือพิมพ์และนิตยสารทั่วไป ปรากฏเป็นจำนวนมากบนอินเทอร์เน็ต การพิจารณาเว็บไซต์ที่จะนำข้อมูลมาสร้างเป็นคลังข้อมูลจะพิจารณาจากความสะดวกในการเก็บรวบรวมข้อมูลคือ เป็นเว็บไซต์ที่สามารถทำการดาวน์โหลด (download) ข้อมูลครั้งละฉบับ

- (4) สัดส่วนของข้อมูลจากแต่ละแหล่งข้อมูลคือ แหล่งข้อมูลละ 1 ฉบับ เช่น หนังสือพิมพ์รายวัน จะนำข้อมูลของฉบับวันใดวันหนึ่งมาสร้างเป็นคลังข้อมูล นิตยสารรายเดือนจะนำเฉพาะข้อมูลของฉบับเดือนใดเดือนหนึ่งมาสร้างเป็นคลังข้อมูล เป็นต้น
- (5) จำนวนของแหล่งข้อมูลที่จะนำข้อมูลมาสร้างคลังข้อมูล ไม่กำหนดไว้ตายตัวสามารถเพิ่มจำนวนได้ ขึ้นอยู่กับจำนวนชื่อเฉพาะที่ปรากฏอยู่ในคลังข้อมูล ซึ่งโดยรวมแล้วมีจำนวนอย่างน้อย 2,000 ชื่อ และจำนวนแหล่งข้อมูลที่เป็นหนังสือพิมพ์จะเท่ากับจำนวนแหล่งข้อมูลที่เป็นนิตยสาร

4.1.2 เก็บรวบรวมข้อมูล

จากการสำรวจเอกสารอิเล็กทรอนิกส์โดยพิจารณาตามโครงสร้างที่ได้ออกแบบไว้ ข้อมูลภาษาไทยที่ได้เก็บรวบรวมมาใช้เป็นคลังข้อมูลในวิทยานิพนธ์ ได้แก่ ข้อความภาษาไทยที่เผยแพร่ไว้ในเว็บไซต์ของหนังสือพิมพ์กรุงเทพธุรกิจ (<http://www.bangkokbiznews.com>) และเว็บไซต์ของนิตยสารสกุลไทย (<http://www.sakulthai.com>) โดยข้อมูลของหนังสือพิมพ์กรุงเทพธุรกิจเป็นข้อมูลของฉบับวันที่ 30 สิงหาคม พ.ศ. 2545 ลักษณะของเนื้อหาเป็นรายงานข่าว บทความ และคอลัมน์ ส่วนข้อมูลของนิตยสารสกุลไทยซึ่งเป็นนิตยสารรายสัปดาห์ได้นำข้อมูลจากฉบับที่ 2511 ประจำวันอังคารที่ 3 ธันวาคม พ.ศ. 2545 เนื้อหาส่วนที่ได้เผยแพร่อยู่ในเว็บไซต์ประกอบด้วย บทความ สารคดี คอลัมน์ และบทสัมภาษณ์ ในการรวบรวมข้อมูลจากเว็บไซต์ทั้งสองผู้วิจัยได้ใช้เครื่องมือตัวหนึ่งในการรวบรวมคือ โปรแกรม WinHTTrack Website Copier (พัฒนาโดย Roche and Other Contributors, 1998-2002)

ข้อมูลที่ได้เก็บรวบรวมมาได้เป็นข้อมูลที่ประกอบด้วย HTML tag ซึ่งไม่สามารถกำกับชื่อเฉพาะได้ จึงทำการแปลงให้มีสภาพเป็นข้อความเพียงอย่างเดียว (plain text) โดยใช้โปรแกรมแปลงข้อมูลที่ชื่อว่า โปรแกรม Dehtml (พัฒนาโดย Wirote Aroonmanakun: Department of Linguistics, Chulalongkorn University) ผลลัพธ์ที่ได้เป็นแฟ้มข้อมูลที่ประกอบด้วยข้อความเพียงอย่างเดียว เป็นข้อความต่อเนื่องเหมือนกับที่ปรากฏในเว็บไซต์และผู้วิจัยยังคงข้อความและเครื่องหมายใดๆภายในข้อความต่อเนื่องโดยไม่ทำการเปลี่ยนแปลงแล้วจัดเก็บเป็นคลังข้อมูล เพื่อทำการวิเคราะห์กำกับข้อมูลเป็นขั้นตอนถัดไป

4.1.3 กำกับข้อมูล

การกำกับข้อมูลภาษาในวิทยานิพนธ์นี้หมายถึง การใส่ข้อมูลทางภาษาศาสตร์บางอย่างเพิ่มเข้าไปในคลังข้อมูลภาษาที่ประกอบด้วยข้อความภาษาไทย เพื่อประโยชน์ใน

การศึกษานหาบริบทบ่งบอกชื่อเฉพาะ ข้อมูลที่กำกับเพิ่มเข้าไปในคลังข้อมูลคือ การระบุรูปภาษาที่เป็นชื่อเฉพาะ การระบุรูปภาษาเพิ่มเติมเพื่อการวิเคราะห์บริบทบ่งบอก โดยใช้สัญลักษณ์ที่เรียกว่า แท็ก (Tag) เป็นสิ่งแสดงถึงข้อมูลดังกล่าว ซึ่งวิธีการกำกับข้อมูลโดยใช้แท็กเป็นไปตามแนวทางการกำกับข้อมูลภาษาของ TEI (Text Encoding Initiative) ที่ได้กล่าวรายละเอียดไว้ในบทที่ 3 สำหรับส่วนนี้จะเป็นการกล่าวถึงรายละเอียดการกำกับข้อมูลภาษาของวิทยานิพนธ์ที่ประยุกต์จากแนวทางการกำกับข้อมูลภาษาของ TEI ได้แก่ 4.1.3.1 การกำกับชื่อเฉพาะ และ 4.1.3.2 การกำกับรูปภาษาอื่นเพื่อการวิเคราะห์บริบทบ่งบอก

4.1.3.1 การกำกับชื่อเฉพาะ

ในกระบวนการกำกับข้อมูลที่เป็นชื่อเฉพาะนั้นมีประเด็นที่เกี่ยวข้องกับกระบวนการด้วยกันสองประการคือ การวิเคราะห์หารูปภาษาที่เป็นชื่อเฉพาะและชุดแท็ก (Tag set) ที่ใช้กำกับชื่อเฉพาะ ซึ่งการพิจารณาถึงรูปภาษาที่เป็นชื่อเฉพาะจะพิจารณาจากเกณฑ์นิยาม (definitional criteria) และสำหรับชุดแท็กเป็นแท็กที่มีรูปแบบตามมาตรฐานกำกับข้อมูลของ TEI โดยจะกล่าวถึงประเด็นการวิเคราะห์หารูปภาษาที่เป็นชื่อเฉพาะในหัวข้อที่ 4.1.3.1.1 และ แท็กสำหรับกำกับชื่อเฉพาะในหัวข้อที่ 4.1.3.1.2

4.1.3.1.1 การวิเคราะห์หารูปภาษาที่เป็นชื่อเฉพาะ

เนื่องจากในภาษาไทยไม่มีลักษณะของการกำหนดตัวบ่งชี้เพื่อใช้แสดงว่ารูปภาษานั้นเป็นชื่อเฉพาะ นอกจากนี้ชื่อเฉพาะในภาษาไทยยังไม่มีลักษณะโครงสร้างทางไวยากรณ์ที่สามารถจะบ่งบอกได้ว่ารูปภาษานั้นเป็นชื่อเฉพาะ ดังนั้นในการวิเคราะห์หาข้อมูลส่วนใดเป็นชื่อเฉพาะจึงต้องวิเคราะห์จากคุณสมบัติเชิงความหมาย กล่าวคือ พิจารณาจากคำนิยามหรือคำอธิบายของชื่อเฉพาะ ซึ่งในวรรณกรรมต่างๆจะมีรายละเอียดบางส่วนแตกต่างกันและขอบเขตของรูปภาษาที่จัดว่าเป็นชื่อเฉพาะก็มีความแตกต่างกัน ทำให้เกิดความสับสนในการตัดสินใจว่ารูปภาษาที่วิเคราะห์เป็นชื่อเฉพาะหรือไม่ใช่ชื่อเฉพาะ ด้วยสาเหตุดังกล่าว จึงได้มีการกำหนดเกณฑ์นิยามที่จะใช้สำหรับการพิจารณาถึงชื่อเฉพาะในการศึกษา โดยเกณฑ์นิยาม* ที่กำหนดเป็นเกณฑ์นิยามทางด้านความหมายที่ประกอบด้วยเกณฑ์ความหมายต่อไปนี้

- (1) เป็นรูปภาษาที่ใช้อ้างอิงสิ่งที่มีอยู่หนึ่งเดียว กล่าวคือ ชื่อเฉพาะกับสิ่งที่อ้างอิงนั้นมี ความสัมพันธ์ในลักษณะหนึ่งต่อหนึ่ง

* ดูรายละเอียดในบทที่ 3, หน้า 27-29

- (2) สิ่งที่มีรูปร่างภาษาอ้างถึงอาจเป็นสิ่งเดียวหรือกลุ่มของสิ่งใด ๆ ที่มีลักษณะของการรวมกันเป็นหน่วยเดียว (Unit)
- (3) ความหมายของชื่อเฉพาะเป็นไปเพื่อการอ้างถึง กล่าวคือ แม้ว่าชื่อเฉพาะนั้นอาจประกอบด้วยคำต่างๆที่มีความหมาย ความหมายของชื่อเฉพาะนั้นไม่ได้มาจากการรวมความหมายย่อยๆ ของส่วนต่างๆ

ในการวิเคราะห์โดยใช้เกณฑ์นิยามนี้ ผู้วิจัยจะพิจารณาคคุณสมบัติทางความหมายของรูปภาษาที่ทำการวิเคราะห์ว่าเป็นไปตามเกณฑ์นิยามหรือไม่เป็นไปตามเกณฑ์นิยาม โดยอาศัยความรู้ภาษาแม่ (native intuition) ของผู้วิจัย ซึ่งบางรูปภาษาผู้วิจัยพิจารณาตัดสินได้จากตัวรูปภาษา แต่บางรูปภาษาผู้วิจัยต้องพิจารณาตัดสินจากการประมวลผลความหมายที่สื่อจากข้อความที่อยู่รอบข้างของรูปภาษานั้น หรืออาศัยข้อความต่อเนื่องที่รูปภาษานั้นปรากฏอยู่ อย่างไรก็ตามในการวิเคราะห์จะพบว่า มีรูปภาษาที่น่าจะเป็นชื่อเฉพาะแต่ไม่สามารถวิเคราะห์ถึงคุณสมบัติตามเกณฑ์นิยามอันเนื่องมาจากข้อความต่างๆที่ปรากฏอยู่ร่วมกับรูปภาษานั้น ไม่แสดงความหมายเชิงอ้างอิงตามเกณฑ์นิยามอย่างชัดเจนที่จะทำให้ผู้วิจัยตัดสินได้ และรูปภาษานั้นอาจจะปรากฏด้วยความถี่น้อย ปรากฏเพียงครั้งเดียวภายในคลังข้อมูล ซึ่งในกรณีนี้จะวิเคราะห์รูปภาษาดังกล่าวโดยใช้ข้อมูลเพิ่มเติมจากรายการผลลัพธ์ของโปรแกรมคอนคอร์แดนซ์ที่เผยแพร่บนอินเทอร์เน็ตที่ประมวลผลจากคลังข้อมูลภาษาขนาดใหญ่และประกอบด้วยข้อมูลภาษาที่เกิดขึ้นจริงในการใช้ภาษา (Online Cu Thai Concordance, Aroonmanakun, 1999-2001) รวมทั้งวิเคราะห์เพิ่มเติมจากข้อความต่อเนื่องในเอกสารอิเล็กทรอนิกส์ที่สืบค้นได้จากบนอินเทอร์เน็ต

4.1.3.1.2 แท็กสำหรับการกำกับชื่อเฉพาะ

เมื่อทำการวิเคราะห์ได้รูปภาษาที่เป็นชื่อเฉพาะตามเกณฑ์นิยามแล้ว ขั้นตอนถัดมาคือการกำกับชื่อเฉพาะที่พบด้วยหน่วยกำกับข้อมูลหรือแท็ก ซึ่งชื่อเฉพาะที่ปรากฏในคลังข้อมูลปรากฏอยู่ในสองลักษณะคือ ปรากฏเป็นรูปภาษาหนึ่งหน่วยที่มีคุณสมบัติเป็นชื่อเฉพาะตามเกณฑ์นิยาม เช่น "ไทย" หรือปรากฏเป็นส่วนประกอบของรูปภาษาอื่นโดยความหมายของทั้งหน่วยอาจเป็นชื่อเฉพาะหรือไม่เป็นชื่อเฉพาะตามเกณฑ์นิยาม เช่น "ธนาคารไทยพาณิชย์" เป็นชื่อเฉพาะ "คนไทย" ขาดคุณสมบัติของการเป็นชื่อเฉพาะตามเกณฑ์นิยาม (ดูรายละเอียดบทที่ 5) แท็กสำหรับกำกับชื่อเฉพาะจึงประกอบด้วยแท็กต่อไปนี้เป็นชื่อ

<rs>

- ใช้กำกับรูปภาพหนึ่งหน่วยที่ปรากฏชื่อเฉพาะตามเกณฑ์นิยาม ประกอบด้วยคุณสมบัติ (Attribute) ตามนี้คือ

type ใช้ระบุลักษณะการปรากฏของชื่อเฉพาะ โดยมีค่าของคุณสมบัติ (value) ต่อไปนี้

pm สำหรับรูปภาพทั้งหมดที่เป็นชื่อเฉพาะตามเกณฑ์นิยาม โดยภายในรูปภาพนั้น อาจปรากฏชื่อเฉพาะรูปอื่นเป็นองค์ประกอบหรือไม่ก็ได้

ตัวอย่างเช่น <rs type=pm>ไทย</rs>

<rs type=pm>ธนาคารไทยพาณิชย์</rs>

comp สำหรับรูปภาพทั้งหมดซึ่งขาดคุณสมบัติของชื่อเฉพาะตามเกณฑ์นิยาม แต่มีชื่อเฉพาะปรากฏเป็นส่วนประกอบหนึ่งภายในรูปภาพนั้น

ตัวอย่างเช่น <rs type=comp>คนไทย</rs>

<name>

- ใช้กำกับสายอักขระของชื่อเฉพาะภายในรูปภาพที่กำกับด้วยแท็ก <rs>

ตัวอย่างเช่น <rs type=pm><name>ไทย</name></rs>

<rs type=pm>ธนาคาร<name>ไทยพาณิชย์</name></rs>

4.1.3.2 การกำกับบริบทบ่งบอก

ในการวิเคราะห์บริบทบ่งบอกชื่อเฉพาะจำเป็นต้องมีการกำกับรูปภาพอื่นนอกเหนือจากชื่อเฉพาะ หรือกำกับบริบทบ่งบอก โดยอาจเป็นการกำกับข้อมูลเฉพาะบางส่วนของข้อมูลทั้งหมดที่ปรากฏในคลังข้อมูล ทั้งนี้ขึ้นอยู่กับประเด็นต่างๆที่ต้องการวิเคราะห์เพื่อให้ได้มาซึ่งผลการศึกษาที่เป็นบริบทบ่งบอกถึงชื่อเฉพาะ ซึ่งการกำกับข้อมูลลักษณะนี้ในระหว่างการศึกษาบริบทบ่งบอกประกอบด้วยประเด็นต่างๆที่มีรายละเอียดปลีกย่อยแตกต่างกัน ในที่นี้จึงจะกล่าวถึงเพียงบางส่วนเพื่อแสดงให้เห็นแนววิธีดำเนินการวิจัยในขั้นตอนนี้

บริบทบ่งบอกถึงชื่อเฉพาะในภาษาไทยบริบทหนึ่ง que พบในการศึกษาคือ รูปภาพที่ปรากฏติดกับชื่อเฉพาะที่จะเรียกว่า คำบ่งบอก สามารถเป็นตัวบ่งบอกถึงชื่อเฉพาะได้ ในการศึกษาบริบทนี้มีประเด็นย่อยที่ต้องวิเคราะห์หลายประเด็น เช่น สรุปรายการคำบ่งบอกวิเคราะห์ขีดความสามารถในการบ่งบอก เป็นต้น (ดูรายละเอียดบทที่ 5) การกำกับรูปภาพที่เป็นคำบ่งบอกจึงเป็นเครื่องมือที่จะเอื้อประโยชน์ต่อการวิเคราะห์ข้อสรุปของบริบทนี้และแท็กที่ใช้สำหรับกำกับคำบ่งบอกคือ <fpm>

<fprn>

- ใช้กำกับคำบ่งบอกที่ปรากฏติดกับชื่อเฉพาะ โดยคำบ่งบอกนี้จะต้องปรากฏติดกับชื่อเฉพาะที่ต่างรูปกันตั้งแต่สองรูปภาษาขึ้นไป

type ระบุกลุ่มคำของคำบ่งบอกแต่ละคำ

ตัวอย่างเช่น <fprn type=gr9O>ธนาคาร</f>ไทยพาณิชย์

<fprn type=gr9O>ธนาคาร</f>ไทยธนาคาร

นอกจากการกำกับข้อมูลเพิ่มโดยกำหนดแท็กเพิ่มเติมแล้ว ในบางประเด็นของการศึกษาบริบทบ่งบอก สามารถจะศึกษาโดยการกำหนดคุณสมบัติให้กับแท็กที่กำหนดใช้อยู่ ดังกรณีของการศึกษาตัวบ่งบอกระดับปริจเฉท (ดูรูปที่ 5) จะกำหนดคุณสมบัติให้กับแท็ก <name> คือคุณสมบัติ key โดยชื่อเฉพาะที่มีความสัมพันธ์กันในระดับปริจเฉทจะมีค่าของคุณสมบัติ (value) เดียวกันดังตัวอย่างในรูปที่ 4-2

บล. กรุงศรีอยุธยา ประเมินว่า จากข่าวที่คณะกรรมการกลางว่าด้วยราคาสินค้าและบริการ ได้อนุมัติให้ บริษัท<name key=11>สหวิริยาстил อินดัสตรี</name> ปรับราคาจำหน่ายเหล็กแผ่นรีดร้อนจากราคาเดิม 11.66 บาทต่อกิโลกรัม เป็น 13.30 บาทต่อกิโลกรัม ตามที่บริษัทได้ยื่นขอไป โดยให้มีผลในทางปฏิบัติในทันทีนั้น จะส่งผลดีต่อแนวโน้มการทำธุรกิจของ <name key=11>สหวิริยาстил</name> ในครั้งหลังของปีนี้สดใสต่อเนื่องจากครึ่งปีแรก...

รูปที่ 4-2 ตัวอย่างการกำกับคุณสมบัติเพิ่มเติมเพื่อการศึกษาบริบทบ่งบอก

การกำกับข้อมูลอีกลักษณะหนึ่งนอกเหนือจากการกำกับชื่อเฉพาะและรูปภาษาที่เป็นบริบทบ่งบอกแล้ว ภายในคลังข้อมูลได้ทำการกำกับรายละเอียดที่เป็นเค้าโครงของข้อมูล ผู้วิจัยได้ทำการกำกับไว้เพื่อใช้สำหรับการอ้างอิงแหล่งข้อมูลอิเล็กทรอนิกส์หรือการอ้างอิงเพื่อความสะดวกในขณะทำการวิเคราะห์ สามารถดูรายการแท็กที่ใช้กำกับเค้าโครงข้อมูลได้ในภาคผนวก ก และตัวอย่างข้อมูลในรูปที่ 4-3 เป็นข้อมูลในคลังข้อมูลของวิทยานิพนธ์ที่แสดงให้เห็นภาพรวมของการกำกับข้อมูลลักษณะต่างๆที่ได้กล่าวผ่านมาแล้ว

เมื่อดำเนินการวิจัยตามลำดับขั้นตอนต่างๆดังกล่าวมาแล้ว จะได้คลังข้อมูลภาษาสำหรับวิทยานิพนธ์ฉบับนี้ หลังจากนั้นผู้วิจัยจะทำการวิเคราะห์บริบทบ่งบอกชื่อเฉพาะจากคลังข้อมูลนี้ ซึ่งรายละเอียดจะกล่าวถึงในส่วนถัดไปคือ หัวข้อที่ 4.2 (ดูรูปที่ 4-1)

```

<CORPUS1 size=87text>
<corpusHeader>
  <eAddress type=www>www.bangkokbiznews.com/2002/08/30/index.html</eAddress>
  <distributor type=newspaper>The Krungthep turakij web site</distributor>
  <pubDate>วันศุกร์ที่ 30 สิงหาคม พ.ศ. 2545</pubDate>
</corpusHeader>

<text1>
<head>ได้กระแส : เวลา พื้นที่ อุตสาหกรรม : ชุมชนเมืองขอนแก่น</head>
<byline>อรรถจักร์ สัตยานุรักษ์</byline>
  :
  ชุมชนแออัดใน<rs type=prn><fprn type=gr13PL>เมือง</f><name key=39 type=Place>ขอนแก่น</name></rs>ได้
  เคลื่อนไหวต่อสู้เพื่อความเป็นธรรมมาเนิ่นนาน ในที่สุดรัฐก็ได้ยอมรับว่าได้ทำงานผิดพลาดและได้คืนสิทธิบนที่ดินให้แก่<rs
  type=comp>ชาวชุมชนบะขาม</rs>
  :
</text1>
  :
<text10>
<head>learning comer</head>
<div n=1>
<divhead>บริติช เคาน์ซิล เปิดอบรมภาษาอังกฤษระดับประถม</divhead>
  :
</div>
<div n=2>
<divhead>สวนสุนันทาจัดมินิคอมพิวเตอร์ปิดเทอม</divhead>
  :
</div>
  :
</text10>
  :
</CORPUS1>

<CORPUS2 size=12text>
<corpusHeader>
  <eAddress type=www>www.sakulthai.com/DsearchdetailbyIssuesql.asp?StIssueID=2511</eAddress>
  <distributor type=magazine>นิตยสารสกุลไทย-บริษัท อักษรโสภณ จำกัด</distributor>
  <pubDate>ฉบับที่ 2511 ปีที่ 49 ประจำวันอังคารที่ 3 ธันวาคม 2002</pubDate>
</corpusHeader>
  :
  :
</CORPUS2>

```

รูปที่ 4-3 ภาพรวมของการกำกับข้อมูลในคลังข้อมูลภาษาของวิทยานิพนธ์

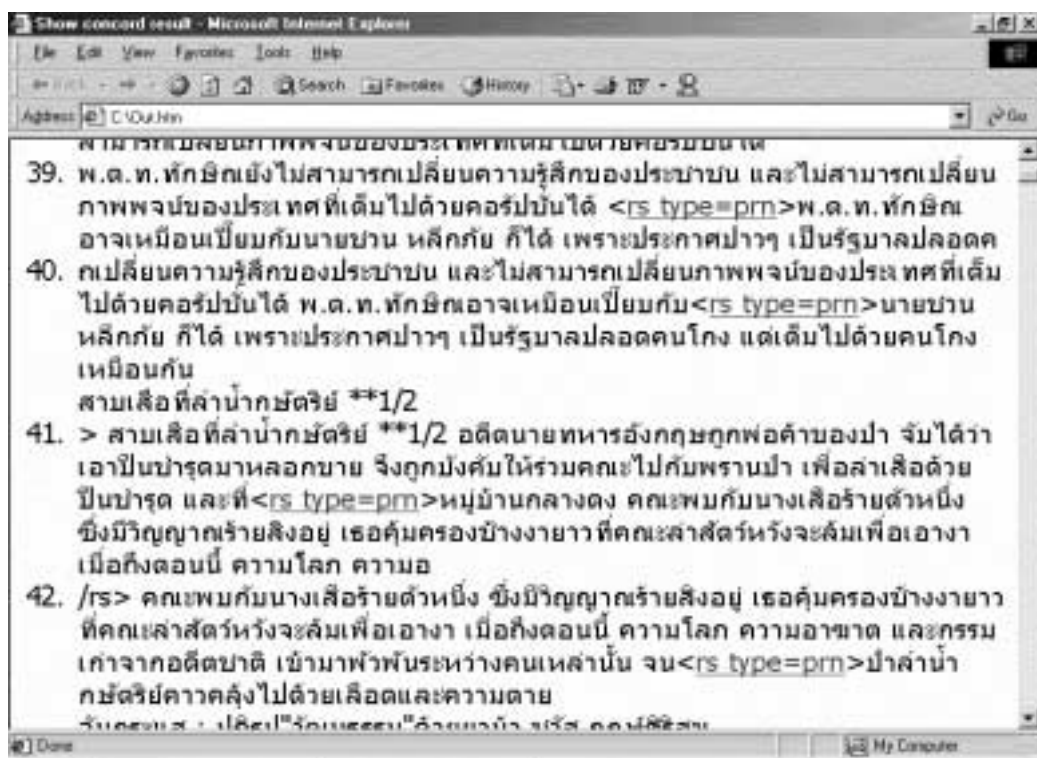
4.2 การวิเคราะห์บริบทบ่งบอก

กระบวนการศึกษาวิเคราะห์ถึงบริบทที่สามารถบ่งบอกชื่อเฉพาะจากคลังข้อมูล จะต้องใช้วิธีวิเคราะห์หลายวิธีการประกอบกัน รวมทั้งวิธีวิเคราะห์บางวิธีการจะอาศัยเครื่องมือตัวอื่นนอกเหนือจากคลังข้อมูล โดยวิธีวิเคราะห์ของขั้นตอนการวิเคราะห์บริบทบ่งบอกประกอบด้วย 4 วิธีการหลักคือ วิเคราะห์บริบทบ่งบอกจากข้อความต่อเนื่องในคลังข้อมูลโดยตรง ถัดมาคือ วิเคราะห์บริบทบ่งบอกจากผลลัพธ์ที่ได้จากโปรแกรมคอนคอร์แดนซ์ ต่อมาคือกำกับข้อมูลเพิ่มเติมเพื่อการวิเคราะห์บริบทบ่งบอก และวิเคราะห์หาผลเฉพาะอย่างโดยเขียนโปรแกรมภาษา Perl ให้ประมวลผลจากคลังข้อมูล ทั้งนี้ในการวิเคราะห์บริบทโดยบริบทหนึ่งเพื่อหาข้อสรุปเกี่ยวกับการบ่งบอกถึงชื่อเฉพาะอาจต้องใช้วิธีวิเคราะห์หลายวิธีการประกอบกัน ซึ่งจะมีรายละเอียดแตกต่างกันในการวิเคราะห์แต่ละบริบท ดังนั้นในส่วนถัดไปจะแสดงให้เห็นแนวทางของกระบวนการวิเคราะห์บริบทบ่งบอกชื่อเฉพาะโดยการยกตัวอย่างขั้นตอนวิธีในการวิเคราะห์บริบทอันหนึ่งทีพบว่าสามารถเป็นบริบทบ่งบอกถึงชื่อเฉพาะได้

เมื่อจัดสร้างคลังข้อมูลแล้ว ขั้นตอนแรกของการวิเคราะห์คือวิเคราะห์รายการข้อมูลซึ่งเป็นผลลัพธ์จากการใช้โปรแกรม CU Thai Concordance ให้ประมวลผลจากคลังข้อมูลของวิทยานิพนธ์ รายการข้อมูล 1 รายการจะแสดงส่วนของข้อมูลที่เป็นชื่อเฉพาะใดๆหนึ่งชื่อเฉพาะและมีข้อความปรากฏข้างเคียงทั้งด้านซ้ายและด้านขวาของชื่อเฉพาะนั้น (ดูรูปที่ 4-4) บริบทหนึ่งที่พบจากการวิเคราะห์ชื่อเฉพาะที่ปรากฏในรายการข้อมูลคือ คำที่ปรากฏติดกับชื่อเฉพาะซึ่งจะเรียกว่าคำบ่งบอก คำบ่งบอกมักปรากฏร่วมกับชื่อเฉพาะจึงวิเคราะห์ว่าคำบ่งบอกเหล่านี้จะสามารถเป็นบริบทบ่งบอกถึงชื่อเฉพาะได้ แต่ในขณะเดียวกันผู้วิจัยก็พบว่าในรายการข้อมูล คำบ่งบอกนอกจากจะปรากฏหน้าชื่อเฉพาะแล้ว ยังปรากฏหน้าคำนามทั่วไปด้วย เช่น "นาย" ใน "นายชวัน หลีกภัย ผู้นำฝ่ายค้านและหัวหน้าพรรคประชาธิปัตย์ กล่าวถึง..." และ "นาย" ใน "...อดีตนายทหารอังกฤษถูกพ่อค้าของป่า..." จึงเกิดประเด็นที่ต้องศึกษาวิเคราะห์ต่อมาคือ คำบ่งบอกแต่ละคำปรากฏร่วมกับชื่อเฉพาะมากน้อยเพียงใด หรือกล่าวได้ว่าคำบ่งบอกแต่ละคำมีขีดความสามารถในการบ่งบอกถึงชื่อเฉพาะแตกต่างกัน จึงต้องวิเคราะห์ถึงขีดความสามารถของคำบ่งบอกแต่ละคำด้วย

ขั้นตอนถัดมา จึงเป็นขั้นตอนการวิจัยเพื่อวิเคราะห์เกี่ยวกับสองประเด็นคือ สรุปรายการคำบ่งบอกตามสมมุติฐานที่ว่าคำบ่งบอกสามารถเป็นตัวบ่งบอกถึงชื่อเฉพาะ และวิเคราะห์ขีดความสามารถของคำบ่งบอกแต่ละคำตามที่พบปรากฏการณ์ทางภาษาของคำบ่งบอกดังกล่าวมาแล้ว ซึ่งการวิเคราะห์ในขั้นตอนนี้จะต้องอาศัยวิธีวิเคราะห์อีกลักษณะหนึ่งคือ การกำกับ

ข้อมูลเพิ่มเติมแล้วใช้โปรแกรมภาษา Perl เป็นตัวช่วยประมวลผลให้ได้ผลการวิเคราะห์ของทั้งสองประเด็น สาเหตุที่ต้องใช้การกำกับข้อมูลและโปรแกรมภาษา Perl เข้ามาเป็นเครื่องมือช่วยวิเคราะห์ก็เนื่องมาจากการศึกษาบริบทของบอกรับนี้เกี่ยวข้องกับการค้นหาคำบงบอกและนับจำนวน ซึ่งชื่อเฉพาะที่ใช้ศึกษามีจำนวนมากกว่า 2,000 ชื่อและปรากฏกระจายอยู่ในคลังข้อมูล โดยกระบวนการวิเคราะห์จะเริ่มจากการกำกับคำบงบอกรูปที่ปรากฏติดกับชื่อเฉพาะตามวิธีการที่กล่าวไว้ในหัวข้อที่ 4.1.3.2 จากนั้นใช้โปรแกรมภาษา Perl ประมวลผลการคำบงบอกทั้งหมดออกมาและใช้รายการคำบงบอกเหล่านี้เป็นแหล่งข้อมูลให้โปรแกรมนับจำนวนคำบงบอกแต่ละคำ โดยนับที่ปรากฏร่วมกับชื่อเฉพาะและนับจำนวนคำบงบอกนั้นที่ปรากฏในทุกลักษณะไม่ว่าจะปรากฏร่วมกับชื่อเฉพาะหรือปรากฏร่วมกับรูปภาษาอื่น แล้วนำผลการนับจำนวนที่ได้มาวิเคราะห์หาขีดความสามารถของคำบงบอกแต่ละคำในรูปของเปอร์เซ็นต์



รูปที่ 4-4 รายการข้อมูลที่แสดงชื่อเฉพาะพร้อมบริบทของโปรแกรม CU Thai Concordance

หลังจากดำเนินการวิเคราะห์ โดยใช้กระบวนการวิเคราะห์วิธีการต่างๆประกอบกันแล้ว จะได้ผลการวิเคราะห์เป็นบริบทต่างๆที่สามารถบ่งบอกถึงชื่อเฉพาะภาษาไทยที่กล่าวไว้ในบทที่ 5

บทที่ 5

ชื่อเฉพาะและบริบทบ่งบอกชื่อเฉพาะ

ในบทนี้จะกล่าวถึงบริบทต่างๆที่สามารถเป็นตัวบ่งบอกหรือเป็นบริบทบ่งบอกถึงชื่อเฉพาะในภาษาไทย บริบทต่างๆศึกษาวิเคราะห์จากชื่อเฉพาะในคลังข้อมูลของวิทยานิพนธ์ที่มีจำนวน 2,855 ชื่อ เมื่อวิเคราะห์ชื่อเฉพาะที่ปรากฏในคลังข้อมูลโดยใช้เกณฑ์นิยามพบว่าชื่อเฉพาะเป็นรูปภาษาที่มีการปรากฏในลักษณะที่แตกต่างกันสองลักษณะใหญ่คือ ปรากฏเป็นรูปภาษาหนึ่งหน่วย และปรากฏเป็นส่วนประกอบของรูปภาษาอื่น กล่าวคือ ชื่อเฉพาะบางรูปนอกจากจะปรากฏเป็นหน่วยภาษาหนึ่งหน่วยที่มีคุณสมบัติเป็นชื่อเฉพาะตามเกณฑ์นิยาม เช่น "ไทย" แล้ว ยังปรากฏประกอบกับรูปภาษาอื่นรวมกันเป็นรูปภาษาใหม่หนึ่งหน่วย โดยความหมายของทั้งหน่วยรูปภาษาอาจอยู่ในขอบข่ายของการเป็นชื่อเฉพาะหรือไม่ก็ได้ เช่น "ธนาคารไทยพาณิชย์" "คนไทย" เป็นต้น ในที่นี้ "ธนาคารไทยพาณิชย์" เป็นชื่อเฉพาะ ส่วน "คนไทย" ขาดคุณสมบัติของการเป็นชื่อเฉพาะ บริบทบ่งบอกที่จะกล่าวถึงต่อไปจึงเป็นบริบทที่ปรากฏสัมพันธ์กับรูปภาษาซึ่งทั้งหน่วยมีคุณสมบัติเป็นชื่อเฉพาะตามเกณฑ์นิยามของวิทยานิพนธ์ฉบับนี้ โดยชื่อเฉพาะดังกล่าวมีจำนวน 2,547 ชื่อ ซึ่งก่อนที่จะกล่าวถึงบริบทบ่งบอกที่พบในการศึกษา จะได้กล่าวถึงลักษณะโครงสร้างของรูปภาษาที่เป็นชื่อเฉพาะไว้ในหัวข้อที่ 5.1 จากนั้นจะเป็นรายละเอียดเกี่ยวกับบริบทบ่งบอกที่พบ ได้แก่ คำที่ปรากฏติดกับชื่อเฉพาะซึ่งจะกล่าวไว้ในหัวข้อที่ 5.2, การเว้นวรรค (Space) ในหัวข้อที่ 5.3, คำบ่งบอกปรากฏร่วมกับการเว้นวรรคในหัวข้อที่ 5.4 และตัวบ่งบอกระดับปริศนาในหัวข้อที่ 5.5 นอกจากนี้ยังมีชื่อเฉพาะที่ไม่ปรากฏบริบทบ่งบอกซึ่งจะกล่าวถึงในหัวข้อที่ 5.6 ส่วนชื่อเฉพาะที่ปรากฏเป็นส่วนประกอบหนึ่งของรูปภาษาซึ่งทั้งหน่วยภาษาไม่มีคุณสมบัติครบถ้วนตามเกณฑ์นิยามที่พบเป็นจำนวน 308 รูปภาษา จะกล่าวรายละเอียดไว้ต่างหากในหัวข้อที่ 5.7

5.1 ลักษณะโครงสร้างของรูปภาษาที่เป็นชื่อเฉพาะ

ในการวิเคราะห์ชื่อเฉพาะจากคลังข้อมูลพบว่า ชื่อเฉพาะมักจะปรากฏร่วมกับคำนามทั่วไปหรือคำนามสามัญ โดยคำนามสามัญดังกล่าวมักปรากฏอยู่ที่ส่วนหน้าชื่อเฉพาะด้วยจำนวนหนึ่งคำหรือสองคำ ดังตัวอย่างในคลังข้อมูล เช่น

- (ก) นายชวน หลีกภัย
- (ข) อาจารย์หม่อมราชวงศ์คึกฤทธิ์
- (ค) กรมบังคับคดี
- (ง) กระทรวงเทคโนโลยีสารสนเทศ
- (จ) สำนักข่าวรอยเตอร์
- (ฉ) สถาบันการแพทย์แผนไทย
- (ช) พรรคไทยรักไทย

โดยคำนามสามัญดังกล่าวสามารถเป็นบริบทบ่งบอกถึงชื่อเฉพาะ ที่ในวิทยานิพนธ์ฉบับนี้จะเรียกว่า ‘คำบ่งบอก’ ซึ่งจะกล่าวถึงรายละเอียดต่อไปในหัวข้อที่ 5.2 และพบว่าชื่อเฉพาะในภาษาไทยเป็นรูปภาพที่มักจะประกอบขึ้นจากคำมากกว่าหนึ่งคำ ซึ่งคำที่นำมาสร้างเป็นชื่อเฉพาะนั้น โดยส่วนใหญ่เป็นคำศัพท์ทั่วไปที่มีการกำหนดขึ้นใช้อยู่แล้วในภาษา เช่น “ชวน” “หลีก” และ “ภัย” ในตัวอย่าง (ก) และชนิดคำที่นำมาสร้างขึ้นเป็นชื่อเฉพาะมักจะเป็นคำนามทั่วไป เช่น “ภัย” “คดี” ในตัวอย่าง (ก) และ (ค) หรืออาจเป็นคำกริยา เช่น “ชวน” “หลีก” “บังคับ” ในตัวอย่าง (ก) และ (ค) นอกจากนี้ชื่อเฉพาะอาจประกอบด้วยคำทับศัพท์ภาษาต่างประเทศ เช่น “เทคโนโลยี” “รอยเตอร์” ในตัวอย่าง (ง) และ (จ) หรือชื่อเฉพาะบางรูปก็เป็นคำหนึ่งในชื่อเฉพาะอื่น เช่น “ไทย” ในตัวอย่าง (ฉ) และ (ช)

นอกจากนี้ยังพบว่า คำที่มาประกอบเป็นชื่อเฉพาะ มักจะมีโครงสร้างในลักษณะเขียนเรียงต่อเนื่องกันและมีการเรียงลำดับคำเหมือนกับการเรียงลำดับคำของหน่วยภาษาระดับสูงกว่าคำ เช่น ตัวอย่าง (ง) (จ) และ (ฉ) มีโครงสร้างเหมือนวลี, ตัวอย่าง (ค) (ช) มีโครงสร้างเหมือนประโยค ยกเว้นในกรณีของชื่อบุคคล (Person) จะมีรูปแบบที่เฉพาะออกไป คือ มีการแทรกช่องว่างอยู่ในชื่อเฉพาะระหว่างชื่อและนามสกุล ดังเช่นในตัวอย่าง (ก) ในกรณีที่ชื่อเฉพาะมีโครงสร้างของรูปภาพเหมือนหน่วยภาษาระดับสูงกว่าคำ โดยปกติแล้วจะปรากฏคำบ่งบอกอยู่ด้วย ดังเช่น ตัวอย่าง (ฉ) เมื่อปรากฏเป็นชื่อเฉพาะจะปรากฏด้วยรูป “สถาบันการแพทย์แผนไทย” เสมอ ในขณะที่ตัวอย่าง (ก) เมื่อปรากฏเป็นชื่อเฉพาะอาจปรากฏด้วยรูป “นายชวน หลีกภัย” หรือ “ชวน หลีกภัย” ทั้งที่รูปภาพนี้ก็ประกอบด้วยคำที่มีใช้ในภาษาเหมือนกับ “สถาบันการแพทย์แผนไทย” แต่ความสามารถในการปรากฏอยู่ในข้อความแตกต่างกัน ทั้งนี้อาจเนื่องมาจากโครงสร้างที่แตกต่างกัน คือ รูปภาพ “สถาบันการแพทย์แผนไทย” ถ้าไม่ปรากฏคำนาม “สถาบัน” อยู่ด้วยนั้น รูปภาพ “การแพทย์แผนไทย” ไม่สามารถบ่งบอกได้ว่ารูปภาพนี้เป็นชื่อเฉพาะ เพราะเขียนในรูปแบบเดียวกับวลีและสื่อความหมายได้เป็นความหมายรวมของส่วนประกอบในวลี แต่

รูปภาพ “นายชวน หลีกภัย” ถ้าไม่ปรากฏ “นาย” ยังสามารถแยกแยะได้ว่ารูปภาพนี้เป็นชื่อเฉพาะอันเนื่องมาจากมีโครงสร้างในรูปแบบที่เฉพาะ

กล่าวโดยสรุป โครงสร้างของชื่อเฉพาะนั้นประกอบจากคำในลักษณะต่างๆ อาจเป็นคำไทยหรือคำทับศัพท์ก็ได้ โดยที่อาจมีคำบ่งบอกปรากฏร่วมด้วยก็ได้ เช่น “อาจารย์หม่อมราชวงศ์คึกฤทธิ์” ประกอบด้วยคำบ่งบอกสองคำต่อเรียงกัน (“อาจารย์” และ “หม่อมราชวงศ์”) อยู่หน้าชื่อเฉพาะ “คึกฤทธิ์” หรือในบางกรณี คำที่เป็นคำบ่งบอกนั้นเป็นส่วนหนึ่งของชื่อเฉพาะเอง เช่น “กระทรวงเทคโนโลยีสารสนเทศ” เป็นชื่อเฉพาะที่มีคำบ่งบอก “กระทรวง” อยู่ภายใน

5.2 คำบ่งบอกชื่อเฉพาะ

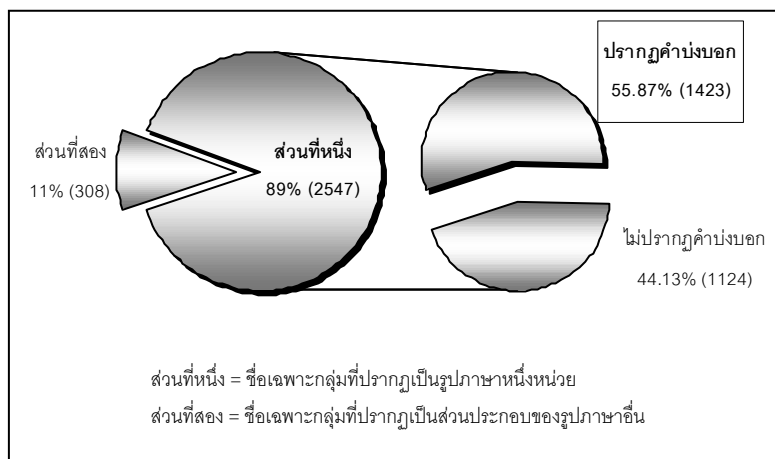
จากการวิเคราะห์ชื่อเฉพาะที่ปรากฏในคลังข้อมูลร่วมกับการวิเคราะห์ชื่อเฉพาะในรายการข้อมูลที่เป็นผลลัพธ์ของโปรแกรมคอนคอร์แดนซ์ เห็นแนวโน้มว่ามีรูปภาพที่เป็นคำปรากฏอยู่กับชื่อเฉพาะในลักษณะที่คำนั้นปรากฏซ้ำๆ อยู่ในชื่อเฉพาะที่มีรูปต่างกัน ตัวอย่างเช่น

"...องค์การศึกษาศาสตร์และวัฒนธรรมแห่งสหประชาชาติ (ยูเนสโก) ได้ยกย่อง..."

"...ผู้บริหารขององค์การสื่อสารมวลชนแห่งประเทศไทย (อ.ส.ม.ท.) ได้ประสานไปยัง..."

"...ที่สอดคล้องตามกฎระเบียบขององค์การการค้าโลกหรือดับบลิวทีโอ และสามารถ..."

เมื่อทำการประเมินความเป็นไปได้ของรูปภาพกลุ่มนี้ในการเป็นตัวบ่งบอก โดยประเมินจากจำนวนชื่อเฉพาะที่ปรากฏรูปภาพกลุ่มนี้ ซึ่งต่อไปจะกล่าวถึงรูปภาพกลุ่มนี้โดยใช้คำว่า 'คำบ่งบอก' ผลการประเมินที่ได้ซึ่งแสดงไว้ในแผนภูมิรูปที่ 5-1 จะพบว่าชื่อเฉพาะที่ปรากฏคำบ่งบอกมีจำนวนถึง 55.87 เปอร์เซ็นต์ ดังนั้นคำกลุ่มนี้ที่ปรากฏอยู่กับชื่อเฉพาะน่าจะสามารถเป็นตัวบ่งบอกถึงชื่อเฉพาะในภาษาไทยได้ ตัวบ่งบอกในกรณีนี้มีลักษณะสำคัญที่น่าสนใจหลายประการที่จะกล่าวถึงอย่างละเอียดในเนื้อหาส่วนถัดไป โดยลำดับแรกเป็นหัวข้อที่ 5.2.1 รายการคำที่ปรากฏเป็นตัวบ่งบอก ต่อมาคือหัวข้อที่ 5.2.2 ลักษณะการปรากฏของคำบ่งบอก หัวข้อที่ 5.2.3 รูปและความหมายของคำบ่งบอก หัวข้อที่ 5.2.4 การจัดกลุ่มคำบ่งบอกและประเภทของชื่อเฉพาะ และหัวข้อที่ 5.2.5 ซีดความสามารถในการบ่งบอกถึงชื่อเฉพาะของคำบ่งบอก



รูปที่ 5-1 สัดส่วนของชื่อเฉพาะที่ปรากฏคำบ่งบอก

5.2.1 คำที่ปรากฏเป็นตัวบ่งบอก (คำบ่งบอก)

รูปภาษาที่ปรากฏให้เห็นซ้ำๆ ในชื่อเฉพาะเป็นรูปภาษาเฉพาะกลุ่มหนึ่งที่มีลักษณะเป็นคำ โดยคำที่ปรากฏเป็นตัวบ่งบอกนี้พบประมาณ 161 คำ ดังรายการรูปภาษาที่ได้สรุปไว้ในรูปที่ 5-2

โค้ง พล.ต. นายกรัฐมนตรี ทะเล ฝ้าย ทวีป สมรภูมิ นคร นาง วง พ.ร.บ. บิ๊ก อาคาร แม่น้ำ หาด ราชวงศ์ นายแพทย์ ซอย โรงแรม ครอบครัว พล.ต.อ. บ้าน อุทยาน สนามบิน กอง รองประธานาธิบดี องค์การ เมือง ถ้ำ แผนก เขตรักษาพันธุ์สัตว์ป่า ลุ่มน้ำ บริษัท วิทยาลัย รศ. ปากน้ำ ทะเลสาบ ครู เชื้อน ร.ร. จังหวัด ร.ต.ต. นาย คาวบอย ราชสำนัก รายการ พระราชวัง เจ้า ต. พล.อ. สังเวียน หน่วย ตระกูล สหพันธ์ คุณที่ สวน เทศบาล ศูนย์ บล. ภูมิภาค กลุ่ม สถาบันราชภัฏ หมู่บ้าน อ่าว ตลาด ป่า สนามกอล์ฟ หม่อมหลวง วุฒิสมาชิก วง ถ. ร.ต.อ. ห้อง เขตห้ามล่าสัตว์ป่า พรรค พระเจ้า วัด ผู้กำกับ สำนักข่าว อ. ฝันป่า นางสาว สภา น.ส. ผศ. เกาะ พล.อ.อ. มหาวิทยาลัย จ. ประเทศ ร.พ. โฆษก องค์การ ประธานาธิบดี อาจารย์ ประธาน คณะกรรมการ น้อง ลูก เจ้าหญิง โรงงาน ตำบล ถนน ที่ เครือ ดร. สถาบัน ดีเจ. อำเภอ ภาษา แพทย์หญิง ผู้พัน สหภาพ ส.ต.ท. ศ. สะพาน หม่อมราชวงศ์ โรงพยาบาล คุณ โรงเรียน มูลนิธิ ม. เผ่า สนาม นายกษ เขต กระทรวง กิ่งอำเภอ ทิวเขา จอมพล รัฐ ย่าน สน. พ.ต.อ. สำนัก ช. แถบ กองทุน กรม หม่อมเจ้า สำนักงาน แบงก์ หนู พล.ต.ต. ค่าย สยามคม ป่า เขา พ.ต.ท. เลี้ยว ฝั่ง เครือข่าย เขตฯ สนิทสัญญา กรุง ธนาคาร ศาล เจ้าชาย จ.ส.ต. กรู๊ป บริษัท...จำกัด

รูปที่ 5-2 รายการคำที่เป็นตัวบ่งบอกถึงชื่อเฉพาะ

5.2.2 ลักษณะการปรากฏของคำบ่งบอก

คำบ่งบอกมีลักษณะปรากฏติดกับชื่อเฉพาะ โดยมีตำแหน่งการปรากฏใน 3 ลักษณะใหญ่ คือ

(1) คำบ่งบอกปรากฏอยู่ส่วนหน้าชื่อเฉพาะ เช่น คำบ่งบอก *กรุง ประเทศ* ในตัวอย่างต่อไปนี้

"...สภากิจการแรงงานได้ในวันได้ออกแถลงการณ์ที่**กรุงเทพฯ**เวลานี้..."
 "...ในวันที่ 1 ก.ย. นี้เวลา 17.00 น.ที่สนามกีฬาแห่งชาติ **กรุงเทพฯ**..."
 "...ดารานักแสดงชายและหญิงที่มีชื่อเสียงใน**ประเทศไทย**จำนวนมาก..."
 "...ผมใฝ่ฝันอยากไปเรียนศิลปะต่อที่**ประเทศเยอรมนี** เพื่อให้ความคิด..."

(2) คำบ่งบอกปรากฏอยู่ส่วนหลังชื่อเฉพาะ ซึ่งพบเพียง 1 คำคือ คำบ่งบอก *กรุ๊ป* ดังตัวอย่างเช่น

"...การทำธุรกิจของ**สามารถ กรุ๊ป** จะมุ่งสร้างมูลค่าเพิ่มให้กับสิ่งที่มีอยู่..."
 "...ประธาน**ไออีซี กรุ๊ป** ผู้ดำเนินธุรกิจร้านอาหารญี่ปุ่นบุฟเฟต์..."

(3) คำบ่งบอกปรากฏอยู่หน้าหรือหลังชื่อเฉพาะ ได้แก่ คำบ่งบอกคำเดียวกันปรากฏให้เห็นทั้งในตำแหน่งส่วนหน้าชื่อเฉพาะและส่วนหลังชื่อเฉพาะ พบ 1 คำคือคำบ่งบอก *มหาวิทยาลัย* โดยพบกรณีที่คำบ่งบอก *มหาวิทยาลัย* ปรากฏอยู่ส่วนหลังกับรูปภาพที่เป็นชื่อเฉพาะเพียงรูปเดียวคือ "จุฬาลงกรณ์มหาวิทยาลัย" และชื่อเฉพาะรูปอื่นที่ปรากฏคำบ่งบอก *มหาวิทยาลัย* จะพบคำบ่งบอกปรากฏอยู่ในส่วนหน้าชื่อเฉพาะทั้งหมดดังตัวอย่างในคลังข้อมูล เช่น

"...**จุฬาลงกรณ์มหาวิทยาลัย**นั้นเป็นมหาวิทยาลัยแห่งแรกของไทย..."
 "...พบว่าได้ออกจาก**มหาวิทยาลัยรามคำแหง**ไปแล้วตั้งแต่เวลา 15.00..."
 "...ได้ไปเรียนปริญญาโท ในสาขาอนุรักษวิทยา ที่**มหาวิทยาลัยลอนดอน**..."

จากข้อมูลแสดงให้เห็นว่า คำบ่งบอกจะปรากฏติดกับชื่อเฉพาะในตำแหน่งส่วนหน้าชื่อเฉพาะหรือส่วนหลังชื่อเฉพาะ โดยคำบ่งบอกจำนวน 158 คำ (98.14 เปอร์เซ็นต์) จากคำบ่งบอก 161 คำ มีลักษณะของการปรากฏอยู่ส่วนหน้าชื่อเฉพาะ จึงสามารถกล่าวได้ว่าคำบ่งบอกส่วนใหญ่จะปรากฏติดกับชื่อเฉพาะในส่วนหน้าชื่อเฉพาะและมีคำบ่งบอกเพียงบางรูปภาพที่ปรากฏติดกับชื่อเฉพาะในลักษณะพิเศษ

5.2.3 รูปและความหมายของคำบ่งบอก

จากลักษณะทางภาษาที่พบในการวิเคราะห์เกี่ยวกับลักษณะการปรากฏของคำบ่งบอกดังที่ได้กล่าวมาในหัวข้อที่ 5.2.2 นั้น เป็นการแสดงให้เห็นข้อสรุปเกี่ยวกับตำแหน่งการปรากฏคำบ่งบอก นอกจากนี้แล้วข้อสรุปเกี่ยวกับลักษณะการปรากฏของคำบ่งบอกอาจพิจารณาได้ในอีกแง่มุมหนึ่งคือ การพิจารณาเรื่องความสัมพันธ์ระหว่างรูปของคำบ่งบอกและความหมายของคำบ่งบอก เนื่องจากมีคำบ่งบอกบางคำปรากฏด้วยรูปที่ต่างกันแต่สื่อความหมายเดียวกัน ดังเช่น คำบ่งบอก *บริษัท* และ *บริษัท...จำกัด* นอกจากนี้ในการวิเคราะห์ก็ยังพบว่าคำบ่งบอกบางคำสามารถปรากฏได้มากกว่าหนึ่งความหมาย เช่น คำบ่งบอก *เจ้า* ซึ่งความสัมพันธ์ระหว่างรูปและความหมายของคำบ่งบอกที่พบในการศึกษาครั้งนี้ สรุปได้ 3 กรณีดังต่อไปนี้คือ

(1) คำบ่งบอกหลายรูป ความหมายเดียวกัน (หลายรูป หนึ่งความหมาย)

ได้แก่ คำบ่งบอก *บริษัท* และ คำบ่งบอก *บริษัท...จำกัด*

ตัวอย่าง "...แนวโน้มการดำเนินธุรกิจของ**บริษัทวิริยะสตีล** ในครึ่งปีหลังสดใส..."
 "...รองประธานกรรมการ**บริษัทสุวรรณชาติ จำกัด** ซึ่งเป็นบริษัทที่ดำเนินงาน..."
 "...ดำเนินการในนาม**บริษัท เนสท์เล่ แดรี่ (ประเทศไทย) จำกัด** ซึ่งเป็นธุรกิจ..."
 "...ผู้จัดการฝ่ายการตลาด **บริษัท อีเลคโทรลักซ์ ประเทศไทย จำกัด** ผู้จำหน่าย..."
 "...กรรมการผู้จัดการ **บริษัท ควอลิตี้ เฮาส์ จำกัด (มหาชน)** เปิดเผยแผนดำเนิน..."

(2) คำบ่งบอกหนึ่งรูป มีความหมายมากกว่าหนึ่งความหมาย (หนึ่งรูป หลายความหมาย)

โดยคำบ่งบอกที่มีการปรากฏความสัมพันธ์นี้ พบอยู่ในคลังข้อมูลของวิทยานิพนธ์นี้จำนวน 4 คำ ได้แก่ คำบ่งบอก *เจ้า* *กลุ่ม* *ศาล* และ *สหภาพ*

ตัวอย่างเช่น คำบ่งบอก *เจ้า* ปรากฏเป็นคำบ่งบอกอยู่ในคลังข้อมูลโดยมีความหมายสองความหมาย ความหมายหนึ่งเป็นความหมายที่แสดงถึงฐานันดร ดังเช่นใน "...เป็นชื่อเมืองไทยใหญ่เมืองหนึ่ง ชนนี้ของ**เจ้าสิมโป**..." และอีกความหมายหนึ่งเป็นความหมายที่แสดงการเป็นคำเรียกขาน ดังเช่นใน "...ที่ผ่านมา "**เจ้าบอล**" ภราดร ศรีชาพันธุ์ นักหวดลูกสักหลาด..."

(ดูผลการวิเคราะห์ในหัวข้อที่ 5.2.4.3)

(3) คำบ่งบอกหนึ่งรูป มีเพียงความหมายเดียว (หนึ่งรูป หนึ่งความหมาย)

ได้แก่ คำบ่งบอกโดยส่วนใหญ่ที่พบในการศึกษาครั้งนี้ นอกเหนือจากที่แจ้งไว้ในข้อ 1 และ 2

5.2.4 การจัดกลุ่มคำบ่งบอกและประเภทชื่อเฉพาะ

ประเด็นที่น่าสนใจของคำบ่งบอกอีกประการหนึ่งซึ่งพบในการวิเคราะห์คือ คำบ่งบอกคำหนึ่งจะมีความหมายบางอย่างร่วมกับคำบ่งบอกคำอื่น เช่น คำบ่งบอก *กรม* มีความหมายว่า "ส่วนราชการในราชการบริหารส่วนกลางรองจากกระทรวงและทบวง" (ราชบัณฑิตยสถาน, พจนานุกรมศัพท์อิเล็กทรอนิกส์, 2525) กับ คำบ่งบอก *กระทรวง* หมายถึง "ส่วนราชการหนึ่งในราชการบริหารส่วนกลางซึ่งมีรัฐมนตรีว่าการกระทรวงเป็นหัวหน้า" (ราชบัณฑิตยสถาน) จะเห็นว่า *กรม* กับ *กระทรวง* ต่างก็หมายถึง หน่วยหนึ่งที่ได้รับผิดชอบงานบริหารราชการโดยแตกต่างกันในเรื่องของขนาดและลำดับความสำคัญ นอกจากนี้ความหมายของคำบ่งบอกยังทำให้ทราบว่าชื่อเฉพาะนั้นอ้างอิงถึงสิ่งใด และชื่อเฉพาะกลุ่มที่อ้างอิงถึงสิ่งที่มีคุณสมบัติเหมือนกันจะปรากฏคำบ่งบอกเฉพาะกลุ่มหนึ่ง ซึ่งคำบ่งบอกเฉพาะกลุ่มนี้จะไม่ปรากฏกับชื่อเฉพาะกลุ่มอื่น เช่น คำบ่งบอก *กรม* กับ คำบ่งบอก *กระทรวง* จะไม่ปรากฏกับชื่อเฉพาะที่อ้างอิงถึงคนในลักษณะปัจเจกบุคคล ดังนั้นคำบ่งบอกต่างๆที่ปรากฏกับชื่อเฉพาะจึงสามารถที่จะนำมาจำแนกออกเป็นกลุ่มคำที่มีคุณลักษณะเฉพาะกลุ่มได้และคำบ่งบอกสามารถบ่งชี้ถึงประเภทของชื่อเฉพาะ โดยรายละเอียดของทั้งสองประเด็นจะกล่าวแยกเป็นหัวข้อที่ 5.2.4.1 และ 5.2.4.2 ตามลำดับ และสุดท้ายจะกล่าวถึงหัวข้อที่ 5.2.4.3 ความคาบเกี่ยวของกลุ่มคำบ่งบอกและประเภทชื่อเฉพาะ ดังนี้

5.2.4.1 การจัดกลุ่มคำบ่งบอก

ในวิทยานิพนธ์ฉบับนี้ ได้จำแนกคำบ่งบอกออกเป็นกลุ่มคำโดยใช้เกณฑ์ทางด้านความหมาย คำบ่งบอกใดที่มีความหมายบางส่วนร่วมกันจะถูกจัดให้อยู่ในกลุ่มเดียวกัน และวิธีที่ใช้ในการศึกษาความหมายของคำบ่งบอกแต่ละคำคือ การวิเคราะห์อรรถลักษณะ (Componential Analysis) ซึ่งเป็นวิธีที่นำความหมายของแต่ละคำมาแจงเป็นส่วนความหมายย่อยที่เรียกว่า อรรถลักษณะ (Semantic Feature) กล่าวคือ คำบ่งบอกที่มีอรรถลักษณะบางตัวร่วมกันจัดเป็นกลุ่มเดียวกัน โดยในการวิเคราะห์อรรถลักษณะของแต่ละคำบ่งบอก วิเคราะห์จากความหมายที่สื่อออกมาจากบริบทของข้อความในคลังข้อมูลและวิเคราะห์จากความหมายตามพจนานุกรม (เจียรชัย เอี่ยมวรเมธ, 2544; นววรรณ, 2544; พรพิมล ธิรคุณโกวิท และ วิโรจน์ ธิรคุณโกวิท, 2527; มานิต มานิตเจริญ, 2538; สำนักพจนานุกรมมติชน, 2543) ในกรณีที่พบปัญหาความกำกวมหรือมีความขัดแย้งกันระหว่างความหมายจากบริบทของข้อมูลกับความหมายตามพจนานุกรม จะถือเอาความหมายที่สื่อออกมาจากบริบทของข้อมูลเป็นหลัก

คำบ่งบอกที่พบจำนวน 161 คำ สามารถจำแนกออกเป็น 18 กลุ่มคำและคำบ่งบอกจำนวน 3 คำไม่พบอรรถลักษณะร่วมกับคำบ่งบอกอื่น จึงจัดให้อยู่ในกลุ่มคำอีกกลุ่มหนึ่งเป็นกลุ่มคำที่ 19 โดยคำบ่งบอกทั้ง 19 กลุ่มคำได้แสดงไว้ในตารางที่ 5-1 ข้างล่างนี้

| อรรถลักษณะร่วมของกลุ่มคำ | คำบ่งบอก |
|--|--|
| กลุ่มคำที่ 1 <ปัจเจกบุคคล> <แสดงฐานะหรือลำดับชั้น> | พ.ต.ท. พล.อ. พ.ต.อ. พล.ต.ต. ร.ต.ต. ส.ต.ท. พล.ต.อ. ร.ต.อ. จอมพล จ.ส.ต. พล.ต. พล.อ.อ. |
| กลุ่มคำที่ 2 <ปัจเจกบุคคล> <แสดงตำแหน่งด้านการบริหารปกครอง> | รองประธานาธิบดี ประธานาธิบดี วุฒิสมาชิก นายกฯ นายกรัฐมนตรี ประธาน |
| กลุ่มคำที่ 3 <ปัจเจกบุคคล> <แสดงตำแหน่งทางวิชาการ> | ผศ. รศ. ศ. |
| กลุ่มคำที่ 4 <ปัจเจกบุคคล> <แสดงระดับการศึกษา> | ดร. |
| กลุ่มคำที่ 5 <ปัจเจกบุคคล> <แสดงฐานันดร> | หม่อมเจ้า หม่อมราชวงศ์ หม่อมหลวง เจ้าหญิง เจ้าชาย พระเจ้า เจ้า * |
| กลุ่มคำที่ 6 <ปัจเจกบุคคล> <บอกอายุ เพศ และสถานภาพการสมรส> | นางสาว นาย นาง น.ส. |
| กลุ่มคำที่ 7 <ปัจเจกบุคคล> <คำเรียกทางวิชาชีพ> | แพทย์หญิง นายแพทย์ ดีเจ. ผู้พัน อาจารย์ ผู้กำกับ ไชยภัก ครู |
| กลุ่มคำที่ 8 <ปัจเจกบุคคล> <คำเรียกเรียกขาน> | ควาบอย บิ๊ก หนู ป้า คุณพี่ พี่ น้อง คุณ เสี่ย ลูก เจ้า * |
| กลุ่มคำที่ 9 <กลุ่มบุคคล> <มีการรวมตัวกัน> <ประกอบขึ้นเป็นหน่วยเดียว> <มีสภาพเป็นนิติบุคคล> <เพื่องานราชการ> | กระทรวง องค์การ กรม ศาล* แผนก คณะกรรมการ กอง |
| กลุ่มคำที่ 10 <กลุ่มบุคคล> <มีการรวมตัวกัน> <ประกอบขึ้นเป็นหน่วยเดียว> <มีสภาพเป็นนิติบุคคล> <เพื่อทำธุรกิจ> | บล. บริษัท...จำกัด กรุ๊ป แแบงก์ บริษัท เครือ ธนาคาร ค่าย กลุ่ม * |

* คำบ่งบอกที่สามารถจัดให้อยู่ในกลุ่มคำที่แตกต่างกัน 2 กลุ่มคำ - ดูหัวข้อที่ 5.2.4.3

| | | |
|---------------|--|---|
| กลุ่มคำที่ 11 | <กลุ่มบุคคล> <มีการรวมตัวกัน> <ประกอบขึ้นเป็นหน่วยเดียว> <มีสภาพเป็นนิติบุคคล> <เพื่อทำกิจการตามวัตถุประสงค์ใด วัตถุประสงค์หนึ่ง> | สหพันธ์ สำนักข่าว มูลนิธิ สำนักงาน สมาคม กองทุน พรรค เครือข่าย สหภาพ* สภา องค์กร สถาบัน สำนัก หน่วย วง กลุ่ม* |
| กลุ่มคำที่ 12 | <กลุ่มบุคคล> <มีความสัมพันธ์ทางเชื้อสายหรือเชื้อชาติ> | ราชสำนัก เผ่า ราชวงศ์ ตระกูล ครอบครัว |
| กลุ่มคำที่ 13 | <พื้นที่หรือบริเวณ> <สร้างหรือกำหนดขึ้น> <สำหรับใช้ประโยชน์หรือทำกิจกรรม> | ร.ร. ถ. สน. ร.พ. ซอย โรงแรม สนามกอล์ฟ สนามบิน เขื่อน ช. สังเวียน มหาวิทยาลัย โรงเรียน โรงพยาบาล สถาบันราชภัฏ พระราชวัง ถนน สนาม สะพาน โรงงาน ศูนย์ ม. อาคาร สวน วัด วิทยาลัย ห้อง ศาล* ตลาด |
| กลุ่มคำที่ 14 | <พื้นที่หรือบริเวณ> <สร้างหรือกำหนดขึ้น> <เป็นเขตปกครอง> | ทวีป กิ่งอำเภอ จ. อำเภอ จังหวัด เทศบาล อ. สหภาพ* ตำบล กรุง ประเทศ เมือง เขต หมู่บ้าน บ้าน ต. นคร รัฐ |
| กลุ่มคำที่ 15 | <พื้นที่หรือบริเวณ> <สร้างหรือกำหนดขึ้น> <ระบุตำแหน่งหรือบริเวณ> | โค้ง ย่าน สมรภูมิ แถบ ภูมิภาค วัง (น้ำ) |
| กลุ่มคำที่ 16 | <พื้นที่หรือบริเวณ> <เกิดขึ้นตามธรรมชาติ> <ส่วนพื้นน้ำ> | ทะเลสาบ ปากน้ำ อ่าว แม่น้ำ ทะเล |
| กลุ่มคำที่ 17 | <พื้นที่หรือบริเวณ> <เกิดขึ้นตามธรรมชาติ> <ส่วนพื้นดิน> | เขตห้ามล่าสัตว์ป่า ลุ่มน้ำ เขตรักษาพันธุ์สัตว์ป่า ถ้ำ ทิวเขา เขตข อุทยาน ผืนป่า เกาะ ฝั่ง ป่า หาด เขา |
| กลุ่มคำที่ 18 | <กฎข้อบังคับหรือข้อตกลง> | สนธิสัญญา พ.ร.บ. |
| กลุ่มคำที่ 19 | ไม่พบอรรถลักษณะร่วม | ภาษา รายการ ฝ่าย |

ตารางที่ 5-1 คำบ่งบอกกลุ่มต่างๆ

* คำบ่งบอกที่สามารถจัดให้อยู่ในกลุ่มคำที่แตกต่างกัน 2 กลุ่มคำ - ดูหัวข้อที่ 5.2.4.3

5.2.4.2 การจำแนกประเภทของชื่อเฉพาะโดยใช้คำบ่งบอก

ความเกี่ยวข้องกับชื่อเฉพาะอีกประการหนึ่งที่คำบ่งบอกแสดงให้เห็นคือ คำบ่งบอกเป็นบริบทที่บอกถึงประเภทของชื่อเฉพาะ กล่าวคือ ความหมายของคำบ่งบอกทำให้ทราบว่าชื่อเฉพาะที่ปรากฏคำบ่งบอกนั้นอ้างอิงถึงสิ่งใดหรือในกรณีชื่อเฉพาะไม่ปรากฏติดกับคำบ่งบอก ความหมายของคำบ่งบอกที่ปรากฏอยู่ในข้อความรอบข้างก็ยังสามารถในการที่จะเป็นตัวบ่งบอกถึงประเภทของชื่อเฉพาะนั้น โดยปรากฏการณ์ทางภาษาที่สนับสนุนข้อสรุปนี้คือ

- (1) เมื่อพิจารณาตัวอย่างชื่อเฉพาะในรูปที่ 5-3 เปรียบเทียบกับชื่อเฉพาะในรูปที่ 5-4 ซึ่งชื่อเฉพาะในรูปที่ 5-3 คำบ่งบอกถูกตัดออกไป ส่วนชื่อเฉพาะในรูปที่ 5-4 ได้ทำการคัดลอกมาจากคลังข้อมูลเช่นกันแต่ไม่มีการเปลี่ยนแปลงใดๆ เป็นการจำลองให้เห็นว่าในกรณีที่ผู้ใช้ภาษาไม่มีความรู้หรือรู้จำเกี่ยวกับรูปภาษาที่เป็นชื่อเฉพาะนั้นมาก่อน บริบทสำคัญที่จะทำให้ผู้ใช้ภาษาทราบโดยง่ายว่ารูปภาษานั้นเป็นชื่อเฉพาะและอ้างอิงถึงสิ่งใดก็คือคำบ่งบอกที่ปรากฏติดกับชื่อเฉพาะ

"...การสร้าง **แลง** เพื่อแก้ไขปัญหาคนว่างงาน..."
 "...เกิดมาคงไม่เคยดูละคร**ศิลป์**เสียด้วยซ้ำ..."
 "...สำหรับนักวิจารณ์เริ่มแรก **เจตนา**แนะว่า..."

รูปที่ 5-3 ตัวอย่างชื่อเฉพาะจากคลังข้อมูลที่ผ่านการดัดแปลง (ตัดคำบ่งบอก)

"...การสร้าง **ตลาดแลง** เพื่อแก้ไขปัญหาคนว่างงาน..."
 "...เกิดมาคงไม่เคยดูละคร**กรรมศิลป์**เสียด้วยซ้ำ..."
 "...สำหรับนักวิจารณ์เริ่มแรก **ดร.เจตนา**แนะว่า..."

รูปที่ 5-4 ตัวอย่างชื่อเฉพาะจากคลังข้อมูลที่ปรากฏเหมือนต้นฉบับ

- (2) กรณีที่ชื่อเฉพาะเป็นรูปที่ประกอบด้วยคำนามทั่วไปหรือคำชนิดอื่นและการเรียงลำดับของคำที่เป็นองค์ประกอบมีโครงสร้างเหมือนวลีหรือประโยค คำบ่งบอกจะเป็นตัวบ่งบอกว่ารูปภาษานั้นอ้างอิงถึงสิ่งใดหรือเป็นชื่อเฉพาะประเภทใดนอกเหนือจากการบ่งบอกว่ารูปนั้นเป็นชื่อเฉพาะ ดังตัวอย่างต่อไปนี้

| <u>ถ้าชื่อเฉพาะถูกตัดคำบ่งบอก</u> | <u>ชื่อเฉพาะตามที่ปรากฏในคลังข้อมูล</u> |
|-----------------------------------|---|
| "การแพทย์แผนไทย" | สถาบันการแพทย์แผนไทย |
| "ปฏิรูปที่ดินเพื่อเกษตรกรรม" | สำนักงานปฏิรูปที่ดินเพื่อเกษตรกรรม |
| "ฟุตบอลอังกฤษ" | สมาคมฟุตบอลอังกฤษ |
| "ทหารไทย" | เบงกัทหารไทย |
| "ลัทธิลัทธิกลาง" | ศาลลัทธิลัทธิกลาง |

- (3) กรณีที่ชื่อเฉพาะบางรูปถูกนำไปใช้ในการอ้างอิงถึงสิ่งใหม่ คำบ่งบอกจะเป็นสิ่งที่ช่วยให้ผู้ใช้ภาษาเข้าใจได้อย่างชัดเจนว่า ชื่อเฉพาะนั้นเป็นชื่อเฉพาะประเภทใด เช่น

| | | | | |
|-------------|---|------------------|----------------------|---------------|
| "กรุงเทพ" | : | จังหวัดกรุงเทพ | มหาวิทยาลัยกรุงเทพ | ธนาคารกรุงเทพ |
| "เชียงใหม่" | : | จังหวัดเชียงใหม่ | มหาวิทยาลัยเชียงใหม่ | |

- (4) ในกรณีที่ชื่อเฉพาะไม่ปรากฏติดกับคำบ่งบอก ในข้อความใกล้เคียงกับชื่อเฉพาะอาจปรากฏคำบ่งบอกซึ่งเป็นบริบทที่สามารถบอกถึงประเภทของชื่อเฉพาะได้เช่นกัน โดยจะยกตัวอย่างข้อมูลบางส่วนที่พบในการศึกษาวิเคราะห์ดังนี้

"...ทั้งนี้ **อาเซียน เรตกาฟูล** เป็นบริษัทร่วมทุนระหว่างบริษัทประกันอิสลาม (ทากาฟูล)..."
 "...โดยกำไรที่เพิ่มขึ้นนี้ รวมถึง 121 ล้านยูโรที่ได้จากการขาย **พิการ์ด** บริษัทค้าปลีกอาหาร..."

ด้วยปรากฏการณ์ทางภาษาที่สนับสนุนว่าคำบ่งบอกสามารถบ่งชี้ประเภทของชื่อเฉพาะดังที่ได้กล่าวมาในข้อ (1) – (4) ผู้วิจัยจึงจำแนกชื่อเฉพาะออกเป็นประเภทต่างๆ โดยพิจารณาจากอรรถลักษณะร่วมของคำบ่งบอกแต่ละกลุ่ม เมื่อพบอรรถลักษณะใดอรรถลักษณะหนึ่งหรือมากกว่านั้นที่แบ่งแยกกลุ่มคำต่างๆ ออกเป็นชุดๆ ได้ จะจัดให้ชื่อเฉพาะที่ปรากฏคำบ่งบอกในกลุ่มคำเหล่านั้นเป็นชื่อเฉพาะประเภทเดียวกันและกำหนดชื่อของแต่ละประเภทตามความหมายของอรรถลักษณะที่พบ ซึ่งภาพรวมของการจำแนกชื่อเฉพาะออกเป็นประเภทต่างๆ ได้แสดงไว้ด้วย

แผนภาพของรูปที่ 5-5 และผลการจำแนกประเภทของชื่อเฉพาะที่ได้นั้นพบว่าจำแนกได้ 6 ประเภท คือ ชื่อคน (Person) ชื่อองค์กร (Organization) ชื่อกลุ่มคนร่วมเชื้อสาย (Affinity) ชื่อสถานที่ (Place) ชื่อกฎหมาย (Law) และ ชื่ออื่นๆ (Miscellaneous) ดังที่ปรากฏอยู่ในแผนภาพของรูปที่ 5-5 ในส่วนถัดไปจะกล่าวถึงตัวอย่างแนวการวิเคราะห์ประเภทชื่อเฉพาะโดยอ้างอิงจากแผนภาพในรูปที่ 5-5

จากผลการจำแนกประเภทชื่อเฉพาะโดยใช้คำบ่งบอกที่ได้ (รูปที่ 5-5) สามารถนำมาใช้วิเคราะห์ประเภทของชื่อเฉพาะได้ตามตัวอย่างต่อไปนี้คือ *

ถ้าเป็นการวิเคราะห์ถึงประเภทของชื่อเฉพาะรูปที่มีคำบ่งบอกปรากฏติดอยู่กับชื่อเฉพาะ สามารถวิเคราะห์ได้โดยตรงจากคำบ่งบอกที่ปรากฏอยู่ นั่น ดังเช่น

ก. "นายชวัน หลีกภัย ผู้นำฝ่ายค้าน และหัวหน้าพรรคประชาธิปัตย์ กล่าวถึง..."

ข. "...ช่วงเดือน ต.ค.ฝนจะมาก ก็จะทำให้หน้าในแม่น้ำเจ้าพระยามีระดับสูงได้..."

ตัวอย่าง ก. "นายชวัน หลีกภัย"

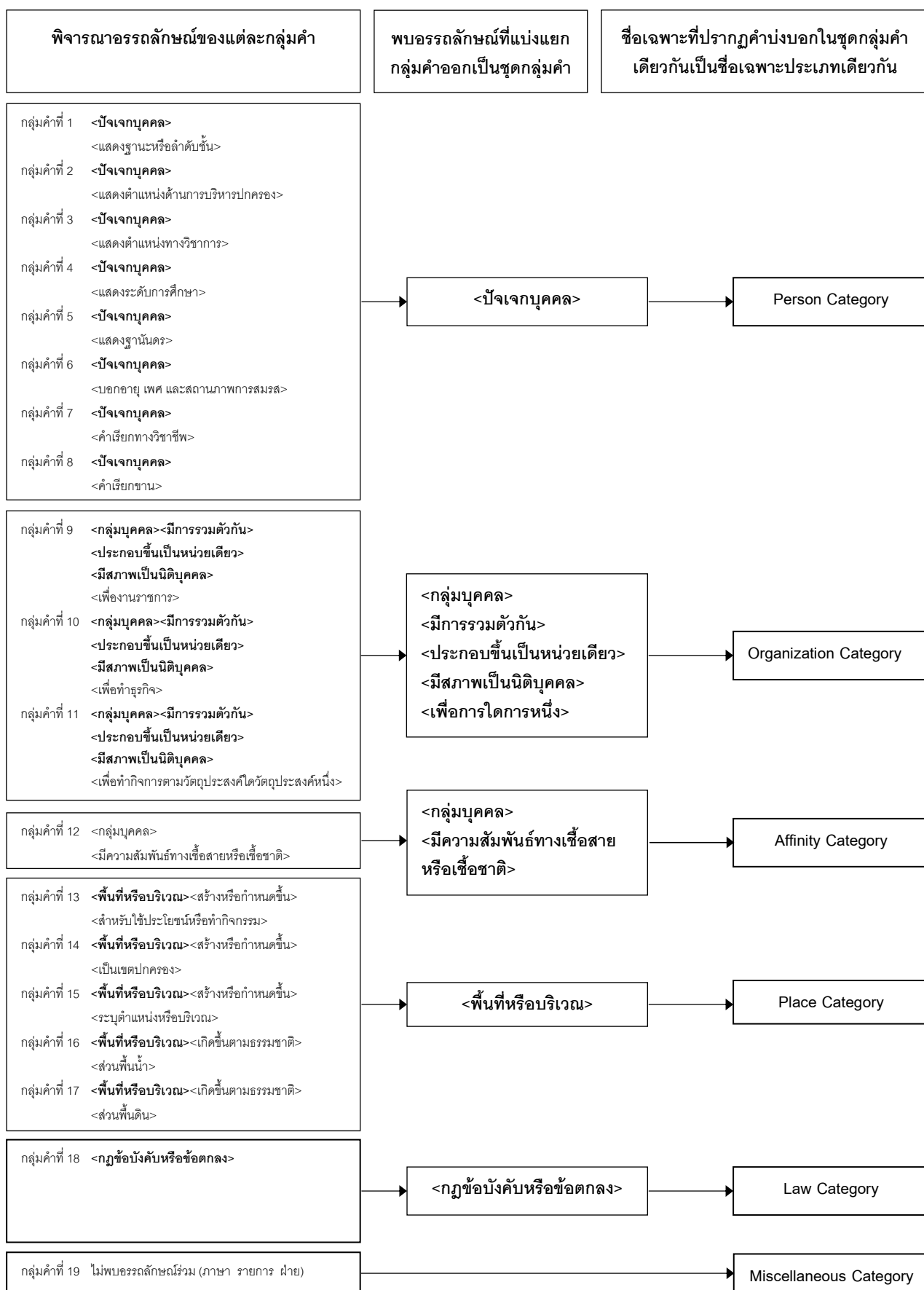
คำบ่งบอกคือ นาย อยู่ในกลุ่มคำที่ 6 ประเภทของชื่อเฉพาะรูปนี้คือ ชื่อคน

ตัวอย่าง ข. "แม่น้ำเจ้าพระยา"

คำบ่งบอกคือ แม่น้ำ อยู่ในกลุ่มคำที่ 16 ประเภทของชื่อเฉพาะรูปนี้คือ ชื่อสถานที่

ส่วนการวิเคราะห์ประเภทของชื่อเฉพาะที่ไม่ปรากฏคำบ่งบอก แนวทางที่เป็นไปได้มี 2 แนวทางหลักคือ วิเคราะห์โดยอาศัยชื่อเฉพาะรูปที่มีคำบ่งบอกและวิเคราะห์โดยใช้บริบทใกล้เคียง เช่น มีคำบ่งบอกปรากฏอยู่ในข้อความใกล้เคียงกับชื่อเฉพาะ

* ขั้นตอนการวิเคราะห์ตัวอย่างที่กล่าวถึงในที่นี้ ผู้วิจัยจำลองเหตุการณ์ที่อาจเป็นไปได้เมื่อให้คอมพิวเตอร์เป็นผู้มองภาษาและเป็นผู้ประมวลผล



รูปที่ 5-5 ผลการจำแนกชื่อเฉพาะออกเป็นประเภทต่างๆ

5.2.4.3 ความคาบเกี่ยวของกลุ่มคำบ่งบอกและประเภทชื่อเฉพาะ

ในการจำแนกคำบ่งบอกและประเภทของชื่อเฉพาะพบลักษณะทางภาษาที่ควร จะนำมากล่าวถึงด้วยกันสองประการ ประการแรกเกี่ยวข้องกับการจำแนกคำบ่งบอกพบว่า คำบ่งบอกบางคำไม่ได้ปรากฏเพียงความหมายเดียว และเมื่อวิเคราะห์อรรถลักษณะก็เป็น อรรถลักษณะชุดที่ต่างกัน เช่น คำบ่งบอก *เจ้า* เมื่อพิจารณาตามตัวอย่างที่แสดงไว้ในตารางที่ 5-2 จะเห็นว่า *เจ้า* ปรากฏสองความหมาย โดยชุดอรรถลักษณะของความหมายที่หนึ่งมีอรรถลักษณะ ร่วมกับคำบ่งบอกในกลุ่มคำที่ 5 และอรรถลักษณะของความหมายที่สองมีอรรถลักษณะร่วมกับ กลุ่มคำที่ 8 ดังนั้นคำบ่งบอก *เจ้า* จึงสามารถจำแนกให้อยู่ในกลุ่มคำสองกลุ่มคำ ซึ่งปรากฏการณ์นี้ เป็นลักษณะทางภาษาที่พบได้ทั่วไปในเรื่องความหมายที่คำๆหนึ่งอาจมีความหมายและชนิดของ คำได้มากกว่าหนึ่งกรณี ผู้วิจัยจึงวิเคราะห์ให้คำบ่งบอกแต่ละคำที่ปรากฏอยู่ในตารางที่ 5-2 อยู่ใน กลุ่มคำสองกลุ่มคำ ประการต่อมาในการจำแนกประเภทของชื่อเฉพาะ ปัญหาหนึ่งที่จะพบในการ วิเคราะห์จำแนกคือ ความหมายอ้างอิงของชื่อเฉพาะประเภทองค์กร กับ ชื่อเฉพาะประเภทสถานที่ มีลักษณะคาบเกี่ยวกัน กล่าวคือ ในบางบริบทรูปภาษาที่เป็นชื่อเฉพาะประเภทองค์กร จะถูก นำมาใช้ในความหมายอ้างอิงถึงสถานที่คือมีอรรถลักษณะ <พื้นที่หรือบริเวณ> ของชื่อสถานที่ (Place Category) เช่น "...ตายไปอย่างดีก็เอาชื่อมาติดที่อนุสาวรีย์หน้ากรมป่าไม้..." ขณะเดียวกันในบางบริบทของการนำชื่อเฉพาะประเภทสถานที่ไปใช้ก็เป็นการนำไปใช้ในความหมายของกลุ่มบุคคลที่มีสภาพเป็นนิติบุคคลหรือมีความหมายของชื่อเฉพาะประเภทองค์กร เช่น "...อาเซียนไม่ควรมองจีนเป็นคู่แข่ง..." การที่ความคาบเกี่ยวกันทางความหมายปรากฏอย่าง ชัดเจนในชื่อเฉพาะสองประเภทนี้ ก็เนื่องมาจากสภาพความเป็นจริงที่ว่า กลุ่มบุคคลที่มาประกอบ รวมกันเป็นหน่วยเพื่อร่วมทำกิจการใดๆ ย่อมมีเรื่องของสถานที่อาจรวมถึงถาวรวัตถุอย่างเช่น อาคารเข้ามาเกี่ยวข้องสำหรับใช้ประกอบกิจการ และในทางกลับกันสถานที่ใดๆย่อมจะต้องมี บุคคลเข้าไปเกี่ยวข้อง ความคาบเกี่ยวกันทางความหมายที่ใช้ในการอ้างอิงจึงเกิดขึ้น ซึ่งกำชาย (2525: 266) กล่าวไว้ว่า เป็นลักษณะของการสมมุติให้เป็นบุคคล โดยมีการเปลี่ยนความหมาย จากการมีความหมายถึงสถานที่หรือองค์การตามรูปศัพท์ให้เป็นความหมายที่อ้างถึงกลุ่มคนผู้มี หน้าที่รับผิดชอบในสถานที่หรือองค์การนั้นๆ

การนำประเด็นเกี่ยวกับความคาบเกี่ยวกันของกลุ่มคำหรือประเภทซึ่งอาจจะ กล่าวได้ว่าเป็นความกำกวมของการจำแนกคำบ่งบอกและประเภทของชื่อเฉพาะมากล่าวไว้ใน วิทยานิพนธ์ เนื่องจากผู้วิจัยต้องการแสดงให้เห็นบางแง่มุมที่อาจจะต้องคำนึงถึงในการ ประมวลผลเกี่ยวกับชื่อเฉพาะที่จะพบอยู่ในการพัฒนางานประมวลผลภาษาธรรมชาติด้วย คอมพิวเตอร์ด้านต่างๆ

| ตัวอย่างข้อมูลจากคลังข้อมูล | อรรถลักษณะที่ต่างกัน | |
|--|---|---------------|
| (1) "...เป็นชื่อเมืองไทยใหญ่เมืองหนึ่ง ชนนี้ของเจ้าสิมโป..." | <แสดงฐานันดร> | กลุ่มคำที่ 5 |
| (2) "...ที่ผ่านมา <u>เจ้าบอล</u> ภาวคร ศรีชาพันธุ์ นักหวดลูกสักหลาด..." | <คำเรียกขาน> | กลุ่มคำที่ 8 |
| (1) "...ในปีนี้ <u>กลุ่มโออิชิ</u> พร้อมที่จะขยายกลุ่มธุรกิจนม..." | <เพื่อทำธุรกิจ> | กลุ่มคำที่ 10 |
| (2) "...ผู้อำนวยการฝ่ายการเมืองระหว่างประเทศของ <u>กลุ่มกรีนพีซ</u> ..." | <เพื่อทำกิจการตามวัตถุประสงค์ใด วัตถุประสงค์หนึ่ง> | กลุ่มคำที่ 11 |
| (1) "...เตรียมส่ง <u>ศาลรัฐธรรมนูญ</u> สั่งห้ามเล่นการเมือง 5 ปี..." | <กลุ่มบุคคล> <มีการรวมตัวกัน> <ประกอบขึ้นเป็นหน่วยเดียว> <มีสภาพเป็นนิติบุคคล> <เพื่องานราชการ> | กลุ่มคำที่ 9 |
| (2) "...ความศักดิ์สิทธิ์ของ <u>ศาลปู่ตา</u> ขยายออกมาครอบคลุม..." | <พื้นที่หรือบริเวณ> <สร้างหรือกำหนดขึ้น> <สำหรับใช้ประโยชน์หรือทำกิจกรรม> | กลุ่มคำที่ 13 |
| (1) "...การขึ้นของ <u>สหภาพพนักงานองค์การโทรศัพท์แห่งประเทศไทย</u> ..." | <กลุ่มบุคคล> <มีการรวมตัวกัน> <ประกอบขึ้นเป็นหน่วยเดียว> <มีสภาพเป็นนิติบุคคล> <เพื่อทำกิจการตามวัตถุประสงค์ใด วัตถุประสงค์หนึ่ง> | กลุ่มคำที่ 11 |
| (2) "...เราจึงมีจากกล่าวได้ว่า <u>สหภาพไซเวียต</u> หลังการปฏิวัติโค่นล้ม..." | <พื้นที่หรือบริเวณ> <สร้างหรือกำหนดขึ้น> <เป็นเขตปกครอง> | กลุ่มคำที่ 14 |

ตารางที่ 5-2 คำบ่งบอกที่มีสองความหมาย

5.2.5 ซีดความสามารถในการบ่งบอกถึงชื่อเฉพาะของคำบ่งบอก

คำบ่งบอกแต่ละคำซึ่งพบว่ามีปรากฏร่วมกับชื่อเฉพาะดังที่ได้กล่าวมาแล้ว ยังพบว่าคำบ่งบอกนั้นอาจจะปรากฏกับคำนามทั่วไปหรือรูปภาษาชนิดอื่น เช่น คำบ่งบอก *นาย* ใน **"นายชวน หลีกภัย** ผู้นำฝ่ายค้าน และหัวหน้าพรรคประชาธิปัตย์ กล่าวถึง..." ที่ปรากฏหน้าชื่อเฉพาะ แต่คำบ่งบอก *นาย* ใน "...อดีตนายทหารอังกฤษถูกพ่อค้าของป่า..." ปรากฏหน้าคำนามทั่วไป การที่คำบ่งบอกอาจปรากฏร่วมกับรูปภาษาอื่นที่ไม่ใช่ชื่อเฉพาะเป็นการแสดงให้เห็นว่า ใน

การใช้คำบ่งบอกเป็นบริบทบ่งบอกถึงชื่อเฉพาะต้องพิจารณาด้วยว่าโอกาสที่คำบ่งบอกแต่ละคำจะปรากฏร่วมกับชื่อเฉพาะมีมากน้อยเพียงใด กล่าวคือ คำบ่งบอกแต่ละคำมีขีดความสามารถในการบ่งบอกถึงชื่อเฉพาะแตกต่างกัน ซึ่งการวิเคราะห์ขีดความสามารถของคำบ่งบอกแต่ละคำ สามารถวิเคราะห์ที่ได้จากการดูว่าในจำนวนทั้งหมดของคำบ่งบอกแต่ละคำที่ปรากฏอยู่ในคลังข้อมูลของวิทยานิพนธ์ พบคำบ่งบอกนั้นปรากฏติดกับชื่อเฉพาะกี่เปอร์เซ็นต์ โดยรายละเอียดเกี่ยวกับขีดความสามารถในการบ่งบอกของคำบ่งบอกจะกล่าวแยกเป็นสองหัวข้อย่อยคือ หัวข้อที่ 5.2.5.1 แนวทางการวิเคราะห์ขีดความสามารถในการบ่งบอก และหัวข้อที่ 5.2.5.2 ขีดความสามารถของคำบ่งบอก

5.2.5.1 แนวทางการวิเคราะห์ขีดความสามารถในการบ่งบอก

ในการวิเคราะห์หาเปอร์เซ็นต์ของคำบ่งบอกแต่ละคำที่เกิดกับชื่อเฉพาะ จะต้องทำการนับจำนวนคำบ่งบอกแต่ละคำในสองลักษณะคือ นับจำนวนที่ปรากฏกับชื่อเฉพาะและนับจำนวนทั้งหมดทั้งที่ปรากฏและไม่ปรากฏกับชื่อเฉพาะ หลังจากนั้นนำค่าทั้งสองมาคิดเป็นเปอร์เซ็นต์ สิ่งที่เป็นปัญหาในการนับจำนวนคำบ่งบอกคือ คำบ่งบอกบางส่วนนอกจากจะปรากฏติดกับชื่อเฉพาะ, ปรากฏในลักษณะเป็นคำนามทั่วไป คำบ่งบอกนั้นอาจปรากฏเป็นองค์ประกอบหนึ่งของคำบ่งบอกรูปอื่นหรือรูปภาษาที่เป็นชื่อเฉพาะ เช่น คำบ่งบอก *นาง* ใน "*นางสาว*" (คำบ่งบอกรูปอื่น) คำบ่งบอก *ประเทศ* ใน "*กระทรวงการต่างประเทศ*" (ชื่อเฉพาะ) ทำให้การนับจำนวนคำบ่งบอกทั้งหมดสามารถทำได้สองกรณีคือ กรณีแรกนับคำบ่งบอกเฉพาะรูปที่ปรากฏติดกับชื่อเฉพาะและปรากฏในลักษณะคำนามทั่วไป ส่วนกรณีที่สองคือ นับคำบ่งบอกที่ปรากฏทุกลักษณะ เมื่อพิจารณาเปรียบเทียบเปอร์เซ็นต์ที่เป็นตัวแสดงขีดความสามารถของคำบ่งบอกตามวิธีการนับจำนวนของแต่ละกรณีจะมีค่าแตกต่างกัน ซึ่งกรณีที่เหมาะสมสำหรับวิทยานิพนธ์ฉบับนี้คือ นับคำบ่งบอกที่ปรากฏในทุกลักษณะ เนื่องจากเป็นวิธีที่สอดคล้องกับการนำไปใช้ในการพัฒนางานทางด้านภาษาศาสตร์คอมพิวเตอร์ ซึ่งเป็นงานที่คอมพิวเตอร์มีบทบาทในการเป็นผู้ดูภาษาและนำผลการวิเคราะห์ขีดความสามารถนี้ไปใช้ประโยชน์ โดยพื้นฐานแล้วการมองดูภาษาของเครื่องคอมพิวเตอร์จะมีลักษณะเหมือนการเทียบดูแต่รูปภาษาไม่พิจารณาความสัมพันธ์ใดๆ ทางภาษา ดังนั้นคอมพิวเตอร์จะมองว่าคำที่มีสายอักขระเหมือนกันคือรูปภาษาเดียวกัน ดังเช่น คำบ่งบอก *นาง* ไม่ว่าจะปรากฏเป็นคำบ่งบอกหน้าชื่อเฉพาะ, ปรากฏใน "*นางสาว*" หรือปรากฏใน "*นางเอก*" คอมพิวเตอร์จะมองว่าเป็นรูปภาษาเดียวกัน และจะทำการนับทั้ง 3 กรณี เป็นจำนวนของคำบ่งบอก *นาง* ที่พบในคลังข้อมูล การนับจำนวนทั้งหมดของคำบ่งบอกแต่ละคำ จึงควรเลียนแบบตามมุมมองของคอมพิวเตอร์คือ นับคำบ่งบอกที่ปรากฏในลักษณะต่างๆ ทุกกรณี อย่างไรก็ตามในการนับคำตามมุมมองของคอมพิวเตอร์ต้องคำนึงถึงเรื่องความถูกต้องตาม

อักขรวิธีด้วย เช่น ไม่ควรรับ นาง ใน "พัฒนางาน" จึงอาจกล่าวได้ว่า สำหรับวิทยานิพนธ์ฉบับนี้ การวิเคราะห์ขีดความสามารถในการบ่งบอกของค่าบ่งบอกได้ใช้แนวการวิเคราะห์ที่คำนึงถึงความถูกต้องตามหลักภาษาพร้อมทั้งสอดคล้องหรือขนานไปกับลักษณะที่เกิดขึ้นในการประมวลผลภาษาด้วยเครื่องมือมากที่สุดเท่าที่เป็นไปได้

5.2.5.2 ขีดความสามารถของค่าบ่งบอก

เมื่อทำการประมวลหาเปอร์เซ็นต์ของค่าบ่งบอกตามแนวทางที่ได้กล่าวมาแล้วนั้น จะได้ค่าขีดความสามารถของค่าบ่งบอกที่สามารถนำมาวิเคราะห์แจกแจงได้ในสองลักษณะคือ วิเคราะห์แจกแจงขีดความสามารถของค่าบ่งบอกแต่ละค่า และวิเคราะห์แจกแจงขีดความสามารถของค่าบ่งบอกตามกลุ่มค่าบ่งบอก โดยค่าขีดความสามารถของค่าบ่งบอกแต่ละค่าจะแสดงค่าขีดความสามารถจากมากไปหาน้อยดังที่ปรากฏในตารางที่ 5-3 ส่วนการวิเคราะห์ขีดความสามารถของค่าบ่งบอกโดยพิจารณาเป็นกลุ่มค่านั้นจะอาศัยข้อมูลทางสถิติอันประกอบด้วย ค่าพิสัย (Range or Min. – Max.), ค่าเฉลี่ย และค่าเบี่ยงเบนมาตรฐาน (Standard Deviation) ของแต่ละกลุ่มค่าบ่งบอก ดังที่แสดงไว้ในตารางที่ 5-4 ซึ่งค่าเฉลี่ยจะเป็นข้อมูลที่เป็นตัวแทนของการบ่งบอกถึงขีดความสามารถในการบ่งบอกของกลุ่มค่าบ่งบอกแต่ละกลุ่ม ค่าพิสัยจะแสดงให้เห็นถึงค่าต่ำสุดและค่าสูงสุดของขีดความสามารถในการบ่งบอกของแต่ละกลุ่มค่า และค่าเบี่ยงเบนมาตรฐานบอกให้ทราบถึงการกระจายของข้อมูล กล่าวคือ แสดงให้เห็นว่าค่าขีดความสามารถของค่าบ่งบอกในกลุ่มค่าหนึ่งมีค่าแตกต่างกันมากหรือมีค่าขีดความสามารถใกล้เคียงกัน

จากตารางที่ 5-3 จะเห็นได้ว่าขีดความสามารถในการบ่งบอกชื่อเฉพาะของค่าบ่งบอกแต่ละค่าจะแตกต่างกันอยู่ในช่วงตั้งแต่ 0.50 เปอร์เซ็นต์ ไปจนถึง 100 เปอร์เซ็นต์ เมื่อพิจารณาต่อไปโดยมองเป็นกลุ่มค่า (ดูตารางที่ 5-4) จะพบว่าค่าเฉลี่ยของทั้ง 19 กลุ่มค่าจะปรากฏอยู่ในช่วงตั้งแต่ 19.65 ถึง 100 เปอร์เซ็นต์ ซึ่งแสดงให้เห็นว่า กลุ่มค่าบ่งบอกจะมีขีดความสามารถในการบ่งบอกต่ำไปจนถึงสามารถที่จะเป็นตัวบ่งบอกที่ดีมีขีดความสามารถสูงมากในการบ่งบอกการปรากฏชื่อเฉพาะ รวมทั้งแสดงให้เห็นว่าขีดความสามารถของกลุ่มค่าบ่งบอกต่างๆไม่ปรากฏแนวโน้มที่จะมีขีดความสามารถในการบ่งบอกสูงอย่างชัดเจนหรือต่ำอย่างชัดเจน เมื่อพิจารณาประกอบกับค่าเบี่ยงเบนมาตรฐานและค่าพิสัยพบว่า โดยส่วนใหญ่การกระจายของข้อมูลในแต่ละกลุ่มค่าบ่งบอกมีการกระจายของข้อมูลอยู่มาก นั้นหมายความว่าในแต่ละกลุ่มค่าบ่งบอกจะประกอบด้วยค่าบ่งบอกที่มีขีดความสามารถในการบ่งบอกต่ำมากไปจนถึงสูงมาก จะมีเพียงกลุ่มค่าที่ 1 กลุ่มค่าที่ 4 และกลุ่มค่าที่ 18 ที่มีขีดความสามารถในการบ่งบอกสูงและมีขีดความสามารถในการบ่งบอกสูงใกล้เคียงกัน แต่ค่าบ่งบอกกลุ่มที่ 4 และ 18 ในข้อมูลที่พบ

เป็นกลุ่มที่มีสมาชิกเพียงหนึ่งหรือสองตัว จึงยังไม่สามารถมองเป็นภาพของกลุ่มได้ เราจึงสามารถกล่าวได้ว่าขีดความสามารถในการบ่งบอกของกลุ่มคำไม่เป็นไปในทิศทางเดียวกัน จึงไม่สามารถสรุปได้ว่าคำบ่งบอกในลักษณะใดที่น่าจะเป็นกลุ่มที่มีขีดความสามารถในการบ่งบอกสูง จะพบก็เพียงข้อสังเกตที่ว่าคำบ่งบอกในรูปอักษรย่อที่มีความหมายแสดงถึงฐานะหรือลำดับชั้นของบุคคล (กลุ่มคำที่ 1) น่าจะเป็นตัวบ่งบอกที่มีความสามารถในการบ่งบอกการปรากฏของชื่อเฉพาะได้อย่างแม่นยำและถูกต้องทุกครั้ง ดังนั้นในการนำขีดความสามารถของคำบ่งบอกไปใช้บ่งบอกถึงการปรากฏของชื่อเฉพาะ ควรจะมองเป็นขีดความสามารถของคำบ่งบอกแต่ละคำ เนื่องจากการมองขีดความสามารถของคำบ่งบอกในภาพรวมของกลุ่มคำนั้น ปรากฏว่ามีการกระจายตัวของค่าขีดความสามารถมาก ขีดความสามารถของกลุ่มคำบ่งบอกต่างๆจึงไม่เป็นไปในทิศทางใดทิศทางหนึ่งอย่างชัดเจน การมองขีดความสามารถของคำบ่งบอกอย่างเจาะจงเป็นขีดความสามารถของคำบ่งบอกในแต่ละคำ จึงน่าจะมีความเหมาะสมมากกว่า

| ขีดความสามารถในการบ่งบอก | คำบ่งบอก |
|--------------------------|---|
| 100.00% - 90.00 % | โค้ง ร.ร. รศ. ทวีป แม่น้ำ นายแพทย์ ร.พ. รองประธานาธิบดี เขตรักษาพันธุ์สัตว์ป่า ถ้ำ กลุ่มน้ำทะเลสาบ ปากน้ำ คาวบอย ราชสำนัก สหพันธ์ ร.ต.ต. อำเภอ หม่อมหลวง พล.ต. พล.อ.อ. เขตห้ามล่าสัตว์ป่า พ.ต.ท. นางสาว พล.ต.ต. ร.ต.อ. สน. ดร. พล.อ. ผศ. แพทย์หญิง จ.ส.ต. หม่อมราชวงศ์ ถ. ส.ต.ท. พล.ต.อ. เผ่า จอมพล ทิวเขา กิ่งอำเภอ หม่อมเจ้า พ.ต.อ. เขตฯ ดีเจ. บล. กรุ๊ป บริษัท...จำกัด |
| 89.99% - 80.00% | กระทรวง สำนักข่าว อุทยาน มูลนิธิ บิ๊ก ซอย จ. |
| 79.99% - 70.00% | องค์การ โรงแรม ผู้พัน ย่าน สนธิสัญญา |
| 69.99% - 60.00% | สำนักงาน พ.ร.บ. สมาคม นาย ราชวงศ์ สนามกอล์ฟ เจ้าหญิง หนู กรม ศาล สนามบิน อำเภอ |
| 59.99% - 50.00% | นาง จังหวัด เขื่อน สมรภูมิ แผนก สั่งเวียน คุณพี่ เทศบาล ป้า วุฒิสมาชิก ประธานาธิบดี ช. สหภาพ น.ส. กองทุน แบงก์ เจ้าชาย |
| 49.99% - 40.00% | มหาวิทยาลัย ผืนป่า ภาษา |
| 39.99% - 30.00% | พี่ เกาะ พรรค บริษัท พระเจ้า โรงเรียน ตระกูล สถาบันราชภัฏ อ. โรงพยาบาล |
| 29.99% - 20.00 % | อาจารย์ เครือ พระราชวัง น้อง ถนน ผู้กำกับ ตำบล สนาม กรุง ประเทศ นายกฯ ธนาคาร เมือง โฆษก โรงงาน สะพาน ศ. ผึ้ง |
| 19.99% - 10.00 % | คุณ แถบ เครือชาย เขต นายกรัฐมนตรี ศูนย์ สภา ม. ภูมิภาค องค์การ ป่า รายการ หมู่บ้าน |
| 9.99% - 0.50 % | อาคาร ทะเล คณะกรรมการ สถาบัน สวน กอง สำนัก หาด บ้าน ค่าย ครู กลุ่ม ต. ฝ่าย วัด เสีย หน่วย วิทยาลัย ห้อง นคร ครอบครัว เจ้า วัง ประธาน ลูก เขา ตลาด วง รัฐ |

ตารางที่ 5-3 ขีดความสามารถในการบ่งบอกของคำบ่งบอกแต่ละคำ

| กลุ่มคำบ่งบอก | จำนวนคำบ่งบอก ในแต่ละกลุ่ม | ขีดความสามารถของคำบ่งบอก (เปอร์เซ็นต์) | | | |
|---------------|-------------------------------|--|--------|-----------|---------------------|
| | | ค่าพิสัย | | ค่าเฉลี่ย | ค่าเบี่ยงเบนมาตรฐาน |
| | | Min. | Max. | | |
| กลุ่มคำที่ 1 | 12 | 100.00 | 100.00 | 100.00 | 0.00 |
| กลุ่มคำที่ 2 | 6 | 1.69 | 100.00 | 39.41 | 35.53 |
| กลุ่มคำที่ 3 | 3 | 20.00 | 100.00 | 73.33 | 46.19 |
| กลุ่มคำที่ 4 | 1 | 100.00 | 100.00 | 100.00 | 0.00 |
| กลุ่มคำที่ 5 | 7 | 0.99 | 100.00 | 64.69 | 38.49 |
| กลุ่มคำที่ 6 | 4 | 50.00 | 100.00 | 69.07 | 21.78 |
| กลุ่มคำที่ 7 | 8 | 5.26 | 100.00 | 56.58 | 41.12 |
| กลุ่มคำที่ 8 | 11 | 0.99 | 100.00 | 39.93 | 33.20 |
| กลุ่มคำที่ 9 | 7 | 6.85 | 89.61 | 51.81 | 32.50 |
| กลุ่มคำที่ 10 | 8 | 4.24 | 100.00 | 42.01 | 36.30 |
| กลุ่มคำที่ 11 | 16 | 0.85 | 100.00 | 35.78 | 34.98 |
| กลุ่มคำที่ 12 | 5 | 2.63 | 100.00 | 60.53 | 42.56 |
| กลุ่มคำที่ 13 | 29 | 1.25 | 100.00 | 39.82 | 33.18 |
| กลุ่มคำที่ 14 | 18 | 0.50 | 100.00 | 35.41 | 32.03 |
| กลุ่มคำที่ 15 | 6 | 1.92 | 100.00 | 42.68 | 39.07 |
| กลุ่มคำที่ 16 | 5 | 9.09 | 100.00 | 81.82 | 40.66 |
| กลุ่มคำที่ 17 | 13 | 1.30 | 100.00 | 62.24 | 42.19 |
| กลุ่มคำที่ 18 | 2 | 68.75 | 75.00 | 71.88 | 4.42 |
| กลุ่มคำที่ 19 | 3 | 4.17 | 44.44 | 19.65 | 21.69 |

ตารางที่ 5-4 ค่าพิสัย ค่าเฉลี่ย และค่าเบี่ยงเบนมาตรฐานของขีดความสามารถในการบ่งบอก
ของแต่ละกลุ่มคำบ่งบอก

5.3 การเว้นวรรค (Space)

บริบทในอีกลักษณะหนึ่งซึ่งพบว่าปรากฏกับชื่อเฉพาะในรูปแบบซ้ำๆกันคือ การปรากฏวรรค (Space) อยู่ติดกับสายอักขระที่เป็นชื่อเฉพาะ นั่นคือ มีการเว้นวรรค (Space) เมื่อปรากฏชื่อเฉพาะ ปรากฏการณ์ที่สนับสนุนให้บริบทลักษณะที่กล่าวถึงนี้สามารถจะเป็นบริบทบ่งบอกถึงชื่อเฉพาะได้ก็คือ ชื่อเฉพาะจะปรากฏโดยมีการเว้นวรรคปรากฏด้วยเป็นจำนวน 1,809 ชื่อ หรือ 71.02 เปอร์เซ็นต์ของชื่อเฉพาะที่ศึกษา และบริบทนี้พบกับชื่อเฉพาะทั้งรูปที่มีคำบ่งบอกและไม่ปรากฏคำบ่งบอก รวมทั้งมีรูปแบบการปรากฏที่เด่นชัด 3 รูปแบบด้วยกันคือ ปรากฏการเว้นวรรคก่อนแล้วจึงปรากฏชื่อเฉพาะ (รูปแบบที่ 1) ปรากฏชื่อเฉพาะแล้วจึงปรากฏการเว้นวรรค (รูปแบบที่ 2) และปรากฏการเว้นวรรคที่หน้าและหลังของสายอักขระชื่อเฉพาะ (รูปแบบที่ 3) ดังตัวอย่างข้างล่างนี้

| | |
|--|--|
| รูปแบบที่ 1 วรรค – ชื่อเฉพาะ | "...เป็นสิ่งที่สะท้อนให้เห็นว่า <u>ประเทศไทย</u> ยังมีปัญหาการคอร์รัปชันรุนแรง..." |
| รูปแบบที่ 2 ชื่อเฉพาะ – วรรค | "...ธนาคารไทยธนาคารมีลักษณะใกล้เคียง <u>ธนาคารกรุงไทย</u> ที่รับซื้อออร์แกนส์..." |
| รูปแบบที่ 3 วรรค – ชื่อเฉพาะ - วรรค | "อย่างไรก็ตาม <u>บล.กรุงศรีอยุธยา</u> ไม่แน่ใจว่ามาตรการตอบโต้การฟุ่มเฟือย..." |

เมื่อนำชื่อเฉพาะทั้ง 1,809 รูป มาวิเคราะห์แจกแจง (ดูตารางที่ 5-5) พบว่าการปรากฏของรูปแบบที่ 2 คือการปรากฏวรรคหลังจากการปรากฏชื่อเฉพาะเป็นรูปแบบที่พบมากที่สุด รองลงมาได้แก่ การปรากฏการเว้นวรรคที่หน้าและหลังของชื่อเฉพาะ และรูปแบบที่ 1 คือการปรากฏการเว้นวรรคก่อนแล้วจึงปรากฏชื่อเฉพาะ นอกจากนี้แล้วถ้าพิจารณาเรื่องของคำบ่งบอกร่วมด้วย จะพบว่าในชื่อเฉพาะรูปที่ปรากฏการเว้นวรรครูปแบบต่าง ๆ นี้เป็นชื่อเฉพาะที่ปรากฏคำบ่งบอกร่วมด้วยจำนวน 1,132 ชื่อ หากทำการเปรียบเทียบกับจำนวนชื่อเฉพาะที่ปรากฏกับคำบ่งบอกตามที่กล่าวไว้ในหัวข้อ 5.2 คือจำนวน 1,423 ชื่อ เราสามารถกล่าวได้ว่า ชื่อเฉพาะที่ปรากฏติดกับคำบ่งบอกและปรากฏการเว้นวรรคร่วมด้วยพบเป็นจำนวนมากถึง 79.55 เปอร์เซ็นต์ ซึ่งรายละเอียดในส่วนนี้จะได้กล่าวถึงในหัวข้อถัดไปคือ หัวข้อที่ 5.4

| จำนวนชื่อเฉพาะ รูปแบบ | จำนวนทั้งหมด ที่ปรากฏการเว้นวรรค | ปรากฏการเว้นวรรค และ ปรากฏคำบ่งบอก | ปรากฏการเว้นวรรค แต่ ไม่ปรากฏคำบ่งบอก |
|----------------------------|-------------------------------------|--|---|
| 1. วรรค - ชื่อเฉพาะ | 189 | 84 | 105 |
| 2. ชื่อเฉพาะ - วรรค | 987 | 718 | 269 |
| 3. วรรค - ชื่อเฉพาะ - วรรค | 633 | 330 | 303 |
| | 1,809 | 1,132 | 677 |

ตารางที่ 5-5 จำนวนชื่อเฉพาะที่ปรากฏการเว้นวรรคในลักษณะต่างๆ

5.4 คำบ่งบอกปรากฏร่วมกับการเว้นวรรค

การที่ชื่อเฉพาะจำนวนมากปรากฏร่วมกับบริบททั้งที่เป็นคำบ่งบอกและการเว้นวรรคเป็นปรากฏการณ์ที่แสดงให้เห็นว่าบริบททั้งสองลักษณะมีความเกี่ยวข้องกันในการบ่งบอกถึงการปรากฏของชื่อเฉพาะในข้อความ ความเกี่ยวข้องประการแรกจะเห็นได้จากสัดส่วนของชื่อเฉพาะที่ปรากฏทั้งการเว้นวรรคและคำบ่งบอกเปรียบเทียบกับจำนวนชื่อเฉพาะที่ปรากฏเฉพาะการเว้นวรรค (ดูตารางที่ 5-5) จะเห็นว่าจำนวนของชื่อเฉพาะที่ปรากฏทั้งการเว้นวรรคและคำบ่งบอกมีจำนวนมากกว่าจำนวนชื่อเฉพาะที่ปรากฏเฉพาะการเว้นวรรคประมาณสองเท่า แต่เมื่อพิจารณาต่อไปถึงสัดส่วนของทั้งสองกรณีในแต่ละรูปแบบ จะพบว่าในรูปแบบที่ 2 (ชื่อเฉพาะ – วรรค) และรูปแบบที่ 3 (วรรค – ชื่อเฉพาะ – วรรค) จะมีจำนวนชื่อเฉพาะที่ปรากฏคำบ่งบอกร่วมด้วยเป็นจำนวนมากกว่า ในขณะที่รูปแบบที่ 1 (วรรค – ชื่อเฉพาะ) จะพบชื่อเฉพาะที่ปรากฏแต่การเว้นวรรคเป็นจำนวนมากกว่า จึงเป็นประเด็นที่น่าจะวิเคราะห์ต่อไปว่า หากพิจารณาชื่อเฉพาะทั้งหมดที่นำมาศึกษาหาบริบทบ่งบอกจะพบสัดส่วนของชื่อเฉพาะที่ปรากฏหรือไม่ปรากฏคำบ่งบอกร่วมกับการปรากฏหรือไม่ปรากฏการเว้นวรรคเป็นไปในลักษณะใด ซึ่งผลการวิเคราะห์ที่ได้ (ดูตารางที่ 5-6) แสดงให้เห็นว่าชื่อเฉพาะโดยส่วนใหญ่จะปรากฏการเว้นวรรคร่วมด้วย ไม่ว่าจะมีความบ่งบอกหรือไม่มีคำบ่งบอก (พบการเว้นวรรค : ไม่พบการเว้นวรรค เท่ากับ 1,809 : 738) นอกจากนี้แล้วจำนวนชื่อเฉพาะที่ไม่ปรากฏทั้งการเว้นวรรคและคำบ่งบอกพบเป็นจำนวนมากกว่าชื่อเฉพาะที่ไม่ปรากฏการเว้นวรรคแต่มีคำบ่งบอก (447 : 291) ซึ่งปรากฏการณ์นี้เป็นประเด็นที่น่าสนใจ เนื่องจากในกรณีที่ชื่อเฉพาะไม่ปรากฏการเว้นวรรคเป็นบริบทบ่งบอกก็คาดว่าจะเป็นชื่อเฉพาะ จึงน่าจะพบปรากฏคำบ่งบอกร่วมอยู่ด้วยเพื่อเป็นบริบทบ่งบอกว่ารูปภาษานั้นเป็นชื่อเฉพาะ จึงน่าจะพบ

จำนวนชื่อเฉพาะที่ไม่ปรากฏการเว้นวรรคแต่มีคำบ่งบอกมากกว่า แต่ผลกลับไม่เป็นเช่นนั้น ซึ่งชื่อเฉพาะจำนวน 447 ชื่อที่ไม่ปรากฏการเว้นวรรคและคำบ่งบอก ผู้วิจัยได้ทำการวิเคราะห์ศึกษาและนำเสนอรายละเอียดรวมทั้งอภิปรายเกี่ยวกับชื่อเฉพาะกลุ่มนี้ไว้ในหัวข้อที่ 5.6

| | ปรากฏการเว้นวรรค | ไม่ปรากฏการเว้นวรรค |
|------------------|------------------|---------------------|
| ปรากฏคำบ่งบอก | 1,132 | 291 |
| ไม่ปรากฏคำบ่งบอก | 677 | 447 |
| รวม | 1,809 | 738 |

ตารางที่ 5-6 จำนวนชื่อเฉพาะที่แจกแจงตามการปรากฏหรือไม่ปรากฏของคำบ่งบอกและการเว้นวรรค

ความเกี่ยวข้องอีกประการหนึ่งของคำบ่งบอกและการเว้นวรรคในการบ่งบอกถึงการปรากฏของชื่อเฉพาะคือ ประเด็นขีดความสามารถในการบ่งบอก การพิจารณาบริบทที่เป็นการเว้นวรรคพร้อมกับการพิจารณาบริบทที่เป็นคำบ่งบอกน่าจะทำให้คำบ่งบอกมีขีดความสามารถในการบ่งบอกถึงชื่อเฉพาะในภาษาไทยได้มากขึ้น ซึ่งในการประเมินขีดความสามารถในการบ่งบอกของบริบททั้งสองที่ปรากฏกับชื่อเฉพาะร่วมกัน ผู้วิจัยจะทำการประเมินขีดความสามารถออกมาในรูปแบบของเปอร์เซ็นต์ โดยวิเคราะห์หาว่าคำบ่งบอกแต่ละคำที่ปรากฏร่วมกับการเว้นวรรคในแต่ละรูปแบบมีขีดความสามารถในการบ่งบอกเป็นจำนวนกี่เปอร์เซ็นต์ (ดูตารางที่ 5-7) เช่น คำบ่งบอก นาย จะทำการวิเคราะห์หาขีดความสามารถของ (1) คำบ่งบอก นาย ที่ปรากฏร่วมกับการเว้นวรรคในรูปแบบที่ 1 วรรค – ชื่อเฉพาะ (จากตาราง ขีดความสามารถเท่ากับ 81 เปอร์เซ็นต์) (2) คำบ่งบอก นาย ที่ปรากฏร่วมกับการเว้นวรรคในรูปแบบที่ 2 ชื่อเฉพาะ – วรรค (จากตาราง ขีดความสามารถเท่ากับ 63 เปอร์เซ็นต์) และ (3) คำบ่งบอก นาย ที่ปรากฏร่วมกับการเว้นวรรครูปแบบที่ 3 วรรค – ชื่อเฉพาะ - วรรค (จากตาราง ขีดความสามารถเท่ากับ 99 เปอร์เซ็นต์)

จากการวิเคราะห์คำบ่งบอกแต่ละคำที่ปรากฏร่วมกับการเว้นวรรคทั้ง 3 กรณีดังกล่าว จะได้ค่าขีดความสามารถทั้งหมดจำนวน 483 รายการ ซึ่งนับว่าเป็นข้อมูลจำนวนมากสำหรับการนำเสนอในบทนี้ ในที่นี้ผู้วิจัยจึงนำเสนอผลการวิเคราะห์ขีดความสามารถไว้เพียงบางส่วนโดยคัดเลือกเฉพาะข้อมูลที่มีผลรวมของจำนวนชื่อเฉพาะที่ปรากฏคำบ่งบอกร่วมกับการเว้นวรรคทั้ง 3 รูปแบบ สูงกว่า 10 รูปภาษา เพื่อแสดงให้เห็นแนวโน้มของผลการวิเคราะห์ที่ได้ไว้ในตารางที่ 5-7 และจะแสดงผลทั้งหมดไว้ในภาคผนวก ข

| คำบ่งบอก | วรรค - ชื่อเฉพาะ | | | ชื่อเฉพาะ - วรรค | | | วรรค - ชื่อเฉพาะ - วรรค | | | ค่าเฉลี่ย |
|-------------|--|-----|------------|---|-----|------------|--|-----|------------|-----------|
| | "...พร้อมกันนี้ นายอิ๋วกล่าวอีกว่า..." | | | "...อาจเหมือนเทียบกับนายชวน หลีกภัย ก็ได้..." | | | "...ขณะที่ นายกวี พงษ์ เจ้าหน้าที่..." | | | |
| | PN | All | Percentage | PN | All | Percentage | PN | All | Percentage | |
| นาย | 108 | 133 | 81% | 100 | 160 | 63% | 92 | 93 | 99% | 81% |
| บริษัท | 25 | 46 | 54% | 18 | 74 | 24% | 22 | 22 | 100% | 60% |
| จ. | 22 | 23 | 96% | 20 | 22 | 91% | 22 | 22 | 100% | 96% |
| ประเทศ | 12 | 23 | 52% | 36 | 161 | 22% | 7 | 10 | 70% | 48% |
| พ.ต.ท. | 18 | 18 | 100% | 13 | 13 | 100% | 17 | 17 | 100% | 100% |
| จังหวัด | 10 | 12 | 83% | 27 | 41 | 66% | 6 | 10 | 60% | 70% |
| กระทรวง | 7 | 8 | 88% | 25 | 29 | 86% | 4 | 4 | 100% | 91% |
| กรม | 4 | 5 | 80% | 24 | 30 | 80% | 3 | 3 | 100% | 87% |
| นาง | 10 | 10 | 100% | 13 | 36 | 36% | 8 | 8 | 100% | 79% |
| มหาวิทยาลัย | 10 | 13 | 77% | 13 | 29 | 45% | 8 | 8 | 100% | 74% |
| ดร. | 14 | 14 | 100% | 8 | 8 | 100% | 7 | 7 | 100% | 100% |
| ธนาकार | 5 | 17 | 29% | 13 | 49 | 27% | 4 | 4 | 100% | 52% |
| พี่ | 6 | 9 | 67% | 10 | 29 | 34% | 1 | 1 | 100% | 67% |
| เมือง | 1 | 2 | 50% | 15 | 71 | 21% | 1 | 1 | 100% | 57% |
| สมาคม | 4 | 5 | 80% | 9 | 12 | 75% | 4 | 4 | 100% | 85% |
| องค์การ | 8 | 8 | 100% | 3 | 7 | 43% | 6 | 6 | 100% | 81% |
| คุณ | 8 | 23 | 35% | 3 | 62 | 5% | 5 | 5 | 100% | 47% |
| โรงเรียน | 2 | 6 | 33% | 13 | 38 | 34% | 1 | 3 | 33% | 34% |
| อ. | 6 | 10 | 60% | 5 | 10 | 50% | 5 | 5 | 100% | 70% |
| กรุง | 4 | 5 | 80% | 7 | 14 | 50% | 4 | 4 | 100% | 77% |
| บล. | 4 | 4 | 100% | 7 | 7 | 100% | 3 | 3 | 100% | 100% |
| อาจารย์ | 6 | 13 | 46% | 3 | 16 | 19% | 5 | 5 | 100% | 55% |
| กองทุน | 3 | 3 | 100% | 11 | 28 | 39% | 0 | 0 | - | 70% |
| พระเจ้า | 3 | 5 | 60% | 8 | 23 | 35% | 2 | 2 | 100% | 65% |
| ม. | 4 | 8 | 50% | 5 | 24 | 21% | 4 | 4 | 100% | 57% |
| ศูนย์ | 5 | 6 | 83% | 4 | 20 | 20% | 4 | 4 | 100% | 68% |
| สำนักงาน | 5 | 5 | 100% | 5 | 7 | 71% | 3 | 3 | 100% | 90% |
| เขื่อน | 2 | 2 | 100% | 8 | 14 | 57% | 2 | 2 | 100% | 86% |
| พรรค | 2 | 9 | 22% | 10 | 24 | 42% | 0 | 1 | 0% | 21% |
| กอง | 6 | 10 | 60% | 3 | 32 | 9% | 2 | 2 | 100% | 56% |
| พล.ต. | 5 | 5 | 100% | 1 | 1 | 100% | 5 | 5 | 100% | 100% |
| สำนัก | 6 | 8 | 75% | 0 | 12 | 0% | 4 | 4 | 100% | 58% |

ตารางที่ 5-7 ขีดความสามารถในการบ่งบอกถึงชื่อเฉพาะเมื่อพิจารณาบริบทเว้นวรรคร่วมกับคำบ่งบอก

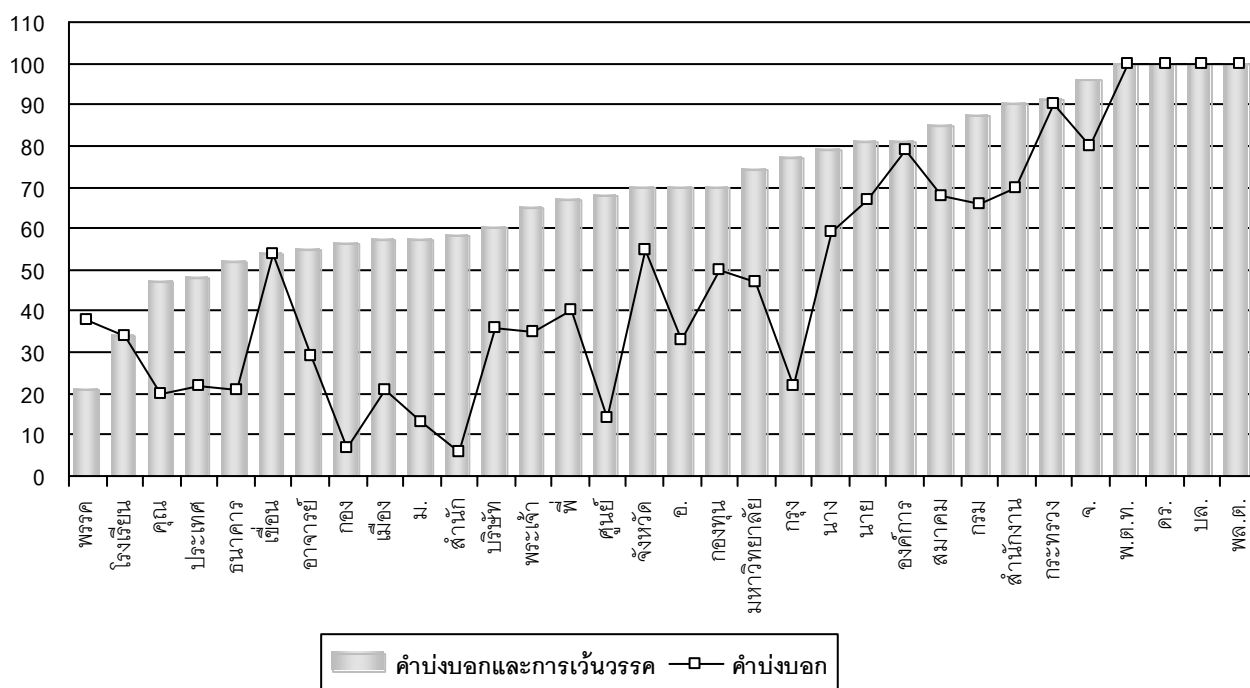
โดยในตารางที่ 5-7 PN, All และ Percentage หมายถึงสิ่งต่อไปนี้

PN หมายถึง จำนวนชื่อเฉพาะที่ปรากฏคำบ่งบอกร่วมกับการเว้นวรรค เช่น จำนวนชื่อเฉพาะที่มีคำบ่งบอก "นาย" และพบการเว้นวรรคที่ส่วนหน้าของชื่อเฉพาะ

All หมายถึง จำนวนรูปภาพทั้งหมดทั้งที่เป็นชื่อเฉพาะและไม่ใช้ชื่อเฉพาะ ซึ่งปรากฏปรากฏคำบ่งบอกร่วมกับการเว้นวรรค ตัวอย่างเช่น จำนวนรูปภาพ "นาย" และปรากฏการเว้นวรรคที่หน้ารูปภาษานั้น

Percentage หมายถึง ชีตความสามารถในการบ่งบอกของคำบ่งบอกที่ปรากฏร่วมกับการเว้นวรรค

เพื่อให้เห็นภาพรวมที่ชัดเจนและสะดวกต่อการนำเสนอ ผู้วิจัยได้ทำการประเมินขีดความสามารถในการบ่งบอกของบริบททั้งสองที่ปรากฏกับชื่อเฉพาะร่วมกัน โดยใช้ค่าเฉลี่ยของขีดความสามารถของคำบ่งบอกแต่ละคำที่ปรากฏร่วมกับการเว้นวรรคทั้ง 3 รูปแบบ (ดูตารางที่ 5-7) มาเปรียบเทียบกับขีดความสามารถในการบ่งบอกเมื่อพิจารณาบริบทคำบ่งบอกเพียงบริบทเดียว ได้ผลดังที่แสดงในแผนภูมิรูปที่ 5-6 ซึ่งจะเห็นแนวโน้มโดยรวมว่าการพิจารณาบริบทเว้นวรรคร่วมไปกับคำบ่งบอก สามารถทำให้คำบ่งบอกส่วนมากมีขีดความสามารถในการบ่งบอกสูงขึ้น



รูปที่ 5-6 ขีดความสามารถของคำบ่งบอกที่ปรากฏร่วมกับการเว้นวรรคในการบ่งบอกถึงชื่อเฉพาะ

5.5 ตัวบ่งบอกระดับปริจเฉท

ตัวบ่งบอกระดับปริจเฉท เป็นอีกบริบทที่มีลักษณะของการบ่งบอกถึงชื่อเฉพาะ แต่เป็นบริบทบ่งบอกถึงชื่อเฉพาะในกรณีชื่อเฉพาะไม่ปรากฏร่วมกับคำบ่งบอก โดยชื่อเฉพาะในคลังข้อมูลปรากฏร่วมกับตัวบ่งบอกระดับปริจเฉทนี้พบประมาณ 193 รูปภาษา หรือกล่าวได้ว่า ในจำนวนชื่อเฉพาะที่ไม่ปรากฏกับคำบ่งบอกซึ่งมีจำนวน 1,124 ชื่อ นั้น (ดูรูป 5.1) จะมีชื่อเฉพาะเพียง 17.17 เปอร์เซนต์ที่ปรากฏตัวบ่งบอกระดับปริจเฉท

ตัวบ่งบอกระดับปริจเฉทที่กล่าวถึงในที่นี้ เป็นบริบทที่บ่งบอกถึงชื่อเฉพาะโดยการพิจารณาความเชื่อมโยงระหว่างชื่อเฉพาะที่ปรากฏในข้อความต่อเนื่อง ความเชื่อมโยงดังกล่าวจะปรากฏในสองลักษณะ โดยในลักษณะแรกชื่อเฉพาะจะปรากฏร่วมกับคำบ่งบอกมาก่อน แต่มาละคำบ่งบอกในข้อความต่อมา ซึ่งการเข้าใจและรู้จำชื่อเฉพาะรูปที่ไม่ปรากฏร่วมกับคำบ่งบอกนี้สามารถทำได้โดยการอ้างอิงย้อนไปถึงชื่อเฉพาะรูปเดียวกันที่ปรากฏติดกับคำบ่งบอกในข้อความที่นำมาข้างหน้า หรือเป็นลักษณะการเชื่อมโยงที่เรียกว่า การอ้างตาม (Anaphoric Relation) โดยชื่อเฉพาะที่ปรากฏตัวบ่งบอกลักษณะนี้มีประมาณ 133 ชื่อ (จาก 193 ชื่อ หรือคิดเป็น 68.91 เปอร์เซนต์) ส่วนการปรากฏในอีกลักษณะหนึ่งมีลักษณะกลับกันคือ รูปที่ปรากฏมาก่อนในข้อความ เป็นชื่อเฉพาะรูปที่ไม่ปรากฏคำบ่งบอกแล้วจึงปรากฏชื่อเฉพาะรูปเดียวกันที่มีคำบ่งบอกปรากฏอยู่ในข้อความต่อมา การรู้จำถึงชื่อเฉพาะที่ถูกละคำบ่งบอกจึงต้องอาศัยการอ้างอิงไปยังชื่อเฉพาะที่ปรากฏร่วมกับคำบ่งบอกในข้อความที่ตามหลังมา เป็นลักษณะการเชื่อมโยงที่เรียกว่า การอ้างก่อน (Cataphoric Relation) ซึ่งลักษณะนี้ปรากฏกับชื่อเฉพาะจำนวน 60 ชื่อ (31.09 เปอร์เซนต์)

ในกรณีของชื่อเฉพาะที่ปรากฏตัวบ่งบอกลักษณะการอ้างก่อน (Cataphoric) พบว่าในชื่อเฉพาะจำนวน 60 ชื่อ มีชื่อเฉพาะจำนวน 42 ชื่อปรากฏอยู่ในข้อความส่วนแรกของข้อความต่อเนื่องซึ่งเป็นข้อความสรุปของข่าวหรือบทความ และมีข้อสังเกตเกี่ยวกับชื่อเฉพาะจำนวน 42 ชื่อนี้คือ มีการปรากฏเครื่องหมายอัฒภาค (" ") ร่วมด้วย โดยปรากฏกับชื่อเฉพาะประเภทบุคคล (หรือชื่อคน) จำนวน 11 ชื่อ, ประเภทองค์กร จำนวน 4 ชื่อ และประเภทสถานที่ จำนวน 1 ชื่อ การที่ชื่อเฉพาะถูกกล่าวถึงเป็นครั้งแรกในข้อความต่อเนื่องโดยไม่มีตัวบ่งบอกถึงการเป็นชื่อเฉพาะ ผู้ส่งสารอาจแสดงความชี้เฉพาะเจาะจงโดยการใส่เครื่องหมายอัฒภาคและใช้เครื่องหมายนี้กับชื่อเฉพาะประเภทสถานที่น้อยที่สุดนั้นอาจเนื่องมาจากชื่อเฉพาะประเภทนี้มักเป็นข้อมูลภาษาที่ผู้ใช้ภาษามีภูมิความรู้ร่วมกันว่าเป็นชื่อเฉพาะมากกว่าข้อมูลที่เป็นชื่อเฉพาะ

ประเภทบุคคลและองค์กร อย่างไรก็ตามประเด็นเกี่ยวกับเครื่องหมายอัฒประกาศนี้เป็นเพียงข้อสังเกตที่พบจากข้อมูลจำนวนไม่มาก ยังไม่สามารถเป็นข้อสรุปได้

ตัวอย่างข้อความต่อเนื่องรูปที่ 5-7 แสดงให้เห็นตัวอย่างของชื่อเฉพาะรูปหนึ่งซึ่งปรากฏลักษณะของตัวบ่งบอกระดับปริจเฉท โดยชื่อเฉพาะรูปนี้ถูกกล่าวถึงในครั้งแรกด้วยรูปภาษาบางส่วนคือ "สหวิริยา สติล" ที่ตำแหน่ง (ก) เมื่อพิจารณาข้อความต่อเนื่องที่ตามมาจะพบชื่อเฉพาะรูปนี้ปรากฏด้วยรูปเต็มที่มีคำบ่งบอกที่ตำแหน่ง (ข) ในการวิเคราะห์รูปภาษา "สหวิริยา สติล" ที่ตำแหน่ง (ก) จึงต้องอาศัยบริบทจากข้อความที่ตามมาคือการปรากฏของรูปภาษา "บริษัท สหวิริยา สติล อินดัสตรี" ที่ตำแหน่ง (ข) ส่วนรูปภาษาที่ตำแหน่ง (ค) (ง) และ (จ) จะมีความเชื่อมโยงในลักษณะที่อ้างไปถึงบริบทที่ปรากฏอยู่ในข้อความที่นำมาข้างหน้า ซึ่งก็คือรูปที่ปรากฏในตำแหน่ง (ข)

บล.กรุงศรีอยุธยาแนะนำทยอยเก็บหุ้น (ก) **"สหวิริยา สติล"** หลังประเมินแนวโน้มธุรกิจครึ่งปีหลังสดใส รับผิดชอบต่อภาระปรับขึ้นราคาขายเหล็กที่ร้อน และมาตรการตอบโต้การทุ่มตลาด คาดกำไรทั้งปี 1.55 พันล้านบาท

บล. กรุงศรีอยุธยา ประเมินว่า จากข่าวที่คณะกรรมการกลางว่าด้วยราคาสินค้าและบริการ ได้อนุมัติให้ (ข) **บริษัท สหวิริยา สติล อินดัสตรี** ปรับราคาจำหน่ายเหล็กแผ่นรีดร้อนจากราคาเดิม 11.66 บาทต่อกิโลกรัม เป็น 13.30 บาทต่อกิโลกรัม ตามที่บริษัทได้ยื่นขอไป โดยให้มีผลในทางปฏิบัติในทันทีนั้น จะส่งผลดีต่อแนวโน้มการทำธุรกิจของ (ค) **สหวิริยา สติล** ในครั้งหลังของปีนี้สดใสต่อเนื่องจากครึ่งปีแรก

.....

บล. กรุงศรีอยุธยา ระบุถึงสาเหตุที่ทำให้มั่นใจว่าแนวโน้มการค้าเงินธุรกิจของ (ง) **สหวิริยา สติล** ในครึ่งปีหลังสดใส มาจากปัจจัยสนับสนุนหลายประการ ประกอบด้วย 1. คาดว่ากำไรสุทธิในครึ่งปีหลัง จะเท่ากับ 557 ล้านบาท เทียบกับครึ่งหลังของปี 2544 ที่มีผลขาดทุนสุทธิ 368 ล้านบาท เนื่องจากบริษัทยังได้รับผลดีจาก 2 ปัจจัย ได้แก่

หลังจากหมดอายุการคุ้มครองด้วยวิธีเซอร์ชาร์จ์ในปลายเดือนก.ค.ที่ผ่านมา บริษัทก็ได้รับการคุ้มครองต่อด้วยวิธีต่อต้านการทุ่มตลาด ทำให้ลูกค้ามีแนวโน้มที่จะซื้อเหล็กที่ร้อนภายในประเทศต่อไป และการที่รัฐบาลอนุมัติปรับราคาขายให้สอดคล้องกับราคาตลาดในขณะนี้เป็นการดีกับ (จ) **สหวิริยา สติล** เพราะจะสามารถรักษาส่วนต่างกำไรระหว่างราคาขายกับวัตถุดิบ slab ได้ เพราะขณะนี้ ราคา slab ได้ปรับขึ้นไปอย่างมากแล้ว...

รูปที่ 5-7 ตัวอย่างตัวบ่งบอกระดับปริจเฉทในข้อความต่อเนื่อง

จากที่กล่าวมานั้น เราสามารถสรุปได้ว่าตัวบ่งบอกในระดับปริจเฉทโดยส่วนใหญ่ มักจะปรากฏในลักษณะอ้างตาม (Anaphoric) นอกจากนี้ชื่อเฉพาะบางรูปภาษาที่ไม่ปรากฏคำบ่งบอกอาจพบร่วมกับชื่อเฉพาะรูปเดียวกันที่มีคำบ่งบอกทั้งในข้อความที่นำมาข้างหน้าและในข้อความที่ตามหลังมาอยู่ภายในข้อความต่อเนื่องเดียวกัน จึงอาจวิเคราะห์ให้เป็นลักษณะการ

เชื่อมโยงแบบอ้างตาม (Anaphoric) หรือ อ้างก่อน (Cataphoric) ก็ได้ ส่วนตัวบ่งบอกระดับปริเฉทแบบอ้างก่อน (Cataphoric) นั้น แม้จะพบเป็นจำนวนน้อยกว่าแต่ก็เป็นการปรากฏของชื่อเฉพาะในลักษณะพิเศษที่พบได้ในการใช้ภาษาที่เกิดขึ้นจริง ดังนั้นในกรณีที่มีการประมวลผลชื่อเฉพาะรูปที่ไม่ปรากฏคำบ่งบอกโดยอาศัยตัวบ่งบอกแบบอ้างตาม (Anaphoric) แล้วไม่สามารถรู้จำชื่อเฉพาะนั้นได้ อาจจะต้องมาคำนึงถึงตัวบ่งบอกแบบอ้างก่อน (Cataphoric) ร่วมด้วย

5.6 ชื่อเฉพาะที่ไม่ปรากฏบริบทบ่งบอก

ในการวิเคราะห์ชื่อเฉพาะส่วนที่ไม่ปรากฏติดกับคำบ่งบอกจำนวน 1,124 ชื่อ ซึ่งพบว่าชื่อเฉพาะจำนวน 193 ชื่อ ปรากฏลักษณะตัวบ่งบอกระดับปริเฉทดังที่กล่าวไปในหัวข้อที่ 5.5 จึงยังมีชื่อเฉพาะอีกจำนวนหนึ่งคือ ชื่อเฉพาะจำนวน 931 ชื่อ ที่ไม่พบบริบทที่เป็นคำบ่งบอกและตัวบ่งบอกระดับปริเฉท เมื่อนำมาวิเคราะห์แจกแจงถึงจำนวนของแต่ละประเภท (ดูตารางที่ 5-8) พบว่าชื่อเฉพาะประเภทที่พบมากที่สุดคือ ประเภทสถานที่ รองลงมาตามลำดับคือ ประเภทบุคคล องค์กร และประเภทอื่นๆ หากพิจารณาต่อไปโดยการเปรียบเทียบจำนวนชื่อเฉพาะที่ไม่พบบริบทบ่งบอกกับชื่อเฉพาะที่ปรากฏร่วมกับคำบ่งบอกประเภทเดียวกัน (ดูตารางที่ 5-8) จะเห็นได้ว่าชื่อเฉพาะประเภทองค์กร ปรากฏในลักษณะที่ไม่พบบริบทบ่งบอกด้วยจำนวนที่น้อยกว่ามาก เมื่อเทียบกับลักษณะการปรากฏร่วมกับคำบ่งบอก คือปรากฏโดยไม่พบบริบทบ่งบอกเพียง 62 ชื่อ แต่ปรากฏโดยมีคำบ่งบอกเป็นจำนวนถึง 421 ชื่อ ในขณะที่ชื่อเฉพาะประเภทสถานที่และบุคคล ปรากฏด้วยสัดส่วนที่ไม่แตกต่างกันมาก และเมื่อทำการวิเคราะห์ชื่อเฉพาะจำนวน 931 ชื่อนี้แยกตามประเภทรวมทั้งพิจารณาเปรียบเทียบกับชื่อเฉพาะที่ปรากฏร่วมกับคำบ่งบอก พบลักษณะทางภาษาบางประการต่อไปนี้

| จำนวน ประเภท | ชื่อเฉพาะที่ไม่ปรากฏบริบทบ่งบอก | ชื่อเฉพาะที่ปรากฏร่วมกับคำบ่งบอก |
|-------------------------|---------------------------------|----------------------------------|
| ชื่อสถานที่ | 444 | 432 |
| ชื่อบุคคล | 410 | 525 |
| ชื่อองค์กร | 62 | 421 |
| ชื่ออื่นๆ | 15 | 23 |
| ชื่อกฎหมาย | 0 | 16 |
| ชื่อกลุ่มคนร่วมเชื้อสาย | 0 | 6 |

ตารางที่ 5-8 จำนวนชื่อเฉพาะแจกแจงตามประเภท

- (1) กลุ่มชื่อเฉพาะประเภทสถานที่ ที่ปรากฏโดยไม่พบบริบทบ่งบอก โดยส่วนใหญ่พบว่าเป็นชื่อเฉพาะที่อ้างถึงสถานที่ที่มีลักษณะเป็นเขตปกครอง ซึ่งชื่อที่พบนั้นได้แก่ ชื่อประเทศ, ชื่อจังหวัด, ชื่ออำเภอ, ชื่อทวีป, ชื่อบ้าน(หมู่บ้าน), ชื่อเมือง และนคร รวมทั้งพบชื่อประเทศเป็นชื่อที่พบเป็นจำนวนมากอย่างเห็นได้ชัดเจนคือปรากฏเป็นจำนวนถึง 339 ชื่อ หรือ 76.35 เปอร์เซนต์ (339 : 444) นอกจากนี้ชื่อเฉพาะส่วนที่เหลือบางรูปภาษามีลักษณะเป็นรูปภาษาที่สามารถใช้อ้างอิงได้อย่างสมบูรณ์ในตัวเองโดยไม่ต้องประกอบกับคำบ่งบอก เช่น "สถานเสาวภา" "กำแพงเมืองจีน" เป็นต้น การที่ชื่อเฉพาะประเภทสถานที่ ที่มีลักษณะดังกล่าว ปรากฏโดยไม่พบบริบทบ่งบอกนั้นอาจเนื่องมาจากเป็นรูปภาษาที่ผู้ใช้ภาษาส่วนใหญ่มีภูมิความรู้เกี่ยวกับรูปภาษาเหล่านี้ร่วมกัน และมักจะพบเห็นได้เป็นประจำในการใช้ภาษา รวมทั้งชื่อเฉพาะที่ใช้เพื่อการอ้างถึงสถานที่นั้น อาจมีแนวโน้มในการที่จะเกิดขึ้นใหม่น้อยกว่าหากเปรียบเทียบกับชื่อเฉพาะประเภทบุคคล หรือ ประเภทองค์กร ดังนั้นในการใช้ภาษาที่เกิดขึ้นทั่วไปในชีวิตประจำวันซึ่งโดยปกติมีแนวโน้มที่จะใช้รูปภาษาที่กระชับและใช้เท่าที่จำเป็นโดยสามารถสื่อความได้อย่างมีประสิทธิภาพ ชื่อเฉพาะกลุ่มนี้จึงเป็นทางเลือกที่เหมาะสมสำหรับผู้ใช้ภาษา ซึ่งหลักฐานที่สนับสนุนจะเห็นได้จากชื่อเฉพาะที่ปรากฏโดยไม่พบบริบทบ่งบอกเกิดกับชื่อสถานที่ (Place Category) เป็นจำนวนมากที่สุดและปรากฏมากกว่ารูปที่มีคำบ่งบอก ในขณะที่ชื่อเฉพาะประเภทอื่นจะปรากฏรูปที่ไม่พบบริบทบ่งบอกน้อยกว่ารูปที่ปรากฏร่วมกับคำบ่งบอก นอกจากนี้ถ้าวิเคราะห์ดูรูปภาษาในกลุ่มที่ปรากฏร่วมกับคำบ่งบอกจะพบว่าเป็นชื่อเฉพาะที่อ้างถึงสถานที่ต่างๆอย่างหลากหลายไม่ได้จำกัดอยู่เพียงชื่อเฉพาะที่อ้างถึงสถานที่บางลักษณะเท่านั้น จึงต้องปรากฏตัวบ่งบอกเพื่อที่จะให้การสื่อความหมายของรูปภาษานั้นมีความชัดเจน
- (2) กลุ่มชื่อเฉพาะประเภทบุคคล ที่ปรากฏโดยไม่พบบริบทบ่งบอก พบว่าเป็นชื่อเฉพาะที่มีลักษณะทางภาษายู่สองลักษณะ คือ เป็นคำไทย หรือ คำทับศัพท์ภาษาต่างประเทศ ตัวอย่างชื่อเฉพาะที่เป็นคำไทยคือ "จิตร ภูมิศักดิ์" "สืบ นาคะเสถียร" "โสภาค สุวรรณ" "ดัสกร ทองเหลา" "ลูกหว่า" "ธนูชัย ดีเทศน์" และตัวอย่างชื่อเฉพาะที่ทับศัพท์ เช่น "ลี จุง มาน" "ซัดดัม ฮุสเซน" "โยชิโร อิรินะ" "เซบาสเตียน โกรสธอม" "เมล กิบสัน" เป็นต้น โดยมีจำนวนสัดส่วนที่เป็นคำไทยกับคำทับศัพท์คือ 40.24 (165 : 410) กับ 59.76 (245 : 410) เปอร์เซนต์ โดยในส่วนของชื่อเฉพาะที่เป็นคำไทยนี้ มักจะเป็นชื่อเฉพาะที่อ้างถึงบุคคลที่มีชื่อเสียงหรือเป็นที่รู้จักกันดีอยู่แล้ว เมื่อพิจารณาเปรียบเทียบกับชื่อเฉพาะประเภทบุคคลที่ปรากฏคำบ่งบอก (525) จะเห็นความแตกต่างอย่างชัดเจนคือ พบชื่อเฉพาะที่เป็นคำไทย 84.95 เปอร์เซนต์ (446) ชื่อเฉพาะที่เป็นคำทับศัพท์ 15.05 เปอร์เซนต์ (79) จึงมองได้ว่าคำบ่งบอกไม่จำเป็นสำหรับชื่อทับศัพท์ เพราะเมื่อเปรียบเทียบกับจำนวนชื่อเฉพาะประเภทบุคคล ที่ทับศัพท์ทั้งหมด (324 ชื่อ) จะมีคำบ่งบอก

ปรากฏด้วยเพียง 24 เปอร์เซนต์ (79 : 324) ส่วนชื่อเฉพาะที่เป็นคำไทย (611 ชื่อ) จะพบคำบ่งบอกถึง 73 เปอร์เซนต์ (446 : 661) ดังนั้นสำหรับชื่อไทย คำบ่งบอกดูจะเป็นบริบทที่จำเป็น ยกเว้นกรณีชื่อเฉพาะนั้นเป็นชื่อที่เป็นรู้จักกันดี การที่ชื่อเฉพาะในลักษณะทับศัพท์โดยส่วนใหญ่ปรากฏโดยไม่มีบริบทบ่งบอก ก็เนื่องมาจากรูปแบบของการทับศัพท์ภาษาต่างประเทศมีลักษณะพิเศษแตกต่างจากรูปศัพท์ในภาษาไทยผสมกับรูปแบบของชื่อบุคคลที่มีรูปแบบเฉพาะตัว จึงมีศักยภาพเพียงพอที่จะนำไปใช้ในลักษณะไม่ต้องอาศัยตัวบ่งบอกโดยไม่เกิดความกำกวมในการสื่อความหมาย แต่ในกรณีของชื่อคนที่มีโครงสร้างเป็นคำไทยจะมีโอกาสเกิดความกำกวมสูงกว่าเพราะมักจะประกอบด้วยรูปศัพท์เดียวกับคำที่มีชื่ออยู่ในภาษาไทยจึงต้องมีตัวบ่งชี้ในการบ่งบอกว่ารูปภาษานั้นเป็นชื่อเฉพาะ นอกจากนี้ปัจจัยในเรื่องที่ชื่อเฉพาะนั้นใช้อ้างถึงบุคคลที่มีชื่อเสียงหรือเป็นที่รู้จัก ซึ่งเป็นเรื่องของกรณีภูมิความรู้ร่วมกันของผู้ใช้ภาษาก็เป็นส่วนหนึ่งที่ทำให้มีการใช้ชื่อเฉพาะนั้นโดยไม่จำเป็นต้องใช้คำบ่งบอกกำกับไว้ด้วย

- (3) ในกรณีของชื่อเฉพาะประเภทองค์กร กลุ่มที่ปรากฏโดยไม่มีบริบทบ่งบอกมีเพียง 62 รูปภาษา และลักษณะของรูปภาษาที่ปรากฏจะมีอยู่สองลักษณะหลักคือ มีโครงสร้างในลักษณะทับศัพท์ภาษาต่างประเทศจำนวน 44 รูปภาษา (70.79 เปอร์เซนต์) เช่น "ซาโลมอน สมิธ บาร์นีย์" (บริษัท) "พิคาร์ด" (บริษัท) "รอยเตอร์" (สำนักข่าว) เป็นต้น ส่วนอีกลักษณะโดยตัวรูปภาษาที่สร้างขึ้นโดยไม่ปรากฏตัวบ่งบอกพบเป็นจำนวน 16 รูปภาษา (25.81 เปอร์เซนต์) เช่น "สหประชาชาติ" "การท่องเที่ยวแห่งประเทศไทย" "การสื่อสารแห่งประเทศไทย" "การรถไฟแห่งประเทศไทย" เป็นต้น เมื่อพิจารณาดูตารางที่ 5-8 จะเห็นได้ว่าชื่อเฉพาะกลุ่มนี้ปรากฏเป็นจำนวนน้อยกว่ามากเมื่อเปรียบเทียบกับชื่อเฉพาะประเภทสถานที่ และประเภทบุคคล รวมทั้งแตกต่างกับชื่อเฉพาะประเภทองค์กรที่ปรากฏร่วมกับคำบ่งบอกอย่างชัดเจน ทั้งนี้อาจเนื่องมาจากชื่อเฉพาะประเภทองค์กรในภาษาไทยโดยปกติแล้วจะมีโครงสร้างของรูปภาษาในลักษณะที่ประกอบด้วยคำนามทั่วไปหรือคำชนิดอื่นและการเรียงลำดับของคำที่เป็นองค์ประกอบอาจมีโครงสร้างเหมือนหน่วยภาษาระดับสูงกว่าคำ เช่น "ธนาคารเพื่อการเกษตรและสหกรณ์" "สถาบันการแพทย์แผนไทย" "สมาคมฟุตบอลอังกฤษ" "แบงก์ทหารไทย" "ศาลล้มละลายกลาง" เป็นต้น หากไม่มีตัวบ่งบอกปรากฏอยู่ด้วยจะทำให้การอ้างถึงของชื่อเฉพาะประเภทนี้เกิดความกำกวมสูง เช่น ถ้าผู้ส่งสารต้องการสื่อความหมายถึงชื่อเฉพาะที่เป็นชื่อธนาคาร "ธนาคารเพื่อการเกษตรและสหกรณ์" แล้วใช้เพียงรูปภาษา "เพื่อการเกษตรและสหกรณ์" ผู้รับสารอาจตีความเป็นหน่วยภาษาระดับวลี หรือเป็นหน่วยภาษาที่ประกอบด้วยหน่วยย่อยหลายๆหน่วย ซึ่งชื่อเฉพาะประเภทองค์กรที่ปรากฏกับคำบ่งบอกในคลังข้อมูลของวิทยานิพนธ์ มีโครงสร้างของรูปภาษาในลักษณะดังกล่าวนี้ 81.95 เปอร์เซนต์ (345 : 421) ส่วนอีก 18.05 เปอร์เซนต์

เป็นรูปภาษาที่มีโครงสร้างในลักษณะทับศัพท์ ดังนั้นชื่อเฉพาะประเภทองค์กร จึงปรากฏร่วมกับคำบ่งบอกเป็นส่วนใหญ่ และปรากฏโดยไม่มีบริบทบ่งบอกเมื่อรูปภาษานั้นมีคุณลักษณะที่สามารถใช้อ้างถึงได้อย่างไม่เกิดความกำกวม

- (4) สำหรับชื่อเฉพาะประเภทชื่ออื่นๆ, ชื่อกฎหมาย และชื่อกลุ่มคนร่วมเชื้อสาย ซึ่งปรากฏด้วยจำนวนไม่มาก จึงไม่พบลักษณะทางภาษาที่เด่นชัดรวมทั้งประเภทชื่อกฎหมาย และชื่อกลุ่มคนร่วมเชื้อสาย ไม่พบรูปภาษาที่ปรากฏโดยไม่มีบริบทบ่งบอก

โดยสรุปแล้ว การที่ชื่อเฉพาะปรากฏโดยไม่มีบริบทบ่งบอกนั้นเกี่ยวข้องกับลักษณะทางภาษาของชื่อเฉพาะบางประการที่เอื้อต่อการนำชื่อเฉพาะนั้นไปใช้โดยไม่ต้องมีตัวบ่งบอกเพื่อแสดงความเจาะจงว่ารูปภาษานั้นมีสภาพเป็นชื่อเฉพาะ ซึ่งเป็นธรรมชาติของขบวนการพิจารณาใช้รูปภาษาต่างๆที่ปรากฏในภาษาด้วยแนวโน้มที่จะเลือกใช้รูปภาษาที่ย่นย่อกว่าหรือมีความกระชับมากกว่าหากรูปภาษานั้นยังคงศักยภาพในการสื่อความหมายอย่างมีประสิทธิภาพ ลักษณะทางภาษาที่พบได้แก่ โครงสร้างของรูปภาษามีลักษณะที่พิเศษไม่ใช่รูปแบบปกติในภาษาไทยซึ่งก็คือรูปแบบการทับศัพท์ภาษาต่างประเทศ, ลักษณะต่อมาคือชื่อเฉพาะถูกสร้างขึ้นด้วยรูปแบบที่ไม่มีคำบ่งบอก และสุดท้ายคือชื่อเฉพาะนั้นเป็นข้อมูลที่ใช้ภาษาที่มีความรู้ร่วมกัน ซึ่งลักษณะทางภาษาทั้งสามประการนี้ทำให้มนุษย์สามารถรู้จำชื่อเฉพาะที่ไม่ปรากฏบริบทบ่งบอกได้ แต่สำหรับคอมพิวเตอร์แล้วอาจต้องหาวิธีการที่มีลักษณะเป็นรูปแบบที่ชัดเจนหรืออาจต้องจำลองลักษณะวิธีการประมวลผลของชื่อเฉพาะกลุ่มนี้ของมนุษย์ให้ออกมาเป็นแบบจำลองสำหรับคอมพิวเตอร์ให้สามารถประมวลผลได้ต่อไป

5.7 ชื่อเฉพาะปรากฏเป็นองค์ประกอบของรูปภาษาอื่น

ชื่อเฉพาะอีกส่วนหนึ่งที่พบในการศึกษาซึ่งมีลักษณะบางประการแตกต่างไปคือชื่อเฉพาะในส่วนนี้ปรากฏเป็นส่วนประกอบหนึ่งของรูปภาษาซึ่งทั้งหน่วยภาษาไม่มีคุณสมบัติครบถ้วนตามเกณฑ์นิยาม ตัวอย่างของชื่อเฉพาะที่กล่าวถึงนี้คือ "คนไทย" "ชาวฮ่องกง" "รัฐบาลสิงคโปร์" โดยประเด็นที่เกี่ยวข้องกับชื่อเฉพาะกลุ่มนี้มีดังนี้

เนื่องจากวิทยานิพนธ์ฉบับนี้ ได้กำหนดขอบเขตที่จะศึกษาบริบทบ่งบอกของชื่อเฉพาะซึ่งมีคุณสมบัติทางความหมายตามเกณฑ์นิยามของวิทยานิพนธ์ (ดูบทที่ 3) โดยเกณฑ์

นิยามเป็นเกณฑ์ด้านความหมายประกอบด้วยคุณสมบัติทางความหมาย 3 ประการ เมื่อวิเคราะห์ชื่อเฉพาะส่วนนี้พบว่า มีลักษณะบางประการแตกต่างกันไปดังจะแสดงให้เห็นจากตัวอย่างต่อไปนี้คือ

| | | |
|--------------|------------------------|--------------|
| ก. คนไทย | ง. ธนาคารไทยพาณิชย์ | ช. ประเทศไทย |
| ข. สินค้าไทย | จ. บริษัทปูนซิเมนต์ไทย | ซ. เมืองไทย |
| ค. ลายไทย | ฉ. พรรคไทยรักไทย | |

เมื่อพิจารณาเฉพาะรูปเขียนของตัวอย่างทั้ง 9 รูป จะเห็นว่า มีสายอักขระ "ไทย" ปรากฏอยู่ร่วมกับคำอื่นเป็นรูปภาษาในลักษณะเดียวกัน แต่ถ้าพิจารณาทั้งรูปภาษาโดยวิเคราะห์ตามเกณฑ์นิยามแล้ว รูปภาษาตัวอย่างนี้สามารถแยกออกเป็นสองส่วนคือ รูปภาษาที่ ง-ช มีคุณสมบัติทางความหมายของเกณฑ์นิยามทั้ง 3 ประการ ส่วนรูปภาษาที่ ก-ค มีคุณสมบัติทางความหมายของเกณฑ์นิยามไม่ครบทั้ง 3 ประการโดยมีคุณสมบัติของการเป็นวลีคือ มีความหมายเกิดจากผลรวมของความหมายที่เป็นองค์ประกอบ (สุริยา, 2544: 182) เช่น "คนไทย" มีความหมายหลักจาก 'คน' แล้วใช้ชื่อเฉพาะ 'ไทย' มาขยายให้มีความเจาะจงมากขึ้น ทำให้ความหมายรวมจากองค์ประกอบมีความหมายถึง คนที่มีสัญชาติไทยและสื่อถึงว่ามีจำนวนมากกว่าหนึ่ง ซึ่งถ้าเปรียบเทียบกับรูปภาษา "ไทยรักไทย" ในการใช้ภาษารูปนี้ไม่ปรากฏการเกิดความหมายที่เป็นผลรวมของความหมายที่เป็นองค์ประกอบจะเกิดแต่ความหมายในเชิงอ้างอิงที่ไม่สัมพันธ์โดยตรงกับความหมายองค์ประกอบรวมทั้งสิ่งที่อ้างถึงมีสิ่งเดียว นอกจากนั้นรูปภาษาที่ ก-ค ยังสามารถใช้คำอื่นแทรกกลาง* โดยยังมีความหมายคงเดิมคือ "คนไทย" เป็น คนสัญชาติไทย "สินค้าไทย" เป็น สินค้าของไทย "ลายไทย" เป็น ลายแบบไทย ซึ่งเป็นการแสดงให้เห็นชัดเจนขึ้นถึงการสามารถแยกรูปภาษานี้เป็นองค์ประกอบย่อยที่อิสระจากกัน ในขณะที่รูปภาษา ง-ช ไม่สามารถทำการแทรกคำโดยที่ยังคงสภาพการสื่อความหมายรูปแบบเดิมของรูปภาษาได้ เช่น "ธนาคารไทยพาณิชย์" ไม่เท่ากับธนาคารของไทยที่เป็นพาณิชย์

จากที่กล่าวมาแสดงให้เห็นว่า ชื่อเฉพาะสามารถนำไปใช้ในรูปแบบอื่นนอกเหนือจากการปรากฏเป็นหน่วยชื่อเฉพาะที่มีคุณสมบัติเฉพาะประจำรูป ซึ่งโดยรวมของชื่อเฉพาะส่วนนี้จำนวน 308 ชื่อ (10.80 เปอร์เซนต์) ที่พบในคลังข้อมูลจะปรากฏในลักษณะเป็นองค์ประกอบหนึ่งของวลี และประเภทของชื่อเฉพาะโดยส่วนใหญ่เป็นชื่อเฉพาะประเภทสถานที่ประมาณ 306 รูป ตัวอย่างเช่น "ครอบครัวไทย" "รัฐบาลจีน" "ชาวมาเลเซีย" "นายทหารอังกฤษ"

* การใช้คำอื่นแทรกกลางเป็นเกณฑ์ทางวากยสัมพันธ์หนึ่งของการพิจารณาเรื่องคำ (นัฐวุฒิ, 2544)

"ขนมโมจิญี่ปุ่น" "ชาวเกาะเสม็ด" "เด็กเซินโยฯ" "เด็กธรรมชาติ" เป็นต้น ส่วนอีก 2 รูปเป็นชื่อเฉพาะประเภทบุคคล คือ "รัฐบาลมาร์คอส" "แนวมาร์กซ์"

การที่ชื่อเฉพาะสามารถเป็นองค์ประกอบของคำอื่นนี้ อาจก่อให้เกิดปัญหาต่อระบบการรู้จำชื่อเฉพาะด้วยคอมพิวเตอร์ได้ โดยเฉพาะระบบที่อาศัยรายการชื่อเฉพาะในการรู้จำ กล่าวคือ วิธีการเทียบหาชื่อเฉพาะโดยเปรียบเทียบจากรายการชื่อเฉพาะที่มีอยู่ ไม่อาจใช้ได้เสมอไป จึงเป็นสิ่งที่ต้องคำนึงสำหรับนักภาษาศาสตร์คอมพิวเตอร์ที่ต้องการพัฒนาระบบการรู้จำชื่อเฉพาะภาษาไทย

บทที่ 6

สรุปและข้อเสนอแนะ

ในบทนี้จะได้กล่าวสรุปประเด็นที่ได้นำเสนอในบทต่างๆที่ผ่านมา โดยจะกล่าวถึงกระบวนการศึกษาและบริบทของข้อเฉพาะภาษาไทยที่พบในการศึกษาของวิทยานิพนธ์ฉบับนี้ ในหัวข้อที่ 6.1 แล้วนำเสนอแนวทางการนำบริบทของข้อเฉพาะไปใช้ประโยชน์และประเด็นที่น่าจะมีการศึกษาเพิ่มเติมในหัวข้อที่ 6.2

6.1 สรุปกระบวนการศึกษาและผลการศึกษหาบริบทของข้อเฉพาะภาษาไทย

วิทยานิพนธ์นี้ทำการศึกษาเกี่ยวกับข้อเฉพาะภาษาไทยตามแนวภาษาศาสตร์คอมพิวเตอร์ เพื่อวิเคราะห์บริบทที่สามารถบ่งบอกการปรากฏของข้อเฉพาะในข้อความภาษาไทย สำหรับใช้เป็นความรู้ทางภาษาในการประมวลผลข้อเฉพาะของงานด้านภาษาศาสตร์คอมพิวเตอร์ หรือการประมวลผลภาษาธรรมชาติ โดยศึกษาบริบทของข้อเฉพาะจากคลังข้อมูลภาษาที่ประกอบด้วยข้อความภาษาไทยที่เกิดขึ้นจริงในการใช้ภาษาและใช้แนวทางภาษาศาสตร์เข้ามาวิเคราะห์ตัวภาษา ในลักษณะที่สอดคล้องกับการนำไปใช้

เนื่องจากคลังข้อมูลภาษาที่ใช้ในวิทยานิพนธ์เป็นคลังข้อมูลภาษาที่มีการกำกับข้อมูลซึ่งผู้วิจัยจัดสร้างขึ้นสำหรับการศึกษาบริบทของข้อเฉพาะโดยตรง จึงต้องทำการกำกับรูปภาษาที่เป็นข้อเฉพาะไว้ในคลังข้อมูล ซึ่งการดำเนินการในขั้นตอนนี้มีประเด็นที่ต้องพิจารณาอยู่สองประการคือ เกณฑ์สำหรับการวิเคราะห์หารูปภาษาที่เป็นข้อเฉพาะและสัญลักษณ์สำหรับการกำกับข้อเฉพาะ ในประการแรก ความรู้พื้นฐานที่มีบทบาทในการพิจารณาคือมโนทัศน์ของข้อเฉพาะ ซึ่งมโนทัศน์ที่กล่าวถึงในวรรณกรรมต่างๆ มีรายละเอียดบางส่วนแตกต่างกันและขอบเขตของรูปภาษาที่จัดว่าเป็นข้อเฉพาะก็มีความแตกต่างกัน จึงต้องมีการกำหนดเกณฑ์นิยามขึ้นเพื่อกำหนดขอบข่ายของข้อเฉพาะที่จะนำมาศึกษาในครั้งนี้ สำหรับประการหลังเกี่ยวข้องกับการใช้หน่วยกำกับข้อมูล (Tags) ที่ประยุกต์จากแนวทางการกำกับข้อเฉพาะของ TEI (Text Encoding Initiative) นอกจากการกำกับข้อเฉพาะแล้วผู้วิจัยยังทำการกำกับรูปภาษาอื่นเพิ่มเติมเพื่อการวิเคราะห์บริบทของข้อเฉพาะ โดยใช้แนวทางการกำกับข้อมูลของ TEI เช่นกัน ซึ่งรายละเอียดของทั้งสองประเด็นอธิบายไว้โดยละเอียดในบทที่ 3

ในการจัดสร้างคลังข้อมูลนั้น ผู้วิจัยได้รวบรวมข้อความภาษาไทยจากข้อมูลที่ได้เผยแพร่ไว้ในเว็บไซต์ของหนังสือพิมพ์กรุงเทพธุรกิจและนิตยสารสกุลไทย โดยมีขนาดของ

คลังข้อมูลเท่ากับคลังข้อมูลที่ปรากฏชื่อเฉพาะจำนวน 2,855 ชื่อ ซึ่งขบวนการขั้นตอนต่างๆในการจัดสร้างคลังข้อมูลประกอบด้วยขั้นตอนหลัก 3 ขั้นตอนดังที่ได้กล่าวไว้ในบทที่ 4 คลังข้อมูลที่ทำ การกำกับข้อมูลแล้วจะนำมาใช้เป็นเครื่องมือหลักในกระบวนการศึกษาวิเคราะห์ถึงบริบทที่ สามารถบ่งบอกชื่อเฉพาะ โดยกระบวนการวิเคราะห์จะต้องใช้วิธีวิเคราะห์หลายวิธีการประกอบกัน รวมทั้งวิธีวิเคราะห์บางวิธีการจะอาศัยเครื่องมือตัวอื่นนอกเหนือจากคลังข้อมูล ซึ่งวิธีวิเคราะห์ หลักๆได้แก่สี่วิธีการต่อไปนี้คือ วิเคราะห์บริบทบ่งบอกจากข้อความต่อเนื่องในคลังข้อมูลโดยตรง, วิเคราะห์บริบทบ่งบอกจากรายการคอนคอร์แดนซ์, กำกับข้อมูลเพิ่มเติมเพื่อการวิเคราะห์บริบท บ่งบอก และวิเคราะห์ผลเฉพาะอย่างโดยเขียนโปรแกรมภาษา Perl ให้ประมวลผลจากคลังข้อมูล รายละเอียดในส่วนกระบวนการวิเคราะห์นี้กล่าวไว้ในบทที่ 4 เช่นกัน

ผลการศึกษาวิเคราะห์เกี่ยวกับบริบทบ่งบอกชื่อเฉพาะภาษาไทยจากคลังข้อมูล ซึ่งนำเสนอไว้ในบทที่ 5 พบว่า ชื่อเฉพาะส่วนหนึ่งประกอบด้วยคำบ่งบอกนำหน้าชื่อเฉพาะ ในขณะที่ชื่อเฉพาะอีกส่วนหนึ่งคำบ่งบอกจะปรากฏเป็นส่วนหนึ่งของชื่อเฉพาะ บริบทที่สามารถ เป็นบริบทบ่งบอกชื่อเฉพาะภาษาไทย ได้แก่ บริบทที่เป็นคำปรากฏติดกับชื่อเฉพาะหรือคำบ่งบอก, การเว้นวรรค (Space) และตัวบ่งบอกระดับปริจเฉท โดยคำบ่งบอกและการเว้นวรรคเป็นบริบทที่ พบกับชื่อเฉพาะจำนวนมากและเป็นบริบทที่มีความเกี่ยวข้องกันในการบ่งบอกถึงชื่อเฉพาะ รวมทั้งพบว่าชื่อเฉพาะจำนวน 36.55 เปอร์เซ็นต์ไม่พบบริบทบ่งบอก โดยข้อสรุปเกี่ยวกับบริบทบ่ง บวกต่างๆและลักษณะของชื่อเฉพาะที่ไม่พบบริบทบ่งบอก มีดังนี้คือ

1. บริบทบ่งบอกชื่อเฉพาะ

1.1 คำบ่งบอก

คำบ่งบอกโดยส่วนใหญ่มีตำแหน่งการปรากฏอยู่ติดกับส่วนหน้าของชื่อเฉพาะ และโดยปกติจะมีความหมายเดียว มีเพียงบางคำที่มีสองความหมาย รวมทั้งคำบ่งบอกต่างๆที่ พบปรากฏความหมายบางประการร่วมกันในลักษณะที่สามารถแยกออกเป็นกลุ่มคำได้ด้วย การวิเคราะห์อรรถลักษณะ (Componential Analysis) ซึ่งอรรถลักษณะเด่นของกลุ่มคำยังแสดง ให้เห็นความหมายที่บ่งบอกถึงสิ่งที่ชื่อเฉพาะอ้างถึง คำบ่งบอกจึงสามารถบ่งบอกถึงประเภท ของชื่อเฉพาะได้ด้วย โดยชื่อเฉพาะประเภทหนึ่งจะปรากฏคำบ่งบอกกลุ่มที่เฉพาะเจาะจงกลุ่ม หนึ่งและมีอรรถลักษณะเด่นเฉพาะของแต่ละประเภท นอกจากนี้ในการใช้คำบ่งบอกเป็นบริบท บ่งบอกถึงชื่อเฉพาะต้องพิจารณาถึงขีดความสามารถในการบ่งบอกของคำบ่งบอกร่วมด้วย เนื่องจากคำบ่งบอกสามารถปรากฏอยู่ติดกับรูปภาษาอื่นนอกเหนือจากชื่อเฉพาะ ซึ่งการ ประเมินขีดความสามารถของคำบ่งบอกในวิทยานิพนธ์ฉบับนี้ ใช้แนวทางที่สอดคล้องกับ

ลักษณะการประมวลผลภาษาด้วยเครื่องคอมพิวเตอร์ โดยพบว่าขีดความสามารถในการบ่งบอกของคำบ่งบอกเมื่อพิจารณาตามกลุ่มคำ ไม่มีแนวโน้มที่ไปในทิศทางเดียวกัน ยกเว้นกลุ่มคำบ่งบอกฐานะลำดับชั้น (เช่น พ.ต.ท. พล.อ. พ.ต.อ.) กล่าวคือ ค่าขีดความสามารถกระจายแตกต่างกันอยู่ในช่วงขีดความสามารถต่ำไปจนถึงขีดความสามารถสูงมาก จนไม่สามารถสรุปได้ว่าคำบ่งบอกลักษณะใดที่สามารถเป็นตัวบ่งบอกที่มีประสิทธิภาพมากที่สุด ดังนั้นในการพิจารณาใช้ขีดความสามารถของคำบ่งบอกควรจะต้องมองเป็นขีดความสามารถของคำบ่งบอกแต่ละคำ น่าจะมีความเหมาะสมกับลักษณะการประมวลผลด้วยเครื่องมากกว่า

1.2 การเว้นวรรค (Space)

ชื่อเฉพาะโดยส่วนใหญ่ปรากฏการเว้นวรรคร่วมด้วยเสมอ และเป็นบริบทที่มีความเกี่ยวข้องกับคำบ่งบอก คือ การเว้นวรรคมักปรากฏร่วมกับคำบ่งบอก การพิจารณาบริบทเว้นวรรคร่วมไปคำบ่งบอกจะทำให้คำบ่งบอกจะมีขีดความสามารถในการบ่งบอกสูงขึ้นอย่างค่อนข้างชัดเจน ซึ่งรูปแบบการปรากฏเว้นวรรคพบอยู่ 3 ลักษณะ เรียงตามลำดับจากที่พบมากที่สุดคือ ปรากฏการเว้นวรรคอยู่ที่ส่วนหลังของชื่อเฉพาะ, การเว้นวรรคพร้อมกันทั้งที่ส่วนหน้าและส่วนหลังของชื่อเฉพาะ และพบน้อยที่สุดคือ ปรากฏการเว้นวรรคอยู่ที่ส่วนหน้าของชื่อเฉพาะ

1.3 ตัวบ่งบอกระดับปริจเฉท

ในกรณีที่ชื่อเฉพาะไม่ปรากฏร่วมกับคำบ่งบอก จะบ่งบอกโดยอาศัยการเชื่อมโยงไปยังชื่อเฉพาะรูปเดียวกันในข้อความต่อเนื่องเดียวกันที่ปรากฏร่วมกับคำบ่งบอก ซึ่งพบว่าส่วนใหญ่จะเป็นการเชื่อมโยงในลักษณะอ้างตาม (Anaphora) และชื่อเฉพาะจำนวนไม่มากในกลุ่มชื่อเฉพาะที่พบตัวบ่งบอกระดับปริจเฉทจะปรากฏการเชื่อมโยงในลักษณะอ้างก่อน (Cataphora) ที่มักจะพบอยู่ในส่วนแรกของข้อความต่อเนื่องที่เป็นข้อความสรุปข่าวหรือสรุปบทความ

2. ลักษณะทางภาษาของชื่อเฉพาะที่ไม่พบบริบทบ่งบอก

ในกลุ่มชื่อเฉพาะที่ไม่ปรากฏบริบทบ่งบอก พบว่าส่วนใหญ่เป็นชื่อเฉพาะประเภทสถานที่ และประเภทบุคคล โดยชื่อเฉพาะเหล่านี้มีลักษณะทางภาษาที่เด่นชัดบางประการได้แก่ (1) เป็นคำทับศัพท์ภาษาต่างประเทศ ซึ่งมีโครงสร้างของรูปภาษาที่ไม่ใช่รูปแบบปกติในภาษาไทย (2) เป็นข้อมูลภาษาที่ผู้ใช้ภาษามีภูมิความรู้ร่วมกัน เช่น ชื่อเฉพาะที่อ้างถึงประเทศ, ชื่อบุคคลที่มีชื่อเสียงหรือเป็นที่รู้จักกันดีในสังคม เป็นต้น และ (3) เป็นชื่อเฉพาะที่ถูกสร้างขึ้นด้วยรูปแบบเฉพาะที่สามารถใช้อ้างอิงโดยไม่ต้องประกอบตัวบ่งบอก ซึ่งลักษณะทาง

ภาษาเหล่านี้เชื่อมต่อการนำชื่อเฉพาะนั้นไปใช้โดยไม่ต้องมีตัวบ่งบอกเพื่อแสดงความเจาะจงว่า รูปภาษานั้นมีสภาพเป็นชื่อเฉพาะ ชื่อเฉพาะเหล่านี้จึงเป็นทางเลือกที่ผู้ใช้ภาษาเห็นว่าเหมาะสมมากกว่า เพราะเป็นธรรมชาติของขบวนการพิจารณาใช้รูปภาษาต่างๆที่ปรากฏใน ภาษา ที่มีแนวโน้มที่จะเลือกใช้รูปภาษาที่ย่นย่อกว่าหรือมีความกระชับมากกว่า หากรูปภาษานั้นยังคงศักยภาพในการสื่อความหมายอย่างมีประสิทธิภาพ จึงพบชื่อเฉพาะเหล่านี้ปรากฏโดยไม่มีบริบทบ่งบอก ในขณะเดียวกันถ้าชื่อเฉพาะรูปใดมีโครงสร้างของรูปภาษาในลักษณะที่มีความกำกวมสูง เช่น ประกอบด้วยคำที่มีใช้อยู่แล้วในภาษาและมีการเรียงลำดับคำเหมือนวลี สามารถตีความหมายเป็นวลีได้ ในกรณีเช่นนี้จึงต้องมีตัวบ่งบอกกำกับไว้ด้วย เพื่อแยกแยะรูปภาษานั้นว่าเป็นชื่อเฉพาะ

ในการศึกษาวิเคราะห์นอกจากพบบริบทบ่งบอกและลักษณะทางภาษาดังที่กล่าวมาแล้ว ยังพบว่า ความหมายอ้างอิงของชื่อเฉพาะประเภทองค์กร กับชื่อเฉพาะประเภทสถานที่มีลักษณะคาบเกี่ยวกัน กล่าวคือ ในบางบริบทรูปภาษาที่เป็นชื่อเฉพาะประเภทองค์กร จะถูกนำมาใช้ในความหมายอ้างอิงถึงสถานที่คือมีอรรถลักษณะ <พื้นที่หรือบริเวณ> ของชื่อเฉพาะประเภทสถานที่ ขณะเดียวกันในบางบริบทของการนำชื่อเฉพาะประเภทสถานที่ไปใช้ก็เป็นการนำไปใช้ในความหมายของกลุ่มบุคคลที่มีสภาพเป็นนิติบุคคลหรือมีความหมายของชื่อเฉพาะประเภทองค์กร การที่ความคาบเกี่ยวกันทางความหมายปรากฏอย่างชัดเจนในชื่อเฉพาะสองประเภทนี้ ก็เนื่องมาจากสภาพความเป็นจริงที่ว่า กลุ่มบุคคลที่มาประกอบรวมกันเป็นหน่วยเพื่อร่วมทำกิจการใดๆ ย่อมมีเรื่องของสถานที่อาจรวมถึงถาวรวัตถุอย่างเช่นอาคาร เข้ามาเกี่ยวข้องสำหรับใช้ประกอบกิจการ และในทางกลับกันสถานที่ใดๆย่อมจะต้องมีบุคคลเข้าไปเกี่ยวข้อง ความคาบเกี่ยวกันทางความหมายที่ใช้ในการอ้างอิงจึงเกิดขึ้น ซึ่งกำชัย (2525 : 266) อธิบายไว้ว่า ปรากฏการณ์ที่พบนี้ เป็นลักษณะของการสมมุติให้เป็นบุคคล โดยมีการเปลี่ยนความหมายจากการมีความหมายถึงสถานที่หรือองค์การตามรูปศัพท์ให้เป็นความหมายที่อ้างถึงกลุ่มคนผู้มีหน้าที่รับผิดชอบในสถานที่หรือองค์การนั้นๆ ดังนั้นในการประมวลผลเกี่ยวกับชื่อเฉพาะที่มีการพิจารณาเกี่ยวกับประเภทของชื่อเฉพาะควรคำนึงถึงแง่มุมในประเด็นดังกล่าวนี้

บริบทที่พบในการศึกษานี้เป็นบริบทที่ผู้วิจัยเห็นว่า สามารถนำไปประยุกต์ใช้กับงานทางด้านภาษาศาสตร์คอมพิวเตอร์ได้จริง เนื่องจากเป็นบริบทที่พบกับชื่อเฉพาะจำนวนมาก และมีรูปแบบที่ชัดเจน เช่น ในกรณีของบริบทที่เป็นคำบ่งบอก รูปแบบการปรากฏโดยปกติของคำบ่งบอกจะปรากฏอยู่ที่ตำแหน่งส่วนหน้าติดกับชื่อเฉพาะ รวมทั้งรูปภาษาที่ปรากฏเป็นรายการคำบ่งบอกก็ปรากฏด้วยจำนวนที่จำกัดอยู่ในปริมาณหนึ่งที่น่าจะเป็นจำนวนน้อย ข้อดีเหล่านี้เป็น

คุณสมบัติที่เหมาะสมสอดคล้องกับการนำไปใช้ในการประมวลผลภาษาด้วยเครื่อง นอกจากนี้บริบทที่เป็นคำบ่งบอกและการเว้นวรรคยังเป็นบริบทที่เราสามารถประเมินความถูกต้องแม่นยำในการใช้บ่งบอกถึงชื่อเฉพาะ การป้องกันความรู้ดังกล่าวนี้ให้กับคอมพิวเตอร์ ทำให้โปรแกรมคอมพิวเตอร์มีความสามารถในการประมวลผลชื่อเฉพาะในเชิงคุณภาพมากขึ้น ส่วนลักษณะทางภาษาประเด็นต่างๆที่กล่าวมาแสดงให้เห็นข้อสรุปทางภาษาบางประการที่น่าจะนำไปศึกษาเพิ่มเติมเพื่อหาบริบทบ่งบอกในลักษณะที่สามารถนำไปประยุกต์ใช้กับงานประมวลผลภาษาธรรมชาติ (จะกล่าวเสนอแนะในส่วนถัดไป) ดังนั้นในการประมวลผลชื่อเฉพาะภาษาไทยในการพัฒนางานทางด้านภาษาศาสตร์คอมพิวเตอร์โดยใช้แนวทางที่อาศัยบริบทบ่งบอกเป็นความรู้หรือข้อสรุปทางภาษาในการรู้จำชื่อเฉพาะ จึงน่าจะเป็นแนวทางที่มีแนวโน้มที่ดีสำหรับการแก้ปัญหาความกำกวมของการประมวลผลชื่อเฉพาะ เพราะนอกจากจะได้รูปแบบที่ชัดเจนสามารถประยุกต์ใช้เป็นแบบจำลองให้กับโปรแกรมคอมพิวเตอร์ได้แล้ว ผู้วิจัยยังเห็นว่าสามารถนำไปใช้ได้โดยไม่ต้องขึ้นอยู่กับข้อจำกัดจากปัจจัยต่างๆมากนัก เช่น ปัจจัยจำกัดจากร่องการกำหนดขอบเขตของรูปคำที่จะนำมาใช้เป็นบริบทว่าควรกำหนดขอบเขตเท่าใดจึงจะเหมาะสม

6.2 ข้อเสนอแนะ

1. ในการนำบริบทบ่งบอกต่างๆไปใช้บ่งบอกการปรากฏของชื่อเฉพาะนั้น ผู้วิจัยเห็นว่าควรใช้ในลักษณะที่นำบริบทต่างๆมาประกอบกัน คือ บริบทที่เป็นคำบ่งบอกและการเว้นวรรค ควรนำมาพิจารณาร่วมกันในการนำไปใช้บ่งบอกถึงชื่อเฉพาะ ดังเช่นถ้าทำการระบุแยกแยะชื่อเฉพาะไปพร้อมกับกระบวนการตัดคำ เราสามารถป้องกันความรู้รายการคำบ่งบอก, ตำแหน่งการปรากฏของคำบ่งบอกและการเว้นวรรค รวมถึงขีดความสามารถของบริบทให้โปรแกรมตัดคำเรียนรู้แล้วให้โปรแกรมตัดคำนำความรู้ดังกล่าวไปใช้สำหรับการตรวจสอบเมื่อพบคำบ่งบอก คำบ่งบอกนั้นปรากฏการเว้นวรรคในลักษณะที่พบกับชื่อเฉพาะหรือไม่ และถ้าตรวจสอบพบให้ทำการประเมินว่ารูปภาษานั้นน่าจะเป็นชื่อเฉพาะหรือไม่เป็นชื่อเฉพาะ โดยอาศัยความรู้เกี่ยวกับขีดความสามารถที่โปรแกรมเรียนรู้มาแล้ว นอกจากนี้แล้วการปรากฏการเว้นวรรคที่ส่วนหลังของชื่อเฉพาะอาจจะใช้เป็นบริบทที่บ่งบอกขอบเขตของชื่อเฉพาะได้ ถ้าโปรแกรมตัดคำพบสายอักขระที่ไม่รู้จักและไม่พบคำบ่งบอก ให้โปรแกรมทำการพิจารณบริบทบ่งบอกตัวอื่นคือตัวบ่งบอกระดับบริจเฉท โดยการนำไปเทียบกับชื่อเฉพาะที่ปรากฏกับคำบ่งบอกซึ่งโปรแกรมรู้จำไว้แล้ว ว่าสายอักขระที่อยู่หลังคำบ่งบอกเป็นรูปภาษาเดียวกันหรือไม่ ในกรณีของตัวบ่งบอกระดับบริจเฉทนี้อาจใช้วิธีการที่ให้โปรแกรมทำการวิเคราะห์หาชื่อเฉพาะรูปเดียวกันที่ปรากฏร่วมกับคำบ่งบอกและไม่ปรากฏคำบ่งบอก แล้วทำการแปลงชื่อเฉพาะรูปเดียวกันที่ไม่

- ปรากฏคำบ่งบอกให้เป็นรูปที่มีคำบ่งบอก อย่างไรก็ตาม แนวทางที่นำเสนอนี้เป็นเพียงแนวทางกว้างๆที่อาจเป็นไปได้แนวทางหนึ่งและประเด็นที่เกี่ยวข้องกับบริบทที่เป็นคำบ่งบอกยังสามารถศึกษาวิเคราะห์เพิ่มเติมได้อีก เนื่องจากวิทยานิพนธ์ฉบับนี้ได้ศึกษาชื่อเฉพาะที่อยู่ในขอบข่ายของเกณฑ์นิยาม ซึ่งเป็นเกณฑ์ที่คัดกรองแต่ชื่อเฉพาะต้นแบบออกมาเท่านั้น
2. ปรากฏการณ์ทางภาษาที่พบในชื่อเฉพาะกลุ่มที่ไม่ปรากฏบริบทบ่งบอกประการหนึ่งที่ปรากฏให้เห็นอย่างชัดเจนคือ ลักษณะของรูปภาษาที่มีโครงสร้างแบบทับศัพท์ภาษาต่างประเทศ ซึ่งเป็นรูปแบบที่มีลักษณะพิเศษแตกต่างจากรูปศัพท์ในภาษาไทย ทำให้ชื่อเฉพาะเหล่านี้สามารถปรากฏโดยไม่ต้องมีตัวบ่งบอกได้ รูปแบบของการทับศัพท์ภาษาต่างประเทศนี้จึงน่าจะเป็นลักษณะทางภาษาที่สามารถบ่งบอกการปรากฏของชื่อเฉพาะภาษาไทยได้ ควรมีการศึกษาเพิ่มเติมต่อไปว่าจะสามารถเป็นลักษณะบ่งบอกถึงการปรากฏของชื่อเฉพาะได้ดีเพียงใด
 3. จากการศึกษาที่พบว่าชื่อเฉพาะประเภทหนึ่งจะปรากฏกับคำบ่งบอกเฉพาะกลุ่มและโดยปกติคำบ่งบอกกลุ่มที่เฉพาะเจาะจงนี้จะไม่ปรากฏกับชื่อเฉพาะประเภทอื่นๆ เป็นการแสดงให้เห็นว่าชื่อเฉพาะภาษาไทยน่าจะเป็นหน่วยภาษาที่มีลักษณะของการเลือกที่จะเกิดร่วมอยู่กับบริบทหรือเหตุการณ์ทางภาษาเฉพาะอย่าง ซึ่งบริบทหรือเหตุการณ์ทางภาษาเฉพาะอย่างนี้น่าจะเป็นบริบทที่สามารถเป็นบริบทบ่งบอกเพิ่มเติมนอกเหนือจากที่ได้กล่าวมา และจากการสำรวจสังเกตเห็นว่า ชื่อเฉพาะในคลังข้อมูลของวิทยานิพนธ์มีการปรากฏคำบุพบทหรือคำเชื่อมอนุภาคบางคำที่มีความหมายในเชิงที่สัมพันธ์กับสิ่งที่ชื่อเฉพาะอ้างอิงถึง เช่น คำว่า “ของ” “โดย” “กับ” มักจะปรากฏกับชื่อเฉพาะประเภทบุคคล และองค์กร หรือคำว่า “ที่” “ใน” “ณ” มักปรากฏกับชื่อเฉพาะประเภทสถานที่ นอกจากนี้ชื่อเฉพาะประเภทบุคคล หรือ องค์กร จะปรากฏให้เห็นคำกริยาบางคำซ้ำๆกัน เช่น “กล่าวว่” “เปิดเผย” บริบททางภาษาดังกล่าวนี้นี้จึงเป็นประเด็นที่น่านำมาศึกษาเพิ่มเติมเพื่อนำมาประกอบกับบริบทที่พบในการศึกษาครั้งนี้ เนื่องจากผู้วิจัยเห็นว่าชื่อเฉพาะภาษาไทยเป็นหน่วยภาษาที่มีความกำกวมสูงมาก การใช้บริบทบ่งบอกมาเป็นกลวิธีในการแก้ปัญหาความกำกวม ควรใช้บริบทบ่งบอกหลายๆบริบทมาประกอบกันและนำบริบทบ่งบอกไปใช้ในลักษณะส่งต่อ กล่าวคือเมื่อบริบทบ่งบอกหนึ่งไม่สามารถรู้จำชื่อเฉพาะรูปหนึ่งหรือรู้จำแต่ยังไม่ถูกต้อง ก็ให้มีการพิจารณาบริบทบ่งบอกอื่นให้เข้ามารับช่วงต่อในการประมวลผลชื่อเฉพาะรูปนั้นต่อไป ทำให้สามารถประมวลผลชื่อเฉพาะภาษาไทยได้อย่างมีประสิทธิภาพและมีความถูกต้องสมบูรณ์มากที่สุด

4. จากการศึกษาที่พบว่า ชื่อเฉพาะสามารถเป็นส่วนประกอบของคำอื่นนั้น อาจก่อให้เกิดปัญหาต่อระบบการรู้จำชื่อเฉพาะด้วยคอมพิวเตอร์ได้ โดยเฉพาะระบบที่อาศัยรายการชื่อเฉพาะในการรู้จำ กล่าวคือ วิธีการเทียบหาชื่อเฉพาะโดยเปรียบเทียบจากรายการชื่อเฉพาะที่มีอยู่ ไม่อาจใช้ได้เสมอไป จึงเป็นสิ่งที่ต้องคำนึงสำหรับนักภาษาศาสตร์คอมพิวเตอร์ที่ต้องการพัฒนาระบบการรู้จำชื่อเฉพาะภาษาไทย

รายการอ้างอิง

ภาษาไทย

- กำชัย ทองหล่อ. 2525. หลักภาษาไทย. พิมพ์ครั้งที่ 6. กรุงเทพมหานคร: บำรุงสาส์น.
- เกียรติชัย เอี่ยมวรเมธ. 2544. พจนานุกรมไทย. กรุงเทพมหานคร: รวมสาส์น.
- นworรรณ พันธุเมธา. 2527. ไวยากรณ์ไทย. กรุงเทพมหานคร: รุ่งเรืองสาส์น.
- นworรรณ พันธุเมธา. 2544. คลังคำ. กรุงเทพมหานคร: อมรินทร์.
- นัฐวุฒิ ไชยเจริญ. 2544. การตัดคำและการกำกับหมวดคำภาษาไทยแบบเบ็ดเสร็จด้วย คอมพิวเตอร์. วิทยานิพนธ์ปริญญาามหาบัณฑิต ภาควิชาภาษาศาสตร์ คณะอักษรศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย.
- เปรมจิต ชนะวงศ์. 2538. หลักภาษาไทย. พิมพ์ครั้งที่ 8. โครงการตำราและเอกสารวิชาการ สถาบันราชภัฏนครศรีธรรมราช.
- พรพิมล ธิรคุณโกวิท และ วิโรจน์ ธิรคุณโกวิท. 2527. พจนานุกรมอักษรย่อ. กรุงเทพมหานคร: แพร่พิทยา.
- มานิต มานิตเจริญ. 2538. พจนานุกรมไทย. พิมพ์ครั้งที่ 15 (ปรับปรุง-เพิ่มเติม). กรุงเทพมหานคร: รวมสาส์น.
- ราชบัณฑิตยสถานและศูนย์เทคโนโลยีอิเล็กทรอนิกส์และคอมพิวเตอร์แห่งชาติ. 2539. พจนานุกรมฉบับราชบัณฑิตยสถาน พ.ศ. 2525 [Computer software]. แหล่งที่มา : <http://www.royin.go.th> [1 พฤศจิกายน 2545]
- วิรัช ศรีเลิศล้ำวาณิช. 2536. การตัดคำในระบบแปลภาษา (Word Segmentation for Thai in Machine Translation System). การแปลภาษาด้วยคอมพิวเตอร์. NECTEC. หน้า 50-55. อ้างถึงใน นัฐวุฒิ ไชยเจริญ. 2544. การตัดคำและการกำกับหมวดคำภาษาไทยแบบเบ็ดเสร็จด้วยคอมพิวเตอร์. วิทยานิพนธ์ปริญญาามหาบัณฑิต ภาควิชาภาษาศาสตร์ คณะอักษรศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย.
- วิโรจน์ อรุณมานะกุล. 2545. ภาษาศาสตร์คลังข้อมูล. กรุงเทพมหานคร: สำนักพิมพ์จุฬาลงกรณ์มหาวิทยาลัย.
- สำนักพจนานุกรมมติชน. 2543. พจนานุกรมนอกราชบัณฑิตยฯ. กรุงเทพมหานคร: มติชน.

สุริยา รัตนกุล. 2544. อรรถศาสตร์เบื้องต้น. นครปฐม: มหาวิทยาลัยมหิดล.

อุปกิตติลปसार, พระยา. 2541. หลักภาษาไทย. พิมพ์ครั้งที่ 4. กรุงเทพมหานคร: ไทยวัฒนาพานิช.

ภาษาอังกฤษ

- Aroonmanakun, W. 1999-2001. CU Thai Concordance [online]. Available from: <http://www.arts.chula.ac.th/~ling>
- Aroonmanakun, W. Dehtml [Computer software]. Available from: <http://www.arts.chula.ac.th/~ling>
- Bowman, R. M., Jr. 1998. Sharp's rule and proper names (chapter II) [online]. rev.ed. Available from: <http://www.atlantaapologist.org/Sharp.html>
- Charoenpornasawat, P., Kijisirikul, B., and Meknavin, S. 1998. Feature-based proper name identification in Thai. In Proceedings of National Computer Science and Engineering Conference: NCSEC'98.
- Chen, H., Huang, S., Ding, Y., and Tsai, S. 1998. Proper name translation in cross-language information retrieval. In Proceedings of 17th International Conference on Computational Linguistics and 36 th Annual Meeting of the Association for Computational Linguistics, pp.232-236. Canada, August 10-14, 1998.
- Fromkin, V., and Rodman, R. 1998. An introduction to language. 6th ed. Harcourt Brace College.
- Kanlayanawat, W., and Prasitjutrakul, S. 1997. Automatic indexing for Thai text with unknown words using trie structure. In Proceedings of the Natural Language Processing Pacific Rim Symposium (NLPRS) 1998, pp115-120. Phuket Thailand, December 2-4, 1997.
- Kawtrakul, A., Thumkanon, C., Poovorawan, Y., Barasrai, P., and Suktarachan, M. 1997. Automatic Thai Unknown Word Recognition. Proceedings of the Natural Language Processing Pacific Rim Symposium (NLPRS) 1997.
- Kawtrakul, A., Suktarachan, M., Varasai, P., and Chanlekha, H. 2002. A state of the art of Thai language resources and Thai language behavior analysis and modeling [online]. Available from: naist.cpe.ku.ac.th/paper/paper%202002/coling2002.
- Lehrer, A. 1994. Proper Names: Linguistic Aspects. The encyclopaedia of language and linguistics 6: 3372-3374.

- Lin, D., Le, S., Yibo, Z., and Yufang, S. 1999. The Chinese-English CLIR based on comparable corpora [online]. Available from: <http://korterm.kaist.ac.kr/nlprs99/mal99-papers/mal-021.doc>. [2002, September 22]
- McArthur, T. 1992. The Oxford companion to the English language [online]. Available from: <http://www.xrefer.com> (the web's reference search engine)
- McDonald, D. D. 1996. Internal and External Evidence in the Identification and Semantic Categorization of Proper Names. In B. Boguraev and J. Pustejovsky (eds.), Corpus processing for lexical acquisition, pp.21-39. London: The MIT Press.
- Meknavin, S., Charoenpornswat, P., and Kijirikul, B. 1997. Feature-based Thai word segmentation. Proceedings of the Natural Language Processing Pacific Rim Symposium (NLPRS) 1997. อ้างถึงใน นัฐวุฒิ ไชยเจริญ. 2544. การตัดคำและการกำกับหมวดคำภาษาไทยแบบเบ็ดเสร็จด้วยคอมพิวเตอร์. วิทยานิพนธ์ปริญญา มหาบัณฑิต ภาควิชาภาษาศาสตร์ คณะอักษรศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย.
- Roche, X., and other contributors. 1998-2002. WinHTTrack Website Copier [Computer software]. Free Software Foundation. Available from: <http://www.httrack.com>
- Sornlertlamvanich, V., Charoenporn, T., and Isahara, H. 1997. ORCHID: Thai part-of-speech tagged corpus. In Technical report Orchid corpus. Bangkok: NECTEC: pp 5-19.
- Sperberg-McQueen, C.M., and Burnard, L. (eds). 1990-1993. TEI Guidelines for electronic text encoding and interchange (P3) [online]. Available from: <http://www.tei-c.org>
- Sukhahuta, R., and Smith, D. 2001. Information Extraction Strategies for Thai Documents. Journal of Computer Processing of Oriental Languages (IJCPOL) 14 (No.2): 153-172 [Online]. Available from: <http://www.math.utah.edu/ftp/pub/tex/bib/toc/ijcpol.html>
- Vaida, C. J. 1996. Liberating grammar. International Edition. Singapore: Prentice Hall, Simon & Shuster (Asia).

ภาคผนวก

ภาคผนวก ก

รายการแท็กสำหรับการกำกับเค้าโครงของข้อมูล

| <ชื่อแท็ก> | คำอธิบาย | คุณสมบัติ |
|--|--|-------------------|
| <CORPUS number> | ระบุคลังข้อมูลย่อยซึ่งมีแหล่งที่มาต่างกัน อันได้แก่ <CORPUS 1> สำหรับหนังสือพิมพ์กรุงเทพธุรกิจ <CORPUS 2> สำหรับนิตยสารสกุลไทย | size |
| <corpusHeader> <eAddress> <distributor> <pubDate> | ระบุรายละเอียดต่างๆเกี่ยวกับแหล่งที่มาของข้อมูล ประกอบด้วยแท็ก <eAddress> <distributor> และ <pubDate> เลขที่อยู่ของเว็บ (web address) ของแหล่งข้อมูล เจ้าของข้อมูล วัน เดือน ปี | type type - |
| <text number> | ระบุแต่ละข้อความต่อเนื่อง (text) หรือบทความ คอลัมน์ ฯลฯ ที่ปรากฏในคลังข้อมูลย่อย | - |
| <head> | หัวข้อเรื่องของบทความ | - |
| <byline> | ผู้เขียนบทความ | - |
| <div> | เรื่องย่อยในบทความ | n |
| <divhead> | หัวข้อเรื่องของเรื่องย่อย | - |
| <shead> | หัวข้อย่อยในบทความ | - |

ภาคผนวก ข

**รายการค่าขีดความสามารถในการบ่งบอก
ของค่าบ่งบอกกับการเว้นวรรคที่ปรากฏกับชื่อเฉพาะร่วมกัน**

คำอธิบายสัญลักษณ์

PN หมายถึง จำนวนชื่อเฉพาะที่ปรากฏค่าบ่งบอกร่วมกับการเว้นวรรค เช่น จำนวนชื่อเฉพาะที่มีค่าบ่งบอก "นาย" และพบการเว้นวรรคที่ส่วนหน้าของชื่อเฉพาะ

All หมายถึง จำนวนรูปภาพทั้งหมดทั้งที่เป็นชื่อเฉพาะและไม่ใช้ชื่อเฉพาะ ซึ่งปรากฏค่าบ่งบอกร่วมกับการเว้นวรรค ตัวอย่างเช่น จำนวนรูปภาพ "นาย" และปรากฏการเว้นวรรคที่หน้ารูปภาพนั้น

Percentage หมายถึง ขีดความสามารถในการบ่งบอกของค่าบ่งบอกที่ปรากฏร่วมกับการเว้นวรรค

| ค่าบ่งบอก | วรรค - ชื่อเฉพาะ | | | ชื่อเฉพาะ - วรรค | | | วรรค - ชื่อเฉพาะ - วรรค | | |
|------------------------|--|-----|------------|--|-----|------------|---|-----|------------|
| | "...พร้อมกันนี้ นายอิวจ๊กกล่าวอีกว่า..." | | | "...อาจเหมือนเปรียบกับนายชวน หลีกภัย ก็ได้..." | | | "...ขณะที่ นายกะวะ ฟางหญู เจ้าหน้าที่..." | | |
| | PN | All | Percentage | PN | All | Percentage | PN | All | Percentage |
| เกาะ | 0 | 1 | 0% | 5 | 11 | 45% | 0 | 0 | - |
| เขต | 0 | 2 | 0% | 3 | 20 | 15% | 0 | 0 | - |
| เขตรักษาพันธุ์สัตว์ป่า | 0 | 0 | - | 0 | 0 | - | 0 | 0 | - |
| เขตห้ามล่าสัตว์ป่า | 0 | 0 | - | 0 | 0 | - | 0 | 0 | - |
| เขตฯ | 0 | 0 | - | 0 | 0 | - | 0 | 0 | - |
| เขา | 0 | 72 | 0% | 0 | 125 | 0% | 0 | 0 | - |
| เขื่อน | 2 | 2 | 100% | 8 | 14 | 57% | 2 | 2 | 100% |
| เครือข่าย | 2 | 5 | 40% | 2 | 19 | 11% | 0 | 0 | - |
| เครือข่าย | 0 | 2 | 0% | 0 | 6 | 0% | 0 | 0 | - |
| เจ้า | 0 | 20 | 0% | 0 | 101 | 0% | 0 | 0 | - |
| เจ้าชาย | 0 | 0 | - | 0 | 0 | - | 0 | 0 | - |
| เจ้าหญิง | 0 | 1 | 0% | 0 | 0 | - | 0 | 0 | - |
| เทศบาล | 0 | 0 | - | 0 | 0 | - | 0 | 0 | - |
| ผ้า | 0 | 0 | - | 0 | 0 | - | 0 | 0 | - |
| เมือง | 1 | 2 | 50% | 15 | 71 | 21% | 1 | 1 | 100% |

| คำบ่งบอก | วรรค - ชื่อเฉพาะ | | | ชื่อเฉพาะ - วรรค | | | วรรค - ชื่อเฉพาะ - วรรค | | |
|------------|---|-----|------------|--|-----|------------|--|-----|------------|
| | "...พร้อมกันนี้ นายอิทธิพลกล่าวอีกว่า..." | | | "...อาจเหมือนเปรียบกับนายชวน หลีกภัย ก็ได้..." | | | "...ขณะที่ นายกวี ฟ้าหยู เจ้าหน้าที่..." | | |
| | PN | All | Percentage | PN | All | Percentage | PN | All | Percentage |
| เสีย | 0 | 0 | - | 0 | 22 | 0% | 0 | 0 | - |
| แถบ | 0 | 1 | 0% | 0 | 2 | 0% | 0 | 0 | - |
| แบงก์ | 0 | 1 | 0% | 1 | 2 | 50% | 0 | 0 | - |
| แผนก | 0 | 1 | 0% | 0 | 8 | 0% | 0 | 0 | - |
| แพทย์หญิง | 0 | 0 | - | 0 | 0 | - | 0 | 0 | - |
| แม่น้ำ | 0 | 0 | - | 2 | 2 | 100% | 0 | 0 | - |
| โค้ง | 0 | 0 | - | 0 | 0 | - | 0 | 0 | - |
| โฆษณา | 0 | 5 | 0% | 0 | 2 | 0% | 0 | 0 | - |
| โรงเรียน | 2 | 6 | 33% | 13 | 38 | 34% | 1 | 3 | 33% |
| โรงแรม | 1 | 1 | 100% | 2 | 2 | 100% | 1 | 2 | 50% |
| โรงงาน | 1 | 1 | 100% | 1 | 4 | 25% | 1 | 2 | 50% |
| โรงพยาบาล | 0 | 1 | 0% | 0 | 3 | 0% | 0 | 0 | - |
| กรม | 4 | 5 | 80% | 24 | 30 | 80% | 3 | 3 | 100% |
| กระทรวง | 7 | 8 | 88% | 25 | 29 | 86% | 4 | 4 | 100% |
| กรุง | 4 | 5 | 80% | 7 | 14 | 50% | 4 | 4 | 100% |
| กรุ๊ป | 0 | 1 | 0% | 0 | 0 | - | 0 | 0 | - |
| กลุ่ม | 2 | 6 | 33% | 2 | 75 | 3% | 0 | 3 | 0% |
| กอง | 6 | 10 | 60% | 3 | 32 | 9% | 2 | 2 | 100% |
| กองทุน | 3 | 3 | 100% | 11 | 28 | 39% | 0 | 0 | - |
| กิ่งอำเภอ | 0 | 0 | - | 1 | 1 | 100% | 0 | 0 | - |
| คณะกรรมการ | 0 | 6 | 0% | 2 | 41 | 5% | 0 | 2 | 0% |
| ครอบครัว | 0 | 0 | - | 0 | 15 | 0% | 0 | 0 | - |
| ครู | 0 | 3 | 0% | 0 | 11 | 0% | 0 | 0 | - |
| ค่าย | 0 | 1 | 0% | 0 | 10 | 0% | 0 | 0 | - |
| ควาบอย | 0 | 0 | - | 0 | 0 | - | 0 | 0 | - |
| คุณ | 8 | 23 | 35% | 3 | 62 | 5% | 5 | 5 | 100% |
| คุณพี่ | 1 | 2 | 50% | 0 | 2 | 0% | 0 | 0 | - |

| คำบ่งบอก | วรรค - ชื่อเฉพาะ | | | ชื่อเฉพาะ - วรรค | | | วรรค - ชื่อเฉพาะ - วรรค | | |
|----------|--|-----|------------|--|-----|------------|---|-----|------------|
| | "...พร้อมกันนี้ นายอิฐจึกสาวอีกว่า..." | | | "...อาจเหมือนเปรียบกับนายชวน หลีกภัย ก็ได้..." | | | "...ขณะที่ นายก๊วย ฟางหยู เจ้าหน้าที่..." | | |
| | PN | All | Percentage | PN | All | Percentage | PN | All | Percentage |
| จ. | 22 | 23 | 96% | 20 | 22 | 91% | 22 | 22 | 100% |
| จ.ส.ต. | 2 | 2 | 100% | 0 | 0 | - | 2 | 2 | 100% |
| จอมพล | 0 | 0 | - | 0 | 0 | - | 0 | 0 | - |
| จังหวัด | 10 | 12 | 83% | 27 | 41 | 66% | 6 | 10 | 60% |
| ช. | 0 | 0 | - | 0 | 0 | - | 0 | 1 | 0% |
| ชอย | 0 | 0 | - | 1 | 1 | 100% | 0 | 0 | - |
| ดร. | 14 | 14 | 100% | 8 | 8 | 100% | 7 | 7 | 100% |
| ดีเจ. | 0 | 0 | - | 0 | 0 | - | 0 | 0 | - |
| ต. | 1 | 11 | 9% | 1 | 22 | 5% | 1 | 1 | 100% |
| ตระกูล | 0 | 0 | - | 0 | 1 | 0% | 0 | 0 | - |
| ตลาด | 1 | 7 | 14% | 1 | 98 | 1% | 1 | 1 | 100% |
| ตำบล | 0 | 0 | - | 2 | 7 | 29% | 0 | 0 | - |
| ถ. | 1 | 1 | 100% | 1 | 1 | 100% | 0 | 0 | - |
| ถนน | 0 | 1 | 0% | 1 | 14 | 7% | 0 | 0 | - |
| ถ้ำ | 0 | 0 | - | 0 | 0 | - | 0 | 0 | - |
| ทวีป | 0 | 0 | - | 2 | 2 | 100% | 0 | 0 | - |
| ทะเล | 0 | 0 | - | 0 | 8 | 0% | 0 | 0 | - |
| ทะเลสาบ | 0 | 0 | - | 0 | 0 | - | 0 | 0 | - |
| ทิวเขา | 0 | 0 | - | 0 | 0 | - | 0 | 0 | - |
| ธนาคาร | 5 | 17 | 29% | 13 | 49 | 27% | 4 | 4 | 100% |
| น.ส. | 2 | 3 | 67% | 0 | 0 | - | 2 | 2 | 100% |
| นคร | 0 | 1 | 0% | 0 | 27 | 0% | 0 | 0 | - |
| น้อง | 0 | 0 | - | 4 | 15 | 27% | 0 | 0 | - |
| นาง | 10 | 10 | 100% | 13 | 36 | 36% | 8 | 8 | 100% |
| นางสาว | 1 | 1 | 100% | 0 | 0 | - | 1 | 1 | 100% |
| นาย | 108 | 133 | 81% | 100 | 160 | 63% | 92 | 93 | 99% |
| นายแพทย์ | 1 | 1 | 100% | 1 | 1 | 100% | 1 | 1 | 100% |

| คำบ่งบอก | วรรค – ชื่อเฉพาะ | | | ชื่อเฉพาะ - วรรค | | | วรรค – ชื่อเฉพาะ - วรรค | | |
|----------------|--|-----|------------|--|-----|------------|--|-----|------------|
| | "...พร้อมกันนี้ นายอิวจ๊กกล่าวอีกว่า..." | | | "...อาจเหมือนเปรียบกับนายชวน หลีกภัย ก็ได้..." | | | "...ขณะที่ นายกะวะ ฟางหญู เจ้าหน้าที..." | | |
| | PN | All | Percentage | PN | All | Percentage | PN | All | Percentage |
| นายกรัฐมนตรี | 0 | 4 | 0% | 0 | 2 | 0% | 0 | 11 | 0% |
| นายฯ | 1 | 3 | 33% | 1 | 3 | 33% | 1 | 2 | 50% |
| บริษัท | 25 | 46 | 54% | 18 | 74 | 24% | 22 | 22 | 100% |
| บริษัท...จำกัด | 0 | 0 | - | 0 | 0 | - | 0 | 0 | - |
| บล. | 4 | 4 | 100% | 7 | 7 | 100% | 3 | 3 | 100% |
| บ้าน | 0 | 2 | 0% | 4 | 107 | 4% | 0 | 0 | - |
| บ๊ิก | 0 | 0 | - | 0 | 2 | 0% | 0 | 0 | - |
| ประเทศ | 12 | 23 | 52% | 36 | 161 | 22% | 7 | 10 | 70% |
| ประธาน | 0 | 0 | - | 0 | 0 | - | 0 | 0 | - |
| ประธานาธิบดี | 0 | 0 | - | 0 | 0 | - | 0 | 0 | - |
| ป่า | 0 | 3 | 0% | 0 | 28 | 0% | 0 | 1 | 0% |
| ป่า | 0 | 0 | - | 0 | 1 | 0% | 0 | 0 | - |
| ปากน้ำ | 0 | 0 | - | 0 | 0 | - | 0 | 0 | - |
| ผศ. | 2 | 2 | 100% | 1 | 1 | 100% | 1 | 1 | 100% |
| ผืนป่า | 0 | 0 | - | 0 | 5 | 0% | 0 | 0 | - |
| ผู้กำกับ | 1 | 5 | 20% | 0 | 5 | 0% | 0 | 0 | - |
| ผู้พัน | 0 | 0 | - | 0 | 1 | 0% | 0 | 0 | - |
| ฝั่ง | 0 | 0 | - | 0 | 0 | - | 0 | 0 | - |
| ฝ่าย | 0 | 11 | 0% | 1 | 34 | 3% | 0 | 2 | 0% |
| พ.ต.ท. | 18 | 18 | 100% | 13 | 13 | 100% | 17 | 17 | 100% |
| พ.ต.อ. | 1 | 1 | 100% | 0 | 0 | - | 1 | 1 | 100% |
| พ.ร.บ. | 0 | 7 | 0% | 0 | 5 | 0% | 0 | 0 | - |
| พรรค | 2 | 9 | 22% | 10 | 24 | 42% | 0 | 1 | 0% |
| พระเจ้า | 3 | 5 | 60% | 8 | 23 | 35% | 2 | 2 | 100% |
| พระราชวัง | 0 | 0 | - | 0 | 2 | 0% | 0 | 0 | - |
| พล.ต. | 5 | 5 | 100% | 1 | 1 | 100% | 5 | 5 | 100% |
| พล.ต.ต. | 0 | 0 | - | 0 | 0 | - | 0 | 0 | - |

| คำบ่งบอก | วรรค - ชื่อเฉพาะ | | | ชื่อเฉพาะ - วรรค | | | วรรค - ชื่อเฉพาะ - วรรค | | |
|-----------------|--|-----|------------|--|-----|------------|---|-----|------------|
| | "...พร้อมกันนี้ นายอิฐจ๊กกล่าวอีกว่า..." | | | "...อาจเหมือนเปรียบกับนายชวน หลีกภัย ก็ได้..." | | | "...ขณะที่ นายก๊วย ฟางหยู เจ้าหน้าที่..." | | |
| | PN | All | Percentage | PN | All | Percentage | PN | All | Percentage |
| พล.ต.อ. | 0 | 0 | - | 0 | 0 | - | 0 | 0 | - |
| พล.อ. | 3 | 3 | 100% | 1 | 1 | 100% | 3 | 3 | 100% |
| พล.อ.อ. | 1 | 1 | 100% | 0 | 0 | - | 1 | 1 | 100% |
| พี | 6 | 9 | 67% | 10 | 29 | 34% | 1 | 1 | 100% |
| ภาษา | 0 | 0 | - | 0 | 11 | 0% | 0 | 0 | - |
| ภูมิภาค | 0 | 0 | - | 0 | 8 | 0% | 0 | 1 | 0% |
| ม. | 4 | 8 | 50% | 5 | 24 | 21% | 4 | 4 | 100% |
| มหาวิทยาลัย | 10 | 13 | 77% | 13 | 29 | 45% | 8 | 8 | 100% |
| มูลนิธิ | 0 | 0 | - | 9 | 12 | 75% | 0 | 0 | - |
| ย่าน | 1 | 1 | 100% | 3 | 3 | 100% | 1 | 1 | 100% |
| ร.ต.ต. | 0 | 0 | - | 0 | 0 | - | 0 | 0 | - |
| ร.ต.อ. | 2 | 2 | 100% | 2 | 2 | 100% | 2 | 2 | 100% |
| ร.พ. | 1 | 1 | 100% | 1 | 1 | 100% | 1 | 1 | 100% |
| ร.ร. | 0 | 1 | 0% | 0 | 1 | 0% | 0 | 0 | - |
| รศ. | 3 | 3 | 100% | 2 | 2 | 100% | 3 | 3 | 100% |
| รองประธานาธิบดี | 0 | 0 | - | 0 | 0 | - | 0 | 0 | - |
| รัฐ | 2 | 30 | 7% | 2 | 183 | 1% | 2 | 3 | 67% |
| ราชวงศ์ | 0 | 0 | - | 0 | 0 | - | 0 | 0 | - |
| ราชสำนัก | 0 | 0 | - | 0 | 0 | - | 0 | 0 | - |
| รายการ | 0 | 6 | 0% | 0 | 14 | 0% | 0 | 5 | 0% |
| ลุ่มน้ำ | 0 | 0 | - | 0 | 0 | - | 0 | 0 | - |
| ลูก | 0 | 11 | 0% | 0 | 69 | 0% | 0 | 0 | - |
| วง | 0 | 6 | 0% | 0 | 151 | 0% | 0 | 0 | - |
| วัง | 0 | 3 | 0% | 0 | 31 | 0% | 0 | 0 | - |
| วัด | 1 | 1 | 100% | 1 | 22 | 5% | 1 | 1 | 100% |
| วิทยาลัย | 0 | 1 | 0% | 0 | 19 | 0% | 0 | 0 | - |
| วุฒิสมาชิก | 0 | 0 | - | 0 | 1 | 0% | 0 | 0 | - |

| คำบ่งบอก | วรรค – ชื่อเฉพาะ | | | ชื่อเฉพาะ - วรรค | | | วรรค – ชื่อเฉพาะ - วรรค | | |
|--------------|--|-----|------------|--|-----|------------|--|-----|------------|
| | "...พร้อมกันนี้ นายอิฐจึกกล่าวอีกว่า..." | | | "...อาจเหมือนเปรียบกับนายชวน หลีกภัย ก็ได้..." | | | "...ขณะที่ นายกวีฯ ฟางหยุด เจ้าหน้าที่..." | | |
| | PN | All | Percentage | PN | All | Percentage | PN | All | Percentage |
| ศ. | 1 | 1 | 100% | 0 | 2 | 0% | 1 | 1 | 100% |
| ศาล | 1 | 1 | 100% | 7 | 16 | 44% | 0 | 0 | - |
| ศูนย์ | 5 | 6 | 83% | 4 | 20 | 20% | 4 | 4 | 100% |
| ส.ต.ท. | 1 | 1 | 100% | 0 | 0 | - | 1 | 1 | 100% |
| สถาบัน | 1 | 4 | 25% | 2 | 21 | 10% | 1 | 1 | 100% |
| สถาบันราชภัฏ | 1 | 1 | 100% | 1 | 1 | 100% | 1 | 1 | 100% |
| สน. | 0 | 0 | - | 0 | 1 | 0% | 0 | 0 | - |
| สนธิสัญญา | 0 | 0 | - | 0 | 0 | - | 0 | 0 | - |
| สนาม | 0 | 1 | 0% | 2 | 8 | 25% | 0 | 0 | - |
| สนามกอล์ฟ | 0 | 0 | - | 2 | 2 | 100% | 0 | 1 | 0% |
| สนามบิน | 0 | 0 | - | 1 | 2 | 50% | 0 | 0 | - |
| สภา | 1 | 3 | 33% | 4 | 56 | 7% | 0 | 0 | - |
| สมรภูมิ | 0 | 0 | - | 0 | 0 | - | 0 | 0 | - |
| สมาคม | 4 | 5 | 80% | 9 | 12 | 75% | 4 | 4 | 100% |
| สวน | 0 | 1 | 0% | 1 | 21 | 5% | 0 | 0 | - |
| สหพันธ์ | 0 | 0 | - | 0 | 1 | 0% | 0 | 0 | - |
| สหภาพ | 0 | 0 | - | 0 | 1 | 0% | 0 | 0 | - |
| สะพาน | 0 | 0 | - | 0 | 1 | 0% | 0 | 0 | - |
| สังเวียน | 0 | 0 | - | 0 | 2 | 0% | 0 | 0 | - |
| สำนัก | 6 | 8 | 75% | 0 | 12 | 0% | 4 | 4 | 100% |
| สำนักข่าว | 1 | 1 | 100% | 6 | 7 | 86% | 1 | 1 | 100% |
| สำนักงาน | 5 | 5 | 100% | 5 | 7 | 71% | 3 | 3 | 100% |
| หน่วย | 0 | 1 | 0% | 0 | 22 | 0% | 0 | 0 | - |
| หนู | 0 | 0 | - | 0 | 0 | - | 0 | 0 | - |
| หม่อมเจ้า | 0 | 0 | - | 0 | 2 | 0% | 0 | 0 | - |
| หม่อมราชวงศ์ | 0 | 0 | - | 0 | 0 | - | 0 | 0 | - |
| หม่อมหลวง | 0 | 0 | - | 0 | 1 | 0% | 0 | 0 | - |

| คำบ่งบอก | วรรค – ชื่อเฉพาะ | | | ชื่อเฉพาะ - วรรค | | | วรรค – ชื่อเฉพาะ - วรรค | | |
|----------|--|-----|------------|--|-----|------------|---|-----|------------|
| | "...พร้อมกันนี้ นายอิฐจึกกล่าวอีกว่า..." | | | "...อาจเหมือนเปรียบกับนายชวน หลีกภัย ก็ได้..." | | | "...ขณะที่ นายกวีระ พงษ์พสุ ผู้อำนวยการ..." | | |
| | PN | All | Percentage | PN | All | Percentage | PN | All | Percentage |
| หมู่บ้าน | 0 | 1 | 0% | 0 | 4 | 0% | 0 | 0 | - |
| ห้อง | 0 | 6 | 0% | 0 | 40 | 0% | 0 | 1 | 0% |
| หาด | 0 | 0 | - | 0 | 7 | 0% | 0 | 0 | - |
| อ. | 6 | 10 | 60% | 5 | 10 | 50% | 5 | 5 | 100% |
| องค์กร | 0 | 1 | 0% | 0 | 11 | 0% | 0 | 3 | 0% |
| องค์การ | 8 | 8 | 100% | 3 | 7 | 43% | 6 | 6 | 100% |
| อาคาร | 0 | 2 | 0% | 0 | 13 | 0% | 0 | 1 | 0% |
| อาจารย์ | 6 | 13 | 46% | 3 | 16 | 19% | 5 | 5 | 100% |
| อ่าว | 0 | 0 | - | 3 | 3 | 100% | 0 | 0 | - |
| อำเภอ | 1 | 1 | 100% | 6 | 7 | 86% | 1 | 2 | 50% |
| อุทยาน | 0 | 0 | - | 0 | 0 | - | 0 | 0 | - |

ประวัติผู้เขียนวิทยานิพนธ์

นางสาวอมรทิพย์ กวินปณิธาน เกิดที่กรุงเทพมหานคร สำเร็จการศึกษาระดับปริญญาตรี วิทยาศาสตร์บัณฑิต มหาวิทยาลัยเชียงใหม่ ในปีการศึกษา 2536 ได้รับราชการอยู่ที่ศูนย์สิรินธรเพื่อการฟื้นฟูสมรรถภาพทางการแพทย์แห่งชาติ ก่อนเข้าศึกษาต่อในหลักสูตรอักษรศาสตรมหาบัณฑิต ภาควิชาภาษาศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย เมื่อปี พ.ศ. 2543