การแยกอนุพากย์ภาษาไทยด้วยการใช้แบบจำลองซัพพอร์ตเวกเตอร์แมชชีน

นางสาวนลินี อินต๊ะซาว

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาอักษรศาสตรมหาบัณฑิต
สาขาวิชาภาษาศาสตร์ ภาควิชาภาษาศาสตร์
คณะอักษรศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย
ปีการศึกษา 2556

Thai Clause Segmentation Using a Support Vector Machine Model

Miss Nalinee Intasaw

A Thesis Submitted in Partial Fulfillment of the Requirements
for the Degree of Master of Arts Program in Linguistics
Department of Linguistics
Faculty of Arts
Chulalongkorn University
Academic Year 2013

หัวข้อวิทยานิพนธ์	การแยกอนุพากย์ภาษาไทยด้วยการใช้แบบจำลองซัพพอร์ตเวกเตอร์แมชชีน
โดย	นางสาวนลินี อินต๊ะซาว
สาขาวิชา	ภาษาศาสตร์
อาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก	รองศาสตราจารย์ ดร. วิโรจน์ อรุณมานะกุล

คณะอักษรศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย อนุมัติให้นับวิทยานิพนธ์ฉบับนี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญามหาบัณฑิต

.....คณบดีคณะอักษรศาสตร์

(ผู้ช่วยศาสตราจารย์ ดร. ประพจน์ อัศววิรุฬหการ)

คณะกรรมการสอบวิทยานิพนธ์

.....ประธานกรรมการ

(รองศาสตราจารย์ ดร. สมชาย ประสิทธิ์จูตระกูล)

.....อาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก

(รองศาสตราจารย์ ดร. วิโรจน์ อรุณมานะกุล)

.....กรรมการภายนอกมหาวิทยาลัย

(ดร. เทพชัย ทรัพย์นิธิ)

นลินี อินต๊ะซาว : การแยกอนุพากย์ภาษาไทยด้วยการใช้แบบจำลองซัพพอร์ตเวกเตอร์แมชชีน. (Thai Clause Segmentation Using a Support Vector Machine Model) อ.ที่ปรึกษาวิทยานิพนธ์หลัก: รศ. ดร. วิโรจน์ อรุณมานะกุล, 98 หน้า.

วัตถุประสงค์ของวิทยานิพนธ์นี้ คือ เพื่อหาลักษณ์ทางภาษาศาสตร์ที่จะนำไปใช้ในการแยกอนุพากย์ภาษาไทยด้วยแบบจำลองซัพพอร์พเวกเตอร์แมชชีน และเปรียบเทียบลักษณ์ทางภาษาศาสตร์ที่ใช้ ว่าส่งผลต่อประสิทธิภาพของระบบการแยกอนุพากย์อย่างไรบ้าง

คลังข้อมูลที่ใช้ในการศึกษานี้เป็นภาษาเขียนทางวิชาการ มีขนาด 76,460 คำ ประกอบไปด้วย 8,102 อนุพากย์ แบบจำลองซัพพอร์ตเวกเตอร์แมชชีนที่ใช้ในการแยกอนุพากย์ในงานนี้คือฟังก์ชั่น SMO ของโปรแกรมวีก้า (Weka) และฟังก์ชั่นเคอร์เนลที่ใช้คือโพลีโนเมียล ระบบทำการแยกอนุพากย์โดยรับข้อมูลเข้าเป็นคำเพื่อให้แบบจำลองตัดสินใจว่าคำนั้นเป็นคำขอบเขตเริ่มต้นอนุพากย์หรือไม่ การตัดสินใจของแบบจำลองอาศัยลักษณ์ทางภาษาศาสตร์ ได้แก่ ลักษณ์หมวดคำปัจจุบัน หมวดคำก่อนหน้า หมวดคำตามหลัง รายการคำเชื่อมอนุพากย์ ความน่าจะเป็นของช่องว่างที่จะเป็นตัวแบ่งอนุพากย์ และเครื่องหมายวรรคตอน การเปรียบเทียบประสิทธิภาพของแต่ละลักษณ์ทำโดยการกำหนดชุดของลักษณ์รูปแบบต่าง ๆ แล้วนำไปทดสอบ รูปแบบของลักษณ์ที่ส่งผลต่อประสิทธิภาพของระบบมากที่สุด คือการใช้ทุกลักษณ์ร่วมกันทั้งหมด สามารถวัดค่าความถูกต้อง (F-measure) ได้ 81.17 เปอร์เซ็นต์ นอกจากนี้ เมื่อปรับค่าพารามิเตอร์ของเคอร์เนลโพลีโนเมียลให้สูงขึ้น พบว่าสามารถช่วยเพิ่มประสิทธิภาพของระบบได้ กล่าวคือ วัดค่าความถูกได้ 84.74 เปอร์เซ็นต์ เมื่อปรับค่าพารามิเตอร์ไว้ที่ D=4

ภาควิชา ภาษาศาสตร์ ลายมือชื่อนิสิต .....

สาขาวิชา ภาษาศาสตร์ ลายมือชื่อ อ.ที่ปรึกษาวิทยานิพนธ์หลัก .....

ปีการศึกษา 2556

5380139422 : MAJOR LINGUISTICS
KEYWORDS: THAI CLAUSE SEGMENTATION / CLAUSE SEGMENTATION / SUPPORT VECTOR MACHINE / SVM / ELEMENTARY DISCOURSE UNIT

NALINEE INTASAW: THAI CLAUSE SEGMENTATION USING A SUPPORT VECTOR MACHINE MODEL. ADVISOR: ASSOC. PROF. WIROTE AROONMANAKUN, Ph.D., 98 pp.

The purposes of this study are to find out linguistic features to be used in Thai clause segmentation using support vector machine (SVM) model as well as to compare efficiency of those features on clause segmentation system.

The corpus used in the study is a 76,460 word collection of Thai academic written language, consisting of 8,102 clauses. SMO, which is one of the functions in Weka, is used for training SVM. The kernel function used with SVM is polynomial kernel. The clause segmentation system uses words as inputs and decides whether a particular word is the beginning of the clause. The system's decision relies on linguistic-based features including the word's present part-of-speech, the word's previous part-of-speech, the word's following part-of-speech, lists of discourse markers, possibility of white space to be a clause separator, and punctuations. The performances of linguistic features are compared by preparing the set of feature patterns and testing those patterns. The feature pattern that performs the best is the mix of all linguistic features which claims the F-measure of 81.17 percent. In addition, when changing the value of the kernel parameter to higher value, it is found that the performance of the system increases. That is, when adjusting the exponent D to the value of 4, the system claims the F-measure of 84.74 percent.

Department: Linguistics Student's Signature .....
Field of Study: Linguistics Advisor's Signature .....
Academic Year: 2013

กิตติกรรมประกาศ

ผู้วิจัยขอขอบพระคุณ รองศาสตราจารย์ ดร.วิโรจน์ อรุณมานะกุล อาจารย์ที่ปรึกษาวิทยานิพนธ์เป็นอย่างสูง ที่ได้เสียสละเวลาในการให้คำแนะนำและความช่วยเหลือผู้วิจัยมาโดยตลอด ทั้งการทำงานวิจัยเพื่อตีพิมพ์ งานวิทยานิพนธ์ ตลอดจนการให้โอกาสรับเงินทุนผู้ช่วยวิจัย จนทำให้ผู้วิจัยสามารถทำงานวิจัยสำเร็จลุล่วงไปด้วยดี

ขอขอบพระคุณ ศาสตราจารย์ ดร.สมชาย ประสิทธิ์จูตระกูล ประธานกรรมการสอบวิทยานิพนธ์ และ ดร.เทพชัย ทรัพย์นิธิ กรรมการสอบวิทยานิพนธ์ ที่เสียสละเวลาช่วยตรวจแก้วิทยานิพนธ์ อีกทั้งยังให้ข้อชี้แนะในการทำงานวิจัย อันทำให้งานวิทยานิพนธ์ฉบับนี้มีความสมบูรณ์ยิ่งขึ้น

ขอขอบพระคุณ ศาสตราจารย์ ดร.ธีระพันธ์ เหลืองทองคำ, รองศาสตราจารย์ ดร.กิ่ง กาญจน์ เทพกาญจนา, ผู้ช่วยศาสตราจารย์ ดร.สุดา รังกุพันธุ์, ผู้ช่วยศาสตราจารย์ ดร.ธีราภรณ์ รติธรรมกุล, ผู้ช่วยศาสตราจารย์ ดร.พิทยาวัฒน์ พิทยาภรณ์ รวมถึงคณาจารย์ภาควิชาภาษาศาสตร์ท่านอื่น ๆ ที่ได้ให้ความรู้ขณะผู้วิจัยศึกษาเล่าเรียน ซึ่งเป็นการเพิ่มพูนความรู้และประสบการณ์ทางด้านภาษาศาสตร์ให้แก่ผู้วิจัย

ขอขอบคุณเพื่อน ๆ ที่ภาควิชาภาษาศาสตร์ ที่นอกจากจะมอบมิตรภาพอันดีแล้ว ยังคอยให้คำแนะนำ กำลังใจ และความช่วยเหลือทางด้านการเรียนแก่ผู้วิจัย และขอขอบคุณเจ้าหน้าที่ภาควิชาภาษาศาสตร์ทุกท่าน ที่คอยให้คำแนะนำและการช่วยเหลือตลอดระยะเวลาที่ผู้วิจัยศึกษาอยู่ ณ จุฬาลงกรณ์มหาวิทยาลัย

สุดท้ายนี้ ผู้วิจัยต้องขอขอบพระคุณบิดา มารดา คุณบุญช่วย และคุณจันทิรา อินต๊ะซาว เป็นอย่างสูง ที่ให้การสนับสนุน และมีความเชื่อมั่นในผู้วิจัยมาโดยตลอด และขอขอบคุณ คุณมาลา ทิพย์ จอห์นสัน, คุณศิริรัตน์ มอร์แกน, และคุณณีรนุช พิริยะสกุลยิ่ง สำหรับมิตรภาพและกำลังใจที่มอบให้เสมอมา

สารบัญ

หน้า

หน้า

หน้า

สารบัญตาราง

ตารางที่ หน้า

ตารางที่ .....หน้า

สารบัญภาพ

ภาพที่ หน้า

บทที่ 1

บทนำ

1.1 ความเป็นมาและความสำคัญของปัญหา

อนุพากย์เป็นหน่วยทางไวยากรณ์ ที่ประกอบไปด้วยภาคประธานและภาคแสดง ในการศึกษาโครงสร้างและความสัมพันธ์ในปริจเฉท อนุพากย์ถือเป็นหน่วยพื้นฐาน (Elementary unit) ที่ใช้เป็นหน่วยการวิเคราะห์ การแยกอนุพากย์จึงเป็นงานที่มีความสำคัญต่องานการประมวลผลภาษาธรรมชาติที่ต้องใช้ข้อมูลจากโครงสร้างปริจเฉท เช่น text summarization, machine translation, text generation, text-to-speech, discourse parsing เป็นต้น นอกจากนี้อนุพากย์ยังสามารถใช้เป็นข้อมูลรับเข้า (Input) ในบางงานได้อีกด้วย เนื่องจากอนุพากย์เป็นหน่วยที่เล็กกว่าประโยคและย่อหน้า แต่ก็เป็นถ้อยความที่ประกอบไปด้วยข้อมูลเนื้อหาอย่างครบถ้วน

สำหรับการประมวลผลภาษาไทยด้วยเครื่องนั้น อนุพากย์ก็น่าจะสามารถเป็นหน่วยของข้อมูลรับเข้าที่มีความเหมาะสมมากกว่าประโยค เนื่องจากการระบุขอบเขตประโยคในภาษาไทยยังมีปัญหา กล่าวคือ ภาษาเขียนของภาษาไทยไม่มีการใช้เครื่องหมายวรรคตอนเมื่อจบประโยค มีเพียงการเว้นวรรคหรือการใช้ช่องว่างเมื่อต้องการขึ้นใจความใหม่ แต่การเว้นวรรคก็มีวิธีการใช้และหน้าที่หลากหลาย เช่น เว้นวรรคระหว่างคำในรายการคำ เว้นวรรคระหว่างชื่อและนามสกุล เว้นวรรคหน้าและหลังตัวเลข ฯลฯ

นอกจากนี้ การระบุขอบเขตของสิ่งที่เรียกว่า “ประโยค” ในภาษาไทยก็ไม่สามารถทำได้โดยง่ายอย่างในภาษาอังกฤษ เนื่องจากภาษาไทยไม่ได้มีธรรมเนียมการเขียนให้เป็นประโยคอย่างในภาษาอังกฤษที่มีลักษณะของการเรียงคำต่างๆให้อยู่ภายในเครื่องหมาย “.” ดังนั้นหน่วยที่อยู่ภายในเครื่องหมายนี้จึงเป็นประโยคเดียวกัน ในขณะที่ภาษาไทย การเขียนเป็นการรวมคำให้เป็นกลุ่มคำหรือวลี จากวลีก็รวมกันเป็นอนุพากย์ และ คำ วลี อนุพากย์ยังสามารถเชื่อมโครงสร้างและความหมายกันได้โดยการใช้ตัวเชื่อม (Connector) ส่วนหน่วยในภาษาไทยที่เทียบเท่ากับ “ประโยค” ในภาษาอังกฤษก็ยังเป็นสิ่งที่ต้องตั้งคำถามว่า ภาษาไทยมีประโยคหรือไม่ และหากเชื่อว่าภาษาไทยมีสิ่งที่เรียกว่า “ประโยค” จริง ก็จะต้องแก้ปัญหาให้ได้ว่า ช่องว่างใดที่ถูกใช้ในการแยกสองประโยคออกจากกัน เพราะช่องว่างก็อาจจะไม่ได้ใช้เพื่อแบ่งประโยคก็ได้ ดังที่ได้กล่าวไว้แล้วข้างต้น จึงเป็นที่มาของปัญหาการระบุขอบเขตของประโยคภาษาไทย ข้อความต่อไปนี้เป็นตัวอย่างที่ชี้ให้เห็นปัญหาการตัดประโยคภาษาไทย เปรียบเทียบกับภาษาอังกฤษ

In Thailand which has the lowest breastfeeding rate in Asia and one of the lowest in the world, only 5.4 per cent of babies are exclusively breastfed during the first six months of life. Napat, who recently visited several flood-ravaged areas in Chainat, Lopburi and Singburi provinces to assess the situation of children and families affected by floods, said most mothers are maintaining the same infant feeding practice as before the floods. Unsurprisingly, infant formula is among the items most requested by mothers affected by the flooding.

ในประเทศไทยซึ่งมีอัตราการเลี้ยงลูกด้วยนมแม่ต่ำที่สุดในเอเชียและต่ำที่สุดประเทศหนึ่งในโลก^มีทารกเพียงร้อยละ^5.4^เท่านั้นที่ได้รับการเลี้ยงด้วยนมแม่เพียงอย่างเดียวในช่วงหกเดือนแรกของชีวิต^ นภัทรซึ่งได้ไปเยี่ยมพื้นที่ที่ได้รับผลกระทบหลายแห่งในจังหวัดชัยนาท^ลพบุรีและสิงห์บุรีเพื่อประเมินสถานการณ์ของเด็กและครอบครัวที่ได้รับผลกระทบจากน้ำท่วมกล่าวว่า^แม่ส่วนใหญ่ยังคงปฏิบัติเหมือนเดิม^คือถ้าให้นมแม่ก็ยังคงให้นมแม่ต่อไป^หรือให้นมผงก็ยังให้นมผงต่อไป^อย่างไรก็ตามเนื่องจากแม่ส่วนใหญ่ในประเทศไทยเลี้ยงลูกด้วยนมผง^จึงไม่น่าแปลกใจที่นมผงเป็นสิ่งที่แม่ต้องการมากที่สุด

(บทความเรื่อง นมแม่ปลอดภัยที่สุดสำหรับภาวะน้ำท่วม ดัดแปลงจาก http://www.unicef.org/thailand/tha/reallives_17802.html)

จากข้อความข้างต้น จะพบว่าข้อความภาษาอังกฤษมีการใช้เครื่องหมายจบประโยค (.) ทั้งหมด 3 แห่ง และขึ้นต้นประโยคใหม่ด้วยอักษรตัวใหญ่ ดังนั้นข้อความนี้ประกอบด้วย 3 ประโยค คือ

1 In Thailand which has the lowest breastfeeding rate in Asia and one of the lowest in the world, only 5.4 per cent of babies are exclusively breastfed during the first six months of life.
2 Napat, who recently visited several flood-ravaged areas in Chainat, Lopburi and Singburi provinces to assess the situation of children and families affected by floods, said most mothers are maintaining the same infant feeding practice as before the floods.
3 Unsurprisingly, infant formula is among the items most requested by mothers affected by the flooding.

ส่วนในข้อความภาษาไทย จะพบว่ามีช่องว่างอยู่ทั้งหมด 10 แห่ง ซึ่งแทนที่โดยเครื่องหมาย ^ ช่องว่างมีการใช้เพื่อเว้นวรรคระหว่างอนุพากย์ ก่อนและหลังตัวเลข และระหว่างรายการคำ ดังนั้น ช่องว่างในภาษาไทยจึงไม่ใช่ตัวบ่งบอกขอบเขตประโยคที่ปราศจากความคลุมเคลือ นอกจากนี้การระบุจำนวนประโยคก็เป็นสิ่งที่มีปัญหา จากข้อความตัวอย่างข้างต้น อาจจะมีคำถามที่ว่า ข้อความหลังคำว่า “กล่าวว่า” ควรจะถือว่าเป็นประโยคเดียวกันหมดภายใต้กริยา “กล่าวว่า” หรือไม่ หรือจะพิจารณาว่ามีมากกว่า 1 ประโยค นอกจากนี้ยังมีการศึกษาของ Aroonmanakun (2007) ที่ได้ทำการทดลองให้กลุ่มตัวอย่างผู้พูดภาษาไทยเป็นภาษาแม่ตัดข้อความเดียวกันออกเป็นประโยค ผลปรากฏว่าผู้พูดแต่ละคนมีการตัดประโยคที่แตกต่างกันออกไป นี่จึงเป็นอีกหลักฐานหนึ่งที่พิสูจน์ให้เห็นถึงปัญหาการระบุขอบเขตประโยคภาษาไทย

จากที่กล่าวมา แสดงให้เห็นแล้วว่า การศึกษาการแยกอนุพากย์ภาษาไทยมีความสำคัญยิ่ง เพราะอนุพากย์เป็นหน่วยที่มีระบุขอบเขตได้ชัดเจนมากกว่า สามารถใช้เป็นหน่วยที่เป็นข้อมูลรับเข้าสำหรับงานการประมวลผลภาษาธรรมชาติภาษาไทย อีกทั้งเป็นหน่วยพื้นฐานที่ใช้ในการศึกษาโครงสร้างปริจเฉทอีกด้วย ดังนั้นในงานนี้ ผู้วิจัยจึงได้ศึกษาการตัดอนุพากย์ภาษาไทยโดยใช้ซัพพอร์ตเวกเตอร์แมชชีน (Support Vector Machines: SVMs) ซึ่งเป็นตัวจำแนกประเภท (Classifier) ที่อาศัยการฝึกฝน (Supervised-learning) และสร้างแบบจำลอง โดยใช้ลักษณ์ต่าง ๆ ในการตัดสินใจและจำแนกข้อมูล

ผู้วิจัยได้เลือกใช้ซัพพอร์ตเวกเตอร์แมชชีน เนื่องจากเป็นตัวจำแนกประเภทที่เป็นที่นิยมใช้ในงานทางด้านการประมวลผลภาษาธรรมชาติ นอกจากจะสามารถกำหนดลักษณ์ที่ใช้ในการฝึกฝนและสร้างแบบจำลองได้แล้ว การใช้เทคนิคซัพพอร์ตเวกเตอร์แมชชีนยังสามารถใช้เคอร์เนลฟังก์ชั่นในการช่วยเพิ่มประสิทธิภาพการทำงานของตัวจำแนกประเภทได้อีกด้วย ทั้งนี้ ยังไม่มีข้อสรุปว่าเคอร์เนลฟังก์ชั่นแบบใดให้ผลลัพธ์ที่ดีที่สุด เนื่องจากเคอร์เนลฟังก์ชั่นหนึ่งอาจให้ผลที่น่าพอใจในงานประเภทหนึ่ง และให้ผลที่ไม่น่าพอใจในงานอีกประเภทหนึ่ง (Nguyen, Ohn et al. 2004, Liberati, Howe et al. 2009) ส่วนเคอร์เนลฟังก์ชั่นที่เป็นที่นิยมและพบในงานการประมวลผลภาษาธรรมชาติ ได้แก่ ฟังก์ชั่นโพลีโนเมียล (Polynomial function) ฟังก์ชั่นเรเดียลเบสิส (Radial basis function) และฟังก์ชั่นซิกมอยด์ (sigmoid function)

งานประมวลผลภาษาธรรมชาติที่ผ่านมาที่ใช้ซัพพอร์ตเวกเตอร์แมชชีน ได้แก่ งานการแจงส่วนประโยค (Pradhan, Ward et al. 2004, Xu and Zhang 2006, Hernault, Prendinger et al. 2010) การกำกับหมวดคำ (Giménez and Márquez 2003, Poel, Stegeman et al. 2007, Antony, Mohan et al. 2010) การจำแนกประเภทเอกสาร (Joachims 1998, Basu, Walters et al. 2003, Al-Saleem 2010) การแก้ปัญหาความกำกวมของคำหลายความหมาย (Lee, Ng et al.

2004, Buscaldi, Rosso et al. 2006) การรู้จำเสียง (Ganapathiraju 2002) และการรู้จำตัวอักษร (Borji and Hamidi 2007) นอกจากนี้วาทินี นุ้ยเพียร, สมชาย ปราการเจริญ et al. (2553) ได้ทำการทดลองเปรียบเทียบประสิทธิภาพของอัลกอริทึมต่างๆที่ใช้ในการจำแนกข้อมูล ได้แก่ โครงข่ายประสาทเทียม (Neural Network) ซัพพอร์ตเวกเตอร์แมชชีน (Support Vector Machines) นาอีฟเบย์ (Naïve Bayes) และเคเนียรเรสต์เนเบอร์ (K-Nearest Neighbor) ผลปรากฏว่าอัลกอริทึมที่ทำงานได้อย่างมีประสิทธิภาพสูงสุดคือซัพพอร์ตเวกเตอร์แมชชีน

1.2 วัตถุประสงค์

1. วิเคราะห์หาลักษณ์ทางภาษาที่จะใช้ในการตัดอนุพากย์ภาษาไทยด้วยซัพพอร์ตเวกเตอร์-แมชชีน
2. พัฒนาระบบการตัดอนุพากย์ด้วยซัพพอร์ตเวกเตอร์แมชชีน
3. เปรียบเทียบประสิทธิภาพของลักษณ์ที่ใช้ในการตัดอนุพากย์ภาษาไทย

1.3 สมมติฐาน

1. ลักษณ์ทางภาษาที่สามารถใช้ในการตัดอนุพากย์ด้วยเครื่อง ได้แก่ คำไวยากรณ์ ตัวเชื่อม ประเภทของคำ ระยะห่างระหว่างตัวเชื่อมที่ 1 และตัวเชื่อมที่ 2 ในตัวเชื่อมที่เป็นคู่ (Correlative conjunctions) เครื่องหมายวรรคตอน อนุภาคท้าย และช่องว่าง
2. ลักษณ์ทางภาษาที่มีประสิทธิภาพดีที่สุดคือตัวเชื่อม รองลงมาได้แก่ ประเภทของคำ ช่องว่าง และอนุพากย์ท้าย

1.4 ขอบเขตของการวิจัย

ใช้ข้อมูลภาษาเขียนจากคลังข้อมูลข่าวที่มีอนุพากย์จำนวนไม่ต่ำกว่า 1,000 อนุพากย์

1.5 วิธีดำเนินการวิจัย

1. ทบทวนวรรณกรรมที่เกี่ยวข้องกับการแยกอนุพากย์ด้วยเครื่อง
2. กำหนดขอบเขตของอนุพากย์ภาษาไทย
3. สร้างคลังข้อมูลที่ใช้ในการฝึกฝนและทดสอบ โดยใช้คลังข้อมูลการฝึกฝน 90 เปอร์เซ็นต์ และคลังข้อมูลทดสอบ 10 เปอร์เซ็นต์
4. วิเคราะห์หาลักษณ์ทางภาษาศาสตร์ที่เหมาะสมสำหรับซัพพอร์ตเวกเตอร์แมชชีน
5. ฝึกฝนแบบจำลองโดยใช้ซัพพอร์ตเวกเตอร์แมชชีน
6. ทดสอบระบบการแยกอนุพากย์ด้วยโดยใช้คลังข้อมูลทดสอบ

7. ประเมินผลการทดสอบ และประเมินลักษณ์ที่มีผลต่อประสิทธิภาพของซัพพอร์ตเวกเตอร์-แมชชีน
8. สรุปผลการวิจัย

1.6 ประโยชน์ที่คาดว่าจะได้รับ

1. เป็นแนวทางในการศึกษาวิเคราะห์หน่วยที่ซับซ้อนกว่าที่เกิดจากการรวมตัวกันของอนุพากย์ ซึ่งได้แก่ ประโยค ปริจเฉท เป็นต้น
2. เป็นแนวทางในการศึกษาการแยกอนุพากย์ในภาษาอื่น ๆ ด้วยเครื่อง

1.7 เครื่องมือที่ใช้ในการวิจัย

1. โปรแกรมภาษา Python ของมูลนิธิซอฟต์แวร์ไพทอน (Python Software Foundation: PSF)
2. โปรแกรมวีก้า (Weka) เวอร์ชั่น 3.6.10 พัฒนาโดย The University of Waikato ประเทศนิวซีแลนด์

บทที่ 2

ทบทวนวรรณกรรม

การศึกษาการแยกอนุพากย์ด้วยซัพพอร์ตเวกเตอร์แมชชีนจำเป็นต้องมีความรู้พื้นฐานในเรื่องอนุพากย์ตามแนวภาษาศาสตร์ และการจะแยกอนุพากย์ได้ต้องมีการกำหนดขอบเขตขอบอนุพากย์ ซึ่งทฤษฎีภาษาแต่ละทฤษฎีก็กล่าวถึงสิ่งที่เรียกว่าอนุพากย์แตกต่างกันทั้งเรื่องขอบเขตและความสัมพันธ์ระหว่างอนุพากย์ จึงมีความจำเป็นต้องทบทวนทฤษฎีเหล่านี้เพื่อเลือกทฤษฎีที่เหมาะสมที่สุดที่จะใช้ในงาน และเนื่องจากงานนี้เป็นการแยกอนุพากย์แบบใช้การเรียนรู้ของเครื่องจึงต้องศึกษาหลักการและการทำงานของแบบจำลองซัพพอร์ตเวกเตอร์แมชชีนด้วย

เนื้อหาในบทนี้จึงจะขอกล่าวถึงหัวข้อต่าง ๆ ที่ได้ทบทวนมา ได้แก่ การศึกษาอนุพากย์ตามแนวภาษาศาสตร์ ทฤษฎีเกี่ยวกับอนุพากย์ที่ใช้ในการแยกอนุพากย์ แนวทางการตัดอนุพากย์ด้วยเครื่อง งานวิจัยที่เกี่ยวข้องกับการตัดอนุพากย์ด้วยเครื่อง ซัพพอร์ตเวกเตอร์แมชชีน และงานประมวลผลภาษาธรรมชาติที่ใช้ซัพพอร์ตเวกเตอร์แมชชีน ตามลำดับ

2.1 การศึกษาอนุพากย์ตามแนวภาษาศาสตร์

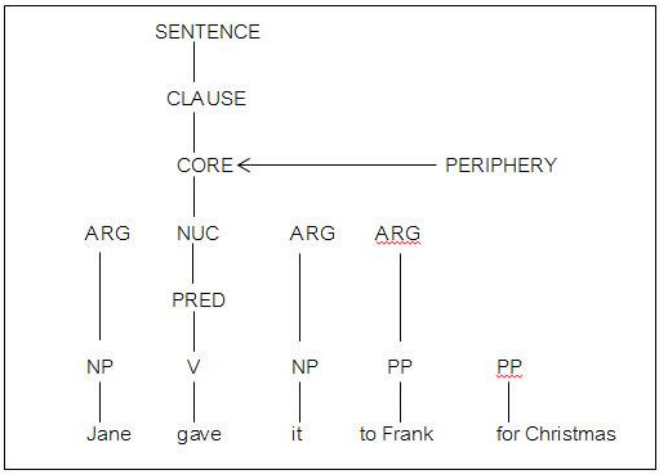
ในไวยากรณ์ภาษาศาสตร์ดั้งเดิม (Traditional Grammar) อนุพากย์ (Clause) คือ หน่วยทางไวยากรณ์ที่ประกอบไปด้วยภาคประธานและภาคแสดง อนุพากย์แบ่งออกเป็นอนุพากย์หลักหรือประโยคหลัก (Main clause) และอนุประโยคหรือประโยคย่อย (Subordinate clause) โดยที่อนุพากย์หลัก 1 อนุพากย์จะมีฐานะเทียบเท่าประโยคความเดียว (Simple sentence) มีใจความสมบูรณ์และสามารถอยู่โดยลำพังได้ ส่วนอนุประโยคจะไม่สามารถอยู่ได้โดยลำพัง มีใจความไม่สมบูรณ์ ต้องพึ่งพาอนุพากย์หลักเสมอ ในตำราหลักภาษาไทยของพระยาอุปกิตศิลปะสาร (2533) อนุประโยคภาษาไทยแบ่งตามเกณฑ์หน้าที่ได้ 3 ชนิด ดังนี้

1 นามานุประโยค คือ อนุประโยคที่ทำหน้าที่แทนคำนาม คำสรรพนาม หรือกริยาสภาวมาลา
2 คุณานุประโยค คือ อนุประโยคที่ทำหน้าที่แทนคำวิเศษณ์ที่ใช้ประกอบนามหรือสรรพนาม
3 วิเศษณานุประโยค คือ อนุประโยคที่ทำหน้าที่เป็นวิเศษณ์ประกอบกริยา หรือวิเศษณ์

ทางด้านไวยากรณ์หน้าที่ (Functional Grammar) Dik (1997) กล่าวถึงอนุพากย์ในลักษณะที่เป็นหน่วยทางอรรถศาสตร์หน่วยหนึ่ง หน่วยทางอรรถศาสตร์ (Semantic units) ในไวยากรณ์หน้าที่ประกอบไปด้วย กริยา (Predicate), คำ (Term), ภาคแสดง (Predication) ,

ประพจน์ (Proposition), และ อนุพากย์ (Clause) โครงสร้างของอนุพากย์จะมีลักษณ์เป็นชั้น (Layers) ของหน่วยต่าง ๆ ทางอรรถศาสตร์ โดยชั้นที่อยู่สูงสุดคืออนุพากย์ ประโยคความเดียวเกิดจากอนุพากย์ 1 อนุพากย์ที่มีใจความสมบูรณ์ ส่วนประโยคซับซ้อน (Complex sentence) เกิดจากอนุพากย์มากกว่า 2 อนุพากย์เชื่อมเข้าด้วยกันด้วยตัวเชื่อม โครงสร้างในประโยคซับซ้อนมี 2 ประเภท คือ โครงสร้างแบบคู่ขนาน (Coordination) และ โครงสร้างซ้อน (Embedding) โดยที่โครงสร้างแบบแรกประกอบไปด้วย 2 อนุพากย์หรือมากกว่าที่มีความเท่าเทียมกันเชิงหน้าที่ อยู่ในโครงสร้างระดับเดียวกัน และเชื่อมเข้าด้วยกันโดยตัวเชื่อม ส่วนโครงสร้างซ้อนเป็นโครงสร้างที่มีอนุพากย์หนึ่งอยู่ภายในอีกอนุพากย์หนึ่ง

ไวยากรณ์หน้าที่อีกไวยากรณ์หนึ่ง คือไวยากรณ์บทบาทและการอ้างอิง (Role and Reference Grammar) ของ Foley and VanValin (1984) อธิบายว่าอนุพากย์ประกอบด้วยส่วนหลัก (Core) และ ส่วนชายขอบ (Periphery) โดยที่ส่วนหลักประกอบไปด้วยนิวเคลียส (Nucleus) และอาร์กิวเมนต์ (Argument) ส่วนชายขอบ ประกอบด้วยส่วนที่ไม่ใช่อาร์กิวเมนต์ของกริยา โครงสร้างของอนุพากย์สามารถแสดงในรูปโครงสร้างต้นไม้ได้ดังภาพที่ 2.1



ภาพที่ 2.1 โครงสร้างต้นไม้ประโยคภาษาอังกฤษตามแนวไวยากรณ์บทบาทและการอ้างอิง ดัดแปลงจากหนังสือ Structure and Function: A Guide to Three Major Structural-Functional Theories (Part 2: From clause to discourse and beyond) ของ Butler (2003)

ในไวยากรณ์นี้ อนุพากย์มากกว่า 1 อนุพากย์สามารถเชื่อมเข้าด้วยกันเป็นประโยคซับซ้อน และความสัมพันธ์ระหว่างอนุพากย์มี 3 ประเภท ได้แก่ coordination, subordination และ cosubordination เรียกความสัมพันธ์เหล่านี้ว่าเน็กซัส (Nexus) แต่ละเน็กซัสจะอยู่ในโครงสร้างระดับใดก็ได้ใน 3 ระดับ คือ นิวเคลียส ส่วนหลัก หรืออนุพากย์ ดังนั้นจึงได้โครงสร้างรวมทั้งสิ้น 9 โครงสร้าง ได้แก่ coordination nucleus, coordination core, coordination clause,

subordination nucleus, subordination core, subordination clause, cosubordination nucleus, cosubordination core, และ cosubordination clause (Butler 2003)

นอกจากนี้ในไวยากรณ์ระบบหน้าที่ (Systemic Functional Grammar) ของ Halliday (1994) อนุพากย์เป็นหน่วยที่สำคัญที่สุดในการวิเคราะห์วากยสัมพันธ์ สำหรับไวยากรณ์ระบบหน้าที่ "ประโยค" เป็นคำที่ใช้อ้างถึงหน่วยที่เกิดจากรูปเขียนของภาษา ซึ่งอยู่ภายในเครื่องหมาย "." Halliday เรียกประโยคว่าเป็นหน่วยสร้างของการเขียน (Constituent of writing) ในขณะที่อนุพากย์เป็นหน่วยสร้างของไวยากรณ์ (Constituent of grammar) ในไวยากรณ์นี้อนุพากย์ประกอบด้วยกลุ่ม (Groups) แต่ละกลุ่มประกอบด้วยคำ (Words) และคำแต่ละคำประกอบด้วยหน่วยคำ (Morphemes) เรียกอนุพากย์ตั้งแต่ 2 อนุพากย์ที่เชื่อมกันด้วยตัวเชื่อมว่าอนุพากย์ซับซ้อน (Clause complex) ซึ่งเทียบได้กับประโยคซับซ้อนในไวยากรณ์อื่น ๆ ความสัมพันธ์ระหว่างอนุพากย์ในอนุพากย์ซับซ้อนมีด้วยกัน 2 มิติ คือ แท็กซิส (Taxis) และ ลอจิกโคซีแมนติก (Logico-semantic) โดยที่แท็กซิสเป็นความสัมพันธ์ระหว่างอนุพากย์ที่แสดงถึงสถานะเชิงโครงสร้างและความหมายของอนุพากย์ภายในอนุพากย์ซับซ้อนแบ่งอออกเป็น 2 สถานะความสัมพันธ์ คือพาราแท็กซิส (Parataxis) และไฮโปแท็กซิส (Hypotaxis) ในความสัมพันธ์แบบพาราแท็กซิส 2 อนุพากย์หรือมากกว่าเชื่อมเข้าด้วยกันโดยมีสถานะที่เท่าเทียมกัน และในความสัมพันธ์แบบไฮโปแท็กซิส 2 อนุพากย์หรือมากกว่าจะมีสถานะที่ต่างกัน โดยที่มีอนุพากย์หนึ่งเป็นอนุพากย์หลัก และอนุพากย์อื่น ๆ เป็นอนุพากย์พึ่งพา ส่วนความสัมพันธ์แบบลอจิกโคซีแมนติก เป็นความสัมพันธ์เชิงตรรกะและความหมาย แบ่งออกเป็น 2 ชนิด คือ การขยายความ (Expansion) และโปรเจ็คชั่น (Projection) แต่ละชนิดจะแบ่งออกเป็นชนิดย่อยอีก การขยายความประกอบไปด้วย 3 ชนิดได้แก่ การซ้ำความ (Elaboration) การเพิ่มความ (Extension) และการให้เหตุผล (Enhancement) ส่วนการโปรเจ็คชั่นประกอบไปด้วย 2 ชนิดได้แก่ คำพูดที่อ้างถึง (Locution) และความคิดที่อ้างถึง (Idea)

จากงานที่ทบทวนมาข้างต้น จะเห็นว่า อนุพากย์เป็นหน่วยปริจเฉทพื้นฐานที่มีองค์ประกอบและขอบเขตของโครงสร้างทางวากยสัมพันธ์ที่ชัดเจน อนุพากย์สามารถประกอบเข้าด้วยกันเป็นโครงสร้างที่ซับซ้อนและใหญ่ขึ้นได้ นั่นคือ ประโยค และปริจเฉทนั่นเอง ในขณะที่ประโยคเป็นหน่วยที่กำหนดจากการเขียน ในภาษาไทยซึ่งไม่มีการกำหนดวิธีเขียนเป็นประโยคที่ชัดเจน อนุพากย์จึงน่าจะเป็นหน่วยที่เหมาะสมมากกว่าประโยค ทั้งสำหรับการศึกษาโครงสร้างปริจเฉท และการศึกษางานทางด้านการประมวลผลภาษาธรรมชาติภาษาไทย

2.2 ทฤษฎีเกี่ยวกับอนุพากย์ที่ใช้ในการแยกอนุพากย์

อนุพากย์เป็นหน่วยพื้นฐานที่เมื่อประกอบหลาย ๆ อนุพากย์เข้าด้วยกันแล้วจะได้โครงสร้างปริจเฉท สำหรับงานทางด้านการประมวลผลภาษาธรรมชาติ อนุพากย์สามารถนำไปใช้เป็นหน่วยพื้นฐานของการประมวลผล การระบุขอบเขตและแยกอนุพากย์จึงเป็นงานหนึ่งที่มีความสำคัญ กรอบทฤษฎีที่เป็นที่ยอมรับและมีการอ้างอิงอย่างแพร่หลายที่สามารถใช้กำหนดขอบเขตของอนุพากย์เพื่อให้ได้หน่วยที่สามารถใช้อธิบายความสัมพันธ์ภายในโครงสร้างปริเฉทได้ ได้แก่ Linguistic Discourse Model หรือ LDM ของ Polanyi (1988) และ Rhetorical Structure Theory หรือ RST ของ Mann and Thompson (1988) ทั้ง 2 ทฤษฎีนี้ต่างมีความเห็นตรงกันที่ว่า อนุพากย์เป็นหน่วยที่มีความสำคัญต่อการศึกษาโครงสร้างปริจเฉท เพราะเป็นหน่วยที่เล็กที่สุดที่มีข้อมูลเชิงเนื้อหาและความหมาย และเรียกหน่วยที่เล็กที่สุดในปริจเฉทนี้ว่า หน่วยปริจเฉทพื้นฐาน (Elementary Discourse Unit: EDU) ซึ่งแต่ละหน่วยจะไม่คาบเกี่ยวกัน (Non-overlapping) การศึกษาโครงสร้างปริจเฉทจึงหมายถึงการศึกษาความสัมพันธ์ระหว่างหน่วยปริจเฉทพื้นฐานว่ามีการเชื่อมโยงความกันอย่างไร และท้ายที่สุดจะสามารถนำโครงสร้างปริจเฉทที่วิเคราะห์มาแสดงในรูปของแผนภูมิต้นไม้ได้ การอธิบายโครงสร้างและขอบเขตอนุพากย์ปริจเฉทของทั้ง 2 ทฤษฎีนี้มีรายละเอียดดังนี้

2.2.1 Linguistic Discourse Model

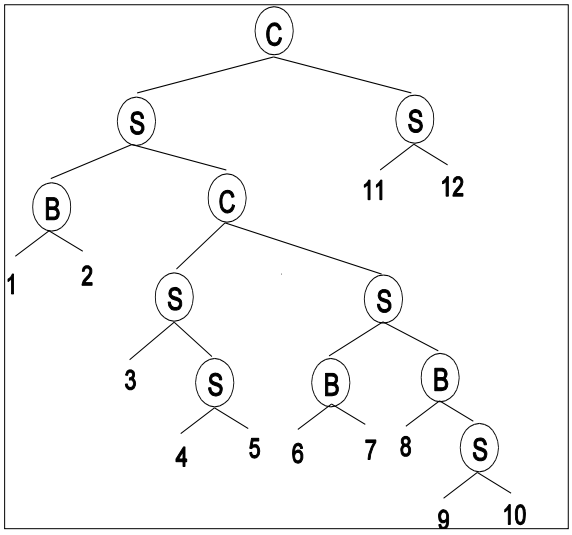
Linguistic Discourse Model หรือ LDM (Polanyi 1988) มีการอธิบายปริจเฉทสัมพันธ์ว่าขึ้นอยู่กับการตีความ (Interpretation) ยึดหลักการทางวากยสัมพันธ์เป็นหลักในการระบุประเภทความสัมพันธ์ระหว่างสองหน่วยว่ามีความสัมพันธ์กันแบบคู่ขนาน, ไม่คู่ขนาน หรือความสัมพันธ์แบบเป็นคู่ (Coordination, Subordination, and Binary) โครงสร้างผิวของปริจเฉทประกอบไปด้วยหน่วยพื้นฐาน 2 หน่วย คือ หน่วยปริจเฉทพื้นฐาน (Elementary Discourse Constituent Units: e-DCUs) และโอเปอร์เรเตอร์ (Discourse Operators: DOs) โดยที่หน่วยปริจเฉทพื้นฐานเป็นหน่วยที่แสดงอาการ หรือเหตุการณ์หนึ่ง มีเนื้อหาเชิงประพจน์ (Propositional content) หน่วยนี้อาจอยู่ในรูปของประโยคความเดียวหรืออนุพากย์ก็ได้ ส่วนโอเปอร์เรเตอร์เป็นหน่วยที่ไม่ได้แสดงเนื้อหาเชิงประพจน์ (Non-propositional content) แต่มีหน้าที่แสดงความสัมพันธ์ระหว่างหน่วยปริจเฉทพื้นฐาน หน่วยที่เป็นโอเปอร์เรเตอร์ ได้แก่ logical operators, vocatives, affirmations/disaffirmations, particles, exclamations, และ connectives การตัดอนุพากย์ตามแนว LDM คือการหาขอบเขตของหน่วยพื้นฐาน 2 หน่วยนี้

ข้อความต่อไปนี้เป็นตัวอย่างข้อความภาษาอังกฤษที่วิเคราะห์ด้วย LDM ซึ่งนำมาจากงานของ Joshi et al. (2006) โดยหน่วยปริจเฉทพื้นฐานจะอยู่ในเครื่องหมายวงเล็บสี่เหลี่ยม และมี

ตัวเลขลำดับของหน่วยปริจเฉทพื้นฐานแสดงไว้ข้างวงเล็บปิด และเมื่อวิเคราะห์โครงสร้างและปริจเฉทสัมพันธ์ของข้อความนี้แล้ว จะสามารถสร้างแผนภูมิต้นไม้ได้ดังภาพที่ 2.2

ข้อความ

[Whatever advances we may have seen in knowledge management,]1 [knowledge sharing remains a major issue.]2 [A key problem is]3 [that documents only assume value]4 [when we reflect upon their content.]5 [Ultimately,]6 [the solution to this problem will probably reside in the documents themselves.]7 [In other words,]8 [the real solution to the problem of knowledge sharing involves authoring,]9 [rather than document management.]10 [This paper is a discussion of several new approaches to authoring and opportunities for new technologies]11 [to support those approaches.]12



ภาพที่ 2.2 แผนภูมิต้นไม้ที่ได้จากการวิเคราะห์ข้อความภาษาอังกฤษตามแนวทฤษฎี LDM ดัดแปลงจาก Discourse Annotation Tutorial ของ Joshi, Prasad et al. (2006)

2.2.2 Rhetorical Structure Theory

ทฤษฎี Rhetorical Structure Theory หรือ RST (Mann and Thompson 1988) เป็นทฤษฎีที่ใช้ในการอธิบายความสัมพันธ์ระหว่างหน่วยต่าง ๆ ในปริจเฉท และแสดงออกมาโดยใช้แผนภูมิโครงสร้างต้นไม้ (Tree diagram) การวิเคราะห์โครงสร้างปริจเฉทจะคำนึงถึงความสัมพันธ์ระหว่างเจตนาของผู้ส่งสารและสาระของตัวสาร ทฤษฎีนี้เสนอว่า หน่วยปริจเฉทพื้นฐานแต่ละหน่วยจะมีสถานะความสำคัญไม่เท่ากัน สถานะความสำคัญ (Nuclearity status) นี้แบ่งออกได้ 2 สถานะ

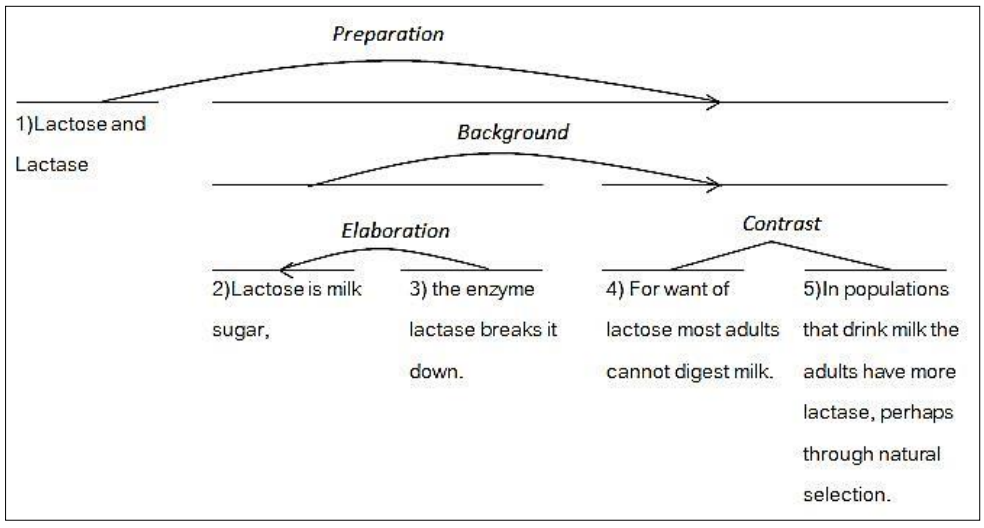
ได้แก่ สถานะนิวเคลียส (Nucleus) และสถานะแซทเทิลไลต์ (Satellite) หน่วยปริจเฉทพื้นฐานที่มีสถานะนิวเคลียสคืออนุพากย์ที่เป็นใจความหลักของถ้อยคำ เป็นส่วนที่จำเป็นและไม่สามารถละทิ้งได้ ส่วนหน่วยปริจเฉทพื้นฐานที่มีสถานะแซทเทิลไลต์ คืออนุพากย์ที่ทำหน้าที่ขยายอนุพากย์สถานะนิวเคลียส เป็นส่วนที่สามารถละทิ้งได้โดยไม่กระทบใจความหลัก หน่วยปริจเฉทพื้นฐานแต่ละหน่วยไม่ว่าจะมีสถานะใดก็ตามเมื่อประกอบเข้าด้วยกันจะแสดงความสัมพันธ์หรือมีปริจเฉทสัมพันธ์ (Rhetorical relation) อย่างใดอย่างหนึ่งต่อกันเสมอ ประเภทปริจเฉทสัมพันธ์ที่เสนอใน RST มีทั้งหมด 78 ประเภท ยกตัวอย่างปริจเฉทสัมพันธ์ ได้แก่ การขัดแย้ง (Contrast), เงื่อนไข (Condition), การสรุป (Summary), และการแสดงเจตนา (Purpose) เป็นต้น

ตัวอย่างข้อความภาษาอังกฤษต่อไปนี้

“Lactose and Lactase

Lactose is milk sugar, the enzyme lactose breaks it down. For want of lactase most adults cannot digest milk. In populations that drink milk the adults have more lactase, perhaps through natural selection.”

ประกอบไปด้วยหน่วยปริจเฉทพื้นฐานจำนวน 5 หน่วย และสามารถเขียนให้อยู่ในรูปโครงสร้างต้นไม้ ซึ่งแสดงปริจเฉทสัมพันธ์ 4 ความสัมพันธ์ ได้แก่ preparation, background, elaboration, และ contrast ได้ดังภาพที่ 2.3



ภาพที่ 2.3 ตัวอย่างการวิเคราะห์โครงสร้างปริจเฉทตามแนวทฤษฎี RST ดัดแปลงจาก Mann and Taboada (2005)

เนื่องจากหน่วยปริจเฉทพื้นฐานในทฤษฎี RST ก็คืออนุพากย์ตามไวยากรณ์ภาษาศาสตร์ทั่วไป ดังนั้นการระบุขอบเขตหน่วยปริจเฉทพื้นฐานก็คือการระบุขอบเขตอนุพากย์ การกำหนดว่า

หน่วยใดบ้างที่เป็นอนุพากย์ปริจเฉทตามทฤษฎี RST ได้มีการเขียนอธิบายไว้ในคู่มือการกำกับหน่วยทางปริจเฉทสำหรับใช้สร้างคลังข้อมูล RST Discourse Treebank ของ Carlson and Marcu (2001) และคู่มือนี้ก็ถูกใช้เป็นแบบอย่างและแนวทางสำหรับการศึกษาการตัดหน่วยปริจเฉทพื้นฐานอย่างแพร่หลาย หน่วยที่กำหนดให้เป็นหน่วยปริจเฉทพื้นฐานจะพิจารณาจากความละเอียดของการกำกับ (Granularity of tagging) และความเป็นไปได้ที่จะใช้ในการศึกษาหน่วยที่ใหญ่ขึ้น โดยใช้โครงสร้างทางวากยสัมพันธ์และคำเชื่อมต่าง ๆ ในการช่วยระบุขอบเขตของหน่วยปริจเฉทพื้นฐาน ตัวอย่างหน่วยที่ถือว่าเป็นหน่วยปริจเฉทพื้นฐานตามการวิเคราะห์ในคู่มือนี้ เช่น superordinate clauses, subordinate clauses, infinitival modifiers, Prepositional phrases with a clausal object, coordinated clauses, syntactic focusing devices (เช่น cleft, extraposition, pseudo-cleft constructions), correlative subordinators, relative clauses, nominal postmodifier with non-finite clause, appositives, parentheticals, phrases with strong discourse markers (เช่น because, in spite of, as a result of, according to) ฯลฯ

ส่วนหน่วยที่ไม่ถือว่าเป็นหน่วยปริจเฉทพื้นฐานได้แก่ clausal subjects and objects of verbs, clausal objects of prepositional phrases, infinitival complements, และ participial complements เป็นต้น อย่างไรก็ตามการนิยามขอบเขตหน่วยปริจเฉทพื้นฐานตามแนวทฤษฎี RST อาจจะแตกต่างกันได้ ขึ้นอยู่กับวัตถุประสงค์ของการวิเคราะห์และลักษณะภาษาที่แตกต่างกันออกไป

เมื่อเปรียบเทียบขอบเขตของอนุพากย์ปริจเฉทตามแนวทฤษฎี RST และ LDM จะพบว่าคำเชื่อมใน RST เป็นส่วนหนึ่งของหน่วยปริจเฉทพื้นฐาน ในขณะที่คำเชื่อมใน LDM ไม่ถือว่าเป็นส่วนหนึ่งของหน่วยปริจเฉทพื้นฐาน แต่เป็นอีกหน่วยที่มีหน้าที่แสดงความสัมพันธ์ระหว่างหน่วยปริจเฉทพื้นฐาน ในงานวิจัยนี้ ผู้วิจัยจะยึดตามแนวทฤษฎี RST อนุพากย์ซึ่งเป็นหน่วยพื้นฐานของการวิเคราะห์ปริจเฉทที่ต้องการแยกในงานนี้ จึงอิงตามหน่วยพื้นฐานที่ต้องใช้ในการวิเคราะห์แบบ RST สาเหตุที่เลือกทฤษฎีนี้ เนื่องจาก RST ให้ความสำคัญต่อความสัมพันธ์ระหว่างเจตนาของผู้ส่งสารและสาระของตัวสาร ทำให้สามารถอธิบายประเภทความสัมพันธ์ได้ละเอียดกว่า เห็นได้จากมีการแบ่งประเภทย่อยของความสัมพันธ์ได้เป็นจำนวนมาก การกำหนดขอบเขตอนุพากย์ตามแนวทฤษฎีนี้ จึงน่าจะสามารถใช้อธิบายความสัมพันธ์และโครงสร้างปริจเฉทได้อย่างครบถ้วน

2.2.3 การใช้ RST วิเคราะห์ภาษาไทย

สำหรับการใช้ RST ในการวิเคราะห์ปริจเฉทนั้น ขั้นแรกจะต้องแยกหน่วยพื้นฐานต่าง ๆ ภายในปริจเฉท ในที่นี่จะขอเรียกหน่วยพื้นฐานสำหรับการวิเคราะห์โครงสร้างปริจเฉท หรือ Elementary discourse unit ว่า “หน่วยปริจเฉทพื้นฐาน” หรือ “EDU” อันประกอบไปด้วยอนุ

พากย์และวลีที่ขึ้นต้นด้วยคำเชื่อมเด่น เมื่อแยก EDU ได้แล้ว ก็ทำการวิเคราะห์และระบุประเภทความสัมพันธ์ระหว่าง EDU ที่อยู่ติดกัน ซึ่งในที่สุดแล้วจะสามารถสร้างแผนภูมิต้นไม้เพื่อแสดงโครงสร้างและความสัมพันธ์ภายในปริจเฉทได้

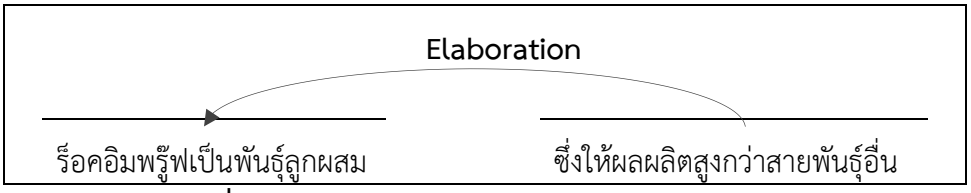
ในส่วนของประเภทความสัมพันธ์ Mann and Thompson (1988) ผู้พัฒนาทฤษฎีนี้ กล่าวไว้ว่า ในการวิเคราะห์ปริจเฉท ไม่จำเป็นต้องใช้ประเภทความสัมพันธ์ทั้งหมด 78 ประเภทตามที่เสนอเอาไว้ใน RST นั่นคือ ผู้ศึกษาปริจเฉทสามารถเลือกความละเอียดของการแบ่งประเภทปริจเฉทสัมพันธ์ได้ โดยสามารถพิจารณาได้จากลักษณะของภาษาที่ต้องการศึกษา และเป้าประสงค์ของการศึกษาหรือการนำไปใช้

งานวิจัยหนึ่งของเมธี วัฒนะเมธานนท์ (2549) ซึ่งศึกษาระบบการรู้จำความสัมพันธ์ระดับปริจเฉทในภาษาไทยโดยใช้แบบจำลองนาอีฟเบย์ ได้นำ RST มาเป็นกรอบในการกำหนดประเภทปริจเฉทสัมพันธ์ในภาษาไทย งานนี้ได้นำเสนอการวิเคราะห์ปริจเฉทภาษาไทยด้วย RST อย่างละเอียด ประเภทปริเฉทสัมพันธ์ที่ใช้ในงานนี้มีจำนวนทั้งหมด 78 ประเภทความสัมพันธ์ ซึ่งในที่นี้ จะขอนำเสนอตัวอย่างการวิเคราะห์ภาษาไทยของงานเมธี ซึ่งผู้วิจัยเห็นว่ามีประโยชน์ ทำให้เห็นภาพการใช้ RST ในการวิเคราะห์ปริจเฉทภาษาไทย ข้อความและรูปภาพต่อไปนี้เป็นตัวอย่างข้อความภาษาไทยที่ใช้ RST ในการวิเคราะห์ โดยเครื่องหมายวงเล็บสี่เหลี่ยมใช้สำหรับแสดงขอบเขตของ EDU และตัวเลขที่กำกับข้างวงเล็บปิดแสดงลำดับที่ของ EDU

ข้อความ 1

[ร็อคอิมพรู๊ฟเป็นพันธุ์ลูกผสม]1[ซึ่งให้ผลผลิตสูงกว่าสายพันธุ์อื่น]2

จากข้อความ 1 แยกได้ 2 EDU คำว่า “ซึ่ง” เป็นคำเชื่อมแสดงการขยายความ นั่นคือ EDU ที่ 2 ขยาย EDU แรก จึงกล่าวได้ว่าทั้ง 2 หน่วยมีปริจเฉทสัมพันธ์แบบขยายความ (Elaboration relation) เมื่อนำข้อความนี้ไปสร้างแผนภูมิต้นไม้ จะเห็นขอบเขตขอบเขตของทั้ง 2 EDU ชัดเจนมากขึ้น และปริจเฉทสัมพันธ์สามารถแสดงโดยการโยงเส้นพร้อมระบุชื่อความสัมพันธ์ไว้เหนือเส้นนั้น แผนภูมิต้นไม้สำหรับข้อความ 1 จะได้ดังภาพที่ 2.4

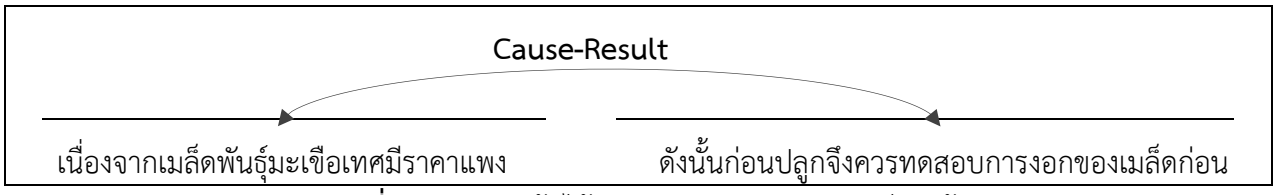


ภาพที่ 2.4 แผนภูมิต้นไม้ RST แสดงปริจเฉทสัมพันธ์ของข้อความ 1

ข้อความ 2

[เนื่องจากเมล็ดพันธุ์มะเขือเทศมีราคาแพง]1[ดังนั้นก่อนปลูกจึงควรทดสอบการงอกของเมล็ดก่อน]2

จากข้อความ 2 แยกได้ 2 EDU คำเชื่อม “เนื่องจาก” และ “ดังนั้น” เป็นคำเชื่อมแสดงความสัมพันธ์แบบให้เหตุและผล (Cause-Result relation) โดย EDU แรกเป็นสาเหตุ และ EDU ที่ 2 เป็นผลจากสาเหตุดังกล่าว สามารถสร้างแผนภูมิต้นไม้แสดงโครงสร้างและปริจเฉทสัมพันธ์ได้ดังภาพที่ 2.5

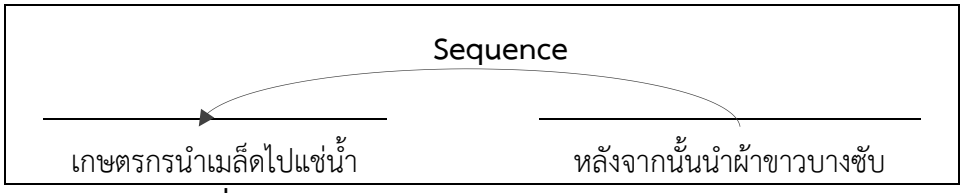


ภาพที่ 2.5 แผนภูมิต้นไม้ RST แสดงปริจเฉทสัมพันธ์ของข้อความ 2

ข้อความ 3

[เกษตรกรนำเมล็ดไปแช่น้ำ]1[หลังจากนั้นนำผ้าขาวบางซับ]2

จากข้อความ 3 แยกได้ 2 EDU คำเชื่อม “หลังจากนั้น” เป็นคำเชื่อมแสดงลำดับของเหตุการณ์ สิ่งที่ตามหลังคำเชื่อมนี้จะเป็นเหตุการณ์ที่เกิดขึ้นทีหลัง ในที่นี้ EDU แรกเป็นเหตุการณ์ที่เกิดขึ้นก่อน และ EDU ที่ 2 เกิดขึ้นในลำดับต่อมา ดังนั้นสามารถกล่าวได้ว่าทั้ง 2 EDU นี้มีปริจเฉทสัมพันธ์แบบเป็นลำดับ (Sequence relation) และสามารถสร้างแผนภูมิต้นไม้ได้ดังภาพที่ 2.6



ภาพที่ 2.6 แผนภูมิต้นไม้ RST แสดงปริจเฉทสัมพันธ์ของข้อความ 3

ในงานนี้ ผู้วิจัยจะไม่วิเคราะห์ปริจเฉทสัมพันธ์ระหว่างอนุพากย์ แต่จะทำการแยกอนุพากย์ภายในปริจเฉทด้วยเครื่องเท่านั้น สาระสำคัญของ RST ที่ผู้วิจัยใช้ในงานนี้คือ การกำหนดหน่วยที่จะเป็นอนุพากย์ภาษาไทย เพื่อนำไปเป็นกรอบสำหรับการแยกอนุพากย์ภาษาไทยด้วยเครื่อง ซึ่งอนุพากย์ที่แยกได้ สามารถนำไปใช้เป็นหน่วยพื้นฐานในการวิเคราะห์ปริจเฉทภาษาไทยตามแนวทฤษฎี RST ได้ ส่วนการกำหนดขอบเขตอนุพากย์ภาษาไทยที่ใช้ในงานนี้ จะกล่าวถึงต่อไปในบทที่ 4

2.3 แนวทางการแยกอนุพากย์ด้วยเครื่อง

ในที่นี้จะใช้คำว่า “การแยกอนุพากย์” หรือ “การแยกหน่วยปริจเฉทพื้นฐาน” หรือ “การแยก EDU” สำหรับการหาขอบเขตระดับอนุพากย์ในปริจเฉท และ “การแยกประโยค” ในกรณีที่เป็นการหาขอบเขตระดับประโยคในปริจเฉท ทั้งการแยกอนุพากย์และการแยกประโยคต่างก็มีแนวคิดเหมือนกัน คือเป็นหาขอบเขตของหน่วยที่ต้องการแยก และระบุประเภทของหน่วยที่แยก การแยกหน่วยเหล่านี้ด้วยเครื่องสามารถแบ่งออกเป็น 2 แนวทางหลัก ได้แก่ แนวทางการใช้กฎ และแนวทางการเรียนรู้ด้วยเครื่อง ดังนี้

2.3.1 แนวทางการใช้กฎ (Rule-based approaches)

แนวทางการใช้กฎ คือ การกำหนดกฎไวยากรณ์ (Regular grammar) หรือนิพจน์ปรกติ (Regular expression) เพื่อใช้ในการหาขอบเขตและระบุประเภทของหน่วยที่ต้องการแยก เช่น กฎการใช้เครื่องหมายวรรคตอนต่าง ๆ กฎการเว้นวรรค กฎการใช้ตัวพิมพ์ใหญ่ตัวพิมพ์เล็ก การกำหนดรายการคำ การใช้กฎในการหารูปแบบของการเรียงตัวกันของอักขระ ฯลฯ ภาพที่ 2.7 แสดงตัวอย่างของการแยกประโยคภาษาอังกฤษด้วยวิธีการใช้กฎ

```
IF ((right context = period + space + capital letter
OR period + quote + space + capital letter
OR period + space + quote + capital letter)
AND (left context != abbreviation))
THEN sentence boundary
```

ภาพที่ 2.7 ตัวอย่างกฎเพื่อใช้ในการแยกประโยคภาษาอังกฤษ ดัดแปลงจากการบรรยายของ Lemnitzer (2006)

อย่างไรก็ตาม วิธีการใช้กฎจะไม่สามารถใช้ได้กับเอกสารทุกประเภท เนื่องจากภาษาแต่ละประเภทมีกฎที่แตกต่างกันออกไป นอกจากนี้ยังจำเป็นต้องสร้างกฎเป็นจำนวนมากเพื่อให้ครอบคลุมและสามารถแยกหน่วยที่ต้องการได้อย่างถูกต้องแม่นยำ

Tofiloski, Brooke et al. (2009) เสนอการแยกหน่วยปริจเฉทพื้นฐานโดยการใช้กฎทางวากยสัมพันธ์ (syntactically-based rules) 12 กฎ และกฎเกี่ยวกับคำ (lexical rules) อีกจำนวนหนึ่ง ได้แก่ วลีระบุนัย และประเภทของคำ ใช้ประโยคเป็นข้อมูลรับเข้าเพื่อแยกหน่วยปริจเฉทพื้นฐาน มีการกำหนดหน่วยที่เป็นหน่วยปริจเฉทพื้นฐานตามแนวทฤษฎี RST

นอกจากนี้ Van der Vliet (2010) ได้เสนอวิธีการระบุขอบเขตหน่วยปริจเฉทพื้นฐานภาษาดัตช์โดยวิธีการใช้กฎ หน่วยที่เขากำหนดให้เป็นหน่วยปริจเฉทพื้นฐาน ได้แก่ clauses, simple sentences, coordinate clauses, non-restrictive relative clauses, embedded non-restrictive relative clauses, fragments functioning as complete utterances, coordinate elliptical clauses การระบุขอบเขตหน่วยปริจเฉทพื้นฐานเริ่มจากการแยกประโยคและสร้างต้นไม้โครงสร้างพึ่งพา (Dependency tree) โดยใช้ Alpino ซึ่งเป็น XML parser และใช้ประโยคซึ่งอยู่ในรูปต้นไม้โครงสร้างพึ่งพาเป็นข้อมูลรับเข้าสำหรับหาขอบเขตอนุพากย์ปริจเฉท จากนั้นเขียนสคริปต์ Xquery ซึ่งใช้ในการคิวรี (Query) ต้นไม้โครงสร้างพึ่งพาที่เป็นภาษา XML สคริปต์ของเขาใช้กฎทางวากยสัมพันธ์และเครื่องหมายวรรคตอนต่างๆในการกำกับขอบเขตอนุพากย์ปริจเฉท ข้อมูลที่ใช้ในการประเมินการทำงานของระบบคือข้อความสารานุกรม (Encyclopedia texts) และจดหมายระดมทุน (Fundraising letters) ได้ค่าความแม่นยำ และค่าความครบถ้วน 73 เปอร์เซ็นต์ และ 67 เปอร์เซ็นต์ สำหรับข้อมูลจากสารานุกรม และ 76 เปอร์เซ็นต์ และ 74 เปอร์เซ็นต์ สำหรับจดหมายระดมทุน

2.3.2 แนวทางการเรียนรู้ด้วยเครื่อง (Machine Learning)

แนวทางการเรียนรู้ด้วยเครื่อง คือ การใช้แบบจำลองทางสถิติมาใช้ในการหาขอบเขตและระบุประเภทของหน่วยที่ต้องการแยก วิธีนี้ไม่ต้องพึ่งพาการเขียนกฎ ทำให้สามารถจัดการกับเอกสารที่กฎไม่ครอบคลุมได้ นอกจากนี้ยังสามารถนำไปใช้กับประเภทภาษาได้หลายประเภทอีกด้วย หลักการคือจะต้องกำหนดลักษณ์ (Features) ต่าง ๆ เพื่อให้แบบจำลองเกิดการเรียนรู้จากชุดข้อมูลการฝึกฝน (Training data) จากนั้นแบบจำลองจะสามารถหาขอบเขตและระบุประเภทของหน่วยที่ต้องการแยกในเอกสารที่เครื่องไม่เคยเห็นมาก่อนได้ ตัวอย่างแบบจำลอง เช่น ฮิดเดนมาคอร์ฟโมเดล (Hidden Markov Models), โครงข่ายประสาทเทียม (Neural Network), ต้นไม้ช่วยตัดสินใจ (Decision Tree), นาอีฟเบย์ (Naïve Bayes), ฯลฯ ตัวอย่างงานวิจัยที่ใช้แบบจำลองทางสถิติมีดังต่อไปนี้

Molina and Pla (2001) ศึกษาการระบุอนุพากย์ภาษาอังกฤษโดยใช้แบบจำลองทางสถิติฮิดเดนมาคอร์ฟโมเดล ใช้ข้อมูลรับเข้าในรูปประโยค โดยทำการตัดคำ ระบุประเภทของคำ และแจงส่วนประโยคแต่ละประโยค ในขั้นตอนการระบุอนุพากย์ประกอบไปด้วย 3 ส่วน ส่วนที่ 1 เป็นการระบุจุดเริ่มต้นของอนุพากย์ ส่วนที่ 2 เป็นการระบุจุดสิ้นสุดของอนุพากย์ และส่วนที่ 3 เป็นการกำกับหน่วยที่เป็นอนุพากย์ ฮิดเดนมาคอร์ฟโมโดลจะทำการคำนวณเพื่อหาค่าความน่าจะเป็นของการเรียงตัวกันของประเภทคำ และการเรียงตัวกันของหน่วยสร้างทางไวยากรณ์ ที่ให้ค่าความน่าจะเป็นสูงสุดที่จะเป็นจุดเริ่มต้นและจุดสิ้นสุดของอนุพากย์ จากการทดสอบระบบทั้ง 3 ส่วน ได้ค่า F-score 86.48

เปอร์เซ็นต์, 78.38 เปอร์เซ็นต์, และ 66.79 เปอร์เซ็นต์ สำหรับส่วนที่ 1, 2, และ 3 ตามลำดับ ทั้งนี้พบว่าการระบุจุดเริ่มต้นของอนุพากย์ทำได้ง่ายกว่าการระบุจุดสิ้นสุด

Sporleder and Lapata (2005) ได้เสนอการแยกหน่วยปริจเฉทพื้นฐานโดยใช้กรอบการวิเคราะห์ขอบเขตหน่วยปริจเฉทพื้นฐานตามแนว RST และทำการกำกับหน่วยปริจเฉทพื้นฐานแต่ละหน่วยด้วยว่ามีสถานะเป็นนิวเคลียสหรือแซทเทิลไลต์ แบบจำลองที่ใช้ในงานนี้คือบูสต์ติง (Boosting) ใช้ข้อมูลรับเข้าเป็นประโยค ลักษณ์ที่ใช้ในขั้นตอนการแยกหน่วยปริจเฉทพื้นฐานเป็นลักษณ์ประเภทข้อมูลเชิงไวยากรณ์และคำ ได้แก่ คำ (Tokens) ประเภทของคำ (Parts of speech) หน่วยสร้างทางไวยากรณ์ (Syntactic chunks) คำเชื่อม (Connectives) และตำแหน่งของคำในประโยค ส่วนในขั้นตอนการกำกับสถานะของหน่วยปริจเฉทพื้นฐานใช้ลักษณ์เพิ่มอีก 2 ลักษณ์ คือความยาวของหน่วยปริจเฉทพื้นฐานโดยวัดจากจำนวนคำ และจำนวนหน่วยปริจเฉทพื้นฐานภายในประโยค ในงานนี้ได้นำเสนอเพียงประสิทธิภาพของการกำกับสถานะของหน่วยปริจเฉทพื้นฐานเท่านั้น เนื่องจากเป็นวัตถุประสงค์หลักของงาน โดยได้ค่า F-score มากกว่า 74 เปอร์เซ็นต์

Subba and Di Eugenio (2007) กำหนดขอบเขตหน่วยปริจเฉทพื้นฐานตามทฤษฎี RST และใช้โครงข่ายประสาทเทียมในการแยกหน่วยปริจเฉทพื้นฐานข้อมูลรับเข้าอยู่ในรูปประโยค ลักษณ์ที่ใช้แบ่งออกเป็น 4 ประเภท ได้แก่ ประเภทของคำ ข้อมูลทางวากยสัมพันธ์ คำระบุนัยปริจเฉท และเครื่องหมายวรรคตอน ใช้ข้อมูลจากคลังข้อมูล RST-DT ในการทดสอบประสิทธิภาพ และเปรียบเทียบผลกับระบบแยกปริจเฉทที่ใช้ต้นไม้ช่วยตัดสินใจ ผลปรากฏว่าระบบที่ใช้โครงข่ายประสาทเทียมได้ค่าความแม่นยำและค่าความครบถ้วนสูงกว่า 80 เปอร์เซ็นต์

Vijay Sundar Ram, Bakiyavathi et al. (2009) ศึกษาการระบุขอบเขตอนุพากย์ในภาษาทมิฬ ในขั้นตอนการประมวลเบื้องต้น ประโยคที่เตรียมเอาไว้แล้วถูกนำมากำกับหมวดคำ และแจงส่วนประโยค จากนั้นใช้แบบจำลองคอนดิชั่นนอลแรนดอมฟิลด์ (Conditional Random Fields: CRFs) ในขั้นตอนการระบุขอบเขตของอนุพากย์ โดยอนุพากย์ที่ต้องการระบุขอบเขตแบ่งออกเป็น 5 ประเภท ได้แก่ Participle clause, conditional clause, infinitive clause, non-finite clause และ main clause ลักษณ์ที่ใช้ในการฝึกฝนแบบจำลอง ได้แก่ ข้อมูลที่ได้จากการกำกับหมวดคำ การแจงส่วนประโยค และโครงสร้างทางวากยสัมพันธ์ ใช้ข้อมูลจากคลังข้อมูลข่าวภาษาทมิฬจำนวน 219 ประโยคในการทดลอง และ 75 ประโยคสำหรับการทดสอบ ได้ผลคือระบบสามารถกำกับจุดเริ่มต้นของอนุพากย์ได้ถูกต้อง 81.58 เปอร์เซ็นต์ และจุดสิ้นสุดของอนุพากย์ 87.37 เปอร์เซ็นต์

2.4 งานวิจัยที่เกี่ยวข้องกับการแยกอนุพากย์ในภาษาไทย

Charoensuk (2005) ได้เสนอวิธีการแยกหน่วยปริจเฉทพื้นฐานด้วยวิธีการผสมระหว่างการเรียนรู้ด้วยเครื่องและการใช้กฎ เรียกหน่วยปริจเฉทพื้นฐานว่า “อนุพากย์ปริจเฉท” วิธีการศึกษาเริ่ม

จากการกำหนดขอบเขตอนุพากย์ปริจเฉทภาษาไทยตามแนวทฤษฎี RST แบ่งอนุพากย์ปริจเฉทออกเป็น 2 ประเภท ได้แก่ อนุพากย์พื้นฐาน และอนุพากย์ซ้อน ตัวอย่างอนุพากย์ปริจเฉทในงานนี้เป็นดังตาราง 2.1

ชนิดของอนุพากย์		**ตัวอย่าง**
Basic	**Simple sentence**	[กะหล่ำปลีสีเขียว]
	Noun phrase	โรคระบาดพบในภาคกลาง [เช่น ปทุมธานี, นครปฐม]
Embedded	**Embedded clause**	กะหล่ำปลี [ที่ถูกทำลาย] จะมีสีเหลือง
	Noun phrase	เกษตรกรควรใส่ปุ๋ยไนโตรเจน [เช่น ปุ๋ยแอมโมเนียมซัลเฟต หรือยูเรีย] ลงในแปลงด้วย

ตารางที่ 2.1 ชนิดและตัวอย่างของอนุพากย์ปริจเฉทในภาษาไทยวิเคราะห์ตามแนวทฤษฎี RST ดัดแปลงจาก Charoensuk (2005)

ในขั้นตอนของการแยกอนุพากย์ด้วยเครื่องของงานนี้ แบ่งออกเป็น 3 ขั้นตอน คือ การประมวลผลเบื้องต้น การแยกอนุพากย์ และการประมวลผลขั้นปลาย ขั้นตอนของการประมวลผลเบื้องต้นประกอบด้วย การตัดคำ การกำกับหมวดคำ การสกัดชื่อเฉพาะ และการสกัดคำนามประสม ขั้นตอนของการแยกอนุพากย์ใช้แบบจำลองต้นไม้ช่วยตัดสินใจ C4.5 ในการระบุจุดเริ่มต้นและสิ้นสุดของอนุพากย์ โดยใช้ลักษณ์ 5 ประเภท ได้แก่ คำระบุนัย ตัวเชื่อมปริจเฉทที่เป็นคู่ ช่องว่างระหว่างคำหรือความ ประเภทของคำ และขอบเขตของวลี ส่วนในขั้นตอนประมวลผลขั้นปลาย ใช้สำหรับแก้ปัญหากรณีที่อนุพากย์ปรากฏรูปตัวเชื่อมมากกว่าหนึ่ง ในขั้นตอนนี้จะใช้กฎในการตัดสินใจว่าตัวเชื่อมที่พิจารณาเป็นตัวเชื่อมปริจเฉทหรือไม่ ผลการทดลองระบบแยกอนุพากย์ พบว่าได้ค่าความแม่นยำ 80 เปอร์เซ็นต์ และค่าความครบถ้วน 81 เปอร์เซ็นต์

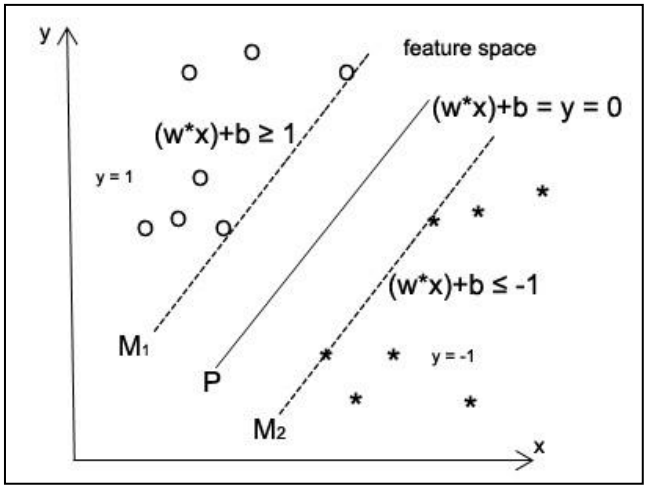
ในงานของ Sinthupoun (2009) เขาได้แยกอนุพากย์ด้วยแบบจำลองทางสถิติ ก่อนที่จะนำไปวิเคราะห์โครงสร้างปริจเฉทในภาษาไทยต่อ เริ่มจากการแยกวลีและระบุประเภทวลีโดยใช้แบบจำลองฮิดเดนมาคอร์ฟโมเดล ในขั้นตอนการแยกอนุพากย์ เขาใช้คุณลักษณะการเรียงลำดับของนามวลีและกริยาวลีประเภทต่าง ๆ ให้แบบจำลองฮิดเดนมาคอร์ฟโมเดลเกิดการเรียนรู้ และระบุขอบเขตอนุพากย์ เฉพาะในส่วนของการแยกอนุพากย์ ได้ค่าความแม่นยำ 94.2 เปอร์เซ็นต์ และค่าความครบถ้วน 85.3 เปอร์เซ็นต์ จากนั้นใช้อนุพากย์ที่ได้จากการแยกมาวิเคราะห์โครงสร้างปริจเฉทต่อไป

อีกงานหนึ่งเป็นการแยกอนุพากย์ภาษาไทยด้วยวิธีการใช้กฎทางไวยากรณ์ของ Ketui, Theeramunkong et al. (2012) โดยเริ่มจากการนิยามหน่วยปริจเฉทพื้นฐาน จากนั้นสร้างชุดของ

กฎไวยากรณ์ไม่พึ่งบริบท (Context free grammar rule) ได้จำนวนกฎ 446 ข้อ ในงานนี้แบ่งการประเมินผลออกเป็น 2 ระดับ ได้แก่ การประเมินกับข้อความที่ผ่านการตัดคำแล้ว และการประเมินกับข้อความต้นฉบับที่ไม่ได้ตัดคำ และมีการทดสอบ 2 แบบ คือ แบบ open test ซึ่งเป็นการใช้คลังข้อมูลฝึกฝนและทดสอบเดียวกัน และแบบ close test ซึ่งเป็นการแยกคลังข้อมูลฝึกฝนและทดสอบ หลังจากผ่านการใช้กฎแล้ว พบว่าได้ค่าความครบถ้วนที่สูงมาก แต่ค่าความแม่นยำต่ำ ทำให้วัดค่า f-measure ได้ประมาณ 36 - 53 เปอร์เซ็นต์ จึงได้นำเทคนิค 2 เทคนิคเปรียบเทียบกัน คือ left-to-right longest matching (L2R-LM) และ maximal longest matching (M-LM) มาใช้เพื่อเพิ่มค่าความแม่นยำ ผลการทดสอบปรากฏว่า การใช้ close test ทดสอบข้อความทั้งที่ผ่านการตัดคำและไม่ผ่านการตัดคำ วัดค่า f-measure ได้ประมาณ 90-100 เปอร์เซ็นต์ ส่วนการใช้ open test ประเมินค่า f-measure ได้ประมาณ 54-67 เปอร์เซ็นต์ โดยรวมแล้ว L2R-LM และ M-LM สามารถเพิ่มประสิทธิภาพของระบบได้ใกล้เคียงกัน

2.5 ซัพพอร์ตเวกเตอร์แมชชีน (Support Vector Machine: SVM)

ซัพพอร์ตเวกเตอร์แมชชีนเป็นแบบจำลองทางคณิตศาสตร์ที่ใช้ในการจำแนกข้อมูล 2 กลุ่มออกจากกัน โดยข้อมูลจะต้องแปลงเป็นค่าลักษณ์ (Features) และวางตัวอยู่ในพื้นที่ลักษณ์ (Feature space) จากนั้น*ซัพพอร์ตเวกเตอร์แมชชีน*จะใช้สมการเชิงเส้นสร้างเส้นแบ่งหรือที่เรียกว่า ระนาบ (Plane) ที่ดีที่สุดที่เป็นเส้นตรงแยกข้อมูลออกจากกัน โดยที่เส้นแบ่งที่ดีที่สุด (OPTimal plane) เกิดจากการสร้างเส้นแบ่งที่ขนานกับเส้นแบ่งเดิมและขยายเส้นขอบ (Margin) ออกจากเส้นแบ่งเดิมไปเรื่อย ๆ จนสัมผัสค่าของข้อมูลที่ใกล้ที่สุด



ภาพที่ 2.8 แสดงสมการเส้นตรงของเส้นขอบและเส้นแบ่ง ดัดแปลงจาก Ivanciuc (2005)

จากภาพที่ 2.8 เส้น M_1 และ M_2 คือเส้นขอบที่ขยายออกจากเส้นแบ่ง P ค่าของข้อมูลที่อยู่บนเส้นขอบของทั้งสองฝั่งเรียกว่า ซัพพอร์ตเวกเตอร์ (Support vectors) *ซัพพอร์ตเวกเตอร์แม*

*ซซีน*จะหาระยะห่างจากเส้นขอบซ้ายไปยังเส้นขอบขวาที่มีค่าระยะความห่างสูงสุด โดยการเปลี่ยนความชันของเส้น P ไปเรื่อย ๆ เพื่อให้ได้ความกว้างสูงสุดของเส้นขอบ

ข้อมูลที่จะนำมาวางบนพื้นที่ลักษณ์จะต้องแปลงให้อยู่ในรูปของเวกเตอร์ โดยกำหนดให้ $(x_1,y_1),\ldots,(x_n,y_n)$ เป็นคู่ของตัวอย่างที่ใช้ในการสอนที่มีจำนวน n และ x คือข้อมูลรับเข้า y คือผลลัพธ์ที่มีค่า 1 หรือ -1 เท่านั้น สมการที่ (1) เป็นสมการเชิงเส้นของเส้นแบ่ง P ตามรูปที่ 4

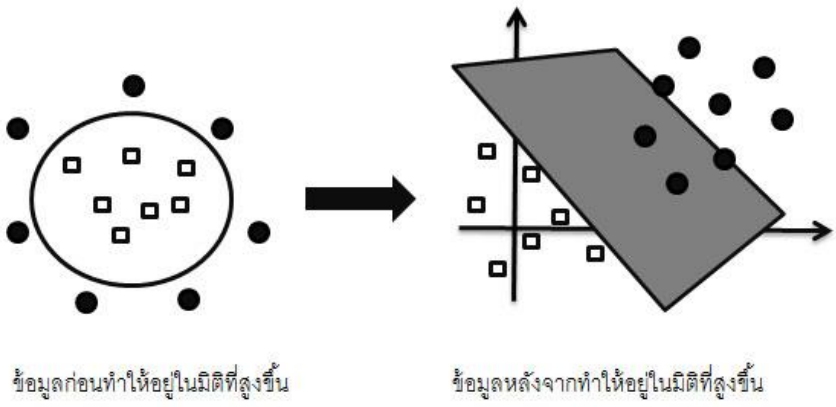
$$(w*x)+b = y = 0 \quad (1)$$

โดยที่ w คือ ค่าน้ำหนัก และ b คือ ค่า bias เส้นขอบ M_1 และ M_2 สามารถกำหนดได้ดังอสมการ (2) และ (3) ตามลำดับ

$$(w*x)+b \geq 1 \quad \text{ถ้า } y_1 = 1 \quad (2)$$

$$(w*x)+b \leq -1 \quad \text{ถ้า } y_1 = -1 \quad (3)$$

ในการแก้ปัญหาการจำแนกข้อมูล เป็นเรื่องที่ยากที่ข้อมูลบนพื้นที่ลักษณ์จะอยู่อย่างเป็นระเบียบและสามารถใช้เส้นตรงแบ่งได้ เพราะมักเป็นข้อมูลไม่เชิงเส้นและกระจัดกระจายอยู่บนพื้นที่ลักษณ์ ไม่สามารถใช้ซัพพอร์ตเวกเตอร์แมชชีนแบบเชิงเส้นได้ ดังภาพด้านล่าง ดังนั้นจึงต้องมีการจัดข้อมูลเสียใหม่ โดยให้อยู่ในพื้นที่มิติที่สูงขึ้น (Higher dimensional space) เพื่อให้สามารถสร้างเส้นแบ่งหรือระนาบแบ่งข้อมูลได้ เรียกระนาบนี้ว่า ระนาบเกิน (Hyperplane) ภาพที่ 2.9 เป็นการจำลองให้เห็นถึงการจัดข้อมูลให้อยู่ในมิติที่สูงขึ้น



ภาพที่ 2.9 แสดงการจัดข้อมูลให้อยู่ในมิติที่สูงขึ้น ดัดแปลงจาก (DTREG)

ในการจัดเรียงข้อมูลให้อยู่ในพื้นที่มิติที่สูงขึ้นนั้น ซัพพอร์ตเวกเตอร์แมชชีนใช้ฟังก์ชั่นการจับคู่ของเคอร์เนล (Kernel mapping function) โดยเคอร์เนลที่นิยมใช้มีอยู่ 3 ประเภท ดังนี้

โพลิโนเมียวฟังก์ชั่น (Polynomial function) คำนวณได้จากสมการที่ (4)

$$K(x_i,x_j) = (1+x_i \cdot x_j)^d \quad (4)$$

เรเดียลเบสสิสฟังก์ชั่น (Radial Basis Function: RBF) คำนวณได้จากสมการที่ (5)

$$K(x_i,x_j) = \exp(-\gamma \| x_i-x_j\|^2) \quad (5)$$

ซิกมอยด์ (Sigmoid)คำนวณได้จากสมการที่ (6)

$$K(x_i,x_j) = \tanh(\gamma \cdot x_i \cdot x_j+r) \quad (6)$$

2.6 ตัวอย่างงานประมวลผลภาษาธรรมชาติที่ผ่านมาที่ใช้ซัพพอร์ตเวกเตอร์แมชชีน

ในงานของ Giménez and Márquez (2003) ได้เสนอการกำกับหมวดคำโดยใช้ SVM light และใช้วิธีกำกับแบบ greedy left-to-right scheme จากการทดลองก็ประเมินผลได้ว่า ซัพพอร์ตเวกเตอร์แมชชีนสามารถฝึกฝนได้อย่างมีประสิทธิภาพโดยไม่ต้องปรับค่าพารามิเตอร์หลายครั้ง นอกจากนี้ยังสามารถกำกับคำได้มากกว่า 1,000 คำต่อวินาที ข้อมูลรับเข้าถูกแบ่งออกเป็นวินโดว์ วินโดว์ละ 7 คำ ลักษณ์ที่ใช้มีลักษณะเป็น n-gram pattern การศึกษานี้ยังได้ทำการทดลองเพื่อหาเคอร์เนลฟังก์ชั่นที่เหมาะสมระหว่างโพลีโนเมียลฟังก์ชั่นและลิเนียร์ฟังก์ชั่น (Linear function) ด้วยการปรับค่าพารามิเตอร์ของเคอร์เนลแล้วเปรียบเทียบประสิทธิภาพการกำกับหมวดคำ ผลปรากฏว่าโพลีโนเมียลฟังก์ชั่น ตั้งค่าดีกรีเท่ากับ 2 ให้ค่าความถูกต้องที่ 93.91 เปอร์เซ็นต์ ซึ่งมากกว่าลิเนียร์ฟังก์ชั่น ตั้งค่าดีกรีเท่ากับ 1 ที่ให้ค่าความถูกต้องที่ 93.84 เปอร์เซ็นต์ อย่างไรก็ตามโมเดลที่ใช้ลิเนียร์ฟังก์ชั่นใช้เวลาในการประมวลผลเร็วกว่าโมเดลที่ใช้โพลีโนเมียลฟังก์ชั่นประมาณ 3 เท่าตัว

งานทางด้านการรู้จำชื่อเฉพาะภาษาบังกาลีและฮินดีของ Ekbal and Bandyopadhyay (2008) ก็ได้นำเสนอการใช้ซัพพอร์ตเวกเตอร์แมชชีน ระบบนี้ทำการจำแนกชื่อเฉพาะออกเป็น 4 ประเภท ได้แก่ ชื่อบุคคล ชื่อสถานที่ ชื่อองค์กร และชื่ออื่นๆ เช่น วันที่ เวลา เปอร์เซ็นต์ จำนวนเงิน ฯลฯ ผลการทดสอบระบบได้ค่าความครบถ้วน ค่าความแม่นยำ และ f-score ที่ 88.61 เปอร์เซ็นต์, 80.12 เปอร์เซ็นต์ และ 84.15 เปอร์เซ็นต์ สำหรับภาษาบังกาลี และ 80.23 เปอร์เซ็นต์, 74.34 เปอร์เซ็นต์ และ 77.17 เปอร์เซ็นต์ สำหรับภาษาฮินดี ลักษณ์ที่ใช้ในการฝึกฝนซัพพอร์ตเวกเตอร์แมชชีนในงานนี้ได้แก่ 1) คำที่อยู่หน้าและหลังคำที่พิจารณา 2) ความยาวของปัจจัย (Suffix) ของคำที่พิจารณาและคำที่ล้อมรอบ 3) ความยาวของอุปสรรค (Prefix) ของคำที่พิจารณาและคำที่ล้อมรอบ 4) พิจาณาว่าเป็นคำที่ขึ้นต้นประโยคหรือไม่ 5) พิจารณาตัวเลข ถ้าเป็นตัวเลขตามด้วยเครื่องหมาย “, . / - เปอร์เซ็นต์” ก็ถือว่าเป็นชื่อเฉพาะ 6) คำที่มีความถี่ต่ำในการปรากฏในคลังข้อมูลมีแนวโน้มที่จะไม่ใช่ชื่อเฉพาะ 7) ความยาวของคำที่พิจารณา หากอักขระของคำที่พิจาณายาวน้อยกว่า 3 ตัว ก็มีแนวโน้มที่จะไม่ใช่ชื่อเฉพาะ 8) ข้อมูลประเภทของคำ

ทางด้านงานวิจัยภาษาไทยก็มีการใช้ซัพพอร์ตเวกเตอร์แมชชีนอย่างแพร่หลาย เช่นในงานของนิเวศ จิระวิชิตชัย, ปริญญา สงวนสัตย์ et al. (2553) ได้ทำการศึกษาการจัดหมวดหมู่เอกสารภาษาไทยแบบอัตโนมัติด้วยซัพพอร์ตเวกเตอร์แมชชีน มีการปรับค่าพารามิเตอร์ของเคอร์เนลฟังก์ชั่น

แบบต่าง ๆ และมีการใช้ Information Gain ในการลดลักษณ์ที่ใช้ทำการฝึกฝนซัพพอร์ตเวกเตอร์แมชชีน ผลปรากฏว่าโมเดลที่ใช้เคอร์เนลฟังก์ชั่นแบบลิเนียร์และโพลีโนเมียล ตั้งค่าดีกรีเท่ากับ 3 ให้ความแม่นยำในการจัดหมวดหมู่ 95.1 เปอร์เซ็นต์ เท่ากัน ส่วนเคอร์เนลฟังก์ชั่นแบบเรเดียลเบสิส ตั้งค่าแกมมา (Gamma) เท่ากับ 0.8 ให้ค่าความแม่นยำที่ 94.9 เปอร์เซ็นต์ นอกจากนี้เมื่อเปรียบเทียบการใช้ซัพพอร์ตเวกเตอร์แมชชีนกับอัลกอริทึม นาอีฟเบย์ และ C4.5 พบว่าซัพพอร์ตเวกเตอร์แมชชีนมีประสิทธิภาพในการจำแนกหมวดหมู่เอกสารมากกว่า

บทที่ 3

คลังข้อมูลและการกำกับข้อมูล

เนื่องจากการแยกอนุพากย์ด้วยซัพพอร์ตเวกเตอร์แมชชีน จะต้องมีการจัดทำคลังข้อมูลเพื่อใช้สำหรับฝึกฝนและทดสอบแบบจำลอง ดังนั้นในบทนี้ ผู้วิจัยจะนำเสนอการรวบรวมข้อมูลภาษาเขียนเพื่อจัดทำคลังข้อมูล จากนั้นในหัวข้อถัดไปจะกล่าวถึงการกำกับข้อมูลในคลังข้อมูลเพื่อใช้เป็นข้อมูลพื้นฐานสำหรับฝึกฝนและทดสอบแบบจำลอง

3.1 การจัดทำคลังข้อมูล

ข้อมูลภาษาที่ผู้วิจัยเก็บรวมรวมเพื่อจัดทำคลังข้อมูลในงานนี้ได้มาจากการรวบรวมข้อมูลภาษาเขียนในคลังข้อมูลภาษาไทยแห่งชาติ (Thai National Corpus: TNC) ซึ่งจัดทำโดยจุฬาลงกรณ์มหาวิทยาลัย คลังข้อมูลนี้ได้รวบรวมงานเขียนภาษาไทยประเภทต่าง ๆ เอาไว้ ได้แก่ งานเขียนวิชาการ งานเขียนกึ่งวิชาการ เรื่องแต่ง และอื่นๆ ข้อมูลภาษาไทยใน TNC เป็นข้อมูลแบบตัดคำและมีการกำกับข้อมูลต่าง ๆ ได้แก่ ขอบเขตของคำ คำอ่านออกเสียง ข้อมูลผู้เขียน และข้อมูลบรรณานุกรม

สำหรับข้อมูลภาษาที่ผู้วิจัยเลือกใช้ในการศึกษานั้นเป็นงานเขียนประเภทบทความวิชาการจากสาขาต่าง ๆ คือ สาขาวิทยาศาสตร์ รัฐศาสตร์ สังคมศาสตร์ และมนุษยศาสตร์ เนื่องจากงานเขียนแต่ละประเภทจะมีลักษณะการใช้ภาษาเขียนที่แตกต่างกันออกไป ซึ่งอาจส่งผลต่อโครงสร้างของหน่วยปริจเฉทพื้นฐาน ผู้วิจัยจึงเลือกเฉพาะงานวิชาการเท่านั้น เพราะงานเขียนวิชาการมีลักษณะการผูกความต่อเนื่องที่ซับซ้อนกว่างานประเภทอื่น มีการใช้ภาษาพูดและภาษาที่ไม่เป็นทางการน้อยกว่างานเขียนประเภทอื่น ข้อมูลงานเขียนที่ใช้จัดทำคลังข้อมูลนี้มีจำนวนไม่ต่ำกว่า 70,000 คำ และสามารถแบ่งหน่วยปริจเฉทพื้นฐานได้ไม่ต่ำกว่า 7,000 หน่วย

3.2 การกำกับคลังข้อมูล

เนื่องจากในการศึกษานี้ ผู้วิจัยต้องการตัดแบ่งอนุพากย์โดยวิธีการเรียนรู้ด้วยเครื่องจำเป็นต้องระบุลักษณ์ต่าง ๆ (Features) เพื่อให้เครื่องเรียนรู้และใช้ในการตัดสินใจ ดังนั้นผู้วิจัยจึงต้องทำการกำกับข้อมูลเพื่อใช้เป็นลักษณ์สำหรับการเรียนรู้ด้วยเครื่อง ได้แก่ การกำกับหมวดคำ (Parts of Speech) และขอบเขตอนุพากย์นอกจากนี้ได้ตัดการกำกับข้อมูลบางประเภทที่มีการกำกับอยู่ใน TNC ออกก่อนจะนำไปฝึกฝนแบบจำลอง ได้แก่ ข้อมูลคำอ่านออกเสียง ประเภทงานเขียน ข้อมูลผู้เขียน และข้อมูลบรรณานุกรม เนื่องจากเห็นว่าข้อมูลดังกล่าวไม่ได้เป็นส่วนหนึ่งของตัวเนื้อหาบทความ เป็นเพียงการให้ข้อมูลที่เกี่ยวข้องกับบทความเท่านั้น

3.2.1 หมวดคำภาษาไทยและการกำกับหมวดคำ

หมวดคำภาษาไทยสามารถจำแนกได้หลายรูปแบบ ขึ้นอยู่กับเกณฑ์ที่นักภาษาศาสตร์ใช้ในการจัดหมวดหมู่ เช่น การใช้เกณฑ์ทางวากยสัมพันธ์จะดูตำแหน่งการปรากฏของคำ คำที่ปรากฏในตำแหน่งเดียวกันได้ก็จะอยู่ในหมวดคำเดียวกัน หรือถ้าใช้เกณฑ์ทางความหมาย ก็จะดูความหมายของคำ คำที่มีความหมายในทำนองเดียวกันก็จะจัดให้อยู่ในหมวดคำเดียวกัน

จากการศึกษาของนัฐวุฒิ ไชยเจริญ (2544) เมื่อพิจารณาวิธีการและเกณฑ์ที่ใช้จัดหมวดคำ พบว่าสามารถแบ่งได้เป็น 2 กลุ่มใหญ่ ได้แก่ การจัดแบ่งหมวดคำโดยใช้ความรู้ทางภาษาของผู้จัดแบ่ง (Intuition-based approach) นักภาษาศาสตร์ที่ใช้วิธีนี้ได้แก่ พระยาอุปกิตศิลปะสาร (2533) บรรจบ พันธุเมธา (2514) กำชัย ทองหล่อ (2515) นววรรณ พันธุเมธา (2527) อุดม วโรตม์สิกขดิตถ์ (2535) และ เรืองเดช ปันเขื่อนขัติย์ (2541) และอีกกลุ่มหนึ่งคือการจัดแบ่งหมวดคำจากการวิเคราะห์คลังข้อมูลหรือประโยคทดสอบ (Corpus based approach) ได้แก่งานของ วิจินตน์ ภาณุพงศ์ (2532) อมรา ประสิทธิ์รัฐสิทธุ์ (2543) และ Sornlertlamvanich, Charoenporn et al. (1997)

สำหรับหมวดคำที่ผู้วิจัยใช้ในการกำกับหมวดคำในครั้งนี้ ประกอบไปด้วย 15 หมวดคำหลัก และสามารถแบ่งออกเป็นหมวดคำย่อยได้ 26 หมวดคำ[1] โดยพิจารณาความหมายและตำแหน่งการปรากฏของคำนั้นในข้อความ รวมไปถึงความสัมพันธ์ของคำนั้นและคำอื่นในข้อความ ดังนั้นคำหนึ่งคำจะไม่มีหมวดคำตายตัว นั่นคือ คำที่มีรูปเดียวกัน เช่นคำว่า “ถูก” สามารถเป็นเป็นคำประเภทใดก็ได้ขึ้นอยู่กับตำแหน่งการปรากฏและความสัมพันธ์กับคำอื่น ๆ เช่น เป็นคำกริยาวิเศษณ์ใน “เขาทำถูก” เป็นคำกริยาใน “ถูกเนื้อต้องตัว” เป็นคำช่วยกริยาใน “ถูกทำโทษ” เป็นต้น

ผู้วิจัยกำหนดสัญลักษณ์ที่ใช้ในการกำกับหมวดคำ คือ ในกรณีที่เป็นหมวดคำที่มีหมวดคำย่อยจะใช้อักษรย่อภาษาอังกฤษของหมวดคำหลักแล้วตามด้วยอักษรย่อชื่อหมวดคำย่อย เช่น หากหมวดคำหลักคือคำนาม หรือ Noun และหมวดคำย่อยคือ คำนามสามัญ หรือ Common Noun ก็จะใช้สัญลักษณ์ในการกำกับหมวดคำคือ NCMN ส่วนในกรณีที่เป็นหมวดคำที่ไม่มีหมวดคำย่อย ก็จะใช้อักษรย่อภาษาอังกฤษของหมวดคำนั้นในการกำกับข้อมูลเลย รายละเอียดของแต่ละหมวดคำพร้อมตัวอย่างภาษาเป็นดังนี้

1. **Noun (N) คำนาม** ประกอบไปด้วย 5 หมวดคำย่อย ดังนี้
   - **1.1 Common noun คำนามทั่วไป** สัญลักษณ์ที่ใช้กำกับหมวดคำคือ **NCMN** คำนามทั่วไปเป็นคำที่ใช้เรียกคน สัตว์ สิ่งของต่าง ๆ ใน

[1] ดัดแปลงจากชุดหมวดคำที่ใช้ในงาน CRSLP

ลักษณะทั่ว ๆ ไป ไม่ได้เฉพาะเจาะจงว่าเป็นคนไหนหรือสิ่งไหน คำนามทั่วไปสามารถปรากฏในตำแหน่งหน้ากริยา หลังกริยา หลังบุพบท และหน้าตัวกำหนด ตัวอย่างเช่น อาคาร, ธรรมชาติ, ศีลธรรม ฯลฯ

1.2 **Proper noun คำนามเฉพาะ** สัญลักษณ์ที่ใช้กำกับหมวดคำคือ **NPRP** คำนามเฉพาะเป็นชื่อเรียกที่ชี้เฉพาะหรืออ้างอิงถึงสิ่งหนึ่งสิ่งใดโดยเจาะจง สามารถปรากฏในตำแหน่งหน้ากริยา หลังกริยา หลังบุพบท และหน้าตัวกำหนด ตัวอย่างคำที่ขีดเส้นใต้ จังหวัดสระแก้ว, อาคารบรมราชกุมารี, สนามกีฬาราชมังคลา ฯลฯ

1.3 **Classifier ลักษณะนาม** สัญลักษณ์ที่ใช้กำกับหมวดคำคือ **NCLS** ลักษณะนามเป็นคำที่บอกหน่วยนับของคำนามทั่วไป ในภาษาไทยลักษณะนามจะแสดงลักษณะหรือชนิดของคำนามทั่วไปให้ชัดเจนยิ่งขึ้น ลักษณะนามสามารถปรากฏในตำแหน่งหลังจำนวนนับหรือคำบอกปริมาณ หน้าตัวกำหนด หน้าตัวนำส่วนเติมเต็ม ยกตัวอย่างคำที่ขีดเส้นใต้ นก 2 ตัว, พระ 2 รูป, รถยนต์คันนี้ ฯลฯ

1.4 **Number นามจำนวนนับ** สัญลักษณ์ที่ใช้กำกับหมวดคำคือ **NNUM** นามจำนวนนับเป็นคำบอกปริมาณของคำนามทั่วไปซึ่งในรูปของจำนวนนับ นอกจากนี้ยังหมายถึงตัวเลขต่าง ๆ ที่ไม่ได้แสดงปริมาณของคำนามทั่วไป นามจำนวนนับสามารถปรากฏในตำแหน่งหน้าลักษณะนาม หรือปรากฏโดด ๆ โดยที่ไม่ได้ประกอบคำนามก็ได้ ตัวอย่างคำที่ขีดเส้นใต้ อาย 30 ปี, มีเงิน 100 บาท, ประการที่สอง, เบอร์โทร 089999998 ฯลฯ

1.5 **Pronoun สรรพนาม** สัญลักษณ์ที่ใช้กำกับหมวดคำคือ **NPRO** คำสรรพนามเป็นคำที่ใช้เรียกแทนคำนาม สามารถปรากฏในตำแหน่งต่าง ๆ เช่นเดียวกันกับคำนามทั่วไปและคำนามเฉพาะ เช่น ฉัน, ดิฉัน, ผม, คุณ, ท่าน ฯลฯ

2. **Determiner ตัวกำหนด** สัญลักษณ์ที่ใช้กำกับหมวดคำคือ **DET** เป็นคำที่ใช้ประกอบคำนามและสรรพนามเพื่อให้แสดงความชี้เฉพาะ โดยจะปรากฏหลังคำที่ประกอบ ตัวอย่างคำที่ขีดเส้นใต้ บ้านหลังนี้, คนเหล่านั้น, สิ่งนั้น ฯลฯ

3. **Verb คำกริยา** สัญลักษณ์ที่ใช้กำกับหมวดคำคือ **VERB** คำกริยาเป็นคำที่แสดงการกระทำ อาการ การปรากฏ และสถานะของสิ่งต่าง ๆ คำกริยาสามารถปรากฏหน้าและ/หรือหลังคำนามหรือสรรพนาม หน้าบุพบท หน้าหรือหลังคำเชื่อม หน้าบุพบท หรือปรากฏโดด ๆ ก็ได้ ใน

วิทยานิพนธ์นี้ไม่ได้แบ่งคำกริยาออกเป็นหมวดย่อย เช่น สกรรมกริยา อกรรมกริยา ฯลฯ เนื่องจากเห็นว่ากริยาทุกประเภทที่มีสถานะเป็นกริยาแท้ สามารถเป็นภาคแสดง (Predicate) ของถ้อยความ (Utterance) ได้เหมือนกันหมด ประเภทย่อยของกริยาจึงไม่ได้เป็นตัวบ่งชี้สถานะความเป็นอนุพากย์ ตัวอย่างคำที่ขีดเส้นใต้เช่น นิดกินก๋วยเตี๋ยว, น้อยเดินออกไปข้างนอก, หนูตายในตาย, หน่อยเป็นหวัดกิน ฯลฯ

4. **Adjective คำคุณศัพท์** สัญลักษณ์ที่ใช้กำกับหมวดคำคือ **ADJ** คำคุณศัพท์เป็นคำที่ใช้ขยายคำนามเพื่อแสดงคุณสมบัติหรือลักษณะของคำนามนั้น ตำแหน่งที่ปรากฏคือหลังคำนาม ตัวอย่างคำที่ขีดเส้นใต้เช่น ความเดิมตอนที่แล้ว, เสื้อขาว, บ้านหลังเบ้อเร่อ, ความรู้เบื้องต้น ฯลฯ

5. **ADVERB กริยาวิเศษณ์** สัญลักษณ์ที่ใช้กำกับหมวดคำคือ **ADVERB** คำกริยาวิเศษณ์เป็นคำที่ใช้ขยายคำกริยาหรือคำคุณศัพท์หรือคำกริยาวิเศษณ์ด้วยกันเอง เพื่อแสดงคุณภาพ ลักษณะ และสถานะ คำกริยาวิเศษณ์จะปรากฏหลังคำที่ต้องการขยาย ตัวอย่างคำที่ขีดเส้นใต้เช่น ดีกว่า, สอบได้, วิ่งเร็ว, ชนะอย่างเด็ดขาด ฯลฯ

6. **Conjunction (C) สันธานหรือคำเชื่อม** แบ่งออกเป็น 4 หมวดคำย่อย ดังนี้

   6.1 Coordinating Conjunction สันธานประสาน สัญลักษณ์ที่ใช้กำกับหมวดคำคือ **CCOR** เป็นคำเชื่อมที่ใช้แสดงความสัมพันธ์ระหว่างปริจเฉทกับปริจเฉท หรืออนุพากย์กับอนุพากย์ หน่วยที่ถูกเชื่อมโยงโดยคำเชื่อมประเภทนี้จะมีสถานะหรือความสำคัญเท่ากัน การตัดคำเชื่อมออกไม่ส่งผลต่อใจความของทั้งสองหน่วย คำเชื่อมประเภทนี้จะปรากฏอยู่ส่วนหน้าสุดของปริจเฉทหรืออนุพากย์เท่านั้น ตัวอย่างเช่น ดังนั้น, อย่างไรก็ตาม, นอกจากนี้, แต่ ฯลฯ

   6.2 Subordinating conjunction อนุสันธาน สัญลักษณ์ที่ใช้กำกับหมวดคำคือ **CSUB** เป็นคำเชื่อมที่ใช้แสดงความสัมพันธ์ระหว่างอนุพากย์กับอนุพากย์หน่วยที่ถูกเชื่อมโยงโดยคำเชื่อมประเภทนี้จะมีสถานะหรือความสำคัญไม่เท่ากัน อนุพากย์ที่มีเนื้อความเด่นและสำคัญกว่าเป็นอนุพากย์หลัก ส่วนอนุพากย์ที่มีเนื้อความขยายอนุพากย์หลักเป็นอนุพากย์รอง การตัดคำเชื่อมออกจะส่งผลต่อใจความของทั้งสองหน่วย คำเชื่อมประเภทนี้มักจะปรากฏตำแหน่งหน้าอนุพากย์ ตัวอย่างเช่น ถึงแม้ว่า, ถ้า, เพราะ, หาก, เมื่อ, เพื่อ ฯลฯ

   6.3 Coordinating conjunction inside clause สันธานประสานภายในอนุพากย์ สัญลักษณ์ที่ใช้กำกับหมวดคำคือ **CCORIN** เป็นคำเชื่อมที่มีตำแหน่งการปรากฏภายนอนุพากย์ แสดงการเชื่อมโยง

ระดับวลีหรือคำ หน่วยที่ถูกเชื่อมโดยคำเชื่อมประเภทนี้จะมีสถานะหรือความสำคัญเท่ากัน ตัวอย่างคำที่ขีดเส้นใต้เช่น น้อยและนิดป่วยหนัก, จะไปเที่ยวสวนสนุกหรือสวนสัตว์ ฯลฯ

6.4 Subordinating conjunction inside clause อนุสันธานภายในอนุพากย์ สัญลักษณ์ที่ใช้กำกับหมวดคำคือ **CSBI** เป็นคำเชื่อมที่มีตำแหน่งการปรากฏภายนอนุพากย์ แสดงการเชื่อมโยงระดับวลีหรือคำ หน่วยที่ถูกเชื่อมโดยคำเชื่อมประเภทนี้จะมีสถานะหรือความสำคัญไม่เท่ากัน ตัวอย่างคำที่ขีดเส้นใต้เช่น เขาจึงเชื่อเช่นนั้น, ปัญหาก็คลี่คลาย ฯลฯ

7. **Preposition บุพบท** สัญลักษณ์ที่ใช้กำกับหมวดคำคือ **PREP** คำบุพบทเป็นคำที่ปรากฏหน้านามวลี ประกอบกันเป็นบุพบทวลีซึ่งมีความสัมพันธ์กับคำกริยาในฐานะที่เป็นส่วนเติมเต็มกริยา (Complement) หรือส่วนขยาย (Adjunct) ตัวอย่างเช่น ใต้, บน, ใน, จาก, เพื่อ ฯลฯ

8. **Auxiliary คำช่วยหน้ากริยา** สัญลักษณ์ที่ใช้กำกับหมวดคำคือ **AUX** คำช่วยหน้ากริยาเป็นคำที่เสริมความหมายและหน้าที่ทางไวยากรณ์ให้กับคำกริยาที่ประกอบ ได้แก่ กาล (Tense) การณ์ลักษณะ (Aspect) ทัศนภาวะ (Modality) วาจก (voice) เป็นต้น ตัวอย่างเช่น ควร, ต้อง, จะ, ยัง, เคย ฯลฯ

9. **Complementizer (COMPF) ตัวนำส่วนเติมเต็ม** แบ่งออกเป็น 2 หมวดคำย่อย ดังนี้

9.1 Finite complementizer ตัวนำส่วนเติมเต็มที่มีกริยาแท้ [2] สัญลักษณ์ที่ใช้กำกับหมวดคำคือ **COMPF** ได้แก่คำว่า ที่, ซึ่ง, อัน, ว่า, ให้

9.2 Non-finite complementizer ตัวนำส่วนเติมเต็มที่ไม่มีกริยาแท้ สัญลักษณ์ที่ใช้กำกับหมวดคำคือ **COMPNF** ตัวนำส่วนเติมเต็มที่ไม่มีกริยาแท้ ได้แก่คำว่า ที่ว่า, ที่จะ

10. **Prefix (PF) คำอุปสรรค** แบ่งออกเป็น 2 หมวดคำย่อย ดังนี้

10.1 Noun prefix อุปสรรคสร้างนาม สัญลักษณ์ที่ใช้กำกับหมวดคำคือ **PFN** อุปสรรคสร้างนาม ใช้ประกอบหน้าคำกริยาหรือคุณศัพท์เพื่อสร้างคำนาม ได้แก่คำว่า การ, ความ, นัก หรือใช้ประกอบหน้าอนุพากย์เพื่อทำให้เป็นนาม (Nominalization) ได้แก่คำว่า การที่, ความที่

[2] เรื่องกริยาแท้และกริยาไม่แท้จะได้กล่าวถึงในบทต่อไป

10.2 **Adverb prefix อุปสรรคสร้างวิเศษณ์** สัญลักษณ์ที่ใช้กำกับหมวดคำคือ **PFAV** อุปสรรคสร้างวิเศษณ์ ใช้ประกอบหน้าคำกริยาหรือคุณศัพท์เพื่อสร้างคำวิเศษณ์ ได้แก่ อย่าง, น่า

11. **Modifier (M) คำขยาย** หมายถึง คำขยายอื่น ๆ นอกเหนือจาก ADJ และ ADVERB แบ่งออกเป็น 3 หมวดคำย่อย

11.1 **Collective noun** สัญลักษณ์ที่ใช้กำกับหมวดคำคือ **MCN** คำขยายหน้านามเป็นคำที่ใช้ประกอบหน้าคำนามทั่วไปหรือคำนามเฉพาะเพื่อแสดงความเป็นกลุ่มของนามนั้น ตัวอย่างคำที่ขีดเส้นใต้เช่น บรรดาข้าราชการ, พวกนักเลง, เหล่านักเรียน ฯลฯ

11.2 **Modifier of number/classifier in front position** สัญลักษณ์ที่ใช้กำกับหมวดคำคือ **MNCF** คำขยายหน้าจำนวนเป็นคำที่ปรากฏหน้านามจำนวนนับหรือลักษณะนาม ตัวอย่างคำที่ขีดเส้นใต้เช่น เหลือเพียง 2 ตัว, แต่ละคน, อีกเรื่องหนึ่ง ฯลฯ

11.3 **Modifier of number/classifier in back position** สัญลักษณ์ที่ใช้กำกับหมวดคำคือ **MNCB** คำขยายหลังจำนวนเป็นคำที่ปรากฏหลังนามจำนวนนับหรือลักษณะนาม ตัวอย่างคำที่ขีดเส้นใต้เช่น อยู่คนเดียว, สองวันรวด, สองประการแรก, 2,000 กว่าบาท ฯลฯ

12. **Negation คำปฏิเสธ** สัญลักษณ์ที่ใช้กำกับหมวดคำคือ **NEG** คำปฏิเสธเป็นคำที่ใช้แสดงการปฏิเสธ ปรากฏตำแหน่งหน้าคำกริยาหรือคำช่วยหน้ากริยา ได้แก่คำว่า ไม่, มิ

13. **Punctuation เครื่องหมายวรรคตอน** สัญลักษณ์ที่ใช้กำกับหมวดคำคือ **PUNC** เครื่องหมายวรรคตอนต่าง ๆ ได้แก่ ไปยาลน้อย, ไปยาลใหญ่, อัญประกาศ, ปรัศนี, จุลภาค, ทับ, วงเล็บ, ไม้ยมก, ยัติภังค์ เป็นต้น

14. **Particle อนุภาคท้าย** สัญลักษณ์ที่ใช้กำกับหมวดคำคือ **PT** อนุภาคเป็นคำที่ปรากฏท้ายถ้อยความเพื่อแสดงความสุภาพหรือคำถาม ตัวอย่างเช่น ทราบหรือไม่, ไม่ใช่เลย, ไม่ว่าใครก็ตาม ฯลฯ

15. **Foreign word คำภาษาต่างประเทศ** สัญลักษณ์ที่ใช้กำกับหมวดคำคือ **FOREIGN** คำที่ใช้อักษรอื่น ๆ ที่ไม่ใช่อักษรภาษาไทย

3.2.2 การกำกับขอบเขตหน่วยปริจเฉทพื้นฐาน

ในส่วนของการกำกับขอบเขตอนุพากย์ ผู้วิจัยได้กำหนดขอบเขตของอนุพากย์ภาษาไทยโดยดัดแปลงหลักการจากคู่มือการแยกอนุพากย์ภาษาอังกฤษของ Carlson และ Marcu (2001) ที่ใช้ในการสร้างคลังข้อมูล RST Discourse Treebank ซึ่งเป็นคลังข้อมูลที่มีการกำกับหน่วยปริจเฉท

พื้นฐานเพื่อใช้ในการศึกษาโครงสร้างปริจเฉท โดยทั่วไปแล้วหน่วยปริจเฉทพื้นฐานได้แก่ อนุพากย์ที่มีกริยาแท้ และนามวลีที่ขึ้นต้นด้วยคำเชื่อมเด่น ส่วนรายละเอียดของการกำหนดขอบเขตหน่วยปริจเฉทพื้นฐานที่ใช้ในวิทยานิพนธ์นี้จะกล่าวถึงในบทที่ 4 สัญลักษณ์ที่ใช้ในการกำกับอนุพากย์ได้แก่ <EDU> สำหรับจุดเริ่มต้น และ </EDU> สำหรับจุดสิ้นสุด

สัญลักษณ์	คำอธิบาย
<w POS>	กำกับหมวดคำ
</w>	ขอบเขตคำ
<EDU>	จุดเริ่มต้น EDU
</EDU>	จุดสิ้นสุด EDU
<s>	ช่องว่าง (space)

ตารางที่ 3.1 แสดงสัญลักษณ์ที่ใช้ในการกำกับคลังข้อมูล

หลังจากกำหนดชุดหมวดคำและสัญลักษณ์ที่จะใช้ในการกำกับข้อมูล ผู้วิจัยได้ทำการกำกับข้อมูลต่าง ๆ ด้วยมือ ข้อมูลที่ได้รับการกำกับหมวดคำและขอบเขตอนุพากย์แล้ว มีรูปแบบดังตาราง 3.2 หรือสามารถดูการกำกับคลังข้อมูลเพิ่มเติมได้ที่ภาคผนวก ก

<EDU><w CCOR>โดยที่</w><s><w FOREIGN>The</w><s><w FOREIGN>Third</w><s><w FOREIGN>Way</w><s><w AUX>ยัง</w><w AUX>คง</w><w VERB>มี</w><w NOUN>สถานะ</w><w VERB>เป็น</w><w NOUN>งาน</w><w NCLS>ชิ้น</w><w ADJ>หลัก</w><w PREP>ของ</w><s><w FOREIGN>Giddens</w><s><w PREP>ใน</w><w NOUN>เรื่อง</w><w DET>นี้</w><w ADVERB>อยู่</w><w ADVERB>ต่อไป</w></EDU><s><EDU><w NCMN>งาน</w><w VERB>เขียน</w><w PREP>ของ</w><s><w FOREIGN>Giddens</w><s><w NEG>ไม่</w><w VERB>แพร่หลาย</w><w PREP>ใน</w><w NCMN>สังคม</w><w NPRP>ไทย</w></EDU>

ตารางที่ 3.2 ตัวอย่างข้อมูลที่กำกับข้อมูลแล้ว

คลังข้อมูลที่สร้างเสร็จสมบูรณ์มีจำนวนคำ (token) ทั้งหมด 76,460 คำ โดยนับเครื่องหมายวรรคตอนและช่องว่างเป็นคำด้วย นับจำนวน EDU ได้ทั้งหมด 8,102 EDU ในส่วนของการกำหนดขอบเขต EDU นั้น จะได้กล่าวถึงรายละเอียดในบทต่อไป

บทที่ 4

การกำหนดขอบเขตอนุพากย์ภาษาไทย

ในบทนี้จะกล่าวถึงการกำหนดขอบเขตเชิงโครงสร้างของอนุพากย์ ซึ่งเป็นหน่วยสร้างที่เล็กที่สุดในปริจเฉท ในที่นี้จะเรียกว่า EDU ทั้งนี้เพื่อที่จะใช้เป็นบรรทัดฐานในการตัดสินว่าหน่วยใดถือว่าเป็นหรือไม่เป็นอนุพากย์ที่ต้องแบ่ง ในงานนี้ ผู้วิจัยได้ใช้หลักการเดียวกันกับที่ Carlson and Marcu (2001) ใช้กำกับอนุพากย์ภาษาอังกฤษในคลังข้อมูล RST (RST Discourse Treebank) ซึ่งเป็นคลังข้อมูลที่ใช้สำหรับศึกษาโครงสร้างปริจเฉทโดยเฉพาะ ทั้งนี้ผู้วิจัยได้ปรับบางหลักการเพื่อให้เหมาะสมกับลักษณะของภาษาไทยโดยสรุปแล้ว อนุพากย์ปริจเฉท ได้แก่ อนุพากย์ที่มีกริยาแท้และกลุ่มของนามวลีที่ขึ้นต้นด้วยคำเชื่อมเด่น ทั้งนี้แต่ละ EDU จะต้องไม่มีขอบเขตคาบเกี่ยวกัน (Non-overlapping spans of text) รายละเอียดและหลักการแยกอนุพากย์ภาษาไทยที่ใช้ในงานวิจัย มีดังต่อไปนี้

1) อนุพากย์ที่มีกริยาแท้ (Finite clause)

กริยาแท้ หมายถึง กริยาที่สามารถทำหน้าที่เป็นภาคแสดงของอนุพากย์ สามารถแสดงข้อมูลทางไวยากรณ์ (Grammatical information) ต่าง ๆ ได้ เช่น กาล, การณ์ลักษณะ, วาจก, ทัศนภาวะ ฯลฯ ในภาษาไทยซึ่งเป็นภาษาโดด (Isolating language) กริยาแท้สามารถให้ข้อมูลทางไวยากรณ์ต่าง ๆ ที่กล่าวมาได้โดยการเติมกริยาช่วยประเภทต่าง ๆ ไว้ข้างหน้ากริยาแท้ เช่น “จะ” แสดงการณ์ลักษณะไม่สมบูรณ์, "เคย" แสดงสถานการณ์สมบูรณ์, "กำลัง" แสดงการณ์ลักษณะที่กำลังดำเนินอยู่, “ถูก” แสดงกรรมวาจก, “ควร” แสดงทัศนภาวะ ฯลฯ

โดยพื้นฐานแล้ว ผู้วิจัยกำหนดให้อนุพากย์ที่ประกอบด้วยกริยาแท้เป็น EDU (ยกเว้นอนุพากย์บางประเภทซึ่งจะกล่าวถึงภายหลัง) อนุพากย์ที่มีกริยาแท้ (Finite clause) สามารถแบ่งออกเป็นอนุพากย์อิสระ (Independent clause) และอนุพากย์ไม่อิสระ (Dependent clause) ในขณะที่อนุพากย์อิสระสามารถอยู่ได้ด้วยตัวเองอย่างมีใจความสมบูรณ์ อนุพากย์ไม่อิสระต้องพึ่งพาอนุพากย์อิสระเสมอ เพราะไม่สามารถอยู่ได้ด้วยตัวเอง อนุพากย์ทั้งสองประเภทสามารถเชื่อมเข้าด้วยกันด้วยหน่วยเชื่อมโยงปริจเฉทหรือคำเชื่อมที่แสดงความสัมพันธ์แบบนิวเคลียร์เดี่ยว (Mononuclear relation) กล่าวคือ อนุพากย์อิสระมีสถานะเป็นข้อมูลหลักหรือนิวเคลียร์ และอนุพากย์ไม่อิสระเป็นข้อมูลสนับสนุน นอกจากนี้อนุพากย์อิสระมากกว่าหนึ่งอนุพากย์สามารถเชื่อมเข้าด้วยกันด้วยหน่วยเชื่อมโยงปริจเฉทหรือคำเชื่อมที่แสดงความสัมพันธ์แบบหลายนิวเคลียร์ (Multinuclear relation) ได้อีกด้วย นั่นคืออนุพากย์อิสระแต่ละอนุพากย์มีสถานะเป็นข้อมูลหลักหรือนิวเคลียร์อย่างเท่าเทียมกัน ตัวอย่างคำเชื่อมที่แสดงความสัมพันธ์แบบนิวเคลียร์เดี่ยว ได้แก่ "เมื่อ”, “ถ้าหาก”, “แม้, “เพราะ"

เป็นต้น โดยอนุพากย์ที่ตามหลังคำเชื่อมเหล่านี้จะมีสถานะเป็นอนุพากย์ไม่อิสระ คำเชื่อมที่แสดงความสัมพันธ์แบบหลายนิวเคลียร์ ได้แก่ “และ”, “แต่”, “หรือ”, “อย่างไรก็ตาม”, “นอกจากนี้” เป็นต้น หลักการกำหนดขอบเขต EDU ที่เป็นอนุพากย์ที่มีกริยาแท้เป็นไปตามรายละเอียดดังต่อไปนี้

1.1) Finite relative clause (คุณานุประโยคที่มีกริยาแท้) เป็นอนุพากย์ไม่อิสระประเภทหนึ่งในทางหน้าที่คุณานุประโยคเป็นหน่วยสร้างที่มีหน้าที่ขยายคำนามหลัก (Head noun) และในทางความหมาย ถือว่าเป็นอนุพากย์ที่ทำให้คำนามหลักที่ถูกอ้างอิงถึงมีความเฉพาะจงเจาะยิ่งขึ้น หน่วยสร้างประเภทนี้จะปรากฏหลังคำนามหลักเท่านั้น และอาจจะมีหรือไม่มีตัวบ่งชี้ “ที่”, “ซึ่ง”, “อัน” นำหน้าก็ได้ ผู้วิจัยกำหนดให้คุณานุประโยคที่มีกริยาแท้เป็นภาคแสดงมีสถานะเป็น EDU คุณานุประโยคในภาษาไทยมีกลวิธีในการสร้าง 2 วิธี คือ กลวิธีคงสรรพนาม (Pronoun retention strategy) ซึ่งใช้สรรพนามอ้างอิงถึงคำนามหลักเป็นประธานของคุณานุประโยค และกลวิธีปล่อยว่าง (Gap strategy) ซึ่งจะละสรรพนามที่อ้างอิงถึงคำนามหลัก ตัวอย่าง (1) และ (2) ดังต่อไปนี้แสดงคุณานุประโยคทั้ง 2 ประเภทดังกล่าว

ตัวอย่าง

(1) [แม่มีสิ่ง]1[ที่เขาไม่ชอบ]2

(2) [คงจะต้องค้นหาจากวาทกรรม]1[ที่___กำลังเปลี่ยนไป]2

จากตัวอย่างที่ยกมา หน่วยที่ 2 ของตัวอย่าง (1) เป็นคุณานุประโยค และมีสรรพนาม “เขา” เป็นประธานในคุณนานุประโยคที่อ้างอิงถึงคำนามก่อนหน้าคือ “สิ่ง” ดังนั้นโครงสร้างคุณานุประโยคนี้เป็นโครงสร้างแบบคงสรรพนาม ส่วนหน่วยที่ 2 ของตัวอย่าง (2) เป็นคุณานุประโยคที่ขึ้นต้นด้วยตัวบ่งชี้ “ที่” และตามด้วยภาคแสดงเลย คือไม่มีการใช้สรรพนามอ้างอิงถึงคำนามข้างหน้าคือคำว่า “วาทกรรม” แบบนี้ถือว่าเป็นคุณานุประโยคที่ใช้กลวิธีปล่อยว่าง

ในบรรดาคำบ่งชี้คุณานุประโยคทั้งหลาย คำว่า “อัน” เป็นคำที่มีการใช้น้อยสุด และมักจะใช้ในภาษาที่มีความเป็นทางการมาก ในขณะที่ “ที่” มีการใช้หลากหลายรูปแบบมากกว่าในปริจเฉท และคำว่า “ซึ่ง” มักใช้ในการพูดหรือเขียนที่ค่อนข้างเป็นทางการมากกว่า Iwasaki and Horie (2005) ได้กล่าวถึงความแตกต่างเชิงหน้าที่ของคำว่า “ที่” และ “ซึ่ง” เอาไว้ว่า คำว่า “ที่” ใช้ในการระบุถึงคำนามหลักอย่างเจาะจง คำนามหลักมักจะเป็นนามรูปธรรม การใช้คำว่า “ที่” จะทำให้คุณานุประโยคมีลักษณะคล้ายกับ restrictive relative clause ในภาษาอังกฤษ นอกจากนี้ "ที่" ยังปรากฏบ่อยกว่าเมื่อทำหน้าที่ขยายคำนามหลัก ในขณะที่คำว่า “ซึ่ง”นอกจากจะมีการใช้น้อยกว่าและมีหน้าที่เหมือนคำว่า “ที่” ยังสามารถใช้นำหน้าอนุพากย์เพื่อขยายปริจเฉทก่อนหน้าได้ทั้ง

ปริจเฉทอีกด้วย นั่นคืออนุพากย์ตามหลัง “ซึ่ง” ไม่จำเป็นต้องทำหน้าที่ขยายเพียงคำนามหลักก่อนหน้าเท่านั้น

อย่างไรก็ตามคุณานุประโยคอาจจะไม่ได้ขึ้นต้นด้วยคำบ่งชี้ ทำให้มีโครงสร้างเหมือนคุณานุประโยคลดรูปซึ่งจะได้กล่าวถึงในหัวข้อต่อไป อีกทั้งยังสามารถมองได้ว่าเป็นนามวลีซับซ้อน แต่ไม่ว่าจะวิเคราะห์ให้เป็นคุณานุประโยคลดรูปหรือนามวลีซับซ้อน ความหมายที่ได้ก็ไม่ต่างกัน ในที่นี้ขอวิเคราะห์เป็นแบบหลังเพราะยึดรูปปรากฏเป็นสำคัญ อีกทั้งการลดรูปของคุณานุประโยคยังทำให้เกิดกริยาไม่แท้อีกด้วย ซึ่งในที่นี่ไม่เป็น EDU อยู่แล้ว

ตัวอย่าง

(3) [เขาชอบขนมปัง]1[ที่มาจากญี่ปุ่น]2

(4) [เขาชอบขนมปังมาจากญี่ปุ่น]1

ตัวอย่าง (3) และ (4) เป็นตัวอย่างสมมติ ที่มีใจความเหมือนหรือใกล้เคียงกัน “ที่มาจากญี่ปุ่น” ในตัวอย่าง (3) เป็นคุณานุประโยคที่ขึ้นต้นด้วยตัวบ่งชี้ “ที่” ทำหน้าที่ขยายคำนามหลัก “ขนมปัง” ในขณะที่ “ขนมปังมาจากญี่ปุ่น” ในตัวอย่าง (4) สามารถมองว่าเป็นนามวลีประเภทหนึ่งได้ โดยที่ “ขนมปัง” เป็นคำนามหลักและที่เหลือเป็นส่วนขยาย ดังนั้นในกรณีนี้ ผู้วิจัยจะแบ่งอนุพากย์ในตัวอย่าง (3) ออกมาเป็น 2 EDU และตัวอย่าง (4) ได้ EDU เดียว

1.2) Adverbial clause (วิเศษณานุประโยค) เป็นอนุพากย์ไม่อิสระประเภทหนึ่งที่ช่วยขยายกริยาหลักในอนุพากย์อิสระเกี่ยวกับเงื่อนไข เหตุผล เวลา เป็นต้น สามารถปรากฏตำแหน่งหน้าหรือหลังอนุพากย์อิสระได้ ผู้วิจัยกำหนดให้อนุพากย์ประเภทนี้เป็น EDU

วิเศษณานุประโยคสามารถแบ่งออกเป็นประเภทต่าง ๆ ตามความหมายและหน้าที่ซึ่งแต่ละประเภทจะใช้คำเชื่อมแสดงปริจเฉทสัมพันธ์ที่แตกต่างกันออกไป เช่น วิเศษณานุประโยคบอกเวลา (Time Adverbial clause) ขึ้นต้นด้วยคำเชื่อม“จนกระทั่ง”, “ในขณะที่”, “ขณะที่”, “ขณะ”, “ในระหว่างที่”, “ในระหว่าง” เป็นต้น วิเศษณานุประโยคบอกเหตุ (Causal Adverbial clause) ขึ้นต้นด้วยคำเชื่อม“เพราะ”, “เพราะว่า”, “เนื่องจาก”, “เนื่องจากว่า”เป็นต้น วิเศษณานุประโยคแสดงเงื่อนไข (Conditional Adverbial clause) ขึ้นต้นด้วยคำเชื่อม“ถ้า”, “หาก”, “ถ้าหาก”, “หากว่า”, “ถ้าหากว่า”เป็นต้น วิเศษณานุประโยคเงื่อนไขเพื่อยืนยัน (Concessive Adverbial clause) ขึ้นต้นด้วยคำเชื่อม“ถึงแม้ว่า”, “แม้ว่า”เป็นต้น คำเชื่อมเหล่านี้สามามารถใช้เป็นตัวบ่งชี้จุดเริ่มต้น EDU ได้

ตัวอย่าง

(5) [ต้องถูกส่งตัวกลับมาก่อน]1[เพราะป่วยเป็นวัณโรค]2

(6) [หากพิเคราะห์ให้ถี่ถ้วนแล้ว]1 [ก็ดูจะมีกลิ่นไอของวิธีคิดเรื่องตัวตนและอัตลักษณ์แบบสำนักคิดหลังสมัยใหม่]2

จากตัวอย่างข้างต้น หน่วยที่ 2 ในตัวอย่าง (3) เป็นคุณานุประโยคแสดงเหตุผล ขึ้นต้นอนุพากย์ด้วยคำเชื่อม “เพราะ” และหน่วยที่ 1 ในตัวอย่าง (6) เป็นคุณานุประโยคแสดงเงื่อนไข ขึ้นต้นด้วยคำเชื่อม “หาก” จึงถือว่าเป็น EDU

1.3) Coordinate clause คือ อนุพากย์มากกว่าหนึ่งอนุพากย์ที่มีความเท่าเทียมกันในเชิงหน้าที่เชื่อมโยงเข้าด้วยกันในโครงสร้างระดับเดียวกันด้วยสันธานประสานหรือคำเชื่อม เช่น “และ” “หรือ” “แต่” ฯลฯ ผู้วิจัยกำหนดให้อนุพากย์ที่เชื่อมด้วยคำเชื่อมเหล่านี้เป็น EDUอย่างไรก็ตามอนุพากย์ประเภทนี้อาจเชื่อมโยงเข้าด้วยกันโดยไม่มีคำเชื่อมเลยก็ได้ ซึ่งทำให้ยากแก่การระบุอนุพากย์ประเภทนี้

อย่างไรก็ตาม coordinate clause อาจมีรูปแบบของโครงสร้างที่คล้ายกับกริยาวลีที่เชื่อมเข้าด้วยกันด้วยสันธานประสานสำหรับ coordinate verb phrase นั้น ผู้วิจัยไม่ถือว่าเป็น EDU เนื่องจากกริยาในกริยาวลีไม่ได้มีปริจเฉทสัมพันธ์ต่อกันวิธีสังเกตว่าเป็น coordinate clause หรือ coordinate verb phrase สามารถพิจารณาได้จากกริยาหลัก กล่าวคือ ใน coordinate clause กริยาของแต่ละอนุพากย์จะไม่ใช้กรรมหรือส่วนขยายหรือส่วนเติมเต็มหรือส่วนเสริมร่วมกัน ส่วนกริยาใน coordinate verb phrase จะใช้กรรมหรือส่วนขยายหรือส่วนเติมเต็มหรือส่วนเสริมร่วมกันได้

ตัวอย่าง

(7) [เกิดในโตเกียว]1 [และมาเติบโตที่โอซากา]2

(8) [แต่หลายส่วนลอกและเพิ่มเติมมาจากกฎหมายตราสามดวง]1

ตัวอย่าง (7) เป็น coordinate clause ที่เชื่อมกันด้วยคำเชื่อม “และ”กริยาแต่ละอนุพากย์ไม่ได้ใช้ส่วนเสริมร่วมกัน นั่นคือ “ในโตเกียว” เป็นส่วนเสริมของกริยา “เกิด” และ “ที่โอซากา” เป็นส่วนเสริมของกริยา “มาเติมโต” จึงแบ่งได้ 2 EDU ส่วนตัวอย่าง (8) เป็น coordinate verb phrase ที่กริยา “ลอก” และ “เพิ่มเติม” เชื่อมเข้าด้วยกันด้วยคำเชื่อม “และ” ใช้ส่วนเติมเต็ม “จากกฎหมายตราสามดวง” ร่วมกัน ดังนั้นจึงไม่แยกออกเป็น 2 EDU

1.4) Subject and object clause เป็นอนุพากย์ไม่อิสระประเภทหนึ่งที่มีหน้าที่ในระดับโครงสร้างอนุพากย์ นั่นคือทำหน้าที่เป็นประธานหรือกรรมของอนุพากย์หลัก ไม่ได้ทำหน้าที่ขยายส่วนใดส่วนหนึ่งของข้อความ ในงานวิจัยนี้จะไม่ถูกแยกออกมาเป็น EDU เนื่องจากเป็นส่วนที่ไม่สามารถละทิ้งหรือแยกออกมาโดด ๆ ได้ อีกทั้งยังไม่ได้แสดงปริจเฉทสัมพันธ์ต่อส่วนใดอีกด้วย ดัง

ตัวอย่าง (9) อนุพากย์ที่ขีดเส้นใต้เป็นอนุพากย์ที่ทำหน้าที่เป็นประธานของอนุพากย์หลัก จึงไม่ถูกแยกออกเป็นเป็น EDU

ตัวอย่าง

(9) [ผู้จบปริญญาเอกด้านวิทยาศาสตร์ต้องมีคุณสมบัติอย่างไรบ้าง]1

2) อนุพากย์ที่ไม่มีกริยาแท้ (Non-finite clauses)

ตรงกันข้ามกับอนุพากย์ที่มีกริยาแท้ อนุพากย์ที่ไม่มีกริยาแท้จะไม่ถูกแยกออกมาเป็น EDU ในภาษาอังกฤษ อนุพากย์ที่ไม่มีกริยาแท้จะใช้กริยาในรูป participle, gerund, หรือ infinitive ซึ่งเป็นกริยาที่ไม่สามารถแสดงข้อมูลเชิงไวยากรณ์ได้[3] ในขณะที่ภาษาไทยไม่มีกริยาในรูปต่าง ๆ ที่กล่าวมา การระบุอนุพากย์ที่ไม่มีกริยาแท้ในภาษาไทยจึงมีความยุ่งยากอยู่บ้าง กล่าวคือ การระบุว่ากริยาในอนุพากย์เป็นกริยาแท้หรือไม่แท้ สามารถทำได้ด้วยวิธีเดียว คือการทดสอบว่ากริยานั้นสามารถปรากฏร่วมกับตัวบ่งชี้ต่าง ๆ ได้หรือไม่ เช่น ตัวบ่งชี้กาล, ตัวบ่งชี้การณ์ลักษณะ, ตัวบ่งชี้วาจก ฯลฯ หากไม่สามารถปรากฏร่วมกับตัวบ่งชี้เหล่านี้ได้ แสดงว่าเป็นกริยาไม่แท้

ในภาษาไทยสามารถพบอนุพากย์ที่ไม่มีกริยาแท้ได้ในคุณานุประโยค Yaowapat and Prasithrathsint (2006) เรียกคุณานุประโยคชนิดนี้ว่า "คุณานุประโยคแบบลดรูป" (Reduced relative clause) อนุพากย์ประเภทนี้จะปรากฏหลังคำนามหลักที่เป็นคำนามทั่วไปหรือไม่ชี้เฉพาะเท่านั้น (Generic or indefinite head noun) นอกจากนี้ยังพบว่าไม่มีการปรากฏหลังคำบ่งชี้ "ที่" เหมือนอย่างในกรณีคุณานุประโยคที่มีกริยาแท้

ตัวอย่าง

(9) [แสดงวิธีการวิเคราะห์**สารสำคัญ**]1

(10) [แสดงวิธีการวิเคราะห์สาร]1[ที่สำคัญ]2

จากตัวอย่าง (9) “สำคัญ” เป็นกริยาไม่แท้และเป็นคุณานุประโยคที่ขยายคำนามหลัก “สาร” ข้างหน้า ดังนั้นจึงไม่แยก “สำคัญ” ออกมาเป็น EDU อย่างไรก็ตามหากเติมตัวบ่งชี้ “ที่” หน้าคำว่า “สำคัญ” ก็จะเป็นดังตัวอย่าง (10) และทำให้เห็นภาพคุณานุประโยคชัดเจนยิ่งขึ้น แต่ก็จะมีคำถามตามมาว่า จะต้องแบ่ง “ที่สำคัญ” ออกเป็น EDU หรือไม่ ในเมื่อมีรูปโครงสร้างเหมือนกับคุณานุประโยคที่มีกริยาแท้ ในกรณีตัวอย่าง (10) ผู้วิจัยตัดสินใจแบ่ง “ที่สำคัญ” ออกมาเป็นอีก EDU เพราะถือว่าการที่ผู้เขียนเลือกใช้คำเชื่อม“ที่” แสดงว่าไม่ต้องการใช้เป็นแบบคุณานุประโยคลดรูปคือมองได้ว่าเป็นลีลาของผู้เขียนที่ใช้โครงสร้างคนละแบบกันในการสื่อสาร นั่นคือ แบบแรกใช้โครงสร้างคุณานุ

[3] English Grammar: An Introductory DescriPTion By MireiaLlinàsiGrau, Alan Reeves(page 74)

ประโยคแบบลดรูปหรือไม่มีกริยาแท้ และแบบที่สองใช้โครงสร้างคุณานุประโยคแบบมีกริยาแท้ ทั้งนี้ทั้งสองแบบไม่ได้ทำให้ความหมายแตกต่างกันเลย นอกจากนี้ การวิเคราะห์ตามรูปนี้ยังสะดวกต่อการแก้ปัญหาการตัดสินใจของเครื่อง

3) อนุพากย์เติมเต็ม (Clausal complements)

อนุพากย์เติมเต็มเป็นอนุพากย์ที่ทำหน้าที่เป็นส่วนเติมเต็มให้กับกริยาหรือคำนามที่ต้องการส่วนเติมเต็ม อนุพากย์เติมเต็มอาจอยู่ในรูปที่มีกริยาแท้หรือไม่มีกริยาแท้ก็ได้ หากมีกริยาแท้ อนุพากย์นั้นก็จะถูกแยกเป็น EDU แต่หากไม่มีกริยาแท้ อนุพากย์นั้นก็จะไม่ถูกแยกให้เป็น EDU

3.1)ส่วนเติมเต็มของกริยาที่มีกริยาแท้ (Finite clausal complement of verb) มักเป็นอนุพากย์ที่เป็นส่วนเติมเต็มของกริยาแสดงการรับรู้ (Cognitive verb) เช่น “คิด”, “เชื่อ”, “รู้”, “จินตนาการ”, “สมมติ”, “หวัง”, “คาด”, “ฝัน” ฯลฯและกริยาที่ใช้ในการรายงานคำพูด (Verb in reported speech) เช่น “พูด”, “ประกาศ”, “ชี้แจง”, “แนะนำ”, “รายงาน”, “อธิบาย”, “ถาม”, “บอก”, “กล่าว” ฯลฯ Carlson and Marcu (2001) เรียกกริยาทั้ง 2 ประเภทนี้ว่า attributive verb นอกจากนี้ อนุพากย์ชนิดนี้มักจะขึ้นต้นด้วยตัวนำหน้าส่วนเติมเต็ม “ว่า” ผู้วิจัยแยกอนุพากย์ประเภทนี้เป็น EDU เนื่องจากเป็นอนุพากย์ที่แสดงปริจเฉทสัมพันธ์ต่ออนุพากย์หลัก นั่นคือ Attributive relation ซึ่งเป็นความสัมพันธ์ที่แสดงรายละเอียดเกี่ยวกับกริยาที่ต้องการส่วนเติมเต็ม ทั้งหลายตัวอย่าง (11) ด้านล่างนี้ แสดงส่วนเติมเติมของกริยาที่มีลักษณะเป็นอนุพากย์ ซึ่งขึ้นต้นด้วยตัวนำหน้าส่วนเติมเต็ม “ว่า” ถูกแยกออกมาเป็น EDU หน่วยที่ 2

ตัวอย่าง

(11) [สรุปได้]1[**ว่า**สามารถจำแนกออกได้เป็น 2 แบบด้วยกัน]2

3.2) ส่วนเติมเต็มของกริยาที่ไม่มีกริยาแท้ (Non-finite clausal complement of verb) คือส่วนเติมเต็มของกริยาที่มีโครงสร้างเป็นอนุพากย์ที่มักปรากฏหลังคำบ่งชี้ “ที่จะ” “จะ” ซึ่ง Jenks (2006) เรียกคำบ่งชี้เหล่านี้ว่า infinitival complementizer หรือตัวนำส่วนเติมเต็มอนุพากย์ที่ไม่มีกริยาแท้ กริยาของอนุพากย์ที่ตามหลังคำบ่งชี้เหล่านี้จะไม่สามารถแสดงข้อมูลทางไวยากรณ์ได้ เพราะมีสถานะเป็นกริยาไม่แท้ คำกริยาหลักที่ต้องการส่วนเต็มเต็มประเภทนี้ ได้แก่ กริยาแสดงความปรารถนา เช่น “อยาก”, “"ชอบ”, “ต้องการ” ฯลฯ กริยาที่สื่อความหมายโดยนัย (Implicative verb) เช่น “พยายาม”, “ลอง” ฯลฯ กริยาช่วย (Modal verb) เช่น “สามารถ”, “ควร” ฯลฯ ผู้วิจัยกำหนดให้ส่วนเติมเต็มประเภทนี้ไม่ต้องถูกแยกออกมาเป็น EDU ตัวอย่าง (12) เป็นตัวอย่างของโครงสร้างที่ประกอบไปด้วยส่วนเติมเต็มประเภทนี้ โดยจะเห็นว่า “ที่จะละเลยเนื้อหา...” เป็นส่วนเติมเต็มของกริยา “เลือก” ดังนั้นจึงเป็นส่วนหนึ่งของ EDU ทั้งหมด

ตัวอย่าง

(12) [แต่ก็เลือก**ที่จะ**ละเลยเนื้อหามาตรฐานของพุทธศาสนา]1

3.3) อนุพากย์เติมเต็มของนาม (Clausal complement of noun) อนุพากย์ชนิดนี้มักจะขึ้นต้นด้วยคำนำหน้า “ที่ว่า” “ที่จะ” ทำหน้าที่เป็นส่วนเติมเต็มให้กับคำนามหลัก ผู้วิจัยจะไม่แยกอนุพากย์ประเภทนี้ออกเป็น EDU ยกตัวอย่าง “ที่จะปรับประยุกต์...” เป็นอนุพากย์เติมเต็มของคำนาม “ประสงค์” ดังนั้นจึงไม่ถูกแยกออกมาเป็น EDU

ตัวอย่าง

(13) [โดยมีวัตถุประสงค์**ที่จะ**ปรับประยุกต์เข้ากับตลาดทุนนิยม]1

4) โครงสร้างกริยาเรียง (Serial verb constructions)

โครงสร้างกริยาเรียงคือโครงสร้างที่มีกริยามากกว่าหนึ่งตัวปรากฏเรียงกันโดยไม่มีอะไรแทรกระหว่างกริยายกเว้นกรรมของกริยาตัวก่อนหน้า กริยาที่เรียงกันทุกตัวเป็นกริยาแท้ ยกเว้นกริยาบางตัวที่ผ่านกระบวนการกลายเป็นคำไวยากรณ์ เช่น “เสีย”, “อยู่”, “ว่า”, “ให้” ฯลฯ แม้กริยาเหล่านี้จะกลายเป็นคำไวยากรณ์แล้ว (Grammaticalized verb) แต่ก็ยังถือว่าเป็นส่วนหนึ่งของกริยาเรียง นอกจากนี้กริยาเรียงจะมีการรวมคุณสมบัติทางความหมายของกริยาทุกตัวเข้าด้วยกันและร่วมกันแสดงถึงเหตุการณ์เพียงเหตุการณ์เดียว (Single event) จึงนับกริยาที่เรียงกันนั้นเป็นหน่วยสร้างเดียวกัน (Foley and Mike 1985, Thepkanjana 1986, Takahashi 2009, Pongsutthi, Ketui et al. 2013) สำหรับงานวิจัยนี้ จะไม่แยกกริยาแต่ละตัวเป็น EDU ด้วยเหตุผลดังกล่าว

ตัวอย่าง

(14) [ขณะเดียวกันก็**รอคอย**โชคชะตา**มาพลิกผัน**ชีวิต**ให้แปรเปลี่ยนไป**]1

(15) [ฉินตะโร**คิด**]1[**ว่า**แม่เริ่มมีสภาพทรุดโทรมทั้งด้านร่างกายและจิตใจ]2

จากตัวอย่าง (14) จะเห็นว่า “รอคอย__มาพลิกผัน__ให้แปรเปลี่ยนไป” เป็นโครงสร้างกริยาเรียง โดยมีคำนาม “โชคชะตา” และ “ชีวิต” แทรกกลางระหว่างกริยาและเป็นกรรมตรงของกริยาที่นำหน้า ตัวอย่าง (14) จึงถือว่ามีเพียง EDU เดียว อย่างไรก็ตาม หากโครงสร้างกริยาเรียงประกอบไปด้วยคำกริยาแสดงการรับรู้หรือคำกริยาที่ใช้ในการรายงาน ก็มักจะปรากฏคำกริยาที่กลายเป็นคำไวยากรณ์หรือตัวนำส่วนเติมเต็ม "ว่า" และตามด้วยอนุพากย์เติมเต็ม ในกรณีนี้ ผู้วิจัยจะแบ่งตั้งแต่ตัวนำส่วนเติมเต็มตามด้วยอนุพากย์เติมเต็มออกเป็นอีก EDU หนึ่งเนื่องจากอนุพากย์เติมเต็มเหล่านั้นแสดงปริจเฉทสัมพันธ์ที่เรียกว่า Attributive relation ซึ่งเป็นความสัมพันธ์ที่แสดงรายละเอียดของคำกริยาก่อนหน้า ดังจะเห็นได้จากตัวอย่าง (15) แม้ว่า “คิดว่า” เป็นโครงสร้างกริยาเรียง แต่เนื่องจาก “คิด” เป็นกริยาแสดงการรับรู้ที่ต้องการอนุพากย์เติมเต็มเพื่อทำให้ใจความสมบูรณ์ และคำ

ว่า “ว่า” เป็นกริยาที่กลายเป็นคำไวยากรณ์แล้ว นั่นคือเป็นตัวนำส่วนเติมเต็ม ตามด้วยอนุพากย์เติมเต็ม “แม่เริ่มมีสภาพทรุดโทรม...” ดังนั้นจึงแบ่ง “ว่าแม่เริ่มมีสภาพทรุดโทรม...” ออกเป็นอีก EDU หนึ่ง ทำให้ตัวอย่าง (15) ประกอบไปด้วย 2 EDU

5) โครงสร้างเคลฟต์ (Clefts)

โครงสร้างเคลฟต์เป็นโครงสร้างภาษาที่แสดงการเน้นส่วน โดยทั่วไปโครงสร้างเคลฟต์จะมีลักษณะเป็นโครงสร้างซับซ้อน นั่นคือประกอบด้วยอนุพากย์มากกว่าหนึ่ง ในที่นี้จะขอยกคำอธิบายเกี่ยวกับโครงสร้างเคลฟในภาษาไทยและตัวอย่างประโยคจากงานวิจัยของ Ruangjaroon (2005) ดังนี้

โครงสร้างเคลฟต์ในภาษาไทยประกอบไปด้วยส่วนที่เรียกว่า“เคลฟที” (cleftee) คำกริยา “เป็น” หรือ “คือ” และส่วนที่เรียกว่า “อนุพากย์เคลฟต์” (cleft clause) ซึ่งมีโครงสร้างเป็นคุณานุประโยค เมื่อพิจารณาประเภทของ copula verb “เป็น” และ “คือ” แล้ว สามารถแบ่งเคลฟต์ในภาษาไทยได้ 2 ประเภท คือ contrastive (or specificational) cleft และ identificational cleft มีรายละเอียดดังต่อไปนี้

ประเภทแรกคือ **contrastive (or specificational) cleft**เรียกชื่อประเภทเคลฟต์ตาม copula verb “เป็น” ซึ่งเป็นกริยาที่เชื่อมประธานและส่วนเติมเต็มประธาน โครงสร้างของเคลฟต์ประเภทนี้ คือ**เคลฟที + definite marker “ที่” + copula “เป็น”+อนุพากย์เคลฟต์** ในส่วนของเคลฟทีจะต้องมีลักษณะของความหมายเป็น [+human] เท่านั้น และในส่วนของอนุพากย์เคลฟต์จะต้องเป็น nominalized clause หรืออนุพากย์ที่ถูกทำให้กลายเป็นคำนาม ทั้งนี้เพราะ copula verb ต้องการส่วนเติมเต็มประธาน (Subject complement) ในรูป predicative nominal เท่านั้น โดยอนุพากย์เคลฟต์ในภาษาไทยจะใช้ nominalizer “คน” ขึ้นต้นอนุพากย์เพื่อทำให้อนุพากย์นั้นกลายเป็นคำนาม

ตัวอย่าง

(16) [นิกที่เป็นคนทำจานแตก]1

(17) [ใครที่เป็นคนทำจานแตก]1

จากตัวอย่าง (16) “นิก” เป็นเคลฟที ซึ่งมีคำว่า “ที่” เป็นตัวบ่งชี้แสดงการชี้เฉพาะ (Definite marker) ว่าหมายถึงนิกไหน อนุพากย์เคลฟต์คือ "คนทำจานแตก" ซึ่งเป็น nominalized clause ที่มีพฤติกรรมเหมือนคำนามทั่วไป คำว่า “คน” ทำหน้าที่เป็น nominalizerเคลฟทีและอนุพากย์เคลฟต์ถูกเชื่อมเข้าด้วยกันด้วย copula "เป็น"ส่วนในตัวอย่าง (17) ก็เหมือนตัวอย่างที่ (16) แตกต่างกันตรงที่เป็น Wh-cleft คืออยู่ในรูปของคำถาม มีเคลฟทีคือ “ใคร”

เคลฟต์อีกประเภทหนึ่งคือ **identificational cleft**หรือเคลฟต์ที่แสดงการชี้เฉพาะเจาะจง โครงสร้างของเคลฟต์คือ**เคลฟที + copula “คือ”+อนุพากย์เคลฟต์** ในส่วนของอนุพากย์เคลฟต์นั้นจะคุณานุประโยค คือประกอบไปด้วยคำนามหลักหรือคำลักษณะนาม ตามด้วยคำบ่งชี้ “ที่” และตามด้วยอนุพากย์ ตัวอย่างของเคลฟต์ประเภทนี้จากคลังข้อมูลเป็นดังตัวอย่าง (18)

ตัวอย่าง

(18) [นั่นแหละคือสิ่ง]1[ที่ทำให้เกิดภาพเหมือน]2

นอกจากนี้ Ruangjaroon (2005) ยังได้กล่าวถึงการใช้ copula “เป็น” ในบริบทโครงสร้างที่เป็น identificational cleft และ copula “คือ” ในบริบทโครงสร้างที่เป็น contrastive cleft ว่าเป็นไปไม่ได้ เพราะไม่สามารถเข้ากันได้และจะทำให้ได้รูปภาษาที่ไม่ถูกต้อง (Ill-formed) ดังตัวอย่างที่ (19b) และ (20b) ซึ่งเป็นรูปภาษาที่ไม่ถูกต้อง ส่วนตัวอย่าง (19a) และ (20a) มีรูปภาษาที่ถูกต้อง

ตัวอย่าง

(19) กรณี copula “เป็น” ใน identificational cleft

(a) นิกคือคนที่ฉันรัก (รูปภาษาที่ถูกต้อง)

(b) *นิกเป็นคนที่ฉันรัก (รูปภาษาที่ไม่ถูกต้อง)

(20) กรณี copula “คือ” ใน contrastive cleft

(a) นิกที่เป็นคนทำ (รูปภาษาที่ถูกต้อง)

(b) *นิกที่คือคนทำ (รูปภาษาที่ไม่ถูกต้อง)

แม้ว่า Carlson and Marcu (2001) จะกำหนดให้ไม่ต้องแยกประโยคเคลฟต์ออกเป็น 2 EDU เนื่องจากทั้งสองอนุพากย์ในโครงสร้างเคลฟต์ไม่ได้มีปริจเฉทสัมพันธ์ต่อกัน แต่สำหรับภาษาไทยผู้วิจัยตัดสินใจแบ่งอนุพากย์ในโครงสร้างเคลฟต์แบบ identificational cleft ออกเป็น 2 EDU เพราะคิดว่าเคลฟต์ประเภทนี้สามารถมองเป็นโครงสร้างที่มีคุณานุประโยคขยายคำนามหลักได้ โดยที่อนุพากย์ที่เป็นคุณานุประโยคมีความสัมพันธ์แบบ elaboration relation กับคำนามหลักที่ขยาย ดังเช่นตัวอย่าง (18) เป็นเคลฟต์ที่ถูกแยกออกเป็น 2 EDU โดยที่หน่วยที่ 2 เป็นคุณานุประโยคที่ขยายคำนามหลัก “สิ่ง”แต่สำหรับ contrastive cleft นั้น ผู้วิจัยจะไม่แยกอนุพากย์ดังตัวอย่าง (16) และ (17) เนื่องจากเห็นว่าอนุพากย์เคลฟต์ “คนทำจานแตก” สามารถมองได้ว่าเป็นคำนามที่เป็นส่วนเติมเต็มประธาน

6) นามวลีที่มีสถานะเป็น EDU (Phrasal EDU)

โดยทั่วไปแล้วนามวลีเป็นหน่วยสร้างระดับที่เล็กกว่าอนุพากย์ ประกอบไปด้วยคำนามหลักและ/หรือส่วนขยายเท่านั้น จึงไม่ถูกแยกให้เป็น EDU ยกเว้นนามวลีที่ขึ้นต้นด้วยคำเชื่อมที่ Carlson

and Marcu (2001) เรียกว่า strong marker หรือคำ เชื่อมเด่น ซึ่งเป็นหน่วยเชื่อมโยงปริจเฉทที่มีความหมายบ่งบอกถึงปริจเฉทสัมพันธ์ระหว่างกลุ่มนามวลีที่ตามหลังคำเชื่อมกับถ้อยความก่อนหน้า และเรียกนามวลีที่ขึ้นต้นด้วยคำเชื่อมเด่นนี้ว่า phrasal EDU หรือ EDU ที่เป็นนามวลี

จากการศึกษาหน่วยเชื่อมโยงในภาษาไทย พบว่ามีคำเชื่อมเด่นอยู่เพียง 2 ประเภท คือ คำเชื่อมแสดงตัวอย่าง ได้แก่คำที่ใช้แสดงตัวอย่างต่าง ๆ เช่น “ได้แก่”, “เช่น”, “ตัวอย่างเช่น ฯลฯ” และคำเชื่อมแสดงวัตถุประสงค์ เช่นคำว่า “เพื่อ” นามวลีที่ขึ้นต้นด้วยคำเชื่อมเด่น 2 ประเภทนี้จะแยกให้เป็น EDU ดังตัวอย่าง (21) กลุ่มนามวลีที่ขึ้นต้นด้วยคำเชื่อมแสดงตัวอย่าง “เช่น” ถูกแยกให้เป็นอีก EDU หนึ่ง มีหน้าที่ระดับปริจเฉทคือการแสดงให้เห็นตัวอย่างของสิ่งที่ถูกกล่าวถึงก่อนหน้า นั่นคือ “ปรากฏการณ์ธรรมชาติ”

ตัวอย่าง

(21) [ตำนานปรัมปราเป็นการอธิบายถึงกำเนิดของจักรวาล โครงสร้าง และระบบของจักรวาล มนุษย์ สัตว์ ปรากฏการณ์ทางธรรมชาติ]1[เช่น ลม ฝน กลางวัน กลางคืน ฟ้าร้อง ฟ้าผ่า]2

นอกจากนี้ ยังมีนามวลีอื่น ๆ ที่ไม่ได้ขึ้นต้นด้วยตัวเชื่อมเด่น แต่ก็ถือว่าเป็น EDU ได้เช่นกัน นามวลีที่กล่าวถึงนี้ ได้แก่ นามวลีที่อยู่ในเครื่องหมายวงเล็บ นามวลีที่เป็นชื่อหัวข้อ/ชื่อเรื่อง/ชื่อผู้เขียนที่แยกอยู่ลำพัง และไม่ได้เป็นทำหน้าที่เป็นประธานหรือกรรมในหน่วยสร้างใด นามวลีประเภทหลังนี้พบได้โดยทั่วไปในงานเขียนและมีหน้าที่ที่ชัดเจนในระดับปริจเฉท เช่น ชื่อเรื่องของงานเขียนกับเนื้อหาทั้งหมดของงานเขียนนั้น หรือชื่อหัวข้อกับเนื้อหาในหัวข้อนั้น แต่ละคู่มีปริจเฉทสัมพันธ์กันแบบที่ Carson and Marcu เรียกว่า Textual Organization relation ดังนั้นในวิทยานิพนธ์นี้จึงแยกนามวลีดังกล่าวเป็น EDU ด้วยเช่นกัน ตัวอย่างข้อมูลจากคลังข้อมูลเป็นดังนี้

ตัวอย่าง

(22) [อพยพมาจากเวียงจันทร์]1 [(ลาวเวียง)]2

(23) กรณีชื่อบทความ ชื่อผู้เขียน และเนื้อหา

[พลวัตของความรู้ชาวบ้านในกระแสโลกาภิวัตน์]1

[อานันท์ กาญจนพันธุ์]2

[เนื้อหาของบทความ]3 [เนื้อหาของบทความ]4 [เนื้อหาของบทความ]5 ....

จากตัวอย่าง จะเห็นว่า “(ลาวเวียง)” ในตัวอย่าง (22) เป็นนามวลีที่อยู่ในเครื่องหมายวรรคเล็บจึงต้องแยกออกมาเป็น EDU และในตัวอย่าง (23) “พลวัตของความรู้ชาวบ้าน...” เป็นชื่อบทความ “อานันท์ กาญจนพันธุ์” เป็นชื่อผู้เขียนบทความ ดังนั้นจึงถือว่าเป็น EDU

7) เครื่องหมายวรรคตอน (Punctuations)

เครื่องหมายวรรคตอนสามารถช่วยในการระบุขอบเขต EDU ได้ เนื่องจากเครื่องหมายวรรคตอนบางตัวจะปรากฏอยู่ในตำแหน่งที่แน่นอน ตัวอย่างเช่น เครื่องหมายวงเล็บมักจะอยู่หน้าและหลังอนุพากย์เสมอ เครื่องหมายคำถามมักจะอยู่ท้ายประโยคภาษาอังกฤษเสมอ เป็นต้น สำหรับภาษาไทยนั้น ราชบัณฑิตยสถานได้กล่าวถึงเครื่องหมายวรรคตอนที่มีการใช้ในภาษาไทยเอาไว้ในหนังสือ “หลักเกณฑ์การใช้เครื่องหมายวรรคตอนและเครื่องหมายอื่น ๆ หลักเกณฑ์การเว้นวรรค หลักเกณฑ์การเขียนคำย่อ” ฉบับราชบัณฑิตยสถาน (2548) พิมพ์ครั้งที่ ๖ ซึ่งสรุปรายละเอียดอย่างย่อได้ดังตารางต่อไปนี้

ชื่อภาษาไทย	**ชื่อภาษาอังกฤษ**	**เครื่องหมาย**	**วิธีใช้**
มหัพภาค	full stop, period	.	ใช้แสดงการจบประโยคหรือจบความ
จุด	dot, point	.	1) ใช้เขียนหลังตัวอักษรเพื่อแสดงว่าเป็นอักษรย่อ 2) ใช้เขียนข้างหลังตัวอักษรหรือตัวเลขที่บอกลำดับข้อ 3) ใช้คั่นระหว่างชั่วโมงกับนาทีเพื่อบอกเวลา 4) ใช้ในเลขทศนิยม 5) ใช้บอกว่าเป็นอักษรนำอักษรควบ ในการบอกคำอ่าน 6) ใช้บอกว่าเป็นอักษรควบหรือเป็นตัวสะกดในการเขียนภาษาบาลีสันสกฤตด้วยอักษรไทย 7) ใช้เขียนเพื่อแสดงการละตัวอักษรหรือข้อความ
จุลภาค หรือจุดลูกน้ำ	comma	,	1) ใช้แยกวลีหรืออนุพากย์ 2) ใช้คั่นคำในรายการตั้งแต่ 3 รายการขึ้นไป 3) ใช้ในการเขียนบรรณานุกรม ดรรชนี และนามานุกรม 4) ใช้คั่นจำนวนเลขนับจากหลักหน่วยไปทีละ 3 หลัก
อัฒภาค	semicolon	;	1) ใช้แยกประโยคเปรียบเทียบออกจากกัน 2) ใช้คั่นระหว่างประโยคที่มีรูปประโยคและใจความสมบูรณ์อยู่แล้วเพื่อแสดงความต่อเนื่องอย่างใกล้ชิดของประโยคนั้น 3) ใช้แบ่งประโยค กลุ่มคำ หรือกลุ่มตัวเลขที่มีเครื่องหมายจุลภาคอยู่แล้วออกเป็นส่วนให้เห็นชัดเจนยิ่งขึ้น เพื่อกันความสับสน 4) ใช้คั่นคำในรายการที่มีจำนวนมาก ๆเพื่อจำแนกรายการออกเป็นพวก ๆ 5) ใช้ในหนังสือประเภทพจนานุกรมเพื่อคั่นบทนิยามของคำที่มีความหมายหลายอย่าง
ทวิภาค	colon	:	1) ใช้ไขความแทนคำว่า "คือ" หรือ "หมายถึง" 2)

			ใช้หลังคำ "ดังนี้" "ดังต่อไปนี้" เพื่อแจกแจงรายการ 3) ใช้คั่นบอกเวลา
ต่อ	-	:	1) ใช้แสดงอัตราส่วนและมาตราส่วน 2) ใช้แสดงสัดส่วน 3) ใช้แสดงปฏิภาค
วิภัชภาค	-	:-	1) ใช้หลังคำ "ดังนี้" "ดังต่อไปนี้" เพื่อแจกแจงรายการโดยรายการที่ตามหลังเครื่องหมายให้ขึ้นย่อหน้าใหม่
ยัติภังค์	hyphen	-	1) ใช้เขียนไว้สุดบรรทัดเพื่อต่อพยางค์ ซึ่งจำเป็นต้องเขียนแยกบรรทัดกันเนื่องจากเนื้อที่จำกัด 2) ใช้เขียนแยกพยางค์เพื่อบอกคำเต็มที่จำเป็นต้องแยกตามฉันทลักษณ์ 3) ใช้แยกพยางค์เพื่อบอกคำอ่าน 4) ใช้แสดงคำที่ละส่วนหน้าหรือส่วนท้ายหรือทั้งส่วนหน้าและส่วนท้ายของคำ 5) ใช้ในความหมายว่า "ถึง" เพื่อแสดงช่วงเวลา จำนวน สถานที่ 6) ใช้เขียนแยกกลุ่มตัวเลขตามรหัสที่กำหนดไว้ เช่น เลขบัญชีธนาคาร 7) ใช้กระจายอักษรเพื่อให้เห็นว่าคำนั้นประกอบด้วยพยัญชนะสระและวรรณยุกต์อะไรบ้าง
ปรัศนี	question mark	?	1) ใช้เขียนท้ายความหรือประโยคที่เป็นคำถาม 2) ใช้เขียนในเครื่องหมายวงเล็บหลังข้อความเพื่อแสดงความสงสัยหรือไม่แน่ใจ
ไปยาลน้อย หรือเปยยาลน้อย	-	ฯ	1) ใช้ละคำ 2) ใช้ในคำ "ฯพณฯ"
ไปยาลใหญ่ หรือเปยยาลใหญ่	-	ฯลฯ	ใช้ละข้อความที่อยู่ในประเภทเดียวกัน
ไมยมกหรือยมก	-	ๆ	ใช้เขียนหลังคำวลี หรือประโยค เพื่อให้อ่านซ้ำอีกครั้ง
วงเล็บ หรือนขลิขิต	parenthesis	( )	1) ใช้กันข้อความที่ขยายหรืออธิบายจากข้อความอื่น 2) ใช้ขยายความให้ชัดเจนยิ่งขึ้น 3) ใช้กันตัวอักษรหรือตัวเลขที่เป็นหัวข้อย่อยอาจใช้เพียงวงเล็บปิดข้างเดียวก็ได้
อัศเจรีย์	exclamation mark	!	1) ใช้เขียนหลังคำ วลี หรือประโยคที่เป็นคำอุทาน 2) ใช้เขียนหลังคำเลียนเสียงธรรมชาติ 3) ใช้เขียนหลังข้อความสั้น ๆ ที่ต้องการเน้น

วงเล็บเหลี่ยม	square brackets	[ ]	ใช้กับสูตรคณิตศาสตร์การเขียนโปรแกรม สัทอักษรสากล
วงเล็บปีกกา	braces	{ }	ใช้กับสูตรคณิตศาสตร์การเขียนโปรแกรม
วงเล็บสามเหลี่ยม	-	<>	ใช้เป็นเครื่องหมายน้อยกว่าและมากกว่าใช้ในการเขียนโปรแกรม
บุพสัญญา	ditto mark	"	ใช้เขียนแทนคำหรือประโยคที่อยู่ในบรรทัดบนเพื่อที่จะไม่ต้องเขียนซ้ำอีก
สัญประกาศ	underscore	_	ใช้ขีดเส้นใต้ข้อความสำคัญหรือข้อความที่ควรสังเกตพิเศษ
เสมอภาค	equal	=	ใช้เพื่อแสดงความเทียบเท่ากับของสิ่งที่อยู่ทางซ้ายและขวา
อัญประกาศ	quotation	“ ”	ใช้กันคำวลี หรือข้อความที่ต้องการเน้นเป็นพิเศษ สามารถใช้ได้ทั้งอัญประกาศคู่ “ ”หรืออัญประกาศเดี่ยว ' '
ทับ	Slash	/	1) ใช้เขียนคั่นระหว่างทางเลือก 2) ใช้คั่นตัวเลข
ยัติภาค	Dash	--	1) ใช้ในความหมาย "และ" หรือ "กับ" 2) ใช้ขยายความข้างหน้า 3) ใช้ในความหมายว่า "ถึง" เช่นเดียวกับยติภังค์ 4) ใช้เป็นสัญลักษณ์นำหัวข้อย่อยที่ไม่ต้องการเรียงลำดับ

ตารางที่ 4.1 แสดงเครื่องหมายวรรคตอนที่ใช้ในภาษาไทย

นอกจากนี้ยังมีเครื่องหมายวรรคตอนอื่น ๆ อีกจำนวนหนึ่งที่ไม่ได้แสดงในตาราง เช่น #,*, &, @, °, ^ ฯลฯ และเครื่องหมายวรรคตอบไทยโบราณ เช่น โคมูตร (๏ิ) อังคั่น (ฯ, ฯ, ฯะ, ฯะ) ฟองมัน (๏) ฯลฯ ซึ่งยังพบได้ในคำประพันธ์ร้อยแก้วและร้อยกรองบางประเภท

จากเครื่องหมายวรรคตอนที่ได้กล่าวมาในตารางข้างต้น พบว่าเครื่องหมายที่มีการใช้มากในงานเขียนภาษาไทย ได้แก่ จุด (.), จุลภาค (,), อัฒภาค (;), ทวิภาค (:), ต่อ (-), ยัติภังค์ (-), ไปยาลน้อย (ฯ), ไปยาลใหญ่ (ฯลฯ), ไม้ยมก (ๆ), วงเล็บ, บุพสัญญา (“), อัญประกาศ (“ ”), ทับ (/) และยติภาค (--) ทั้งนี้เครื่องหมายที่มักเป็นจุดเริ่มต้น EDU เสมอ ได้แก่ วงเล็บเปิด อัญประกาศเปิด และยติภาคที่ทำนำหัวข้อย่อย ส่วนเครื่องหมายที่มักอยู่ท้าย EDU ได้แก่ วงเล็บปิด อัญประกาศปิด และไปยาลใหญ่ เครื่องหมายที่มักแทรกกลางภายใน EDU ได้แก่ จุด จุลภาค ยัติภังค์ ต่อ และทับ ส่วนเครื่องหมายอื่น ๆ เป็นเครื่องหมายที่สามารถปรากฏได้หลายตำแหน่งใน EDU เช่น ไม้ยมกและไปยาลน้อย สามารถปรากฏภายในหรือท้าย EDU ก็ได้ จากที่กล่าวมา ผู้วิจัยจึงได้ใช้การปรากฏของเครื่องหมายวรรคตอนในการช่วยระบุขอบเขต EDU

8) โครงสร้างหน่วยเดียวกัน (Same unit construction)

โครงสร้างหน่วยเดียวกัน หมายถึง โครงสร้าง EDU ที่ถูกแยกออกเป็น 2 ส่วนเนื่องจากมีหน่วยอื่นแทรกกลาง เช่น คุณานุประโยค ข้อความในวงเล็บ ฯลฯ ทั้งนี้ทั้ง 2 ส่วนที่ถูกแทรกยังถือว่าเป็นโครงสร้างหน่วยเดียวกัน Carlson and Marcu (2001) เรียกปริจเฉทสัมพันธ์ระหว่างสองส่วนที่เป็นโครงสร้างหน่วยเดียวกันนี้ว่า multinuclear pesudo-relation นั่นคือเป็นความสัมพันธ์เทียมแบบที่ทั้งสองข้างมีสถานะเท่าเทียมกันยกตัวอย่างคำภาษาอังกฤษในวงเล็บดังตัวอย่าง (24) แทรกกลางระหว่างหน่วยที่ 1 และ 3 แบบนี้ถือว่าทั้ง 2 หน่วยเป็นโครงสร้างหน่วยเดียวกัน

ตัวอย่าง

(24) [ต่อมาในสมัยหลังสมัยใหม่]1 [(Post-modern)]2 [ได้เกิดวรรณกรรมแนวทดลอง] 3

นอกเหนือจากหลักเกณฑ์การกำหนดขอบเขตอนุพากย์ภาษาไทยที่ได้กล่าวมาข้างต้นแล้ว ยังมีหลักเกณฑ์อื่น ๆ อีกที่ Carlson and Marcu (2001) ได้กล่าวถึง แต่ผู้วิจัยไม่ได้นำหลักเกณฑ์เหล่านี้มาอภิปรายและใช้กำหนดขอบเขตอนุพากย์ภาษาไทย เพราะเห็นว่าหลักเกณฑ์บางอย่างอิงลักษณะทางภาษาของภาษาอังกฤษมากเกินไป และไม่เหมาะสมกับลักษณะทางภาษาของภาษาไทย ได้แก่ หน่วยสร้างที่ประกอบด้วยกริยาในรูป gerund หรือ participle เป็นต้น

บทที่ 5

การแยกอนุพากย์ภาษาไทยด้วยแบบจำลองซัพพอร์ตเวกเตอร์แมชชีน

ในบทนี้ ผู้วิจัยจะกล่าวถึงระบบการแยกอนุพากย์ภาษาไทยด้วยแบบจำลองซัพพอร์ตเวกเตอร์แมชชีน ลักษณ์ (Features) สำหรับฝึกฝนและทดสอบแบบจำลอง การเตรียมไฟล์ข้อมูลสำหรับฝึกฝนและทดสอบแบบจำลอง เคอร์เนลฟังก์ชั่นและการตั้งค่าพารามิเตอร์ วิธีการประเมินประสิทธิภาพของแบบจำลอง และผลการทดสอบแบบจำลอง ดังนี้

5.1 ระบบการแยกอนุพากย์ภาษาไทยด้วยแบบจำลองซัพพอร์ตเวกเตอร์แมชชีน

การแยกอนุพากย์ภาษาไทยด้วยเครื่องในงานวิจัยนี้ หมายถึง การระบุคำขอบเขตเริ่มต้นอนุพากย์ภาษาไทย หรือ EDU โดยการใช้แบบจำลองทางสถิติซัพพอร์ตเวกเตอร์แมชชีน ซึ่งผู้วิจัยใช้ตัวแบบจำลองซัพพอร์ตเวกเตอร์แมชชีนจากฟังก์ชั่น SMO (Sequential Minimal Optimization) ที่มีอยู่ในโปรแกรมวีก้า 3.6.10 (Weka 3.6.10) ในส่วนของ SMO นั้น เป็นอัลกอริทึมที่พัฒนาโดย Platt (1998) ใช้ในการจำแนกประเภทข้อมูลด้วยซัพพอร์ตเวกเตอร์แมชชีน ตัวอัลกอริทึมนี้ได้รับการพัฒนาขึ้นเพื่อแก้ปัญหาการโปรแกรมเชิงกำลังสอง (Quadratic programming) ซึ่งเป็นปัญหาที่พบในการจำแนกประเภทที่ไม่เป็นเส้นตรง (Nonlinear classification) ปัญหาที่ว่านี้ก็คือปัญหาของการหาค่าที่ดีที่สุดสำหรับการแยกข้อมูลออกเป็นประเภทต่าง ๆ ตามเป้าหมาย

ขั้นตอนของการแยกอนุพากย์ภาษาไทยด้วยแบบจำลองซัพพอร์ตเวกเตอร์แมชชีน เริ่มจากการจัดทำคลังข้อมูลและกำกับหมวดคำในคลังข้อมูล จากนั้นนำคลังข้อมูลฝึกฝนไปทำการฝึกฝนแบบจำลองโดยใช้ตัวจำแนกประเภทซัพพอร์ตเวกเตอร์แมชชีน เมื่อสร้างแบบจำลองได้แล้วก็นำไปทดสอบกับคลังข้อมูลทดสอบ เพื่อประเมินประสิทธิภาพของแบบจำลอง สำหรับงานวิจัยนี้ ผู้วิจัยทำการทดลองแบบ 10-fold cross-validation กล่าวคือ ได้ทำการฝึกฝนและทดสอบทั้งหมด 10 ครั้ง โดยแบ่งคลังข้อมูลออกเป็น 10 ส่วน แต่ละส่วนจะถูกใช้เป็นคลังข้อมูลทดสอบ และอีก 9 ส่วนที่เหลือจะถูกใช้เป็นคลังข้อมูลฝึกฝน นั่นคือผู้วิจัยใช้คลังข้อมูลฝึกฝน 90 เปอร์เซ็นต์ และคลังข้อมูลทดสอบ 10 เปอร์เซ็นต์ ด้วยวิธีการนี้ ข้อมูลทุกส่วนจะได้รับการฝึกฝนและทดสอบหมด

5.2 การกำหนดลักษณ์ที่ใช้ในการฝึกฝนและทดสอบแบบจำลอง

เนื่องจากแบบจำลองซัพพอร์ตเวกเตอร์แมชชีนเป็นแบบจำลองทางสถิติที่จำแนกประเภทข้อมูลโดยตัดสินใจจากลักษณ์ต่าง ๆ ที่เตรียมเอาไว้ให้ การกำหนดลักษณ์จึงมีความสำคัญและมีผลต่อการตัดสินใจของเครื่อง ประเภทของลักษณ์ที่สามารถใช้ในฟังก์ชั่นการจำแนกประเภทต่าง ๆ ของโปรแกรมวีก้า ได้แก่ ลักษณ์ที่มีสองค่า (Binary), ลักษณ์ที่มีลักษณะเป็นการจัดประเภทต่าง ๆ ที่เกิดจากการแบ่งข้อมูลออกเป็นกลุ่ม ๆ (Nominal), ตัวเลขที่แสดงค่าต่าง ๆ (Numeric), และข้อความ

(String) ลักษณ์ที่ใช้ในงานนี้มีลักษณะเป็นลักษณ์ทางภาษา ซึ่งได้จากการพิจารณาโครงสร้างทางวากยสัมพันธ์ของอนุพากย์ในการกำหนดลักษณ์ ในส่วนนี้จะนำเสนอลักษณะโครงสร้าง EDU ภาษาไทยในหัวข้อย่อย 5.2.1 และลักษณ์ที่ใช้ในการฝึกฝนและทดสอบแบบจำลองในงานนี้ในหัวข้อย่อย 5.2.2

5.2.1 ลักษณะทางโครงสร้าง EDU ภาษาไทย

โดยทั่วไปแล้ว อนุพากย์และกลุ่มนามวลีที่ขึ้นต้นด้วยตัวเชื่อมเด่นจะใช้เป็นหน่วยพื้นฐานในการศึกษาโครงสร้างและความสัมพันธ์ภายในปริจเฉท ในที่นี้ ผู้วิจัยจึงแบ่ง EDU ออกเป็น 2 ประเภท ตามระดับโครงสร้างภาษา คือ EDU ที่มีโครงสร้างระดับอนุพากย์ และ EDU ที่มีโครงสร้างต่ำกว่าระดับอนุพากย์ ทั้งสองประเภทนี้มีรูปแบบและองค์ประกอบของโครงสร้างดังนี้

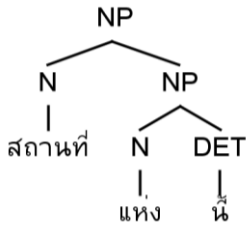
5.2.1.1 EDU ที่มีโครงสร้างระดับอนุพากย์

อนุพากย์ที่เป็น EDU ในที่นี้จะหมายถึงเป็นอนุพากย์ที่มีกริยาแท้เท่านั้น โดยทั่วไปแล้ว อนุพากย์จะประกอบไปด้วยประธานและภาคแสดง (Subject and predicate) หากพิจารณาหน่วยสร้างระดับอนุพากย์ ก็จะพบว่าอนุพากย์ประกอบไปด้วยหน่วยสร้างนามวลี (Noun phrase) ในตำแหน่งประธาน และกริยาวลี (Verb phrase) ซึ่งจะปรากฏในตำแหน่งภาคแสดงของอนุพากย์ โดยหน่วยสร้างกริยาวลีเป็นหน่วยสร้างที่จำเป็นและขาดไม่ได้ในโครงสร้างอนุพากย์ เพราะกริยาวลีประกอบไปด้วยคำกริยาหลัก ซึ่งเป็นแกนหลักของโครงสร้างอนุพากย์ มีหน้าที่แสดงข้อมูลเกี่ยวกับประธาน ได้แก่ การกระทำ ลักษณะ สภาวะ และอาการ

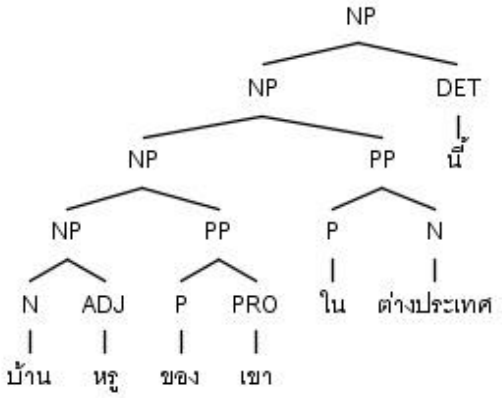
ในตำแหน่งประธานของอนุพากย์จะเป็นหน่วยสร้างนามวลี ซึ่งมีรูปแบบโครงสร้าง **Head noun + (Complement) + (Modifier) + (Determiner)** โดยที่ **Head noun** หมายถึง คำนามหลัก ซึ่งเป็นส่วนที่สำคัญและจำเป็นต้องมีเสมอ **Complement** หมายถึง ส่วนเติมเต็ม ได้แก่ บุพบทวลี และอนุพากย์เติมเต็ม **Modifier** หมายถึง ส่วนขยายคำนาม ได้แก่ คำคุณศัพท์, คำนาม, บุพบทวลี **Determiner** หมายถึง ตัวกำหนด เช่น “นี้”, “เหล่านี้”, “นั้น”, “เหล่านั้น” ฯลฯ องค์ประกอบของนามวลีในเครื่องหมายวงเล็บ ได้แก่ complement, modifier, **Determiner** หมายถึงองค์ประกอบที่สามารถละได้

ตัวอย่างนามวลีภาษาไทย

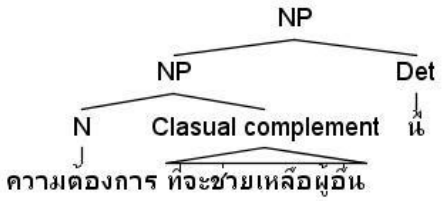
(1) “สถานที่แห่งนี้” ประกอบไปด้วย คำนามหลัก “สถานที่” + ส่วนขยาย “แห่งนี้” สามารถแยกองค์ประกอบและแสดงในรูปแผนภูมิต้นไม้ได้ดังนี้



(2) “บ้านหรูของเขาในต่างประเทศนี้” ประกอบไปด้วย คำนามหลัก “บ้าน” + ส่วนขยาย “หรู” + ส่วนเติมเต็ม “ของเขา” + ส่วนขยาย “ในต่างประเทศ” + ตัวกำหนด “นี้” สามารถแยกองค์ประกอบและแสดงในรูปแผนภูมิต้นไม้ได้ดังนี้



(3) “ความต้องการที่จะช่วยเหลือผู้อื่นนี้” ประกอบไปด้วย คำนามหลัก “ความต้องการ” + อนุพากย์เติมเต็ม “ที่จะช่วยเหลือผู้อื่น” + ตัวกำหนด “นี้” สามารถแยกองค์ประกอบและแสดงในรูปแผนภูมิต้นไม้ได้ดังนี้

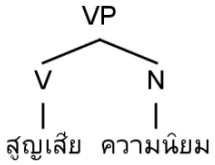


ในส่วนของภาคแสดงของอนุพากย์จะเป็นกริยาวลี มีรูปแบบโครงสร้างคือ **Head verb + (Object) + (Complement) + (Modifier)** โดยที่ **Head verb** หมายถึง คำกริยาหลัก ซึ่งเป็นส่วนที่สำคัญและขาดไม่ได้ **Object** หมายถึง กรรมของกริยา ซึ่งอาจเป็นกรรมตรงหรือกรรมรองขึ้นอยู่กับลักษณะของกริยา **Complement** หมายถึง ส่วนเติมเต็มกริยา ได้แก่ บุพบทวลี และอนุพากย์เติมเต็ม **Modifier** หมายถึง ส่วนขยายคำกริยา ได้แก่ คำกริยาวิเศษณ์ และบุพบทวลี

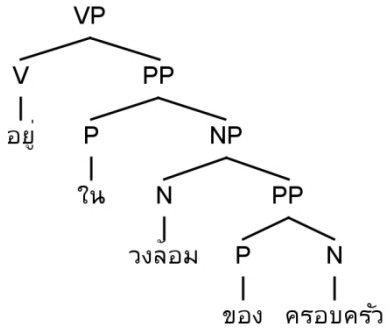
องค์ประกอบของโครงสร้างนามวลีในเครื่องหมายวงเล็บได้แก่ object, complement, modifier หมายถึงองค์ประกอบที่ละได้

ตัวอย่างกริยาวลีภาษาไทย

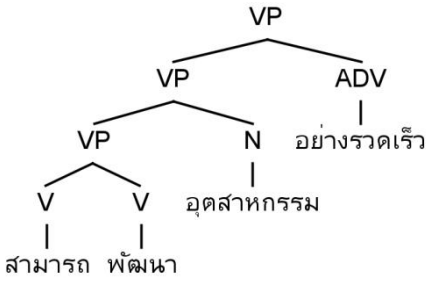
(4) “สูญเสียความนิยม” ประกอบไปด้วย คำกริยาหลัก “สูญเสีย” + กรรมตรง “ความนิยม สามารถแยกองค์ประกอบและแสดงในรูปแผนภูมิต้นไม้ได้ดังนี้



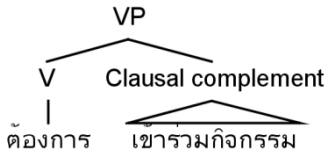
(5) “อยู่ในวงล้อมของครอบครัว” ประกอบไปด้วย คำกริยาหลัก “อยู่” + ส่วนเติมเต็ม “ในวงล้อมของครอบครัว” สามารถแยกองค์ประกอบและแสดงในรูปแผนภูมิต้นไม้ได้ดังนี้



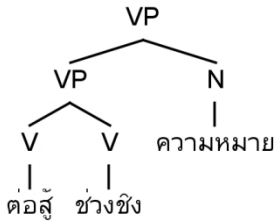
(6) “สามารถพัฒนาอุตสาหกรรมอย่างรวดเร็ว” ประกอบไปด้วย คำกริยาช่วย “สามารถ” + คำกริยาหลัก “พัฒนา” + กรรมตรง “อุตสาหกรรม” + ส่วนขยาย “อย่างรวดเร็ว” สามารถแยกองค์ประกอบและแสดงในรูปแผนภูมิต้นไม้ได้ดังนี้



(7) “ต้องการเข้าร่วมกิจกรรม” ประกอบไปด้วย คำกริยาหลัก “ต้องการ” + อนุพากย์เติมเต็ม “เข้าร่วมกิจกรรม” สามารถแยกองค์ประกอบและแสดงในรูปแผนภูมิต้นไม้ได้ดังนี้



(8) “ต่อสู้ช่วงชิงความหมาย” เป็นโครงสร้างกริยาเรียง ประกอบไปด้วย คำกริยาหลัก “ต่อสู้” + “ช่วงชิง” + กรรมตรง “ความหมาย” สามารถแยกองค์ประกอบและแสดงในรูปแผนภูมิต้นไม้ได้ดังนี้



เมื่ออนุพากย์มากกว่าหนึ่งอนุพากย์ประกอบเข้าด้วยกัน ก็จะได้โครงสร้างปริจเฉทที่ใหญ่ขึ้น โดยอาจใช้กลวิธีการเชื่อมโยงความด้วยคำเชื่อมหรือไม่ก็ได้ โดยคำเชื่อมจะมีหน้าที่เป็นตัวเชื่อมโยงอนุพากย์ และยังสามารถบ่งบอกถึงลักษณะความสัมพันธ์ระหว่างอนุพากย์ได้ด้วย ซึ่งอาจเป็นความสัมพันธ์แบบสองข้างซ้ายขวาเท่ากันหรือสองข้างไม่เท่ากัน อีกทั้งยังสามารถช่วยระบุขอบเขตเริ่มต้นอนุพากย์ได้อีกด้วย คำเชื่อมอนุพากย์ ได้แก่ อนุสันธานซึ่งจะปรากฏหน้าอนุพากย์ไม่อิสระ และคำเชื่อมสันธานประสานซึ่งจะปรากฏหน้าอนุพากย์อิสระ

นอกเหนือจากคำเชื่อมอนุพากย์แล้ว ในกรณีคุณานุประโยค ซึ่งเป็นอนุพากย์ที่ทำหน้าที่ขยายนาม มักจะขึ้นต้นด้วยตัวนำส่วนเติมเต็ม “ที่” “ซึ่ง” “อัน” และกรณีอนุพากย์เติมเต็มของกริยามักจะขึ้นต้นด้วยตัวนำส่วนเติมเต็ม “ว่า” ดังนั้นสามารถเรียกคำเชื่อมและตัวนำส่วนเติมเต็มที่ปรากฏหน้าอนุพากย์ว่า คำบ่งชี้ (Marker) เพราะเป็นคำที่สามารถใช้บ่งชี้ขอบเขตอนุพากย์ได้

ตัวอย่างจากคลังข้อมูล

(9) “ซึ่งเรียกว่าไมโครเซลล์” เป็นอนุพากย์ที่มีรูปแบบ คำบ่งชี้ “ซึ่ง” + ภาคแสดง “เรียกว่าโมโครเซลล์”

(10) “เพราะได้รับอิทธิพลจากแม่” เป็นอนุพากย์ที่มีรูปแบบ คำบ่งชี้ “เพราะ” + ภาคแสดง

“ได้รับอิทธิพลจากแม่”

(11) “และเนื้อหาจากวรรณกรรมทำให้งานจิตกรรมของเขาโดดเด่น” เป็นอนุพากย์ที่มีรูปแบบ คำบ่งชี้ “และ” + ประธาน “เนื้อหาจากวรรณกรรม” + ภาคแสดง “ทำให้งานจิตกรรมของเขาโดดเด่น”

นอกจากนี้ยังมีข้อสังเกตประการหนึ่งเกี่ยวกับการประกอบเข้าด้วยกันของอนุพากย์ก็คือ มักจะมีการเว้นวรรคหรือใช้ช่องว่างคั่นระหว่างอนุพากย์ ทำให้สามารถระบุขอบเขตอนุพากย์โดยอาศัยการปรากฏของช่องว่างได้ โดยเฉพาะในกรณีที่อนุพากย์ไม่ได้ขึ้นต้นด้วยคำบ่งชี้แต่มีช่องว่างปรากฏอยู่ข้างหน้า ดังนั้นรูปแบบของปริจเฉทที่เกิดจากการเชื่อมโยงความกันของอนุพากย์สามารถเป็นได้ดังนี้

อนุพากย์ที่ 1 + (ช่องว่าง) + (คำบ่งชี้ +(ช่องว่าง)) + อนุพากย์ที่ 2+ (ช่องว่าง) + (คำบ่งชี้ +(ช่องว่าง)) + อนุพากย์ที่ 3

แต่อย่างไรก็ตาม ช่องว่างในภาษาไทยไม่ได้มีหน้าที่เพียงการเป็นตัวคั่นอนุพากย์ ทำให้เกิดความกำกวมในการตัดสินว่าช่องว่างที่ปรากฏในข้อความเป็นช่องว่างที่เป็นตัวคั่นอนุพากย์หรือไม่ จากการสังเกตการปรากฏของช่องว่างในคลังข้อมูล พบว่ามีจำนวนทั้งหมด 9,938 ครั้ง และพบว่าเป็นช่องว่างที่คั่นอนุพากย์จำนวน 4,606 ครั้ง นั่นคือเกือบครึ่งของช่องว่างทั้งหมดทำหน้าที่เป็นตัวคั่นอนุพากย์ นอกนั้นทำหน้าที่อื่น ๆ เช่น คั่นระหว่างคำในรายการคำ คั่นระหว่างคำเชื่อมและอนุพากย์ที่ตามมา คั่นระหว่างเครื่องหมายวรรคตอนและข้อความ เป็นต้น

เมื่อนำช่องว่างที่ทำหน้าที่คั่นอนุพากย์จำนวน 4,606 ครั้งมาพิจารณาดู โดยอาศัยหมวดคำรอบข้างช่องว่างนั้นประกอบ ก็พบว่าช่องว่างที่เป็นตัวคั่นอนุพากย์มักจะตามมาด้วยหมวดคำบางหมวดคำ ที่พบจำนวนมากที่สุดคือช่องว่างแล้วตามด้วยสันธานประสานหรืออนุสันธานหรือตัวนำส่วนเติมเต็ม รวมกันทั้งหมดแบ่งอนุพากย์จำนวน 2,166 ครั้ง ซึ่งเป็นจำนวนเกือบครึ่งของช่องว่างที่เป็นตัวคั่นอนุพากย์ แต่ก็ไม่ใช่เรื่องแปลกมากนัก เพราะตัวเชื่อมและตัวนำส่วนเติมเต็มนี้มักปรากฏหน้าอนุพากย์ และธรรมเนียมการเขียนภาษาไทยก็มักจะเว้นวรรคก่อนขึ้นต้นอนุพากย์ใหม่

เมื่อพิจารณาช่องว่างที่ทำหน้าที่คั่นอนุพากย์จำนวนที่เหลือ คือ 2,440 ครั้ง พบว่าช่องว่างที่เป็นตัวคั่นอนุพากย์มักจะตามด้วยหมวดคำกริยา กริยาช่วย บุพบท คำบอกปฏิเสธ กริยาวิเศษณ์ สรรพนาม เครื่องหมายวรรคตอน และคำบอกจำนวนหน้านาม ดังนั้นความเป็นไปได้ที่ช่องว่างจะเป็นตัวคั่นอนุพากย์อาจสามารถอาศัยหมวดคำที่ตามหลังพิจารณาประกอบได้

5.2.1.2 EDU ที่มีลักษณะทางโครงสร้างต่ำกว่าระดับอนุพากย์

โครงสร้างของ EDU ที่ต่ำกว่าอนุพากย์และเป็นหน่วยพื้นฐานในการศึกษาโครงสร้างปริจเฉท มีเพียงนามวลีอย่างเดียว ซึ่งจะต้องเป็นนามวลีที่ตามหลังคำเชื่อมเด่น หรือไม่ก็เป็นนามวลีที่มีหน้าที่ในระดับปริจเฉทเท่านั้น

กรณีนามวลีตามหลังคำเชื่อม จะต้องเป็นคำเชื่อมใดคำเชื่อมหนึ่งในสองประเภทเท่านั้น คือ คำเชื่อมแสดงตัวอย่าง เช่น “เช่น” “ตัวอย่างเช่น” “อาทิ” ฯลฯ และคำเชื่อมแสดงวัตถุประสงค์ ซึ่งมีคำเดียวคือ “เพื่อ” ส่วนกรณีที่ไม่ได้นำหน้าด้วยคำเชื่อม ต้องเป็นนามวลีที่มีหน้าที่ในระดับปริจเฉท เช่น นามวลีที่อยู่ในเครื่องหมายวงเล็บ ทำหน้าที่แสดงข้อมูลเพิ่มเติมให้กับข้อความส่วนหน้า หรือนามวลีที่เป็นชื่อหัวข้อข้อ/ชื่อเรื่องที่ปรากฏในบรรทัดโดด ๆ ในงานเขียน ซึ่งเป็นการแสดงให้เห็นว่าเนื้อหาที่กำลังจะกล่าวถึงเป็นเรื่องเกี่ยวกับอะไร ดังนั้นรูปแบบโครงสร้าง EDU ระดับต่ำกว่าอนุพากย์ได้จึงมีเพียง 2 รูปแบบ คือ

รูปแบบที่ 1 คำเชื่อมเด่น + นามวลีหรือกลุ่มนามวลี
รูปแบบที่ 2 นามวลีที่มีหน้าที่ในระดับปริจเฉท

ทั้งนี้ จากการสังเกตข้อมูลพบว่านามวลีที่เป็น EDU มีความแตกต่างจากนามวลีที่ทำหน้าที่ประธานอนุพากย์ที่กล่าวไปข้างต้น คือ มักจะไม่พบตัวกำหนดท้ายคำนาม นั่นคือมีรูปแบบ **Head noun + (Complement) + (Modifier)** และนอกจากนี้ยังพบว่า ในกรณีการใช้คำเชื่อมแสดงตัวอย่าง นามวลีที่ตามหลังคำเชื่อมมักจะมีจำนวนมากกว่าหนึ่ง และมีการใช้ช่องว่างคั่นระหว่างนามวลีแต่ละตัว และในบางครั้งพบว่ามีการใช้เครื่องหมายจุลภาค (,) คั่นระหว่างนามวลีด้วย ในกรณีที่มีรายการที่เป็นตัวอย่างเป็นจำนวนมาก มักจะมีการละตัวอย่างอื่น ๆ ในกลุ่มเดียวกันโดยการใช้เครื่องหมายไปยานใหญ่ (ฯลฯ) หรือคำกริยาวิเศษณ์ “เป็นต้น” ท้ายตัวอย่างสุดท้าย ซึ่งเป็นขอบเขตสิ้นสุดของ EDU

ตัวอย่างนามวลีที่เป็น EDU

(12) **$**_เช่น คะแนนเสียงในสภา, คำพิพากษาของศาลรัฐธรรมนูญ, คำวินิจฉัยของคณะกรรมการตรวจเงินแผ่นดิน, ลายเซ็นของสำนักงานนโยบายและแผนสิ่งแวดล้อม ฯลฯ
(13) **$**_เช่น หอก แหลน หลาว ไม้ตะบอง เป็นต้น
(14) **$**_กู้หนี้ยืมสินมา**$**_เพื่อการต่อสู้คดี
(15) **$**_พลวัตของความรู้ชาวบ้านในกระแสโลกาภิวัตน์ (ชื่อบทความ)
(16) **$**_1.1 ความเป็นมาและความสำคัญของปัญหา (ชื่อหัวข้อย่อยของบทความ)

5.2.2 ลักษณ์ที่ใช้ในการฝึกฝนและทดสอบแบบจำลอง

หลังจากวิเคราะห์ลักษณะทางโครงสร้างของ EDU ภาษาไทยแล้ว ผู้วิจัยจึงได้กำหนดรายการต่อไปนี้เป็นลักษณ์ที่ใช้ในการฝึกฝนและทดสอบแบบจำลอง

1) ประเภทของคำ (Parts of speech: POS) เป็นข้อมูลเชิงโครงสร้างภาษาที่มีความสำคัญต่อการแยกอนุพากย์เป็นอย่างมาก เนื่องจากหากมองอนุพากย์ในเชิงโครงสร้างทางวากยสัมพันธ์แล้วก็จะพบว่าอนุพากย์ประกอบไปด้วยการเรียงตัวกันอย่างเป็นระบบของ POS ต่าง ๆ ที่กล่าวว่ามีความเป็นระบบนั้นหมายความว่า POS ที่เรียงตัวกันนั้นจะมีการเรียงลำดับการปรากฏร่วมกันอย่างมีแบบแผน เช่น ตัวกำหนดจะปรากฏหลังคำนามเท่านั้น ไม่สามารถปรากฏหน้าคำนามได้ ดังนั้น “นี้โรงเรียน” หรือ “นั้นการปฏิบัติ” จึงเป็นการเรียงตัวกันที่ผิด และจะไม่เกิดขึ้นในภาษาที่ใช้กันปกติ โดยทั่วไปแล้ว อนุพากย์จะต้องมีกริยาแท้ซึ่งเป็นกริยาหลักของอนุพากย์ 1 ตัว หรือมากกว่านั้นหากเป็นโครงสร้างกริยาเรียง POS บางประเภทสามารถช่วยในการระบุของอนุพากย์ได้ เช่น coordinator ซึ่งเป็นคำเชื่อมที่จะปรากฏหน้าอนุพากย์เสมอ สามารถใช้ระบุขอบเขตเริ่มต้นอนุพากย์ได้

ในการศึกษานี้ ผู้วิจัยใช้ POS เป็นลักษณ์สำหรับให้แบบจำลองใช้ในการตัดสินใจ ได้แก่ POS ปัจจุบันของคำที่พิจารณาอยู่ (POS-P), POS ก่อนหน้าคำที่พิจารณา (POS-B) และ POS ที่ตามหลังคำที่พิจารณา (POS-A) ค่าของลักษณ์ (Feature value) จะเป็นแบบ nominal เช่น NCNM, DET, CCOR, PREP,... ฯลฯ ซึ่งก็คือ POS ของแต่ละคำที่ได้กำกับมาแล้วในคลังข้อมูล ตัวอย่างค่าลักษณ์ POS ของข้อมูลเป็นดังต่อไปนี้

คำ(token)	POS-P	POS-B	POS-A
กิจกรรม	NCMN	SPACE	NCMN
นันทนาการ	NCMN	NCMN	PREP
สำหรับ	PREP	NCMN	NCMN
ชาว	NCMN	PREP	NCMN
ชนบท	NCMN	NCMN	AUX
ควร	AUX	NCMN	AUX
จะ	AUX	AUX	VERB
เน้น	VERB	AUX	NCMN

ตารางที่ 5.1 แสดงตัวอย่างค่าลักษณ์ POS ของคลังข้อมูล

2) รายการคำเชื่อม (Discourse markers: DM) หมายถึงรายการคำที่ทำหน้าที่เชื่อมระดับอนุพากย์หรือระดับที่ใหญ่กว่า เช่นคำว่า “อย่างไรก็ตาม” “และ” “หรือไม่ก็” ฯลฯ นอกจากนี้ยังหมายถึง strong marker หรือคำเชื่อมเด่น เช่นคำว่า “เช่น” “ยกตัวอย่างเช่น” “อย่างเช่น” “ได้แก่” ฯลฯ คำเชื่อมเหล่านี้ทำหน้าที่เชื่อมโยงปริจเฉทก่อนหน้ากับกลุ่มของวลีที่ตามหลังคำเชื่อมเหล่านี้

ผู้วิจัยได้จัดทำรายการคำเชื่อม โดยรวบรวมรายการคำเชื่อมจากวิทยานิพนธ์ เรื่องหน่วยเชื่อมโยงปริจเฉทภาษาไทยตั้งแต่สมัยสุโขทัยจนถึงปัจจุบัน ของเทพี จรัสจรุงเกียรติ (2543) ซึ่งได้ศึกษาและรวบรวมรายการคำเชื่อมภาษาไทยตั้งแต่สมัยสุโขทัยจนถึงปัจจุบัน โดยแบ่งคำเชื่อมออกเป็น 2 กลุ่มตามรูปภาษา คือ คำเชื่อมที่มีรูปเป็นคำ เช่นคำว่า “ก็” “เพราะ” และคำเชื่อมที่มีรูปเป็นกลุ่มคำ เช่นคำว่า “ก็เพราะ” “เพราะว่า” พบว่าคำเชื่อมที่มีการใช้กันอยู่ในปัจจุบันมีดังตารางต่อไปนี้

คำเชื่อมแบบคำ	**คำเชื่อมแบบกลุ่มคำ**
ก็, เกลือก, กว่า, กับ, ครั้น, คือ, จน, จึง, ฉะนั้น, เช่น, ด้วย, ดัง, โดย, ต่อ, ถ้า, ถึง, ทั้ง, ทำนอง, เท่า, เผื่อ, เพราะ, เพื่อ, แม้น, แม้, เมื่อ, แล, และ, แล้ว, สำหรับ, เหมือน, หรือ, ฤา, หาก, เหตุ, อนึ่ง, กระนั้นถ้า, ก็เพราะ, ก็เพื่อ, กับทั้ง, กับอนึ่ง, ครั้นเมื่อ, จนกระทั่ง, จนกว่า	จนถึง, ต่อเมื่อ, แต่ก็, แต่ทว่า, แต่อย่างไรก็ดี, แต่อย่างไรก็ตาม, ถ้าแล, ถ้าหาก, ถึงกระนั้น, ถึงกระนั้นก็ดี, ถึงกระนั้นถ้า, ถึงมาทว่า, ถึงแม้, เท่าเมื่อ, เพราะฉะนั้น, เพราะฉะนั้นก็, เพราะด้วย, เพราะเหตุ, เพื่อสำหรับ, แม้กระนั้น, แล้วก็, แล้วจึง, หรืออย่างไรก็ดี, หากแต่, หากทว่า, เหตุฉะนั้น, เหตุฉะนี้, เหตุด้วย, เหมือนเช่น, เหมือนอย่าง, อีกทั้ง, กระนั้นก็ตาม, กล่าวคือ, ก็เพราะเหตุว่า, ขณะที่, ขณะนั้น, ครั้งต่อมา, คือว่า, ด้วยเป็นเพราะ, ด้วยเหตุที่, ด้วยเหตุว่า, ดุจหนึ่ง, โดยที่ในขณะเดียวกันก็, โดยที่ในขณะเดียวกันก็, ตรงกันข้ามด้วยซ้ำ, ต่อนั้นมา, ต่อมา, ตั้งแต่, ตัวอย่างเช่น, แต่ตรงกันข้าม, แต่ทั้งนี้ทั้งนั้นก็, แต่ที่จริง, แต่ที่แท้ก็, แต่ที่แท้จริง, แต่ว่า, ถึงอย่างไร, ทั้งนี้, ทั้ง ๆ ที่, ทั้งนี้คงเนื่องมาจาก, ทั้งนี้ด้วยเหตุผลที่ว่า, ทั้งนี้เนื่องจาก, ทำนองเดียวกันกับที่, ที่จริง, ที่จริงก็, เท่าที่, นอกจากนี้, นั่นก็คือ, เนื่องจาก, เนื่องด้วย, เนื่อง

	ด้วยเหตุผลที่ว่า, เนื่องเพราะ, เนื่องมาจาก, เนื่องมาแต่เมื่อ, ในขณะเดียวกัน, ในขณะที่, ในทางตรงกันข้าม, ในทำนองเดียวกัน, ในทำนองเดียวกันกับ, ในที่สุด, ในระหว่างนั้น, บัดนี้, ประการหนึ่ง, เป็นต้นแต่, เป็นต้นว่า, เพราะว่า, เพราะเหตุที่, เพราะเหตุว่า, เพราะอย่างน้อยก็, แม้ว่า, ระหว่างนี้, ราวกับว่า, หรือกล่าวอีกนัยหนึ่ง, หรือที่ถูกก็, หรือมิฉะนั้นอย่างสูงกว่านั้นขึ้นไป, หรือไม่ก็, หรือว่า, หรือว่าอีกอย่างหนึ่ง, หรืออีกนัยหนึ่ง, หรืออีกอย่างหนึ่ง, หากแต่ว่า, เหตุดังนั้น, เหตุนี้, เหมือนดังว่า, อย่างไรก็ดี, อย่างไรก็ตาม, อนึ่งคือว่า, อีกประการหนึ่ง, อีกอย่างหนึ่ง

ตารางที่ 5.2 แสดงรายการคำเชื่อม แบ่งตามรูปภาษา

เนื่องจากรายการคำเชื่อมที่แสดงในตาราง 5.2 ไม่มีคำเชื่อมเด่นอยู่ด้วย ผู้วิจัยจึงได้เพิ่มรายการคำเชื่อมเด่นเข้าไป เนื่องจากเป็นตัวเชื่อมความที่มีความหมายบ่งบอกถึงความสัมพันธ์ทางปริจเฉท ได้แก่คำ “เช่น” “อาทิ” “อาทิเช่น” “อย่างเช่น” “ตัวอย่างเช่น” “ยกตัวอย่างเช่น” “ได้แก่”

สำหรับลักษณ์นี้ ผู้วิจัยใช้ค่าลักษณ์แบบ binary คือกำหนดให้มีสองค่า คือ Y (เป็นคำเชื่อม) และ N (ไม่เป็นคำเชื่อม) วิธีการกำหนดค่าของลักษณ์นี้ทำโดยการเทียบคำในคลังข้อมูลกับคำในรายการคำเชื่อมที่เตรียมไว้ หากคำที่พิจารณาปรากฏอยู่ในรายการคำเชื่อม ก็จะมีค่าเป็น Y แต่หากไม่ปรากฏอยู่ในรายการคำเชื่อม ก็จะให้ค่าเป็น N

อย่างไรก็ตาม คลังข้อมูลที่ใช้ในงานนี้ เป็นคลังข้อมูลแบบตัดคำ ดังนั้นคำเชื่อมที่มีรูปเป็นกลุ่มคำ เช่นคำว่า “เหตุฉะนี้” “ถึงแม้ว่า” ฯลฯ จะถูกแบ่งคำเป็น “เหตุ+ฉะนี้” และ “ถึง+แม้+ว่า” หากนำคำว่า “เหตุ” และ “ถึง” ไปค้นในรายการคำเชื่อม ก็จะไม่พบ ผู้วิจัยจึงแก้ปัญหานี้โดยการเทียบคำในคลังข้อมูลทีละ 4 คำเรียงติดกัน หรือสายของคำ 4 คำ (token1|token2|token3|token4) หาก token1 หรือ token1|token2 หรือ token1|token2|token3 หรือ token1|token2|token3|token4 ปรากฏอยู่ในรายการคำเชื่อม ก็จะให้คำแรกของคำเชื่อมนั้นมีค่าลักษณ์เป็น Y คำถัดไปที่จะนำเทียบกับรายการคำเชื่อมคือคำที่อยู่ถัดจาก token สุดท้ายของ

คำเชื่อม แต่ถ้าหากว่าคำในรายการเชื่อมตั้งแต่ token1 ไม่ปรากฏในรายกายคำเชื่อม ก็จะให้ token นั้นมีค่าเป็น N และใช้สายของคำ 4 คำถัดจาก token1 เทียบกับรายการคำเชื่อมต่อไป ยกตัวอย่าง

|ถึง|กระนั้น|ก็|ดี| |การ|หา|ทาง|ออก|ของ|ปัญหา|...

จากตัวอย่าง จะเทียบสายของคำ |ถึง|กระนั้น|ก็|ดี| ก่อน และพบว่า 4 คำเรียงกันนี้พบในรายการคำเชื่อม จึงกำหนดให้มีค่าลักษณ์เป็น Y ตรง token แรกเท่านั้น คือ |ถึง| ส่วนอีก 3 token คือ |กระนั้น|ก็|ดี| จะมีค่าเป็น N สายของคำที่จะใช้เทียบต่อไปคือ | |การ|หา|ทาง| และพบว่าช่องว่าง | | ไม่พบในรายการคำเชื่อม จึงมีค่าเป็น N สายของคำถัดไปที่จะใช้เทียบคือ |การ|หา|ทาง|ออก| ซึ่งไม่พบ |การ| ในรายการคำเชื่อม จึงมีค่าเป็น N สายของคำถัดไปที่จะใช้เทียบคือ |หา|ทาง|ออก|ของ| ซึ่งไม่พบ |หา| ในรายการคำเชื่อม จึงมีค่าเป็น N ทำเช่นนี้ไปเรื่อย ๆ ก็จะได้ค่าของลักษณ์ของแต่ละคำดังนี้

คำ (tokens)	**DM**
ถึง	Y
กระนั้น	N
ก็	N
ดี	N
<s>	N
การ	N
หา	N
ทาง	N
ออก	N
ของ	N
ปัญหา	N

ตารางที่ 5.3 แสดงตัวอย่างจากคลังข้อมูลและลักษณ์รายการคำเชื่อม

ทั้งนี้ ผู้วิจัยไม่ได้ใช้ประโยชน์จากระยะห่างระหว่างคำเชื่อมคู่ เช่น “แม้...แต่”, “เมื่อ...จึง”, “เพราะ...จึง” เป็นต้น ตามที่ได้กล่าวเอาไว้ในสมมติฐานว่าระยะห่างระหว่างคำเชื่อมคู่จะสามารถช่วยในการแยกอนุพากย์ภาษาไทย แต่เนื่องจากเห็นว่า การกำกับหมวดคำ CCOR และ CSUB ก็สามารถช่วยระบุขอบเขตเริ่มต้น EDU ได้แล้ว จึงไม่มีความจำเป็นต้องพิจารณาระยะห่างระหว่างคำเชื่อมคู่อีก

3) ช่องว่าง (Space) แม้ว่าช่องว่างจะไม่ได้ทำหน้าที่แบ่งอนุพากย์ออกจากกันทุกครั้งไป แต่เมื่อพิจารณาจากสถิติการปรากฏร่วมกันของ POS หลังช่องว่าง หรือ space|POS ก็จะพบว่า POS บางประเภทที่ตามหลังช่องว่างมีโอกาสที่จะเป็นตัวแบ่งอนุพากย์มากกว่า POS ประเภทอื่น ดังนั้นการพิจารณา space|POS จึงน่าจะสามารถบอกได้ว่าช่องว่างนั้นเป็นตัวแบ่งอนุพากย์หรือไม่

SPACE|POS	เกิดร่วมกัน (ครั้ง)	แบ่งอนุพากย์ (ครั้ง)	คิดเป็นเปอร์เซ็นต์
SPACE SPACE	7	0	0
SPACE CCOR	1074	1066	99.25512104
SPACE FOREIGN	902	19	2.106430155
SPACE VERB	899	380	42.26918799
SPACE CSUB	620	574	92.58064516
SPACE PUNC	1006	495	49.20477137
SPACE NNUM	552	9	1.630434783
SPACE NPRP	631	137	21.71156894
SPACE NCMN	1472	529	35.9375
SPACE CCORIN	342	21	6.140350877
SPACE DET	29	1	3.448275862
SPACE COMPF	527	526	99.81024668
SPACE PREP	543	256	47.14548803
SPACE PFAV	8	2	25
SPACE ADJ	3	0	0
SPACE CSBI	112	35	31.25
SPACE NCLS	110	4	3.636363636
SPACE MNCB	4	0	0
SPACE AUX	296	133	44.93243243
SPACE NEG	51	37	72.54901961
SPACE PT	8	1	12.5
SPACE MNCF	56	32	57.14285714
SPACE PFN	332	119	35.84337349

SPACE ADVERB	224	146	65.17857143
SPACE NPRO	127	84	66.14173228
รวม	**9935**	**4606**	

ตารางที่ 5.4 แสดงจำนวนครั้งที่ POS ต่าง ๆ ปรากฏหลังช่องว่าง และจำนวนครั้งที่ช่องว่างจะเป็นตัวแบ่งอนุพากย์

จากตารางดังกล่าว พบว่าช่องว่างทำหน้าที่เป็นตัวแบ่งอนุพากย์ทั้งหมด 4606 ครั้ง โดยช่องว่างที่นำหน้า CCOR หรือ space|CCOR ทำหน้าที่เป็นตัวแบ่งอนุพากย์มากถึง 1066 ครั้ง คิดเป็น 99 เปอร์เซ็นต์ของ space|CCOR ที่เกิดร่วมกันทั้งหมด รองลงมาคือ space|CSUB และ space|COMPF โดยช่องว่างนั้นเป็นตัวแบ่งอนุพากย์ทั้งหมด 574 และ 526 ครั้ง หรือคิดเป็น 92 และ 99 เปอร์เซ็นต์ ตามลำดับ ในขณะที่ space|SPACE, space|ADJ, space|MNCB เป็นคู่ที่ไม่ใช่ตัวแบ่งอนุพากย์เลย

แต่อย่างไรก็ตาม ปรากฏการณ์ที่ช่องว่างนำหน้าคำเชื่อมอนุพากย์ CCOR , CSUB และตัวนำส่วนเติมเต็ม COMPF แล้วช่องว่างนั้นทำหน้าที่เป็นตัวแบ่งอนุพากย์ เป็นปรากฏการณ์ที่ปกติ เนื่องจากการเริ่มอนุพากย์หรือถ้อยความใหม่โดยมีคำเชื่อมขึ้นต้น การใช้ช่องว่างคั่นเป็นสิ่งที่พบได้ทั่วไปอยู่แล้วในภาษาเขียนภาษาไทย ดังนั้นจึงไม่แปลกที่จำนวนช่องว่างตามด้วย POS เหล่านั้นจะเป็นตัวแบ่งอนุพากย์มากถึง 90 กว่าเปอร์เซ็นต์

ผู้วิจัยได้ใช้ประโยชน์จากข้อมูลดังตารางดังกล่าวในการกำหนดค่าของลักษณ์ โดยนำเฉพาะคู่ space|POS ที่เกิดร่วมกันและช่องว่างนั้นเป็นตัวแบ่งอนุพากย์มากกว่า 40 เปอร์เซ็นต์ของของการเกิด space|POS นั้น ๆ มาพิจารณา นั่นคือกำหนดให้ช่องว่างที่ตามด้วย POS ได้แก่ CCOR, VERB, CSUB, PUNC, COMPF, PREP, AUX, NEG, MNCF, ADVERB, NPRO เป็นช่องว่างที่มีความเป็นไปได้ที่จะทำหน้าที่เป็นตัวแบ่งอนุพากย์ และมีผลทำให้คำที่มี POS เหล่านี้มีความเป็นไปได้ที่จะเป็นขอบเขตเริ่มต้นอนุพากย์เช่นกัน ดังนั้นผู้วิจัยจึงกำหนดค่าของลักษณ์เป็นแบบสองค่า คือ Y และ N โดยคำที่พิจารณาจะมีค่าเป็น Y เมื่อคำ ๆ นั้นปรากฏหลังช่องว่างที่มีความเป็นไปได้ที่จะเป็นตัวแบ่งอนุพากย์ ในขณะที่จะมีค่าเป็น N เมื่อคำ ๆ นั้นไม่ได้ปรากฏหลังช่องว่างที่มีความเป็นไปได้ที่จะเป็นตัวแบ่งอนุพากย์ ตัวอย่างคลังข้อมูลและค่าของลักษณ์เป็นดังนี้

คำ (tokens)	**Space**
<s>	N
<w PUNC>(</w>	Y
<w FOREGIN>Impressionist</w>	N

<w PUNC>)</w>	N
<s>	N
<w VERB>มา</w>	Y
<w VERB>ใช้</w>	N
<w PREP>ใน</w>	N
<w NCMN>ลักษณะ</w>	N
<w NCMN>ประติมากรรม</w>	N

ตารางที่ 5.5 แสดงตัวอย่างจากคลังข้อมูลและการกำหนดลักษณ์ช่องว่าง

จากตารางจะเห็นว่า เครื่องหมายวงเล็บเปิด และคำว่า “มา” มีค่าลักษณ์เป็น Y เนื่องจากทั้งสองมี POS เป็น PUNC และ VERB ตามลำดับ และปรากฏหลังช่องว่างที่มีความเป็นไปได้ที่จะเป็นตัวแบ่งอนุพากย์ ในขณะที่คำอื่น ๆ มีค่าเป็น N เนื่องจากไม่ได้ปรากฏหลังช่องว่าง

4) เครื่องหมายวรรคตอน (Punctuations) จากการสำรวจคลังข้อมูล คำที่ถูกกำกับหมวดคำเป็น PUNC หรือเครื่องหมายวรรคตอนมีจำนวนทั้งหมด 2,665 ครั้ง เครื่องหมายที่สามารถใช้ช่วยระบุขอบเขตอนุพากย์ได้มีเพียงเครื่องหมายวงเล็บเท่านั้น โดยพบว่ามีเครื่องหมายวงเล็บเปิดและวงเล็บปิดรวมกันได้ 900 ครั้ง นอกนั้นเป็นเครื่องหมายวรรคตอนอื่น ๆ ที่มีการปรากฏแบบกระจายคือปรากฏได้หลายตำแหน่ง ในที่นี้จึงกำหนดให้มองเฉพาะเครื่องหมายวงเล็บเป็นลักษณ์เท่านั้น การกำหนดค่าของลักษณ์คือ กำหนดให้เฉพาะคำ หรือ token ที่เป็นเครื่องหมายวงเล็บเปิด และคำถัดจากช่องว่างที่ปรากฏหลังเครื่องหมายวงเล็บปิดมีค่าลักษณ์เป็น Y คือเป็นขอบเขตเริ่มต้นอนุพากย์นั่นเอง นอกเหนือจากนี้ กำหนดให้มีค่าลักษณ์เป็น N คือเป็นคำที่ไม่ใช่ขอบเขตเริ่มต้นอนุพากย์ ตัวอย่างข้อมูลและค่าของลักษณ์เป็นดังนี้

คำ (tokens)	**PUNC**
ไบรอน	N
แมคเคน	N
(	Y
BYRON	N
<s>	N
MCCAIN	N
)	N
<s>	N

พ่อค้า	Y
ขาย	N
ม้า	N

ตารางที่ 5.6 แสดงตัวอย่างจากคลังข้อมูลและการกำหนดลักษณ์เครื่องหมายวรรคตอน

จากตางราง เครื่องหมายวงเล็บเปิด และคำว่า “พ่อค้า” มีค่าเป็น Y เนื่องจากมีความเป็นไปได้ที่จะเป็นจุดเริ่มต้นอนุพากย์ ในขณะที่คำอื่น ๆ มีค่าเป็น N เนื่องจากไม่ตรงกับข้อกำหนดที่ได้กล่าวมา

5.3 การเตรียมไฟล์ข้อมูลสำหรับฝึกฝนและทดสอบแบบจำลอง

ข้อมูลที่ต้องการจะจำแนกประเภทโดยโปรแกรมวีก้าจะต้องจัดให้อยู่ในรูปแบบไฟล์ CSV (Comma Separated Value) หรือ ARFF (Attribute-Relation File Format) ในที่นี้จะใช้ไฟล์นามสกุล ARFF ซึ่งมีโครงสร้างของไฟล์ที่ประกอบไปด้วย 2 ส่วนหลัก คือ ส่วนหัว (Header) และชุดข้อมูล (Data)

ส่วนหัวเป็นส่วนที่แสดงรายละเอียดต่าง ๆ ของข้อมูล ประกอบไปด้วยชื่อ relation ซึ่งเป็นชื่อเรียกตารางข้อมูลเชิงสัมพันธ์ สามารถใช้ชื่อเรียกอะไรก็ได้ ไม่มีผลต่อการตัดสินใจของแบบจำลอง อีกส่วนหนึ่งคือ attribute ซึ่งจะแสดงชื่อและประเภทของลักษณ์ การตั้งชื่อ attribute จะต้องไม่ซ้ำกัน ส่วนประเภทของลักษณ์จะต้องกำหนดให้ถูกต้องตรงตามชุดข้อมูล เช่น numeric, string, nominal ฯลฯ ในส่วนของ attribute สุดท้าย จะเป็นคำตอบที่ต้องการให้เครื่องตอบ ไม่ใช่ลักษณ์ที่ต้องการให้เครื่องพิจารณา ตัวอย่างรูปแบบของส่วนหัวจะมีลักษณะดังนี้

```
@relation weather

@attribute outlook {sunny, overcast, rainy}
@attribute temperature numeric
@attribute humidity numeric
@attribute windy {TRUE, FALSE}
@attribute class {yes, no}
```

ตัวอย่างดังกล่าวนำมาจากงานของ Witten and Frank (2005) แสดงรูปแบบไฟล์ส่วนหัว ชื่อเรียกตารางมีชื่อว่า weather บรรทัดถัดมาเป็นรายละเอียด attribute หรือลักษณ์ที่ใช้ในการจำแนกประเภท ลักษณ์แรกมีชื่อว่า outlook เป็นลักษณ์ประเภท nominal กำหนดให้มีค่าของลักษณ์ 3 แบบ คือ sunny, overcast, rainy ลักษณ์ต่อมามีชื่อว่า temperature ซึ่งเป็นประเภท

numeric หรือมีค่าเป็นตัวเลข บรรทัดสุดท้ายของส่วนหัวคือคำตอบที่ต้องการให้เครื่องเลือก ในที่นี้ตั้งชื่อว่า class กำหนดให้ตอบได้เพียง 2 แบบเท่านั้น คือ yes กับ no

ในส่วนข้อมูล จะขึ้นต้นด้วย @data บรรทัดถัดมาจะแสดงข้อมูลแถวละหนึ่งตัวอย่าง (Instance) แต่ละตัวอย่างจะประกอบไปด้วยค่าของลักษณ์ เรียงลำดับตาม attribute ของส่วนหัว แต่ละค่าคั่นด้วยเครื่องหมายจุลภาค (Comma) โดยค่าสุดท้ายจะเป็นคำตอบเสมอ ตัวอย่างรูปแบบส่วนข้อมูลเป็นดังนี้

@data

sunny, 85, 85, FALSE, no
sunny, 80, 90, TRUE, no
overcast, 83, 86, FALSE, yes
rainy, 70, 96, FALSE, yes
rainy, 68, 80, FALSE, yes
rainy, 65, 70, TRUE, no
overcast, 64, 65, TRUE, yes
sunny, 72, 95, FALSE, no
sunny, 69, 70, FALSE, yes
rainy, 75, 80, FALSE, yes
sunny, 75, 70, TRUE, yes
overcast, 72, 90, TRUE, yes
overcast, 81, 75, FALSE, yes
rainy, 71, 91, TRUE, no

ในวิทยานิพนธ์นี้ ผู้วิจัยใช้ลักษณ์ในการฝึกฝนและทดสอบแบบจำลองทั้งหมด 6 ลักษณ์ ได้แก่ หมวดคำของคำปัจจุบัน (POS-P) หมวดคำก่อนหน้า (POS-B) หมวดคำที่ตามหลัง (POS-A) รายการคำเชื่อมอนุพากย์ (DM) ช่องว่าง (Space) และเครื่องหมายวรรคตอน (Punc) โดยมีรูปแบบคำตอบเป็นแบบสองคำตอบ คือ เป็นขอบเขตเริ่มต้นอนุพากย์ (Boundary) และไม่เป็นขอบเขตเริ่มต้นอนุพากย์ (NonBoundary) ในงานนี้จะไม่มีคำตอบว่าเป็นขอบเขตสิ้นสุดอนุพากย์ เนื่องจากเห็นว่า หากสามารถระบุขอบเขตเริ่มต้นได้แล้ว ก็จะทำให้เห็นขอบเขตขอบของอนุพากย์แล้ว นั่นคือ เมื่อเครื่องระบุจุดใดให้เป็นขอบเขตเริ่มต้นแล้ว จุดที่อยู่ก่อนหน้าก็คือจุดสิ้นสุดนั่นเอง จึงไม่มีความจำเป็นที่จะต้องระบุจุดสิ้นสุดอีก

ในการนำข้อมูลที่กำกับในคลังข้อมูลมาจัดเรียงโครงสร้างใหม่ให้เป็นไฟล์ ARFF นี้ ผู้วิจัยเขียนโปรแกรมเพื่อดึงข้อมูลที่จะใช้เป็นลักษณ์ ข้อมูลที่สามารถดึงได้ทันที ได้แก่ POS ของคำนั้น POS

ของคำก่อนหน้า และ POS ของคำที่ตามหลัง ส่วนลักษณ์อื่น ๆ ได้แก่ ลักษณ์รายการคำเชื่อม ลักษณ์ช่องว่าง และลักษณ์เครื่องหมายวรรคตอน จะแทนค่าลักษณ์ด้วย Y หรือ N หากเข้าเงื่อนไขที่กำหนดไว้ในหัวข้อการกำหนดลักษณ์ ซึ่งได้กล่าวไปแล้วในหัวข้อก่อนหน้านี้ สำหรับคำที่กำกับ <EDU> หรือเป็นคำขอบเขตเริ่มต้นอนุพากย์ ก็จะดึงออกมาและใช้คำว่า Boundary แทน ส่วนคำที่ไม่ได้กำกับ <EDU> ก็จะดึงออกมาและใช้คำว่า NonBoundary ทั้ง 2 นี้จะใช้เป็นคำตอบสำหรับฝึกฝนแบบจำลอง

เมื่อดึงข้อมูลทุกอย่างจากคลังข้อมูลครบแล้ว ก็สร้างไฟล์ขึ้นมาใหม่ ใช้เครื่องหมายจุลภาค (,) คั่นระหว่างทุกอย่างที่ดึงออกมา โดยแต่ละบรรทัดจะประกอบไปด้วย POS ของคำนั้น POS ของคำก่อนหน้า POS ของคำที่ตามหลัง ลักษณ์รายการคำเชื่อม ลักษณ์ช่องว่าง ลักษณ์เครื่องหมายวรรคตอน และคำตอบว่าเป็นขอบเขตเริ่มต้นอนุพากย์หรือไม่ ตามลำดับ จากนั้นเพิ่มส่วนหัวของไฟล์ ARFF เข้าไป ข้อมูลไฟล์ ARFF ที่ใช้ในงานนี้ สามารถดูเพิ่มเติมได้ที่ภาคผนวก ข

5.4 เคอร์เนลฟังก์ชั่นและการตั้งค่าพารามิเตอร์

ในการจำแนกประเภทข้อมูลด้วยเทคนิคซัพพอร์ตเวกเตอร์แมชชีน จะต้องมีการเลือกใช้เคอร์เนลฟังก์ชั่น (Kernal Fuction) ซึ่งมีให้เลือกใช้หลายประเภท ได้แก่ เคอร์เนลแบบเชิงเส้น (Linear) เคอร์เนลแบบโพลีโนเมียล (Polynomial) เรเดียลเบสิสฟังก์ชั่น (Radial Basis Function) เป็นต้น การเลือกเคอร์เนลฟังก์ชั่นที่เหมาะสม ย่อมส่งผลต่อประสิทธิภาพของตัวจำแนกประเภท ในงานนี้ผู้วิจัยเลือกใช้เคอร์เนลแบบโพลีโนเมียล ซึ่งมีการใช้อย่างแพร่หลายในงานทางด้านการประมวลผลภาษาธรรมชาติ แม้จะใช้เวลาในการฝึกฝนแบบจำลองนานกว่าเคอร์เนลแบบเชิงเส้น แต่ก็ให้ผลที่ดีกว่า ในขณะที่เมื่อนำไปเปรียบเทียบกับเรเดียลเบสิสฟังก์ชั่น เคอร์เนลแบบโพลีโนเมียลใช้เวลาในการฝึกฝนน้อยกว่าและให้ผลที่ใกล้เคียงกัน

นอกจากการเลือกเคอร์เนลฟังก์ชั่นแล้ว ยังจำเป็นต้องปรับค่าพารามิเตอร์ต่าง ๆ ให้เหมาะสมเพื่อเพิ่มประสิทธิภาพการทำงานของแบบจำลอง พารามิเตอร์สำคัญที่เกี่ยวข้องกับการจำแนกประเภทด้วยเทคนิคซัพพอร์ตเวกเตอร์แมชชีนของฟังก์ชั่น SMO ในโปรแกรมวีก้ามีดังนี้

1) ค่า C (Soft-Margin Constant) เป็นค่าการควบคุมข้อผิดพลาดที่เกิดจากการฝึกฝนแบบจำลอง (Training error) การเพิ่มค่า C จะทำให้ข้อผิดพลาดของการฝึกฝนแบบจำลองน้อยลง แต่จะทำให้ตัวจำแนกหรือซัพพอร์ตเวกเตอร์แมชชีนสูญเสียคุณสมบัติทั่วไป เนื่องจากค่า C ที่สูงขึ้นจะทำให้เกิดการบังคับให้สร้างแบบจำลองที่สมบูรณ์ที่สุด นอกจากนี้ยังทำให้ต้องใช้เวลาในการฝึกฝนแบบจำลองนานขึ้นอีกด้วย ในงานวิจัยนี้ ผู้วิจัยตั้งค่า C เท่ากับ 1 ซึ่งเป็นค่า default

2) พารามิเตอร์ของเคอร์เนล ในงานนี้ใช้เคอร์เนลแบบโพลีโนเมียล ค่าพารามิเตอร์ที่ต้องปรับคือ ค่า D (Degree of polynomial kernel) หรือ ค่า Exponent D การปรับค่า D ให้เหมาะสม จะช่วยให้แบบจำลองทำงานมีประสิทธิภาพสูงสุด โดย default ของโปรแกรม ค่านี้จะเท่ากับ 1 ซึ่งจะทำให้ได้ตัวจำแนกซัพพอร์ตเวกเตอร์แบบเชิงเส้น (Linear SVM) และหากปรับค่าพารามิเตอร์ตัวนี้ให้มากกว่า 1 ก็จะได้ตัวจำแนกซัพพอร์ตเวกเตอร์แบบไม่เชิงเส้นแทน (Nonlinear SVM) ในงานนี้ผู้วิจัยได้ทำการทดสอบพารามิเตอร์ของเคอร์เนลเพื่อหาค่าพารามิเตอร์ที่เหมาะสมที่สุด โดยการทดลองครั้งแรกจะตั้งค่าพารามิเตอร์ D=1 ก่อน เพื่อหารูปแบบลักษณ์ที่ส่งผลต่อประสิทธิภาพของแบบจำลองมากที่สุด จากนั้นจะนำแบบจำลองที่ใช้รูปแบบลักษณ์ที่ดีที่สุดมาทำการทดลองกับค่าพารามิเตอร์ D=2, D=3 และ D=4 เพื่อดูว่าพารามิเตอร์ของเคอร์เนลค่าใดที่จะทำให้ผลการทดสอบแบบจำลองออกมาดีที่สุด ผู้วิจัยจะไม่ทำการทดสอบค่า D ที่สูงกว่า 4 เนื่องจากค่าที่สูงอาจทำให้เกิดปัญหา overfitting นั่นคือการที่แบบจำลองมีประสิทธิภาพสูงมากเมื่อใช้กับข้อมูลฝึกฝน แต่จะมีประสิทธิภาพต่ำเมื่อใช้กับข้อมูลทดสอบ นอกจากนี้ การตั้งค่า D สูงยังทำให้ต้องใช้เวลาในการฝึกฝนแบบจำลองนานขึ้นอีกด้วย

5.5 การประเมินประสิทธิภาพของแบบจำลอง

การประเมินประสิทธิภาพของแบบจำลองทำได้โดยการหาค่าความแม่นยำ (Precision) ค่าความครบถ้วน (Recall) และค่า F-measure

ค่าความแม่นยำ เป็นการวัดความสามารถของระบบในการขจัด token ที่ไม่ใช่ขอบเขตเริ่มต้นอนุพากย์ออกไป โดยแสดงค่าออกมาในรูปของสัดส่วนของจำนวนขอบเขตอนุพากย์ที่เครื่องตอบถูกต้องเมื่อเปรียบเทียบกับจำนวนของขอบเขตอนุพากย์ที่เครื่องตอบมาทั้งหมด สามารถคำนวณได้จากสมการต่อไปนี้

$$Precision(\text{ เปอร์เซ็นต์}) = \left(\frac{\text{จำนวน boundary ที่เครื่องตอบถูก}}{\text{จำนวน } boundary \text{ ที่เครื่องตอบทั้งหมด}}\right) \cdot 100$$

ค่าความครบถ้วน เป็นการวัดความสามารถของระบบในการระบุ token ที่เป็นขอบเขตเริ่มต้นอนุพากย์ โดยแสดงค่าออกมาในรูปของสัดส่วนของจำนวนขอบเขตอนุพากย์ที่เครื่องตอบถูกต้องเมื่อเปรียบเทียบกับจำนวนขอบเขตอนุพากย์ที่มีทั้งหมดในคลังข้อมูล สามารถคำนวณได้จากสมการต่อไปนี้

$$Recall(\text{ เปอร์เซ็นต์}) = \left(\frac{\text{จำนวน boundary ที่เครื่องตอบถูก}}{\text{จำนวน } boundary \text{ ในคลังข้อมูลทั้งหมด}}\right) \cdot 100$$

F-measure คือค่าที่แสดงความสัมพันธ์ระหว่างค่าความแม่นยำและค่าความครบถ้วน เพื่อหาค่าความถูกต้องโดยรวม สามารถคำนวณได้จากสมการต่อไปนี้

$$F - measure(\text{ เปอร์เซ็นต์}) = \left(\frac{\text{precision } \cdot \text{recall}}{\text{precision } + \text{recall}}\right) \cdot 2$$

ในงานนี้ ผู้วิจัยได้ประเมินประสิทธิภาพของการระบุขอบเขตเริ่มต้นอนุพากย์ภาษาไทยด้วยแบบจำลองซัพพอร์ตเวกเตอร์แมชชีนที่ใช้กับคลังข้อมูลทดสอบ มีการทดสอบทั้งหมด 10 ครั้ง และรายงานผลการทดสอบแต่ละครั้งโดยใช้ค่าความแม่นยำ ค่าความครบถ้วน และ F-measure ตามที่ได้กล่าวมา ผลการทดสอบจะได้กล่าวถึงในลำดับต่อไป

5.6 ผลการทดสอบ

ในการทดสอบการทำงานของแบบจำลองซัพพอร์ตเวกเตอร์แมชชีนที่ใช้ในการระบุขอบเขตหรือจุดเริ่มต้นของอนุพากย์ ผู้วิจัยได้ทดลองใช้ลักษณ์รูปแบบต่าง ๆ ร่วมกัน แล้วนำแต่ละรูปแบบไปทดสอบทั้งหมด 10 ครั้ง และเปรียบเทียบผลการทดสอบ เพื่อดูว่าลักษณ์ใดบ้างที่ส่งผลต่อประสิทธิภาพของแบบจำลอง โดยตั้งค่าพารามิเตอร์ C=1 และ D=1 หลังจากได้ผลการทดลองออกมาแล้ว ผู้วิจัยจะเลือกรูปแบบของลักษณ์ที่ได้ผลการทดสอบออกมาที่สุดมาทำการทดลองอีกครั้งกับค่าพารามิเตอร์ D= 2, D=3 และ D=4 ตามลำดับ เพื่อดูว่าการปรับค่า D ของเคอร์เนลโพลีโนเมียลจะส่งผลต่อประสิทธิภาพแบบจำลองได้มากน้อยเพียงใด ในส่วนของรูปแบบต่าง ๆ ของลักษณ์ มีดังต่อไปนี้

รูปแบบที่ 1 ใช้ลักษณ์เดียว คือ POS ของคำ

รูปแบบที่ 2 ใช้ 3 ลักษณ์ ได้แก่ POS ของคำ, POS คำก่อนหน้า และ POS คำตามหลัง

รูปแบบที่ 3 ใช้ 4 ลักษณ์ ได้แก่ POS ของคำ, POS คำก่อนหน้า, POS คำตามหลัง และคำเชื่อม

รูปแบบที่ 4 ใช้ 4 ลักษณ์ ได้แก่ POS ของคำ, POS คำก่อนหน้า, POS คำตามหลัง และช่องว่าง

รูปแบบที่ 5 ใช้ 4 ลักษณ์ ได้แก่ POS ของคำ, POS คำก่อนหน้า, POS คำตามหลัง และเครื่องหมายวรรคตอน

รูปแบบที่ 6 ใช้ 6 ลักษณ์ ได้แก่ POS ของคำ, POS คำก่อนหน้า, และ POS คำตามหลัง, Discourse Marker, Space, Punctuation

ในการทดสอบรูปแบบลักษณ์ต่าง ๆ ผู้วิจัยกำหนดให้รูปแบบลักษณ์ที่ 1 เป็นการใช้ลักษณ์เพียงลักษณ์เดียว คือ POS ของคำ รูปแบบที่ 2 เป็นการดู POS ของคำที่อยู่ข้างหน้าและข้างหลังร่วมด้วย ส่วนลักษณ์รูปแบบที่ 3–6 เป็นการใช้ลักษณ์ทางภาษาศาสตร์ต่าง ๆ ร่วมกับการใช้ POS ของ 3 คำ เพื่อที่จะได้เห็นว่าลักษณ์เหล่านั้นช่วยเพิ่มประสิทธิภาพให้กับแบบจำลองหรือไม่

5.6.1 ผลการทดสอบของลักษณ์รูปแบบต่าง ๆ ของลักษณ์ พารามิเตอร์ C=1 และ D=1

รูปแบบลักษณ์ในแบบต่าง ๆ จะได้รับการทดสอบรูปแบบละ 10 ครั้ง ทุกส่วนของข้อมูลจะถูกใช้เป็นข้อมูลทดสอบซึ่งคิดเป็น 10 เปอร์เซ็นต์ของคลังข้อมูลทั้งหมด ประเมินผลการทดสอบโดยการคำนวณค่าความแม่นยำ ค่าความครบถ้วน และค่า F-measure ผลการทดสอบเป็นดังตาราง 5.7–5.13

ครั้งที่	1	2	3	4	5	6	7	8	9	10	mean
P	87.6	91.4	92.8	91.3	94.8	91.6	92.7	94.9	88.8	94.7	92.06
R	50.7	40.4	66.6	63.4	70.6	61.5	50.9	53.3	59.5	57.4	57.43
F	64.3	56	77.5	74.9	80.9	73.6	65.7	68.3	71.3	71.5	70.4
จำนวน	461	319	461	420	478	523	416	505	531	496	461

ตารางที่ 5.7 แสดงค่าความแม่นยำ ค่าความครบถ้วน F-measure และจำนวนครั้งที่แบบจำลองระบุขอบเขตเริ่มต้น EDU ได้ถูกต้อง ของแบบจำลองที่ใช้ลักษณ์รูปแบบที่ 1 (POS-P)

ครั้งที่	1	2	3	4	5	6	7	8	9	10	mean
P	94.6	96.2	96.1	95.3	98	96.1	97.7	96.1	94.8	98.9	96.38
R	56.1	55.1	70.4	70.7	80.2	72.2	56.6	59.8	61	61.9	64.4
F	70.4	70	81.2	81.2	88.2	82.4	71.7	73.7	74.2	76.2	76.92
จำนวน	510	435	487	468	543	614	463	566	544	535	516.5

ตารางที่ 5.8 แสดงค่าความแม่นยำ ค่าความครบถ้วน F-measure และจำนวนครั้งที่แบบจำลองระบุขอบเขตเริ่มต้น EDU ได้ถูกต้อง ของแบบจำลองที่ใช้ลักษณ์รูปแบบที่ 2 (POS-P, POS-B, POS-A)

ครั้งที่	1	2	3	4	5	6	7	8	9	10	mean
P	97.5	97.8	97.4	97.2	98.3	97.9	79.8	99.3	98.9	98.4	96.25
R	56.4	55.2	70.7	73.3	83.2	75.6	59.8	65.2	61.1	60.8	66.13
F	71.5	70.6	81.9	83.5	90.1	75.6	74.2	78.7	75.5	75.1	77.67
จำนวน	513	436	489	485	563	643	566	563	545	497	530

ตารางที่ 5.9 แสดงค่าความแม่นยำ ค่าความครบถ้วน F-measure และจำนวนครั้งที่แบบจำลองระบุขอบเขตเริ่มต้น EDU ได้ถูกต้อง ของแบบจำลองที่ใช้ลักษณ์รูปแบบที่ 3 (POS-P, POS-B, POS-A, DM)

ครั้งที่	1	2	3	4	5	6	7	8	9	10	mean
P	94.6	96.2	96.1	95.3	98	96.1	97.7	96.1	94.8	98.9	96.38
R	56.1	55.1	70.4	70.7	80.2	72.2	56.6	59.8	61	61.9	64.4
F	70.4	70	81.2	81.2	88.2	82.4	71.7	73.7	74.2	76.2	76.92
จำนวน	510	435	487	468	543	614	463	566	544	535	516.5

ตารางที่ 5.10 แสดงค่าความแม่นยำ ค่าความครบถ้วน F-measure และจำนวนครั้งที่แบบจำลองระบุขอบเขตเริ่มต้น EDU ได้ถูกต้อง ของแบบจำลองที่ใช้ลักษณ์รูปแบบที่ 4 (POS-P, POS-B, POS-A, Space)

ครั้งที่	1	2	3	4	5	6	7	8	9	10	mean
P	94.9	95.9	96.1	95.5	98.1	96.3	97.7	96.3	94.9	98.9	96.46
R	59.2	62	71.5	73.4	83	77.2	57.3	66.8	64.3	65	67.97
F	72.9	75.3	82	83	89.9	85.7	72.3	78.9	76.7	78.5	79.52
จำนวน	538	490	495	486	562	657	469	633	574	562	546.6

ตารางที่ 5.11 แสดงค่าความแม่นยำ ค่าความครบถ้วน F-measure และจำนวนครั้งที่แบบจำลองระบุขอบเขตเริ่มต้น EDU ได้ถูกต้อง ของแบบจำลองที่ใช้ลักษณ์รูปแบบที่ 5 (POS-P, POS-B, POS-A, Punc)

ครั้งที่	1	2	3	4	5	6	7	8	9	10	mean
P	97.7	97.2	97.5	97.3	98.3	98	98.4	97.8	98.8	99.3	98.03
R	59.5	62.2	71.8	76	85.8	80.4	61.5	66.8	64.5	68.3	69.68
F	74	75.8	82.7	85.3	91.6	88.3	75.7	79.4	78	80.9	81.17
จำนวน	541	491	497	503	581	684	503	633	575	590	559.8

ตารางที่ 5.12 แสดงค่าความแม่นยำ ค่าความครบถ้วน F-measure และจำนวนครั้งที่แบบจำลองระบุขอบเขตเริ่มต้น EDU ได้ถูกต้อง ของแบบจำลองที่ใช้ลักษณ์รูปแบบที่ 6 (POS-P, POS-B, POS-A, DM, Space, Punc)

	รูปแบบ 1	รูปแบบ 2	รูปแบบ 3	รูปแบบ 4	รูปแบบ 5	รูปแบบ 6
P	92.06	96.38	96.25	96.38	96.46	98.03
R	57.43	64.4	66.13	64.4	67.97	69.68
F	70.4	76.92	77.67	76.92	79.52	81.17
จำนวน	461	516.5	530	516.5	546.6	559.8

ตารางที่ 5.13 แสดงค่าเฉลี่ยของผลการทดสอบแบบจำลองที่ใช้ลักษณ์รูปแบบที่ 1 - 6 ในการฝึกฝนและทดสอบ ประกอบไปด้วยค่าความแม่นยำ ค่าความครบถ้วน F-measure และจำนวนครั้งที่แบบจำลองระบุขอบเขต EDU ได้ถูกต้อง

จากตาราง 5.7 ที่รายงานผลการทดสอบ 10 ครั้งของแบบจำลองที่ฝึกฝนด้วยลักษณ์รูปแบบที่ 1 ซึ่งเป็นการใช้หมวดคำของแต่ละคำเพียงอย่างเดียว พบว่าคำนวณค่าความแม่นยำออกมาได้ค่อนข้างสูง คืออยู่ระหว่าง 94.1–97.7 เปอร์เซ็นต์ แสดงให้เห็นว่าคำที่ถูกดึงออกมาส่วนใหญ่จะเป็นคำขอบเขตเริ่มต้นที่ถูกต้อง มีที่ผิดพลาดประมาณ 3-6 เปอร์เซ็นต์เท่านั้น ส่วนค่าความครบถ้วนคำนวณได้ประมาณ 40.4–70.6 เปอร์เซ็นต์ แสดงให้เห็นว่าแบบจำลองยังไม่มีความสามารถที่จะดึงคำในคลังทดสอบที่เป็นขอบเขตเริ่มต้น EDU ได้มากนัก ส่งผลให้คำนวณค่า F-measure ออกมาได้ระหว่าง 56–80.9 เปอร์เซ็นต์ หรือเฉลี่ยได้ 70.4 เปอร์เซ็นต์

เมื่อทดสอบแบบจำลองที่ใช้ลักษณ์รูปแบบที่ 2 ซึ่งเป็นการเพิ่มหมวดคำของคำข้างเคียงเข้าไปได้แก่ หมวดคำก่อนหน้าและหมวดคำที่ตามหลัง ก็จะได้ผลการทดสอบดังตารางที่ 5.8 ซึ่งพบว่าแบบจำลองสามารถตัดสินใจได้ดียิ่งขึ้น คำนวณค่าความแม่นยำ ค่าความครบถ้วน และค่า F-measure ได้มากกว่าการใช้ลักษณ์รูปแบบแรก แต่ยังถือว่าได้ค่าความครบถ้วนที่ไม่สูงมาก ดังจะเห็นได้ว่า ผลการทดสอบจำนวน 6 ครั้ง ได้ค่าความครบถ้วนระหว่าง 55.1–61.9 เปอร์เซ็นต์ และส่งผลให้

ค่าความแม่นยำเฉลี่ยแล้วได้เพียง 64.4 เปอร์เซ็นต์ แสดงว่าแบบจำลองที่ใช้ลักษณ์รูปแบบนี้ยังไม่สามารถดึงคำที่ถูกต้องจากคลังข้อมูลได้มากนัก

ส่วนลักษณ์รูปแบบที่ 3 ได้ผลการทดสอบดังตารางที่ 5.9 เป็นการใช้ลักษณ์รายการคำเชื่อมร่วมกับลักษณ์หมวดคำของ 3 คำ พบว่าได้ค่าความแม่นยำที่ค่อนข้างสูง คืออยู่ระหว่าง 97.2–99.3 เปอร์เซ็นต์ ค่าความครบถ้วนสูงกว่าการใช้ลักษณ์ 2 แบบแรกเล็กน้อย คืออยู่ระหว่าง 55.2–83.2 เปอร์เซ็นต์ แสดงว่าแบบจำลองมีความสามารถในการดึงคำขอบเขตเริ่มต้น EDU ที่ถูกต้องจากคลังข้อมูลเพิ่มขึ้นเล็กน้อยเท่านั้น อย่างไรก็ตามภาพรวมของแบบจำลองที่ใช้ลักษณ์รูปแบบนี้มีประสิทธิภาพที่สูงกว่าการใช้ลักษณ์รูปแบบที่ 2 เล็กน้อย ซึ่งเห็นได้จากค่า F-measure ที่วัดเฉลี่ยแล้วได้ 77.67 เปอร์เซ็นต์ หรือเพิ่มขึ้นไม่ถึง 1 เปอร์เซ็นต์

สำหรับลักษณ์รูปแบบที่ 4 ได้ผลการทดสอบดังตารางที่ 5.10 เป็นการใช้ลักษณ์ช่องว่างร่วมกับลักษณ์หมวดคำของ 3 คำ พบว่าวัดค่าความแม่นยำ ค่าความครบถ้วน และค่า F-measure ได้เท่ากันกับลักษณ์รูปแบบที่ 2 คือได้ 96.38, 64.4 และ 76.92 เปอร์เซ็นต์ตามลำดับ แสดงว่าการใช้ลักษณ์ช่องว่างร่วมกันกับลักษณ์หมวดคำของคำ 3 คำไม่ได้ช่วยเพิ่มประสิทธิภาพของแบบจำลองเลยแม้แต่น้อย อย่างไรก็ตามลักษณ์ช่องว่างกลับมีบทบาทช่วยเพิ่มประสิทธิภาพการตัดสินใจของแบบจำลองได้ หากปรับค่าพารามิเตอร์เคอร์เนลให้สูงขึ้น ซึ่งประเด็นนี้จะอภิปรายต่อไปในบทที่ 6

เมื่อทดสอบแบบจำลองที่ฝึกฝนด้วยลักษณ์รูปแบบที่ 5 ได้ผลการทดสอบดังตารางที่ 5.11 ลักษณ์รูปแบบที่ 5 นี้เป็นการใช้ลักษณ์เครื่องหมายวรรคตอนร่วมกับลักษณ์หมวดคำของ 3 คำ ก็พบว่าคำนวณค่าความแม่นยำได้ใกล้เคียงกับการใช้ลักษณ์หมวดคำของ 3 คำ คือเฉลี่ยได้ 96.46 เปอร์เซ็นต์ แต่ได้ค่าความครบถ้วนเฉลี่ย 67.97 เปอร์เซ็นต์ ซึ่งเพิ่มขึ้นประมาณ 4 เปอร์เซ็นต์ และคำนวณค่า F-measure เฉลี่ยได้ 79.52 เปอร์เซ็นต์

สำหรับแบบจำลองที่ฝึกฝนด้วยลักษณ์รูปแบบที่ 6 ซึ่งประกอบไปด้วยลักษณ์หมวดคำปัจจุบัน หมวดคำก่อนหน้า หมวดคำตามหลัง รายการคำเชื่อม ช่องว่าง และเครื่องหมายวรรคตอน ผลการทดสอบเป็นดังตารางที่ 5.12 พบว่าการรวมทุกลักษณ์เข้าด้วยกัน ทำให้แบบจำลองมีประสิทธิภาพมากที่สุด กล่าวคือ คำนวณค่าความแม่นยำได้ 97.2–99.3 เปอร์เซ็นต์ ค่าความครบถ้วนอยู่ระหว่าง 59.5–85.8 เปอร์เซ็นต์ และคำนวณค่า F-measure เฉลี่ยได้ 81.17 เปอร์เซ็นต์ ซึ่งถือว่าสูงกว่าแบบจำลองที่ใช้ลักษณ์รูปแบบอื่น ๆ ประมาณ 2-11เปอร์เซ็นต์

5.6.2 ผลการทดสอบที่ปรับค่าพารามิเตอร์ C=1 และ D=2, D=3, D=4

จากการประเมินประสิทธิภาพของแบบจำลองที่ฝึกฝนและทดสอบด้วยลักษณ์รูปแบบต่าง ๆ จะพบว่าการใช้ลักษณ์รูปแบบที่ 6 ซึ่งประกอบไปด้วยลักษณ์ POS ของคำปัจจุบัน, POS ของคำก่อน

หน้า, POS ของคำหลัง, รายการคำเชื่อมอนุพากย์, ช่องว่าง, และเครื่องหมายวรรคตอน เป็นรูปแบบของลักษณ์ที่ทำให้แบบจำลองมีประสิทธิภาพสูงสุด ทั้งนี้เป็นการฝึกฝนและทดสอบแบบจำลองโดยใช้ตัวจำแนกประเภทซัพพอร์ตเวกเตอร์แมชชีนที่ตั้งค่าพารามิเตอร์ของเคอร์เนลโพลีโนเมียล D=1 ดังนั้นผู้วิจัยจึงนำลักษณ์เหล่านี้ไปทดสอบต่อ เพื่อหาค่าพารามิเตอร์ของเคอร์เนลที่เหมาะสมต่อไป เนื่องจากการใช้ค่าพารามิเตอร์ของเคอร์เนลที่เหมาะสม จะส่งผลต่อประสิทธิภาพการตัดสินใจของแบบจำลอง ในการทดสอบนี้ ผู้วิจัยจะปรับค่าพารามิเตอร์จากเดิม D=1 เป็น D=2, D=3 และ D=4 ส่วนค่า C ยังใช้ค่าเดิม คือ C=1 ทำการทดสอบทั้งหมด 10 ครั้ง ผลการทดสอบเป็นดังตารางต่อไปนี้

ครั้งที่	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**mean**
P	92.1	91	89.9	89.3	91.7	92.4	90.3	97.8	92.2	93	91.97
R	71.7	74.2	78.6	80.8	88.3	85.5	77.4	66.8	75.1	77.1	77.55
F	80.6	81.7	83.9	84.9	90	88.8	83.3	79.4	82.8	84.3	83.97
B	652	586	544	535	598	728	633	633	670	666	624.5

ตารางที่ 5.14 แสดงค่าเฉลี่ยของผลการทดสอบทั้งหมด 10 ครั้ง ที่มีการปรับค่าพารามิเตอร์ของเคอร์เนล D=2 ให้กับตัวจำแนกประเภทซัพพอร์ตเวกเตอร์แมชชีน

ครั้งที่	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**mean**
P	91.7	91.8	89.8	89.6	92.5	93.6	91.9	92.3	91.4	92.1	91.67
R	73.8	73.4	77.9	81.7	87.9	85.7	77.4	77	75.1	78.5	78.84
F	81.8	81.6	83.4	85.5	90.2	89.4	84	83.9	82.5	84.7	84.7
B	671	580	539	541	595	729	633	729	670	678	636.5

ตารางที่ 5.15 แสดงค่าเฉลี่ยของผลการทดสอบทั้งหมด 10 ครั้ง ที่มีการปรับค่าพารามิเตอร์ของเคอร์เนล D=3 ให้กับตัวจำแนกประเภทซัพพอร์ตเวกเตอร์แมชชีน

ครั้งที่	1	2	3	4	5	6	7	8	9	10	mean
P	91.7	92.4	89.8	90	92.5	93.6	91.9	92.7	91.7	93.4	91.97
R	73.7	73.4	77.9	81.7	87.9	85.8	77.4	76.7	75.1	77.3	78.69
F	81.7	81.8	83.4	85.7	90.2	89.5	84	83.9	82.6	84.6	84.74
B	670	580	539	541	595	730	633	726	670	668	635.2

ตารางที่ 5.16 แสดงค่าเฉลี่ยของผลการทดสอบทั้งหมด 10 ครั้ง ที่มีการปรับค่าพารามิเตอร์ของเคอร์เนล D=4 ให้กับตัวจำแนกประเภทซัพพอร์ตเวกเตอร์แมชชีน

	D=1	D=2	D=3	D=4
P	98.03	91.97	91.67	91.97
R	69.68	77.55	78.84	78.69
F	81.17	83.97	84.70	84.74

ตารางที่ 5.17 สรุปค่าเฉลี่ยของผลการทดสอบทั้งหมด 10 ครั้ง ที่มีการปรับค่าพารามิเตอร์ของเคอร์เนลต่าง ๆ ให้กับตัวจำแนกประเภทซัพพอร์ตเวกเตอร์แมชชีน

จากตาราง 5.14 ซึ่งปรับค่าพารามิเตอร์ D=2 เมื่อทดสอบแบบจำลองที่ฝึกฝนด้วยลักษณ์รูปแบบที่ 6 นั่นคือใช้ทุกลักษณ์ร่วมกัน ทดสอบทั้งหมด 10 ครั้ง คำนวณค่าความแม่นยำออกมาได้ไม่ต่ำกว่า 90 เปอร์เซ็นต์ทุกครั้ง ความครบถ้วนคำนวณได้ต่ำกว่า 70 เปอร์เซ็นต์แค่ครั้งเดียว คือในการทดสอบครั้งที่ 8 นอกนั้นวัดค่าได้ระหว่าง 71.1–88.3 เปอร์เซ็นต์ ค่าความครบถ้วนนี้ชี้ให้เห็นว่าแบบจำลองสามารถดึงคำขอบเขตเริ่มต้น EDU จากคลังข้อมูลได้ถูกต้องจำนวนมากขึ้น เมื่อเปรียบเทียบกับการทดสอบแบบจำลองที่ปรับค่าพารามิเตอร์ D=1 ส่วนค่า F-measure เฉลี่ยวัดได้ 83.97 เปอร์เซ็นต์ ซึ่งสูงกว่าการปรับค่าพารามิเตอร์ D=1 เช่นกัน

ส่วนการปรับค่าพารามิเตอร์ D=3 ได้ผลการทดสอบ 10 ครั้งดังตาราง 5.15 ซึ่งแสดงให้เห็นว่าแบบจำลองระบุขอบเขตเริ่มต้น EDU ได้ผลดียิ่งขึ้น ดังจะเห็นได้จากค่าเฉลี่ย F-measure ที่สูงกว่าการปรับ D=1 และ D=2 คือวัดได้ 84.70 เปอร์เซ็นต์ แม้ว่าค่าความแม่นยำจะตกลงมาเล็กน้อย นั่นคืออยู่ระหว่าง 89.6–92.5 เปอร์เซ็นต์ แต่วัดค่าความครบถ้วนได้มากขึ้นเล็กน้อย คือเฉลี่ยได้ 78.84 เปอร์เซ็นต์ โดยรวมแล้ว แบบจำลองนี้สามารถดึงคำขอบเขตเริ่มต้น EDU จากคลังข้อมูลได้ถูกต้องเพิ่มขึ้นเล็กน้อย เมื่อเปรียบเทียบกับแบบจำลองที่ปรับค่าพารามิเตอร์ D=2

สำหรับการปรับค่าพารามิเตอร์ D=4 นั้น ได้ผลการทดสอบ 10 ครั้งดังตาราง 5.16 พบว่าคำนวณค่าความแม่นยำ ค่าความครบถ้วน และค่า F-measure ได้สูงกว่าการปรับ D=3 เพียง

เล็กน้อยเท่านั้น นั่นคือวัดค่าเฉลี่ยได้ 91.97, 78.69 และ 84.74 เปอร์เซ็นต์ตามลำดับ เมื่อเปรียบเทียบผลการทดสอบ 10 ครั้งของแบบจำลองที่ปรับค่า D=4 และ D=3 แล้ว จะเห็นว่าแต่ละครั้งที่ทดสอบจะได้ผลการทดสอบใกล้เคียงกันทุกครั้ง แสดงให้เห็นว่า มีความเป็นไปได้ว่า หากมีการปรับค่าพารามิเตอร์ของเคอร์เนลให้สูงไปมากกว่านี้ ก็อาจจะไม่ทำให้ได้ผลที่ดีขึ้นมากแล้ว

โดยสรุปแล้วจะเห็นได้ว่า การปรับค่าพารามิเตอร์ของเคอร์เนลให้สูงขึ้น ส่งผลให้ตัวจำแนกประเภทซัพพอร์ตเวกเตอร์แมชชีนมีประสิทธิภาพสูงขึ้นตามไปด้วย ทั้งนี้ การปรับค่าพารามิเตอร์ของเคอร์เนลแบบโพลีโนเมียลมากกว่า 1 จะทำให้ได้ตัวจำแนกประเภทซัพพอร์ตเวกเตอร์แมชชีนแบบไม่เชิงเส้น หรือ nonlinear SVM และการปรับค่าให้เท่ากับ 1 จะได้ซัพพอร์ตเวกเตอร์แมชชีนแบบเชิงเส้น หรือ linear SVM ซึ่งเมื่อพิจารณาผลการทดสอบดังตารางข้างต้นก็จะพบว่าตัวจำแนกประเภท nonlinear SVM มีประสิทธิภาพในการระบุขอบเขตเริ่มต้นอนุพากย์มากกว่า linear SVM

นอกจากนี้ ผู้วิจัยมีความเห็นว่า การปรับค่าพารามิเตอร์ D=2, D=3 และ D=4 ให้ผลที่ค่อนข้างใกล้เคียงกัน คือวัดค่า F-measure ได้ 83.97, 84.70 และ 84.74 เปอร์เซ็นต์ ตามลำดับ โดยการปรับค่า D=2 จะใช้เวลาในการฝึกฝนแบบจำลองที่น้อยกว่า ในขณะที่การปรับค่า D=4 ใช้เวลามากที่สุด ดังนั้นผู้วิจัยจึงคิดว่าค่าพารามิเตอร์ที่เหมาะสมที่สุดสำหรับการทดลองนี้ควรจะอยู่ที่ D=3 เพราะนอกจากจะใช้เวลาในการฝึกฝนแบบจำลองไม่มากนัก ยังได้ผลการทดสอบที่วัดค่าออกมาแล้วได้สูงกว่าค่า D=2 และในขณะเดียวกันก็ใกล้เคียงมากกับการใช้ค่า D=4 อีกด้วย

จากที่ได้กล่าวมาทั้งหมดนี้ ชี้ให้เห็นว่าการใช้ POS ของ 3 คำติดกัน หรือ POS-B|POS-P|POS-A ร่วมกับรายการคำเชื่อม ช่องว่างที่มีความเป็นไปได้ที่จะเป็นตัวแบ่งขอบเขตอนุพากย์ และเครื่องหมายวรรคตอนที่มักจะอยู่หัวและท้ายอนุพากย์ เหล่านี้สามารถช่วยฝึกฝนแบบจำลองสำหรับใช้ระบุขอบเขตเริ่มต้นอนุพากย์ภาษาไทยได้ นอกจากเรื่องลักษณ์แล้ว การปรับค่าพารามิเตอร์ของเคอร์เนลที่ใช้กับตัวจำแนกประเภทซัพพอร์ตเวกเตอร์แมชชีนก็สามารถช่วยเพิ่มประสิทธิภาพของแบบจำลองได้เป็นอย่างดี อย่างไรก็ตามการใช้ค่าพารามิเตอร์ที่สูงมาก ๆ อาจส่งผลเสียต่อแบบจำลองได้ เพราะอาจจะมีความเสี่ยงที่จะเกิดปัญหา overfitting หรือปัญหาการมีประสิทธิภาพในการฝึกฝนแบบจำลองที่ดีเยี่ยม แต่เมื่อนำไปใช้กับคลังข้อมูลทดสอบ กลับให้ผลการทดสอบที่แย่ อีกทั้งยังทำให้ต้องใช้เวลานานเกินไปในการฝึกฝนแบบจำลองอีกด้วย

บทที่ 6

ลักษณ์ทางภาษาที่มีผลต่อประสิทธิภาพของแบบจำลอง

เนื่องจากการใช้ตัวจำแนกซัพพอร์ตเวกเตอร์แมชชีนต้องอาศัยลักษณ์ ในการฝึกฝนแบบจำลอง การใช้ลักษณ์ที่เหมาะสมจึงเป็นการช่วยเพิ่มประสิทธิภาพของตัวจำแนก ในบทนี้ ผู้วิจัยจะอภิปรายประสิทธิภาพของลักษณ์ทางภาษาที่ใช้ในการฝึกฝนแบบจำลองสำหรับระบุคำขอบเขตเริ่มต้น EDU และประสิทธิภาพของแบบจำลองเมื่อมีการปรับค่าพารามิเตอร์ของเคอร์เนลให้สูงขึ้น

6.1 ประสิทธิภาพของลักษณ์ทางภาษาที่ใช้

จากรูปแบบโครงสร้างอนุพากย์และนามวลีที่เป็น EDU ที่ได้กล่าวถึงไปแล้วในบทก่อนหน้า นำมาสู่การกำหนดลักษณ์ทางภาษาสำหรับใช้ในการฝึกฝนแบบจำลอง ได้แก่ การใช้หมวดคำ การใช้รายการคำเชื่อม การใช้ความน่าจะเป็นที่ช่องว่างจะเป็นตัวแบ่งอนุพากย์ และการใช้เครื่องหมายวรรคตอนมักปรากฏในตำแหน่งที่เป็นขอบเขต EDU ลักษณ์แต่ละตัวมีบทบาทในการช่วยระบุขอบเขตเริ่มต้น EDU ในระดับที่แตกต่างกันออกไป ประสิทธิภาพของลักษณ์ที่จะอภิปรายในหัวข้อนี้ ผู้วิจัยได้วิเคราะห์จากการพิจารณาผลการทดสอบแบบจำลองที่มีการปรับค่าพารามิเตอร์ไว้ต่ำสุด ซึ่งเป็นค่า default ของโปรแกรม รายละเอียดมีดังต่อไปนี้

6.1.1 หมวดคำ

เมื่อพิจารณาโครงสร้างภายในของอนุพากย์ ซึ่งประกอบไปด้วยหน่วยสร้างนามวลีในตำแหน่งประธานตามด้วยหน่วยสร้างกริยาวลีในตำแหน่งภาคแสดง ก็จะพบว่ามีการเรียงตัวกันของหมวดคำต่าง ๆ (POS sequence) เช่นข้อความจากคลังข้อมูล “แวน โก๊ะห์ ก็ได้นำเก้าอี้ถักสีเหลืองที่เคยวางแจกันใบนั้นมาเป็นแบบ” มีการเรียงลำดับหมวดคำคือ NPRP-NPRP-CSBI-AUX-VERB-NCMN-VERB-ADJ-COMPF-AUX-VERB-NCMN-NCLS-DET-VERB-VERB-NCMN ทั้งนี้การเรียงตัวกันของหมวดคำของอนุพากย์มักจะปรากฏเป็นรูปแบบ (Pattern) ที่ซ้ำ ๆ กัน เช่นรูปแบบ นาม+คำปฏิเสธ+กริยา+นาม เป็นต้น ดังนั้นการกำกับหมวดคำในคลังข้อมูลจึงมีประโยชน์ต่อการแยกอนุพากย์อย่างมาก เพราะนอกจากจะสามารถช่วยแยกความกำกวมของคำได้แล้ว เช่นคำว่า “ป้าย” เป็นได้ทั้งคำนามและกริยา สายของหมวดคำที่เรียงกันยังช่วยทำให้เห็นโครงสร้างทางวากยสัมพันธ์อีกด้วย ดังเช่นตัวอย่างที่กล่าวไปแล้วข้างต้น

ในเบื้องต้น ผู้วิจัยได้ทำการทดสอบแบบจำลองโดยใช้ลักษณ์หมวดคำเพียงลักษณ์เดียวก่อน ผลจากการทดสอบ พบว่าสามารถระบุคำขอบเขตเริ่มต้น EDU ได้มากกว่าครึ่งหนึ่งของ EDU ทั้งหมดในคลังข้อมูลทดสอบ EDU ส่วนใหญ่ที่แบบจำลองระบุขอบเขตเริ่มต้นได้ถูกต้องจะเป็น EDU ที่นำหน้าด้วยคำบ่งชี้ ได้แก่ คำเชื่อมและตัวนำส่วนเติมเต็ม ความผิดพลาดที่เกิดขึ้นส่วนใหญ่เกิดจาก

การไม่สามารถระบุขอบเขตอนุพากย์ที่มีช่องว่างปรากฏด้านหน้าและไม่มีคำบ่งชี้อนุพากย์ขึ้นต้น เช่น “$_ซึ่งเป็นผู้ช่วยแม่บ้าน<s>$_จอร์จเล่าเรื่องนี้ให้มิชิโยะฟัง” ประกอบไปด้วยขอบเขตเริ่มต้น EDU 2 แห่ง คือตรงคำว่า “ซึ่ง” และ “จอร์จ” เครื่องหมาย <s> แสดงการใช้ช่องว่าง และเครื่องหมาย $_ แสดงขอบเขตเริ่มต้น EDU ในข้อความนี้ แบบจำลองไม่สามารถระบุคำขอบเขตเริ่มต้น EDU ที่ 2 ได้ ดังนั้นจึงเป็น “$_ซึ่งเป็นผู้ช่วยแม่บ้าน<s>จอร์จเล่าเรื่องนี้ให้มิชิโยะฟัง” ทั้งนี้เกิดจากการที่แบบจำลองมองเห็นเพียงหมวดคำของคำของที่กำลังพิจารณาเท่านั้น ไม่ได้พิจารณาหมวดคำรอบข้าง ทำให้แบบจำลองไม่สามารถตัดสินใจได้ในกรณีที่เป็น token ขอบเขตเริ่มต้น EDU ที่มีช่องว่างด้านหน้าได้ จากตัวอย่างดังกล่าวจึงเห็นได้ว่า แบบจำลองไม่สามารถระบุให้คำ “จอร์จ” เป็นคำขอบเขตเริ่มต้น EDU ใหม่ได้ เพราะแบบจำลองไม่ได้สนใจว่า “จอร์จ” ปรากฏติดกับช่องว่างด้านหน้าหรือไม่

นอกจากนี้แบบจำลองยังไม่สามารถระบุคำขอบเขตเริ่มต้น EDU ที่ไม่มีคำบ่งชี้นำหน้า และเป็น EDU ที่ติดกับ EDU ก่อนหน้าโดยไม่มีช่องว่างคั่น เช่น “$_สำนึกทางจริยธรรมบางอย่าง$_ที่เคยมีมาเมื่อก่อนสมัยใหม่ในญี่ปุ่นนั้น$_จางหายไป” มีทั้งหมด 3 EDU แต่แบบจำลองระบุขอบเขตเริ่มต้นได้เพียง 2 แห่งเท่านั้น คือ “$_สำนึกทางจริยธรรมบางอย่าง$_ที่เคยมีมาเมื่อก่อนสมัยใหม่ในญี่ปุ่นนั้นจางหายไป”

สำหรับ EDU ที่ขึ้นต้นด้วยเครื่องหมายวงเล็บเปิด พบว่า การใช้ POS ของ 3 คำ สามารถระบุวงเล็บเปิดเหล่านี้ให้เป็นขอบเขตเริ่มต้น EDU ได้เป็นจำนวนมาก แต่สำหรับคำในตำแหน่งหลังวงเล็บปิด ซึ่งควรจะต้องได้รับการระบุให้เป็นขอบเขตเริ่มต้น EDU ด้วย แบบจำลองกลับไม่สามารถตัดสินใจได้ถูกต้อง ดังนั้นควรต้องมีการใช้ลักษณ์อื่นช่วยเพื่อแก้ปัญหาตรงนี้ ในงานนี้ ผู้วิจัยจึงมีการใช้ลักษณ์เครื่องหมายวรรคตอนในการแก้ปัญหานี้ ซึ่งจะได้กล่าวถึงต่อไปในเรื่องของลักษณ์เครื่องหมายวรรคตอน

อีกประเด็นคือคำเชื่อมที่มีลักษณะเป็นวลีซึ่งจะถูกแยกออกเป็นหลาย token เช่น “แต่เมื่อ” ถูกแยกเป็น แต่|เมื่อ และแต่ละ token ก็จะมี POS ประจำคำ ในที่นี้ “แต่” ถูกกำกับให้เป็นคำเชื่อม CCOR และ “เมื่อ” กำกับให้เป็นคำเชื่อม CSUB แบบจำลองจะมองว่าทั้ง 2 คำนี้สามารถเป็นขอบเขตเริ่มต้น EDU ได้ จึงระบุทั้งคู่เป็นคำขอบเขตเริ่มต้น เช่น “$_แต่เมื่อปรับตัวโดยปราศจากฐานคิด” มีจุดเริ่มต้น EDU ตรงคำว่า “แต่” เพียงแห่งเดียว แต่แบบจำลองระบุขอบเขตเริ่มต้น EDU ถึง 2 แห่งด้วยกัน คือ “$_แต่$_เมื่อปรับตัวโดยปราศจากฐานคิด”

ทั้งหมดนี้คือปัญหาของการให้แบบจำลองมอง POS ประจำคำเพียงอย่างเดียวโดยไม่อาศัยบริบทรอบข้าง ทำให้แบบจำลองตัดสินใจโดยดูความเป็นไปได้ของข้อมูลที่ฝึกฝนทั้งหมด ซึ่งจะพบว่าคำบ่งชี้ต่าง ๆ มักเป็นขอบเขตเริ่มต้น EDU เสมอ ดังนั้นทุกครั้งที่แบบจำลองพบคำบ่งชี้ก็จะระบุให้

เป็นขอบเขตเริ่มต้น EDU และถ้าคำไหนไม่ใช่คำบ่งชี้ ก็จะตัดสินใจไม่ให้คำนั้นเป็นขอบเขตเริ่มต้น EDU

ทั้งนี้ ความผิดพลาดเรื่องคำเชื่อมที่มีเป็นรูปแบบวลีและมี POS เป็น CCOR และ CSUB ได้รับการแก้ไขเมื่อใช้ลักษณ์หมวดคำของ 3 token ได้แก่ คำปัจจุบัน คำก่อนหน้า และคำที่ตามหลัง ซึ่งเมื่อทดสอบแบบจำลอง พบว่าแบบจำลองจะระบุขอบเขต EDU ให้กับ token แรกของคำเชื่อมที่เป็นวลีเท่านั้น เช่น จากตัวอย่างเดิม “**$**_แต่เมื่อปรับตัวโดยปราศจากฐานคิด” แบบจำลองระบุขอบเขต EDU ตรงคำว่า “แต่” เท่านั้น ส่วนคำเชื่อมเด่นซึ่งจะมี POS เป็น ADVERB นั้น แบบจำลองยังไม่สามารถตัดสินใจได้ถูกต้อง ตัวอย่างจากคลังข้อมูล แสดงการเปรียบเทียบการใช้ POS ของ 3 คำ และการใช้ POS เดียว เป็นดังนี้

ขอบเขตเริ่มต้น EDU ที่ถูกต้อง	**ใช้ POS ของ 3 token**	**ใช้ POS เดียว**
(17) **$**_แต่ทั้งนี้ภาครัฐหรือภาคเอกชนจะต้องให้การสนับสนุน	**$**_แต่ทั้งนี้ภาครัฐหรือภาคเอกชนจะต้องให้การสนับสนุน	**$**_แต่**$**_ทั้งนี้ภาครัฐหรือภาคเอกชนจะต้องให้การสนับสนุน
(18) **$**_และเมื่อเข้าร่วมกิจกรรมแล้ว	**$**_และเมื่อเข้าร่วมกิจกรรมแล้ว	**$**_และ**$**_เมื่อเข้าร่วมกิจกรรมแล้ว
(19) ...ความสัมพันธ์ในแง่ต่างๆ<s>**$**_เช่น<s>ความสัมพันธ์เชิงอำนาจ<s>ตามชาติพันธุ์<s>ตามเพศ...	...ความสัมพันธ์ในแง่ต่างๆ<s>เช่น<s>ความสัมพันธ์เชิงอำนาจ<s>ตามชาติพันธุ์<s>ตามเพศ...	...ความสัมพันธ์ในแง่ต่างๆ<s>เช่น<s>ความสัมพันธ์เชิงอำนาจ<s>ตามชาติพันธุ์<s>ตามเพศ...

ตารางที่ 6.1 ตัวอย่างข้อความจากคลังข้อมูล แสดงผลการใช้ลักษณ์ POS ของ 3 คำ

จากตัวอย่าง 19 ข้างต้น จะเห็นได้ว่า แม้ว่าผลการทดสอบ POS ของคำ 3 คำ จะทำให้สามารถระบุคำขอบเขตเริ่มต้น EDU ได้ถูกต้องมากขึ้น แต่การใช้ POS ของ 3 คำติดกันไม่สามารถระบุคำเริ่มต้น EDU ที่เป็นคำเชื่อมเด่น หรือ strong marker ได้ สาเหตุน่าจะเนื่องจากคำเชื่อมเด่นถูกกำกับหมวดคำเป็น ADVERB นอกจากนี้ปัญหาคำเชื่อมอนุพากย์ที่เป็นวลีก็ยังไม่สามารถขจัดได้ทั้งหมด อีกทั้งยังไม่ได้ช่วยแก้ปัญหาอนุพากย์ที่มีช่องว่างปรากฏหน้าและไม่ได้ขึ้นต้นด้วยคำบ่งชี้ ดังนั้นการใช้ข้อมูล POS เพียงอย่างเดียวไม่สามารถทำให้แบบจำลองระบุขอบเขตเริ่มต้น EDU ได้อย่างมีประสิทธิภาพสูงสุด

6.1.2 คำเชื่อมหน้าอนุพากย์

จากการสำรวจคลังข้อมูลซึ่งประกอบไปด้วย 8,102 EDU พบว่า เฉพาะอนุพากย์ที่ขึ้นต้นด้วยอนุสันธานและสันธานประสานรวมกันมีจำนวน 2,349 EDU คิดเป็นเกือบ 30 เปอร์เซ็นต์ของคำ

ขอบเขตเริ่มต้น EDU ทั้งคลังข้อมูล ดังนั้นหากใช้คำเชื่อมเหล่านี้ในการช่วยระบุขอบเขตเริ่มต้น EDU ก็จะทำให้แบบจำลองสามารถแยก EDU ได้จำนวนไม่น้อย ผู้วิจัยจึงใช้คำเชื่อมหน้าอนุพากย์เป็นลักษณ์ที่ใช้ในการฝึกฝนแบบจำลอง โดยจัดทำรายการคำเชื่อมแล้วนำไปเทียบกับคำในคลังข้อมูล ด้วยวิธีการเช่นนี้ สามารถขจัดปัญหาคำเชื่อมที่เป็นวลีได้ อีกทั้งยังสามารถจัดการกับคำเชื่อมเด่นที่ขึ้นต้น EDU เช่นคำว่า “ยกตัวอย่าง” “ตัวอย่างเช่น” “เช่น” ฯลฯ ได้อีกด้วย เพราะคำเชื่อมเด่นถือว่าเป็นตัวเชื่อมอนุพากย์ จึงอยู่ในรายการคำเชื่อมที่จัดทำขึ้นด้วย

เมื่อทดสอบว่าลักษณ์รายการคำเชื่อมช่วยเพิ่มประสิทธิภาพให้กับแบบจำลองที่ใช้ POS ของ 3 คำติดกันได้มากน้อยเพียงใด ก็พบว่าแบบจำลองสามารถระบุจำนวนคำขอบเขตเริ่มต้น EDU ได้ถูกต้องเพิ่มขึ้นประมาณ 3 เปอร์เซ็นต์ ในจำนวนนี้ส่วนใหญ่พบว่าเป็นคำเชื่อมเด่นที่ขึ้นต้น EDU ต่อไปนี้เป็นตัวอย่างจากคลังข้อมูล แสดงการเปรียบเทียบการระบุคำขอบเขตเริ่มต้น EDU ของแบบจำลองที่ใช้ลักษณ์รายการคำเชื่อมและแบบจำลองที่ไม่ได้ใช้รายการคำเชื่อม

ขอบเขตเริ่มต้น EDU ที่ถูกต้อง	**POS 3 คำ และคำเชื่อม**	**POS ของ 3 คำ**
(20) **$_**ที่เป็นคู่ตรงกันข้าม<s>**$_**เช่น<s>สุกกับดิบ<s>วัฒนธรรมกับธรรมชาติ<s>เป็นต้น	**$_**ที่เป็นคู่ตรงกันข้าม<s>**$_**เช่น<s>สุกกับดิบ<s>วัฒนธรรมกับธรรมชาติ<s>เป็นต้น	**$_**ที่เป็นคู่ตรงกันข้าม<s>เช่น<s>สุกกับดิบ<s>วัฒนธรรมกับธรรมชาติ<s>เป็นต้น
(21) **$_**และยังเสนอให้เน้นถึงความสำคัญของกระบวนการผสมผสานความคิด<s>**$_**อาทิเช่น<s>กระบวนการสร้างอัตลักษณ์แบบลูกผสม<s>และกระบวนการผสมผสานความรู้ในสถานการณ์เฉพาะ	**$_**และยังเสนอให้เน้นถึงความสำคัญของกระบวนการผสมผสานความคิด<s>**$_**อาทิเช่น<s>กระบวนการสร้างอัตลักษณ์แบบลูกผสม<s>และกระบวนการผสมผสานความรู้ในสถานการณ์เฉพาะ	**$_**และยังเสนอให้เน้นถึงความสำคัญของกระบวนการผสมผสานความคิด<s>อาทิเช่น<s>กระบวนการสร้างอัตลักษณ์แบบลูกผสม<s>และกระบวนการผสมผสานความรู้ในสถานการณ์เฉพาะ

ตารางที่ 6.2 ตัวอย่างข้อความจากคลังข้อมูล แสดงผลการใช้ลักษณ์คำเชื่อม

จะเห็นได้ว่าการใช้รายการคำเชื่อมร่วมกับหมวดคำของคำ 3 คำ ทำให้คำ “เช่น” ในตัวอย่าง (20) และคำ “อาทิเช่น” ในตัวอย่าง (21) สามารถระบุให้เป็นคำขอบเขตเริ่มต้น EDU ได้ถูกต้อง ในขณะที่การใช้หมวดคำของคำ 3 คำโดยไม่มีรายการคำเชื่อมไม่สามารถทำได้

6.1.3 ช่องว่าง

การใช้ช่องว่างในภาษาไทยมีความน่าสนใจเป็นอย่างยิ่ง เนื่องจากช่องว่างในภาษาเขียนภาษาไทยมีหลายหน้าที่ ทั้งนี้ ผู้วิจัยเชื่อว่าช่องว่างที่ทำหน้าที่คั่นระหว่าง EDU มีความเป็นไปได้ที่จะปรากฏในบางบริบทเท่านั้น นั่นคือ ช่องว่างแล้วตามด้วยหมวดคำบางประเภทมีโอกาสที่จะเป็นตัวคั่น

EDU มากกว่าเมื่อตามด้วยบางหมวดคำ กล่าวคือ จากการศึกษาคลังข้อมูล พบว่าช่องว่างแล้วตามด้วยตัวเลขหรือตามด้วยคำภาษาต่างประเทศมีความเป็นไปได้น้อยที่ช่องว่างนั้นจะเป็นตัวคั่นอนุพากย์ ในขณะที่ช่องว่างแล้วตามด้วยคำเชื่อมมีความเป็นไปได้สูงมากที่ช่องว่างนั้นจะเป็นตัวคั่นอนุพากย์ ด้วยเหตุนี้ผู้วิจัยจึงใช้หมวดคำที่ปรากฏหลังช่องว่าง เพื่อให้แบบจำลองใช้ประกอบการตัดสินใจว่าช่องว่างนั้นจะเป็นตัวคั่นอนุพากย์และทำให้คำที่อยู่ติดกับช่องว่างนั้นเป็นขอบเขต EDU หรือไม่

ผลการทดสอบแบบจำลองที่ใช้ลักษณ์ POS ของ 3 คำติดร่วมกับลักษณ์ช่องว่าง โดยปรับค่าพารามิเตอร์ของเคอร์เนล D=1 พบว่าการใช้ลักษณ์ช่องว่างไม่ได้ช่วยเพิ่มประสิทธิภาพการตัดสินใจของแบบจำลองได้ คำที่เป็นขอบเขตเริ่มต้น EDU และนำหน้าด้วยช่องว่างส่วนใหญ่ยังไม่สามารถถูกดึงออกมาได้ ตัวอย่างการระบุคำที่เป็นขอบเขตเริ่มต้น EDU ที่มีช่องว่างข้างหน้าเมื่อเทียบกับการใช้ POS ของ 3 คำติดกันเป็นดังนี้

ขอบเขตเริ่มต้น EDU ที่ถูกต้อง	**POS 3 คำ ร่วมกับ space**	**POS ของ 3 คำ**
(22) **$**_ซึ่งถือเป็นเงื่อนไขหรือบริบทของเรื่องนั้น<s>**$**_ประกอบด้วยบริบทของวัฒนธรรม<s>บริบททางประวัติศาสตร์<s>บริบทของระบบนิเวศ<s>และบริบททางสังคมของความสัมพันธ์	**$**_ซึ่งถือเป็นเงื่อนไขหรือบริบทของเรื่องนั้น<s>ประกอบด้วยบริบทของวัฒนธรรม<s>บริบททางประวัติศาสตร์<s>บริบทของระบบนิเวศ<s>และบริบททางสังคมของความสัมพันธ์	**$**_ซึ่งถือเป็นเงื่อนไขหรือบริบทของเรื่องนั้น<s>ประกอบด้วยบริบทของวัฒนธรรม<s>บริบททางประวัติศาสตร์<s>บริบทของระบบนิเวศ<s>และบริบททางสังคมของความสัมพันธ์
(23) **$**_จะต้องเข้าใจความหมายของคนชายชอบให้ดี<s>**$**_ไม่ใช่มองแต่เฉพาะความเป็นชายขอบในแง่ภูมิศาสตร์	**$**_จะต้องเข้าใจความหมายของคนชายชอบให้ดี<s>ไม่ใช่มองแต่เฉพาะความเป็นชายขอบในแง่ภูมิศาสตร์	**$**_จะต้องเข้าใจความหมายของคนชายชอบให้ดี<s>ไม่ใช่มองแต่เฉพาะความเป็นชายขอบในแง่ภูมิศาสตร์

ตารางที่ 6.3 ตัวอย่างข้อความจากคลังข้อมูล แสดงผลการใช้ลักษณ์ช่องว่าง

จากตารางจะเห็นว่า ทั้ง 2 ตัวอย่างมีช่องว่างคั่นระหว่าง EDU โดยตัวอย่างแรกเป็นช่องว่างแล้วตามด้วยคำกริยา “ประกอบด้วย” เป็นคำเริ่มต้นขอบเขต EDU และอีกตัวอย่างเป็นช่องว่างแล้วตามด้วยคำบอกปฏิเสธ “ไม่” เป็นคำเริ่มต้นขอบเขต EDU แบบจำลองที่ใช้ลักษณ์ช่องว่างไม่สามารถระบุจุดที่เป็นขอบเขตเริ่มต้น EDU ที่มีช่องว่างปรากฏอยู่ด้านหน้าได้ เช่นเดียวกันกับในกรณีแบบจำลองที่ใช้เพียงลักษณ์ POS ของ 3 คำ

อย่างไรก็ตาม เมื่อมีการปรับค่าพารามิเตอร์เคอร์เนลให้มากกว่า 1 พบว่าลักษณ์ช่องว่างช่วยแบบจำลองในการตัดสินใจคำขอบเขตเริ่มต้น EDU ที่มีช่องว่างนำหน้าได้ ซึ่งประเด็นนี้จะกล่าวถึงต่อไปในหัวข้อ 6.2 เรื่องประสิทธิภาพแบบจำลองเมื่อปรับพารามิเตอร์เคอร์เนลให้สูงขึ้น

6.1.4 เครื่องหมายวรรคตอน

จากการใช้ลักษณ์เครื่องหมายวรรคตอน ผู้วิจัยกำหนดให้มองเฉพาะเครื่องหมายวงเล็บเปิดและวงเล็บปิดเท่านั้น เนื่องจากเป็นเครื่องหมายวงเล็บสามารถช่วยระบุขอบเขตของ EDU ได้ ในงานนี้กำหนดให้ token ที่เป็นวงเล็บเปิด และ token ที่ไม่ใช่ช่องว่างและอยู่ถัดจากเครื่องหมายวงเล็บปิดเป็นขอบเขตเริ่มต้น EDU การกำหนดลักษณ์นี้จึงมีลักษณะเป็นกฎ ผู้วิจัยทดสอบประสิทธิภาพของลักษณ์เครื่องหมายวรรคตอนนี้โดยการใช้ผลการทดสอบชุดของลักษณ์รูปแบบที่ 2 นั่นคือประกอบไปด้วย POS ของคำ 3 คำ เปรียบเทียบกับผลการทดสอบชุดลักษณ์รูปแบบที่ 5 คือเป็นการเพิ่มลักษณ์เครื่องหมายวรรคตอนเข้าไป พบว่า ค่าความครบถ้วนและค่า F-measure มีค่าเพิ่มขึ้นเมื่อใช้ลักษณ์รูปแบบที่ 5 และเมื่อพิจารณาผลลัพธ์ของการทดสอบ ก็พบว่าค่าที่เพิ่มขึ้นมานี้ เกิดจากการที่แบบจำลองสามารถระบุให้คำที่ตามหลังวงเล็บปิดเป็นคำขอบเขตเริ่มต้น EDU ได้ โดยสรุปแล้วลักษณ์นี้มีบทบาทในการช่วยแบบจำลองให้สามารถตัดสินใจคำขอบเขตเริ่มต้น EDU ที่อยู่ตำแหน่งหลังวงเล็บปิดได้ ตัวอย่างข้อความในตารางต่อไปนี้ เป็นการเปรียบเทียบผลการระบุคำขอบเขตเริ่มต้น EDU ระหว่างการใช้ POS ของคำ 3 คำ และการใช้ลักษณ์เครื่องหมายวรรคตอนร่วมด้วย

ขอบเขตเริ่มต้น EDU ที่ถูกต้อง	**POS 3 คำ ร่วมกับ Punc**	**POS ของ 3 คำ**
(24) **$_** (Salary<s>man)<s> **$_**สามีหรือพ่อบ้านมักจะใช้เวลาส่วนใหญ่อยู่ในที่ทำงาน	**$_** (Salary<s>man)<s>**$_** สามีหรือพ่อบ้านมักจะใช้เวลาส่วนใหญ่อยู่ในที่ทำงาน	**$_** (Salary<s>man)<s>สามีหรือพ่อบ้านมักจะใช้เวลาส่วนใหญ่อยู่ในที่ทำงาน
(25) ...สีเอกรงค์<s>**$_** (Monochrome)<s>**$_** องค์ประกอบของภาพมีเส้นโค้ง ...	...สีเอกรงค์<s>**$_** (Monochrome)<s>**$_** องค์ประกอบของภาพมีเส้นโค้ง ...	...สีเอกรงค์<s>**$_** (Monochrome)<s> องค์ประกอบของภาพมีเส้นโค้ง ...

ตารางที่ 6.4 ตัวอย่างข้อความจากคลังข้อมูล แสดงผลการใช้ลักษณ์เครื่องหมายวรรคตอน

6.1.5 การใช้ทุกลักษณ์ร่วมกัน

เมื่อใช้ลักษณ์ต่าง ๆ ร่วมกันทั้งหมด ได้แก่ หมวดคำปัจจุบัน หมวดคำก่อนหน้า หมวดคำตามหลัง รายการคำเชื่อม ช่องว่าง และเครื่องหมายวรรคตอน แล้วนำไปฝึกฝนแบบจำลอง พบว่าสามารถช่วยเพิ่มประสิทธิภาพของแบบจำลองได้ กล่าวคือ แบบจำลองสามารถระบุขอบเขตเริ่มต้น EDU ที่มีคำเชื่อมอนุพากย์และตัวนำส่วนเติมเต็มได้เกือบ 100 เปอร์เซ็นต์ ความผิดพลาดที่พบ อย่างแรกคือ คำขอบเขตเริ่มต้น EDU ที่ปรากฏติดกับอนุพากย์ก่อนหน้าโดยไม่มีคำเชื่อมหรือช่องว่างคั่น ดังตัวอย่างข้อความ 24 ความผิดพลาดอีกประการคือ คำขอบเขตเริ่มต้น EDU ที่ปรากฏหลังช่องว่าง ทั้งนี้เพราะลักษณ์ช่องว่างไม่มีประสิทธิภาพในการช่วยแบบจำลองในการตัดสินใจคำที่อยู่หลังช่องว่าง ดังตัวอย่างข้อความในตารางต่อไปนี้

การแบ่ง EDU ที่ถูกต้อง	ใช้ทุกลักษณ์รวมกัน	ปัญหาที่พบ
(26) **$_**หรือไม่ยึดติดกรอบคิดนั้น<s>**$_**คงจะต้องปรับวิธีวิทยาด้วยเช่นเดียวกัน	**$_**หรือไม่ยึดติดกรอบคิดนั้น<s>คงจะต้องปรับวิธีวิทยาด้วยเช่นเดียวกัน	ไม่สามารถระบุขอบเขตเริ่มต้น EDU ที่ปรากฏหลังช่องว่างได้
(27) **$_**ที่เลียนแบบการบริโภคสินค้าแบบชนชั้นกลาง**$_**จะสรุปได้อย่างไร	**$_**ที่เลียนแบบการบริโภคสินค้าแบบชนชั้นกลางจะสรุปได้อย่างไร	ไม่สามารถระบุขอบเขตเริ่มต้น EDU ที่ปรากฏติดกันกับ EDU ก่อนหน้าโดยไม่มีคำเชื่อมหรือช่องว่างคั่นได้

ตารางที่ 6.5 ตัวอย่างข้อความจากคลังข้อมูล แสดงผลการใช้ทุกลักษณ์ร่วมกัน

จากตัวอย่างข้อความ 26 และ 27 สามารถมองได้ว่า ปัญหาอาจไม่ได้เกิดจากช่องว่างเพียงอย่างเดียว แต่อาจเกิดจากคำที่ปรากฏหลังช่องว่าง นั้นก็คือ คำว่า “คง” และ “จะ” ซึ่งเป็น AUX ตรงนี้อาจเป็นส่วนหนึ่งที่ทำให้แบบจำลองตัดสินใจออกมาเป็นเช่นนั้น กล่าวคือ หากพิจารณาการปรากฏร่วมกันของช่องว่างแล้วตามด้วย AUX ในคลังข้อมูล ก็พบว่า มีข้อความอื่น ๆ ในคลังข้อมูลอีกจำนวนหนึ่งที่ AUX หลังช่องว่าง ไม่ได้เป็นขอบเขตเริ่มต้น EDU เหมือนในกรณีตัวอย่าง 26 และ 27 ตัวอย่างจากคลังข้อมูล เช่น “**$_**สำหรับขั้นตอนที่หก<s>จะเป็นแนวทางการอธิบาย” จะเห็นได้ว่าคำว่า “จะ” ซึ่งเป็น AUX ไม่ใช่คำขอบเขตเริ่มต้น EDU จึงเป็นไปได้ว่า เหตุผลที่แบบจำลองไม่ดึง AUX ที่ตามหลัง space ออกมาได้ จะเป็นเพราะการปรากฏร่วมของคู่นี้จำนวนมากจะไม่ใช่ขอบเขตเริ่มต้น EDU

โดยสรุปแล้ว หากตัวจำแนกซัพพอร์ตเวกเตอร์แมชชีนตั้งค่าพารามิเตอร์เคอร์เนล D=1 แล้วลักษณ์ที่มีประสิทธิภาพมากที่สุดคือลักษณ์หมวดคำ ซึ่งต้องพิจารณาหมวดคำข้างเคียงด้วย ส่วนลักษณ์อื่น ๆ ที่นำมาใช้ประกอบกับลักษณ์หมวดคำสามารถช่วยเพิ่มประสิทธิภาพแบบจำลองได้ในระดับที่แตกต่างกันออกไป โดยที่ลักษณ์ที่มีส่วนช่วยเพิ่มความสามารถให้กับแบบจำลองมากที่สุดคือลักษณ์เครื่องหมายวรรคตอน ส่วนลักษณ์รายการคำเชื่อมช่วยเพิ่มค่าความครบถ้วนได้เล็กน้อย และลักษณ์ช่องว่างไม่ช่วยเพิ่มประสิทธิภาพแบบจำลองเลย ทั้งนี้ ผลการทดสอบลักษณ์ที่กล่าวมาทั้งหมดนี้เป็นการทดสอบโดยตั้งค่าพารามิเตอร์ของเคอร์เนลไว้ต่ำสุด หากเพิ่มค่าพารามิเตอร์ให้สูงกว่านี้ก็จะได้ผลที่ดีกว่านี้

6.2 ประสิทธิภาพของแบบจำลองเมื่อปรับค่าพารามิเตอร์ของเคอร์เนลให้สูงขึ้น

เนื่องจากซัพพอร์ตเวกเตอร์แมชชีนมีเคอร์เนลฟังก์ชั่นที่สามารถใช้ได้หลายตัว ในงานนี้ ผู้วิจัยใช้เคอร์เนลโพลีโนเมียลฟังก์ชั่น ในขั้นแรก ผู้วิจัยทำการทดสอบแบบจำลองโดยปรับค่าพารามิเตอร์ของเคอร์เนลไว้ที่ต่ำสุด คือ D=1 ซึ่งเป็นค่า default ของโปรแกรมวีก้า จากนั้นผู้วิจัยได้ปรับค่าพารามิเตอร์นี้ให้สูงขึ้นเป็น D=2, D=3, และ D=4 พบว่าค่าพารามิเตอร์ที่สูงขึ้นสามารถเพิ่ม

ประสิทธิภาพของแบบจำลองได้ เมื่อดูคลังข้อมูลทดสอบที่ผ่านการระบุขอบเขตเริ่มต้น EDU แล้ว พบว่าการเพิ่มขึ้นของคำขอบเขตเริ่มต้น EDU ที่แบบจำลองดึงออกมา ส่วนใหญ่จะเป็นคำที่ขึ้นต้นด้วยช่องว่างและไม่มีคำเชื่อมอนุพากย์ และอีกจำนวนหนึ่งเป็นคำที่ไม่ได้ขึ้นต้นด้วยช่องว่างหรือคำเชื่อมอนุพากย์เลย

ผลของการการปรับค่าพารามิเตอร์ของเคอร์เนลให้มากกว่า 1 ทำให้แบบจำลองสามารถระบุคำขอบเขตเริ่มต้น EDU ที่นำหน้าด้วยช่องว่างแต่ไม่มีคำเชื่อมได้จำนวนหนึ่ง ผู้วิจัยพบว่าหมวดคำที่ตามหลังช่องว่างที่แบบจำลองตัดสินใจให้เป็นคำขอบเขตเริ่มต้น EDU ได้แก่ คำนาม คำกริยา คำปฏิเสธ คำบุพบท คำสรรพนาม และคำนามชี้เฉพาะ

ขอบเขตเริ่มต้น EDU ที่ถูกต้อง	**ใช้ลักษณ์ POS ร่วมกับ space ค่าพารามิเตอร์ D=1**	**ใช้ลักษณ์ POS ร่วมกับ space ค่าพารามิเตอร์ D=3**
(28) **$_**ซึ่งสะท้อนชีวิตความเป็นอยู่ของผู้หญิงเหล่านี้ออกมาได้อย่างชัดเจนที่สุด<s>**$_**ภาพโสเภณี<s>หญิงขายบริการ<s>และนักเต้นเหล่านี้ถูกบันทึกไว้นับร้อยภาพ	**$_**ซึ่งสะท้อนชีวิตความเป็นอยู่ของผู้หญิงเหล่านี้ออกมาได้อย่างชัดเจนที่สุด<s>ภาพโสเภณี<s>หญิงขายบริการ<s>และนักเต้นเหล่านี้ถูกบันทึกไว้นับร้อยภาพ	**$_**ซึ่งสะท้อนชีวิตความเป็นอยู่ของผู้หญิงเหล่านี้ออกมาได้อย่างชัดเจนที่สุด<s>**$_**ภาพโสเภณี<s>หญิงขายบริการ<s>และนักเต้นเหล่านี้ถูกบันทึกไว้นับร้อยภาพ
(29) **$_**ที่น่าสนใจไม่ยิ่งหย่อนไปกว่ากัน<s>**$_**บั้นปลายของชีวิต<s>คัสแสทท์เกือบสูญเสียการมองเห็น	**$_**ที่น่าสนใจไม่ยิ่งหย่อนไปกว่ากัน<s>บั้นปลายของชีวิต<s>คัสแสทท์เกือบสูญเสียการมองเห็น	**$_**ที่น่าสนใจไม่ยิ่งหย่อนไปกว่ากัน<s>**$_**บั้นปลายของชีวิต<s>คัสแสทท์เกือบสูญเสียการมองเห็น

ตารางที่ 6.6 ตัวอย่างข้อความจากคลังข้อมูล แสดงผลการใช้ลักษณ์ช่องว่างเมื่อปรับค่าพารามิเตอร์ต่างกัน

ตัวอย่างจากคลังข้อมูล (28) และ (29) เป็นการเปรียบเทียบผลการระบุคำขอบเขตเริ่มต้น EDU ของแบบจำลองที่ฝึกฝนด้วย SVM ที่มีการตั้งค่าพารามิเตอร์เคอร์เนล D=1 และค่าพารามิเตอร์ที่สูงขึ้น ในที่นี้คือเปรียบเทียบกับพารามิเตอร์ D=3 ซึ่งจะเห็นว่า คำ “ภาพ” ในตัวอย่าง (28) และคำ “บั้นปลาย” ในตัวอย่าง (29) ถูกระบุให้เป็นคำขอบเขตเริ่มต้น EDU ได้ถูกต้องเมื่อปรับค่าพารามิเตอร์ D=3

อย่างไรก็ตาม พบว่ามีความผิดพลาดของการตัดสินใจระบุคำขอบเขตเริ่มต้น EDU ที่มีช่องว่างนำหน้าอยู่พอสมควร ยกตัวอย่าง (30) พบว่าตัวจำแนกที่ตั้งค่าพารามิเตอร์ D=3 สามารถระบุคำ “สตรี” ให้เป็นคำขอบเขตเริ่มต้น EDU ได้ แต่กลับไม่สามารถระบุคำ “ประเภท” และ “มอง” ให้

เป็นคำขอบเขตเริ่มต้น EDU ได้ แม้ว่าจะเป็นคำที่ปรากฏหลังช่องว่างก็ตาม และในตัวอย่าง (31) การตั้งค่าพารามิเตอร์ D=3 ก็ไม่สามารถระบุคำ “ใช้” ให้เป็นคำขอบเขตเริ่มต้น EDU ได้เช่นกัน

ขอบเขตเริ่มต้น EDU ที่ถูกต้อง	**ใช้ลักษณ์ POS ร่วมกับ space ค่าพารามิเตอร์ D=3**
(30) the basic important of the line<s>**$_**สตรีในงานของโลเทรค<s>แบ่งอย่างหยาบๆ<s>ได้เป็นสองประเภท<s>**$_** ประเภทแรกมีใบหน้าแบบสุภาพชน<s>**$_**มองโลกด้านดี	the basic important of the line<s>**$_**สตรีในงานของโลเทรค<s>**$_**แบ่งอย่างหยาบๆ<s>ได้เป็นสองประเภท<s>ประเภทแรกมีใบหน้าแบบสุภาพชน<s>มองโลกด้านดี
(31) **$_**ที่หยาบกร้าน<s>**$_**ใช้ชีวิตในโลกของผู้ชายทั้งสถานเริงรมย์และสำนักโสเภณี	**$_ที่**หยาบกร้าน<s>ใช้ชีวิตในโลกของผู้ชายทั้งสถานเริงรมย์และสำนักโสเภณี

ตารางที่ 6.7 ตัวอย่างข้อความจากคลังข้อมูล แสดงความผิดพลาดที่เกิดขึ้นกับการใช้ลักษณ์ช่องว่างเมื่อปรับค่าพารามิเตอร์ D=3

อย่างไรก็ตาม ความสามารถที่เพิ่มขึ้นของการดึงคำขอบเขตเริ่มต้น EDU ที่ไม่ใช่คำเชื่อม และเป็นคำที่ตามหลังช่องว่าง ชวนให้สงสัยว่า เป็นผลอันเนื่องมาจากการใช้ลักษณ์ใด เพราะหากดูจากผลการทดสอบที่เสนอไปในบทที่ 5 ที่แสดงให้เห็นว่า การใช้ลักษณ์รูปแบบที่ 2 ซึ่งเป็นการใช้ POS ของคำ 3 คำ และลักษณ์รูปแบบที่ 4 ซึ่งเป็นการใช้ลักษณ์ช่องว่างร่วมด้วย ได้ผลการทดสอบออกมาเหมือนกัน ทั้งค่าความแม่นยำ ค่าความครบถ้วน และค่า F-measure ซึ่งแสดงให้เห็นว่าลักษณ์ช่องว่างไม่ได้ช่วยเพิ่มประสิทธิภาพการตัดสินใจของแบบจำลองเลย ดังนั้นเมื่อมีการปรับพารามิเตอร์ให้สูงขึ้น ลักษณ์ช่องว่างก็ไม่น่าจะมีส่วนช่วยในการตัดสินใจเช่นกัน

ในการไขข้อสงสัยนี้ ผู้วิจัยจึงได้ทำการทดสอบแบบจำลอง โดยใช้ชุดของลักษณ์ 2 ชุด เปรียบเทียบผลกัน ชุดแรกเป็นการใช้ลักษณ์ร่วมกันทั้งหมด ซึ่งผลการทดสอบก็เป็นตามที่เสนอไว้แล้วในบทที่ 5 ส่วนลักษณ์อีกชุดหนึ่งประกอบไปด้วยลักษณ์ทุกลักษณ์เหมือนชุดแรก ยกเว้นลักษณ์ช่องว่าง ทำการทดสอบ 10 ครั้งเช่นเคย และมีการปรับค่าพารามิเตอร์ D=1, D=2, D=3, และ D=4 ทั้งนี้จะเปรียบเทียบผลเพื่อดูว่า ความสามารถที่เพิ่มขึ้นของแบบจำลองที่มีการปรับค่าพารามิเตอร์ให้สูงขึ้น ที่ชี้ให้เห็นว่าสามารถระบุคำที่ตามหลังช่องว่างให้เป็นคำขอบเขตเริ่มต้น EDU ได้นั้น เกิดจากการใช้ลักษณ์ช่องว่างหรือลักษณ์อื่น ผลการทดสอบเป็นดังตารางต่อไปนี้

	D=1		D=2		D=3		D=4	
	ชุด 1	**ชุด 2**	**ชุด 1**	**ชุด 2**	**ชุด 1**	**ชุด 2**	**ชุด 1**	**ชุด 2**
P	98.03	98.03	91.97	91.56	91.67	91.62	91.97	91.82
R	69.68	69.68	77.55	76.88	78.84	78.54	78.69	78.54
F	81.17	81.17	83.97	83.4	84.7	84.49	84.74	84.57

ตารางที่ 6.8 ตัวอย่างข้อความจากคลังข้อมูล เปรียบเทียบผลการทดสอบลักษณ์ชุดที่ 1 และ 2

จากตารางจะเห็นว่า หากปรับค่าพารามิเตอร์ D=1 ผลการทดสอบลักษณ์ชุด 1 และชุด 2 ให้ผลออกมาเหมือนกัน ชี้ให้เห็นว่าลักษณ์ช่องว่างไม่ได้ช่วยเพิ่มประสิทธิภาพการตัดสินใจของแบบจำลอง เมื่อปรับค่าพารามิเตอร์ D=2, D=3, และ D=4 จะเห็นว่าลักษณ์ชุดที่ 2 ได้ค่าความแม่นยำที่ต่ำลง แต่ค่าความครบถ้วนและค่า F-measure เพิ่มขึ้นเล็กน้อย แสดงให้เห็นว่า การปรับค่าพารามิเตอร์ ทำให้เครื่องพยายามดึงคำออกมามากขึ้น ผลก็คือ เกิดข้อผิดพลาดมากขึ้น ค่าความแม่นยำจึงลดลง ในขณะเดียวกัน คำที่เป็นขอบเขตเริ่มต้น EDU ก็สามารถดึงออกมาจากคลังข้อมูลได้เพิ่มมากขึ้น ค่าความครบถ้วนจึงสูงขึ้น อย่างไรก็ดี ความสามารถที่เพิ่มขึ้นอันเนื่องมาจากการใช้ลักษณ์ช่องว่างนี้ ถือว่าน้อยมาก เพราะค่า F-measure เพิ่มขึ้นไม่ถึง 1 เปอร์เซ็นต์

ดังนั้นตอบของข้อสงสัยข้างต้นจึงสามารถตอบได้ว่า ลักษณ์ช่องว่างมีส่วนช่วยเพิ่มความสามารถในการระบุคำขอบเขตเริ่มต้น EDU จริง แต่ก็น้อยมาก จึงอาจกล่าวได้ว่า ลักษณ์ช่องว่างมีความโดดเด่นชัดเจนน้อยกว่าลักษณ์อื่น ๆ ซึ่งอาจเป็นเพราะข้อมูลของลักษณ์นี้เป็นข้อมูลที่ซ้ำซ้อนกับข้อมูลของลักษณ์อื่น ๆ พอปรับค่าพารามิเตอร์ไว้ต่ำ จึงไม่ส่งผลต่อการตัดสินใจแบบจำลองเลย และเมื่อปรับค่าพารามิเตอร์ให้สูงขึ้น ก็ส่งผลต่อการตัดสินใจของแบบจำลองได้เพียงเล็กน้อยเท่านั้น และหากลักษณ์ช่องว่างมีบทบาทน้อยมากต่อการเพิ่มประสิทธิภาพแบบจำลองแล้ว แสดงว่าการเพิ่มขึ้นของประสิทธิภาพของแบบจำลองเมื่อมีการปรับค่าพารามิเตอร์สูงขึ้นนั้น เป็นผลมาจากลักษณ์ต่าง ๆ มากกว่าลักษณ์ช่องว่าง

อีกประเด็นหนึ่งที่ยังเป็นปัญหาอยู่ แม้จะปรับค่าพารามิเตอร์ให้สูงขึ้นแล้วก็ตาม ก็คือปัญหาของคำเริ่มต้นขอบเขต EDU ที่ไม่มีช่องว่างและคำเชื่อมใด ๆ ปรากฏข้างหน้า กล่าวคือแบบจำลองไม่สามารถจัดการกับคำเหล่านี้และระบุให้เป็นคำขอบเขตเริ่มต้น EDU ได้เลย ดังนั้นจึงน่าจะเป็นเพราะไม่มีลักษณ์ใดที่ใช้ในการฝึกฝนแบบจำลองที่จะช่วยในการตัดสินใจคำขอบเขตเริ่มต้น EDU ในบริบทเช่นนี้

บทที่ 7

สรุปผลการศึกษา ปัญหา และข้อเสนอแนะ

ในบทที่ 3 ถึงบทที่ 6 ได้กล่าวไปแล้วถึงกระบวนการต่าง ๆ ในการแบ่งอนุพากย์ภาษาไทย ซึ่งถือเป็นหน่วยที่เล็กที่สุดสำหรับการศึกษาโครงสร้างปริจเฉท ด้วยตัวจำแนกประเภทซัพพอร์ตเวกเตอร์แมชชีน ตั้งแต่ขั้นตอนการสร้างคลังข้อมูล การกำกับคลังข้อมูล ไปจนถึงการรายงานผลการทดสอบแบบจำลองและการอภิปรายลักษณ์ทางภาษาที่ใช้ในการดำเนินการทดสอบด้วย ในบทนี้ ผู้วิจัยจะได้สรุปผลการศึกษาโดยภาพรวมอีกครั้ง และอภิปรายปัญหาที่พบในการศึกษาครั้งนี้ พร้อมทั้งเสนอข้อเสนอแนะสำหรับการศึกษาการแบ่งอนุพากย์ภาษาไทยด้วยเครื่องเพื่อเป็นประโยชน์ต่องานทางด้านการประมวลผลภาษาธรรมชาติ

7.1 สรุปผลการศึกษา

การแบ่งอนุพากย์ภาษาไทยด้วยเครื่องสามารถมองว่าเป็นการจำแนกประเภทอย่างหนึ่งได้ กล่าวคือ เมื่อให้เครื่องรับข้อมูลเข้าทีละคำ เครื่องมีหน้าที่ตัดสินใจหรือแยกประเภทว่า คำที่พิจารณาอยู่นั้นเป็นขอบเขตอนุพากย์ภาษาไทยหรือไม่ ซึ่งในที่นี้ ผู้วิจัยให้เครื่องระบุขอบเขตเริ่มต้นของอนุพากย์เท่านั้น เนื่องจากสามารถมองได้ว่า คำสุดท้ายก่อนที่จะเป็นคำเริ่มต้นขอบเขตอนุพากย์ใหม่ ก็คือคำขอบเขตท้ายอนุพากย์นั่นเอง ในการศึกษานี้ ผู้วิจัยใช้ฟังก์ชั่น SMO ของโปรแกรมวีก้าในการสร้างตัวจำแนกประเภทซัพพอร์ตเวกเตอร์แมชชีนเพื่อระบุขอบเขตเริ่มต้นอนุพากย์ ตัวฟังก์ชั่น SMO นี้ทำให้ง่ายต่อการเตรียมไฟล์ข้อมูลที่จะใช้ในการฝึกฝนและทดสอบ เพราะ SMO สามารถอ่านลักษณ์ที่ใช้ฝึกฝน หรือ feature ได้หลายรูปแบบ ได้แก่ ค่าตัวเลข, ข้อความ, และข้อมูลแบบแบ่งกลุ่ม

ลักษณ์ที่ใช้ในการฝึกฝนแบบจำลองเป็นลักษณ์ทางภาษา ได้แก่ หมวดคำปัจจุบัน หมวดคำก่อนหน้า หมวดคำตามหลัง รายการคำเชื่อม ช่องว่างที่เป็นตัวแบ่งอนุพากย์ และเครื่องหมายวรรคตอนที่มักปรากฏหน้าหรือท้ายอนุพากย์ ทั้งนี้ผู้วิจัยได้ทำการทดสอบเพื่อเปรียบเทียบประสิทธิภาพการทำงานของลักษณ์แต่ละตัว โดยการทำการทดสอบลักษณ์รูปแบบต่าง ๆ ทั้งหมด 6 รูปแบบ คือ

รูปแบบที่ 1 ใช้ลักษณ์หมวดคำของคำปัจจุบันเพียงลักษณ์เดียว
รูปแบบที่ 2 ใช้ลักษณ์หมวดคำปัจจุบัน หมวดคำก่อนหน้า และหมวดคำที่ตามหลัง
รูปแบบที่ 3 ใช้รายการคำเชื่อมร่วมกับลักษณ์หมวดคำปัจจุบัน หมวดคำก่อนหน้า และหมวดคำที่ตามหลัง
รูปแบบที่ 4 ใช้ความน่าจะเป็นของช่องว่างที่เป็นตัวคั่น EDU ร่วมกับลักษณ์หมวดคำปัจจุบัน หมวดคำก่อนหน้า และหมวดคำที่ตามหลัง
รูปแบบที่ 5 ใช้เครื่องหมายวรรคตอนที่มักเป็นของเขต EDU ร่วมกับลักษณ์หมวดคำปัจจุบัน หมวด

คำก่อนหน้า และหมวดคำที่ตามหลัง

รูปแบบที่ 6 ใช้ลักษณ์หมวดคำปัจจุบัน หมวดคำก่อนหน้า หมวดคำที่ตามหลัง รายการคำเชื่อม ลักษณ์ช่องว่าง และลักษณ์เครื่องหมายวรรคตอน

จากการทดสอบ พบว่าการใช้ลักษณ์รูปแบบที่ 1 หรือการใช้หมวดคำปัจจุบันเพียงลักษณ์เดียวไม่สามารถทำให้แบบจำลองระบุคำขอบเขตเริ่มต้น EDU ได้มากนัก คำขอบเขตเริ่ม EDU ส่วนใหญ่ที่แบบจำลองดึงออกมาก็มักจะเป็นพวกคำบ่งชี้ ได้แก่ คำเชื่อมอนุพากย์ และตัวนำส่วนเติมเต็ม ส่วนคำขอบเขตเริ่มต้น EDU ที่ไม่ได้ขึ้นต้นด้วยคำบ่งชี้ก็ไม่ถูกดึงออกมา ทั้งนี้เนื่องจากการใช้เพียงลักษณ์เดียวเช่นนี้เป็นการให้ข้อมูลแก่ตัวจำแนกประเภทน้อยเกินไป ทำให้แบบจำลองไม่สามารถตัดสินใจนอกเหนือไปจากข้อมูลที่มีในลักษณ์

การใช้ลักษณ์รูปแบบที่ 2 หรือการใช้หมวดคำปัจจุบัน หมวดคำก่อนหน้า และหมวดคำที่ตามหลังในการฝึกฝนแบบจำลอง พบว่าให้ผลที่ดียิ่งขึ้น สามารถดึงคำขอบเขตเริ่มต้น EDU ได้ถูกต้องคิดเป็น 60 กว่าเปอร์เซ็นต์ของคำขอบเขตเริ่มต้น EDU ทั้งหมดในคลังทดสอบ และวัดค่า F-measure เฉลี่ยได้เกิน 75 เปอร์เซ็นต์

จากผลการทดสอบลักษณ์รูปแบบที่ 2 ผู้วิจัยเห็นว่าคำขอบเขตเริ่มต้น EDU ที่ไม่สามารถดึงออกมาได้ ส่วนใหญ่จะเป็นคำที่ขึ้นต้นด้วยคำเชื่อมที่เป็นวลี คำที่ไม่มีคำบ่งชี้นำหน้า และคำที่มีช่องว่างปรากฏด้านหน้า ดังนั้นจึงใช้หมวดคำปัจจุบัน หมวดคำก่อนหน้าและตามหลัง ร่วมกับลักษณ์อื่น ๆ ที่คาดว่าจะมีส่วนช่วยในการระบุขอบเขตเริ่มต้น EDU ได้แก่ ลักษณ์รายการคำเชื่อม ลักษณ์ช่องว่าง และลักษณ์เครื่องหมายวรรคตอน และเพื่อให้ทราบว่าลักษณ์ต่าง ๆ เหล่านี้สามารถเพิ่มความสามารถให้กับแบบจำลองได้ในระดับที่มากน้อยต่างกันอย่างไรบ้าง จึงได้ทำการทดสอบรูปแบบลักษณ์ที่ 3, 4 และ 5

จากการทดสอบแบบจำลองที่ฝึกฝนด้วยลักษณ์รูปแบบที่ 3, 4 และ 5 พบว่าการใช้ลักษณ์รูปแบบที่ 4 ซึ่งเป็นการใช้หมวดคำปัจจุบัน หมวดคำก่อนหน้าและตามหลัง ร่วมกับลักษณ์ช่องว่าง ไม่สามารถทำให้แบบจำลองระบุคำขอบเขตเริ่มต้น EDU ได้เพิ่มมากขึ้นจากเดิมเลย ทั้งนี้เพราะข้อมูลของลักษณ์ช่องว่างซ้ำซ้อนกับข้อมูลลักษณ์หมวดคำไปแล้ว เมื่อไม่มีข้อมูลใหม่ แบบจำลองจึงไม่สามารถดึงคำที่ถูกต้องออกมาได้มากขึ้น ส่วนการใช้ลักษณ์รูปแบบที่ 3 และ 5 ทำให้แบบจำลองสามารถระบุคำขอบเขตเริ่มต้น EDU ได้จำนวนเพิ่มมากขึ้น

การใช้ลักษณ์รูปแบบที่ 6 เป็นการใช้ทุกลักษณ์ร่วมกัน เมื่อปรับค่าพารามิเตอร์ตัวจำแนกประเภทอย่างต่ำที่สุด พบว่าสามารถดึงคำขอบเขตเริ่มต้น EDU จากคลังข้อมูลได้ถูกต้องเกือบ 70 เปอร์เซ็นต์ ส่วนคำขอบเขตเริ่มต้น EDU ที่เหลืออีกประมาณ 30 เปอร์เซ็นต์ที่แบบจำลองไม่สามารถ

ดึงออกมาได้ ได้แก่ คำที่ปรากฏตามหลังช่องว่าง และคำขอบเขตเริ่มต้นอนุพากย์ที่ไม่มีทั้งช่องว่างและคำบ่งชี้นำหน้า

อย่างไรก็ตาม ความผิดพลาดเรื่องคำขอบเขตเริ่มต้น EDU ที่ปรากฏหลังช่องว่างได้รับการแก้ไขเมื่อมีการปรับพารามิเตอร์ของเคอร์เนลให้สูงขึ้น กล่าวคือ เมื่อปรับค่าพารามิเตอร์เคอร์เนล D=3 พบว่าแบบจำลองสามารถดึงคำขอบเขตเริ่มต้น EDU จากคลังทดสอบได้ถูกต้องคิดเป็น 78.5 เปอร์เซ็นต์ ในจำนวนคำขอบเขตเริ่มต้น EDU ที่เพิ่มขึ้นมานี้ ส่วนใหญ่เป็นคำที่ปรากฏหลงช่องว่างทั้งสิ้น

แม้การปรับค่าพารามิเตอร์ให้สูงขึ้นจะมีส่วนช่วยเพิ่มประสิทธิภาพของแบบจำลองในการระบุคำขอบเขตเริ่มต้น EDU ได้ ผู้วิจัยพบว่า ก็ยังมีความผิดพลาดเกี่ยวกับความกำกวมของช่องว่างอยู่ไม่น้อย นั่นคือแบบจำลองไม่สามารถตัดสินใจได้อย่างถูกต้องในกรณีคำขอบเขตเริ่มต้น EDU ที่ปรากฏหลังช่องว่าง อีกความผิดพลาดหนึ่งที่พบและไม่สามารถแก้ได้ด้วยการปรับค่าพารามิเตอร์ก็คือปัญหา EDU ที่ปรากฏติดกันโดยไม่มีทั้งช่องว่างและคำเชื่อมคั่น

7.2 ปัญหาที่พบในการศึกษา

ผู้วิจัยพบว่ามีลักษณะบางประการของภาษาไทยที่ทำให้เป็นปัญหาต่อการระบุขอบเขตเริ่มต้น EDU ปัญหาที่ใหญ่ที่สุดของการศึกษาครั้งนี้ คือปัญหาการใช้ช่องว่าง ผู้วิจัยพบว่า เกือบครึ่งหนึ่งของช่องว่างที่ปรากฏในข้อเขียน เป็นช่องว่างที่คั่นระหว่างอนุพากย์ ที่เหลือจะเป็นช่องว่างที่ทำหน้าที่อื่น ๆ เช่น คั่นระหว่างคำนามที่อยู่ในชุดเดียวกัน คั่นระหว่างตัวเลขกับตัวอักษร คั่นระหว่างตัวอักษรภาษาไทยกับตัวอักษรภาษาต่างประเทศ คั่นระหว่างเครื่องหมายวรรคตอนและตัวอักษร เป็นต้น ในการศึกษาครั้งนี้ สามารถจัดการกับความกำกวมของช่องว่างที่ปรากฏทั้งหมดได้ประมาณ 60-65 เปอร์เซ็นต์ ในจำนวนนี้ได้แก่ช่องว่างที่ตามด้วยคำบ่งชี้และช่องว่างที่ตามด้วยหมวดคำบางประเภท ส่วนเกือบ 40 เปอร์เซ็นต์ที่เหลือเป็นช่องว่างที่เป็นตัวคั่นอนุพากย์แต่แบบจำลองไม่สามารถตัดสินใจได้

อีกประเด็นที่เป็นปัญหาเช่นกัน ก็คือปัญหาการละคำบ่งชี้อนุพากย์ ในกรณีที่อนุพากย์ไม่ได้ขึ้นต้นด้วยคำบ่งชี้ แบบจำลองอาจจะไปพิจารณาลักษณ์ช่องว่างร่วมด้วย เพื่อดูว่าคำนั้นปรากฏหลังช่องว่างที่มีความเป็นไปได้ที่จะเป็นตัวคั่นอนุพากย์หรือไม่ แต่หากไม่มีช่องว่างปรากฏหน้าขอบเขตเริ่มต้นอนุพากย์เลย บวกกับไม่มีคำบ่งชี้อนุพากย์ด้วย เช่นนี้แล้วแบบจำลองจะไม่สามารถตัดสินใจได้เลยว่าจุดไหนเป็นขอบเขตเริ่มต้นอนุพากย์ ในการศึกษาครั้งนี้พบว่าปัญหาเรื่องความกำกวมของช่องว่างและการละคำบ่งชี้อนุพากย์ ทำให้มีคำขอบเขตเริ่มต้น EDU ที่แบบจำลองไม่สามารถดึงออกมาได้ประมาณ 20 เปอร์เซ็นต์ของคำขอบเขตเริ่มต้นทั้งหมดในคลังทดสอบ

นอกจากนี้ยังมีปัญหาเกี่ยวกับลักษณ์ที่ใช้ในการฝึกฝนแบบจำลอง การกำหนดลักษณ์ที่เหมาะสมมีผลอย่างมากต่อการตัดสินใจของแบบจำลอง จากผลการทดสอบการใช้ลักษณ์ทางภาษาที่ได้เสนอไป ก็พบว่า การตัดสินใจของแบบจำลองเป็นผลอันเนื่องมาจากการใช้ลักษณ์หมวดคำเป็นหลัก ส่วนลักษณ์อื่น ๆ ที่ใช้ ก็ให้ข้อมูลที่ความซ้ำซ้อนกับลักษณ์หมวดคำ ทำให้ช่วยเพิ่มความสามารถในการตัดสินใจของแบบจำลองได้ไม่มากนัก

7.3 ข้อเสนอแนะ

จากผลการศึกษา จะพบว่าปัญหาหลักที่เกิดจากลักษณะทางภาษาจะเป็นเรื่องความกำกวมของช่องว่าง แต่ลักษณ์ที่ใช้ในการจัดการกับปัญหาของงานวิจัยนี้เป็นเพียงการให้ข้อมูลว่าช่องว่างที่ปรากฏติดกับหมวดคำใดมีโอกาสสูงที่จะเป็นตัวคั่นอนุพากย์ เช่น ช่องว่างแล้วตามด้วยคำกริยา หรือตามด้วยคำสรรพนาม เป็นต้น และมีการกำหนดค่าของลักษณ์เป็นแบบสองค่า คือ Y และ N โดยพิจารณาความเป็นไปได้ที่คำตามหลังช่องว่างจะเป็นคำขอบเขตเริ่มต้น EDU แต่ในความเป็นจริง แม้ช่องว่างจะตามด้วยคำบางประเภทที่มีโอกาสเป็นคำขอบเขตเริ่มต้น EDU แต่ไม่ได้หมายความว่า คำ ๆ นั้นที่ปรากฏร่วมกับช่องว่างจะเป็นคำขอบเขตเริ่มต้น EDU เสมอไป เช่น ช่องว่างแล้วตามด้วยคำกริยา ก็ไม่จำเป็นว่าคำกริยานั้นจะเป็นคำขอบเขตเริ่มต้น EDU ดังนั้นหากเพิ่มจำนวนหมวดคำที่ปรากฏร่วมกับช่องว่างให้มากขึ้นก็จะทำให้เห็นรูปแบบของการปรากฏร่วมที่แคบลง ซึ่งอาจช่วยจัดการกับความกำกวมช่องว่างที่เป็นตัวคั่นอนุพากย์ได้

จากการสำรวจคลังข้อมูล พบว่าช่องว่างเกิน 50 เปอร์เซ็นต์ทำหน้าที่แยกอนุพากย์ออกจากกัน การปรากฏของช่องว่างจึงน่าจะยังสามารถใช้กำหนดเป็นลักษณ์ในการฝึกฝนแบบจำลองได้ นอกจากการดูว่าคำประเภทไหนที่ปรากฏร่วมกับช่องว่างที่แยกอนุพากย์แล้ว อาจจะใช้วิธีนับคำจากคำหนึ่งไปยังช่องว่างที่ใกล้สุด เพื่อใช้ระยะห่างระหว่าง 2 จุดนี้เป็นข้อมูลให้เครื่องตัดสินใจว่า ช่องว่างนั้นเป็นช่องว่างแยกอนุพากย์หรือไม่ และหากช่องว่างนั้นเป็นตัวแยกอนุพากย์ คำที่ปรากฏติดกับช่องว่างนั้นก็จะเป็นขอบเขตอนุพากย์

เมื่อพิจารณาการใช้ลักษณ์เครื่องหมายวรรคตอนในงานนี้ จะพบว่ามีลักษณะเป็นการกำหนดกฎตายตัวว่าจะให้เครื่องมองเฉพาะเครื่องหมายวงเล็บ ดังนั้นหากเปลี่ยนรูปแบบการใช้ลักษณ์นี้เป็นการใช้เครื่องหมายวรรคตอนตามรูป คือกำหนดค่าของลักษณ์ตามรูปเลย ดังนั้นแทนที่จะให้เครื่องมองว่าเครื่องหมายวรรคตอนทุกประเภทเป็น PUNC เหมือนกันหมด ก็ให้เครื่องมองว่าเครื่องหมายวรรคตอนแต่ละตัวแตกต่างกันออกไปตามรูปที่ปรากฏแทน ซึ่งเครื่องหมายวรรคตอนแต่ละตัวก็จะมีบทบาทช่วยระบุขอบเขตอนุพากย์แตกต่างกันออกไป ดังนั้นเครื่องน่าจะสามารถใช้ข้อมูลนี้ในการตัดสินใจหาขอบเขตอนุพากย์ได้

อีกประเด็นคือเรื่องของชุดของหมวดคำที่ใช้กำกับคลังข้อมูล เนื่องจากงานนี้ใช้ข้อมูลจากหมวดคำมีบทบาทสำคัญมากในการฝึกฝนแบบจำลอง การกำหนดชุดหมวดคำให้เหมาะสมจึงเป็นสิ่งที่ต้องพิจารณาให้ดียิ่งขึ้น เช่น ในงานนี้จะพบว่า ไม่ว่าจะกำกับหมวดคำ CCOR (Coordinating conjunction) และ CSUB (Subordinating conjunction) แทบไม่มีบทบาทที่ต่างกันเลย เพราะทั้งคู่ต่างก็เป็นคำเชื่อมอนุพากย์ นอกจากนี้คำเชื่อมเด่นที่มีหมวดคำเป็น ADVERB หรือกริยาวิเศษณ์ก็ถือว่าเป็นคำเชื่อมระดับอนุพากย์เช่นกัน ดังนั้นอาจจะรวมทั้ง 3 หมวดคำนี้เข้าด้วยกันเป็นหมวดคำคำเชื่อมก็ได้ เพื่อที่จะได้ไม่ต้องใช้ลักษณ์รายการคำเชื่อมอีก

นอกจากนี้ หากไม่นับอนุภาคท้าย (Final particle) ที่ผู้วิจัยได้กำกับหมวดคำเป็น PT แล้ว อาจสำรวจคลังข้อมูลเพื่อดูคำที่มักจะลงท้ายอนุพากย์และไม่ได้กำกับหมวดคำ PT แล้วดูว่าคำเหล่านั้นมีความน่าจะเป็นที่จะปรากฏท้ายอนุพากย์มากกว่าปรากฏในตำแหน่งอื่น ๆ ในอนุพากย์หรือไม่ หากมีสัดส่วนการปรากฏที่ท้ายอนุพากย์มากกว่า ก็อาจจะจัดทำรายการคำเหล่านี้แล้วนำไปเทียบกับคำในคลังข้อมูล เพื่อให้ตัวจำแนกประเภทพิจารณาคำเหล่านี้ในฐานะเป็นขอบเขตท้ายอนุพากย์ ซึ่งจะส่งผลให้คำที่อยู่ถัดไปข้างหลังมีโอกาสเป็นคำขอบเขตเริ่มต้น EDU ได้ ทั้งนี้ จากการสำรวจคลังข้อมูลที่ใช้ในการศึกษานี้ ผู้วิจัยพบคำที่ไม่ได้กำกับหมวดคำ PT และลงท้ายอนุพากย์จำนวนไม่มากนัก และส่วนใหญ่มีการปรากฏแบบกระจาย คือสามารถปรากฏได้หลายตำแหน่ง ดังนั้นจึงไม่ได้ใช้ประโยชน์จากคำที่มักจะลงท้ายอนุพากย์ในงานนี้ แต่เป็นไปได้ว่า หากขยายคลังข้อมูลให้ใหญ่ยิ่งขึ้น อาจจะเห็นความน่าจะเป็นที่คำบางคำมักจะลงท้ายอนุพากย์

รายการอ้างอิง

Al-Saleem, S. M. (2010). "Associative Classification to Categorize Arabic Data Sets " The International Journal of CM Jordan (ISSN 2078-7952) **1**(3).

Antony, P. J., et al. (2010). SVM Based Part of Speech Tagger for Malayalam. International Conference on Recent Trends in Information, Telecommunication and Computing (ITC).

Aroonmanakun, W. (2007). Thoughts on Word and Sentence Segmentation in Thai. The Seventh International Symposium on Natural Language Processing.

Basu, A., et al. (2003). Support Vector Machines for Text Categorization. The 36th Annual Hawaii International Conference.

Borji, A. and M. Hamidi (2007). Support Vector Machine for Persian Font Recognition. World Academyof Science, Engineering and Technology 28.

Buscaldi, D., et al. (2006). Verb Sense Disambiguation Using Support Vector Machines: Impact of WordNet-Extracted Features. International Conference on Intelligent Text Processing and Computational Linguistics (CICLing 2006).

Butler, C. S. (2003). Structure and Function: A Guide to Three Major Structural Functional Theories. Part 1: Approaches to the Simplex Clause. Part 2: From Clause to Discourse and Beyond. Amsterdam and Philadelphia: John Benjamins.

Carlson, L. and D. Marcu (2001). Discourse Tagging Reference Manual. ISI Technical Report, ISI-TR-545.

Charoensuk, J. (2005). Thai Elementary Discourse Unit Segmentation by Discourse Segmentation Cues and Syntactic Information, Kasetsart University, Bangkok.

Dik, S. C. (1997). The Theory of Functional Grammar,Part 1: The Structure of the Clause (2 nd ed.). Berlin: Mouton de Gruyter.

DTREG. "SVM-Support Vector Machines: Introduction to Support Vector Machine (SVM) Models." 2012, from http://www.dtreg.com/svm.htm.

Ekbal, A. and S. Bandyopadhyay (2008). Named Entity Recognition using Support Vector Machine: A Language Independent Approach. International Journal of Computer, Systems Sciences and Engg (IJCSSE).

Foley, W. A. and O. Mike (1985). Clausehood and verb serialization. Grammar Inside and Outside the Clause: Some Approaches to Theory from the Field. J. a. A. C. W. Nichols. Cambridge, Cambridge University Press**:** 17-60.

Foley, W. A. and R. D. VanValin, Jr. (1984). Functional syntax and universal grammar, Cambridge: Cambridge University Press.

Ganapathiraju, A. (2002). Support Vector Machines for Speech Recognition. (Ph. D dissertation), Mississippi State University, Mississippi.

Giménez, J. and L. Márquez (2003). Fast and Accurate Part-of-Speech Tagging: The SVM Approach Revisited. The International Conference RANLP – 2003.

Halliday, M. A. K. (1994). An introduction to functional grammar (2nd ed.). London: Edward Arnold.

Hernault, H., et al. (2010). "A Discourse Parser Using Support Vector Machine Classification." Dialogue and Discourse, 1(3): 1–33.

Ivanciuc, O. (2005). "SVM-Support Vector Machines OPTimum Separation Hyperplane." from http://www.support-vector-machines.org/SVM_osh.html.

Iwasaki, S. and I. P. Horie (2005). A Reference Grammar of Thai. Cambridge, UK, Cambridge University Press.

Jenks, P. (2006). Control in Thai. Variation in control structures. S. D. University of California. University of California, San Diego.

Joachims, T. (1998). Text Categorization with Support Vector Machines: Learning with Many Relevant Features. The European Conference on Machine Learning (ECML).

Joshi, A., et al. (2006). Discourse Annotation: Discourse Connectives and Discourse Relations. In: Tutorial at COLING/ACL.

Ketui, N., et al. (2012). A Rule-Based Method for Thai Elementary Discourse Unit Segmentation (TED-Seg). The 2012 Seventh International Conference on Knowledge, Information and Creativity Support Systems (KICSS 2012).

Lee, Y. K., et al. (2004). Supervised Word Sense Disambiguation with Support Vector Machines and multiple knowledge sources. Senseval-3.

Lemnitzer, L. (2006). "Sentence Segmentation." from http://www.cl.uni-heidelberg.de/courses/archiv/ss06/textech/slidesSentSeg.pdf.

Liberati, C., et al. (2009). "Data AdaPTive Simultaneous Parameter and Kernel Selection in Kernel Discriminant Analysis (KDA) Using Information Complexity." Journal of Pattern Recognition Research 1: 119-132.

Mann, W. C. and M. Taboada (2005). "Intro to RST.". from http://www.sfu.ca/rst/01intro/intro.html.

Mann, W. C. and S. Thompson (1988). "Rhetorical Structure Theory: Toward a Functional Theory of Text Organization." Text8(3): 243-281.

Molina, A. and F. Pla (2001). Clause DETection using HMM. The 5th Conference on Computational Natural Language Learning (CoNLL-2001), Toulouse, France.

Nguyen, H., et al. (2004). Combined Kernel Function for Support Vector Machine and Learning Method Based on Evolutionary Algorithm. ICONIP 2004, Calcutta, India.

Platt, J. (1998). Fast Training of Support Vector Machines using Sequential Minimal OPTimization. Advances in Kernel Methods - Support Vector Learning. B. Schoelkopf, B. C. and S. A., The MIT Press.

Poel, M., et al. (2007). A Support Vector Machine Approach to Dutch Part-of-Speech Tagging. the 7th International Symposium on Intelligent Data Analysis.

Polanyi, L. (1988). "A formal Model of the Structure of Discourse." Journal of Pragmatics 12: 601–638.

Pongsutthi, M., et al. (2013). Thai Serial Verb Constructions: A Corpus Based Study. SNLP 2013.

Pradhan, S., et al. (2004). Shallow Semantic Parsing Using Support Vector Machines. HLT-NAACL, Boston.

Ruangjaroon, S. (2005). The syntax of WH-expressions as variables in Thai. Vancouver, University of British Columbia.

Sinthupoun, S. (2009). Thai Rhetorical Structure Analysis. (PhD dissertation). National Institute of Development Administration, Bangkok.

Sornlertlamvanich, V., et al. (1997). ORCHID: Thai part-of-speech tagged corpus. , In Technical report Orchid corpus**:** 5-19.

Sporleder, C. and M. Lapata (2005). Discourse Chunking and its Application to Sentence Compression. The Human Language Technology Conference and the Conference on Empirical Methods in Natural Language Processing.

Subba, R. and B. Di Eugenio (2007). Automatic Discourse Segmentation Using Neural Networks. The 11th Workshop on the Semantics and Pragmatics of Dialogue.

Takahashi, K. (2009). "Basic Serial Verb Constructions in Thai." Journal of the Southeast Asian Linguis-tics Society 1.

Thepkanjana, K. (1986). Serial Verb Constructions in Thai, University of Michigan.

Tofiloski, M., et al. (2009). A Syntactic and Lexical-based Discourse Segmenter. The ACL-IJCNLP 2009 Conference Short Papers Association for Computational Linguistics.

Van der Vliet, N. (2010). Syntax-based Discourse Segmentation of Dutch Text. The 15th Student Session, ESSLLI.

Vijay Sundar Ram, R., et al. (2009). Tamil Clause Identifier. The 3rd National Conference on Recent Advances and Future Trends in IT-2009(RAFIT2009)

Witten, I. H. and E. Frank (2005). Data Mining: Practical Machine Learning Tools and Techniques San Francisco, Morgan Kaufmann.

Xu, Y. and F. Zhang (2006). "Using SVM to construct a Chinese dependency parser." Journal of Zhejiang University-Science A (7)2: 199-203.

Yaowapat, N. and A. Prasithrathsint (2006). Reduced relative clauses in Thai and Vietnamese. Paper presented at The Sixteenth Annual Meeting of the Southeast Asian Linguistics Society (SEALS XVI), Jakarta, Indonesia, SePTember 20-21, 2006.

กำชัย ทองหล่อ (2515). หลักภาษาไทย. กรุงเทพมหานคร, รวมสาส์น.

เทพี จรัสจรุงเกียรติ (2543). หน่วยเชื่อมโยงปริจเฉทภาษาไทยตั้งแต่สมัยสุโขทัยจนถึงปัจจุบัน, จุฬาลงกรณ์มหาวิทยาลัย.

นววรรณ พันธุเมธา (2527). ไวยากรณ์ไทย. กรุงเทพมหานคร, รุ่งเรืองสาส์น.

นัฐวุฒิ ไชยเจริญ (2544). การตัดคำและการกำกับหมวดคำภาษาไทยแบบเบ็ดเสร็จด้วยคอมพิวเตอร์, จุฬาลงกรณ์มหาวิทยาลัย.

นิเวศ จิระวิชิตชัย, et al. (2553). "การพัฒนาประสิทธิภาพการจัดหมวดหมู่เอกสารภาษาไทยแบบอัตโนมัติ." วารสารพัฒนบริหารศาสตร์. สถาบันบัณฑิตพัฒนบริหารศาสตร์.

บรรจบ พันธุเมธา (2514). ลักษณะภาษาไทย. กรุงเทพมหานคร, มหาวิทยาลัยรามคำแหง.

พระยาอุปกิตศิลปะสาร (2533). หลักภาษาไทย. กรุงเทพมหานคร, ไทยวัฒนาพานิช.

เมธี วัฒนะเมธานนท์ (2549). การรู้จำความสัมพันธ์ระดับปริจเฉทในเอกสารภาษาไทยโดยใช้โมเดลการคัดแยกแบบเนอีฟเบย์, จุฬาลงกรณ์มหาวิทยาลัย.

ราชบัณฑิตยสถาน (2548). หลักเกณฑ์การใช้เครื่องหมายวรรคตอนและเครื่องหมายอื่น ๆ หลักเกณฑ์การเว้นวรรค หลักเกณฑ์การเขียนคำย่อ ฉบับราชบัณฑิตยสถาน.พิมพ์ครั้งที่ 6 (แก้ไขเพิ่มเติม). กรุงเทพฯ, ราชบัณฑิตยสถาน.

เรืองเดช ปันเขื่อนขัติย์ (2541). ภาษาศาสตร์ภาษาไทย. นครปฐม, มหาวิทยามหิดล.

วาทินี นุ้ยเพียร, et al. (2553). การเปรียบเทียบประสิทธิภาพและวิเคราะห์การจำแนกข้อมูลโดยใช้โครงข่ายประสาทเทียมซัพพรอร์ตเวกเตอร์แมชชีน นาอีฟเบย์ และเคเนียรต์เนเบอร์. การประชุมทางวิชาการเสนอผลงานวิจัยระดับบัณฑิตศึกษา ครั้งที่ 11, มหาวิทยาลัยขอนแก่น.

วิจินตน์ ภาณุพงศ์ (2532). โครงสร้างของภาษาไทย: ระบบไวยากรณ์. กรุงเทพมหานคร, มหาวิทยาลัยรามคำแหง.

อมรา ประสิทธิ์รัฐสิทธุ์ (2543). ชนิดของคำในภาษาไทย: การวิเคราะห์ทางวากยสัมพันธ์โดยอาศัยฐานข้อมูลภาษาไทยปัจจุบันสองล้านคำ, รายงานวิจัยเสนอสำนักงานกองทุนสนับสนุนการวิจัย (ทุนวิจัยองค์ความรู้ใหม่ที่เป็นพื้นฐานต่อการพัฒนา).

อุดม วโรตม์สิกขดิตถ์ (2535). ความรู้เบื้องต้นเกี่ยวกับภาษา. กรุงเทพมหานคร, มหาวิทยาลัยรามคำแหง.

ภาคผนวก

ภาคผนวก ก ตัวอย่างคลังข้อมูลและการกำกับข้อมูล

<EDU><w CCOR>อย่างไรก็ตาม</w><w SPACE>space</w><w NCMN>อุดมการณ์</w><w NCMN>สากล</w></EDU><EDU><w COMPF>ที่</w><w VERB>ครอบงำ</w><w DET>นั้น</w></EDU><w SPACE>space</w><EDU><w PREP>ใน</w><w MNCF>หลาย</w><w PUNC>ๆ</w><w SPACE>space</w><w NCLS>กรณี</w><w CSBI>ก็</w><w NEG>ไม่</w><w ADVERB>เพียง</w><w VERB>ช่วย</w><w VERB>เปิด</w><w VERB>ให้</w><w VERB>เกิด</w><w PFN>การ</w><w VERB>สร้าง</w><w NCMN>อุดมการณ์</w><w VERB>ต่อต้าน</w><w ADVERB>เท่านั้น</w></EDU><w SPACE>space</w><EDU><w CCOR>แต่</w><w AUX>ยัง</w><w AUX>สามารถ</w><w VERB>นำ</w><w VERB>มา</w><w VERB>ใช้</w><w PFAV>อย่าง</w><w VERB>มี</w><w NCMN>พลวัต</w></EDU><w SPACE>space</w><EDU><w CSUB>เพื่อ</w><w VERB>ปรับ</w><w VERB>เปลี่ยน</w><w VERB>ให้</w><w NCMN>ประเพณี</w><w NCMN>ท้องถิ่น</w><w VERB>เป็น</w><w NCMN>กฎหมาย</w><w PREP>ของ</w><w NCMN>ชาติ</w><w SPACE>space</w><w CCORIN>หรือ</w><w PREP>แม้กระทั่ง</w><w PFN>การ</w><w VERB>สร้าง</w><w NCMN>ความหมาย</w><w VERB>ใหม่</w><w PREP>ใน</w><w NCMN>เรื่อง</w><w PREP>ของ</w><w NCMN>สิทธิ</w><w NCMN>ชุมชน</w></EDU><w SPACE>space</w><EDU><w CCOR>ดังนั้น</w><w SPACE>space</w><w NCMN>พลวัต</w><w PREP>ของ</w><w NCMN>ความรู้</w><w NCMN>ชาวบ้าน</w></EDU><EDU><w COMPF>ที่</w><w VERB>เกิด</w><w VERB>ขึ้น</w><w PREP>ใน</w><w NCMN>ประเทศ</w><w NPRP>ไทย</w></EDU><EDU><w VERB>บ่งชี้</w><w PFAV>อย่าง</w><w VERB>ชัดเจน</w></EDU><EDU><w COMPF>ว่า</w><w SPACE>space</w><w NCMN>กระแส</w><w NCMN>โลกาภิวัตน์</w><w DET>นั้น</w><w VERB>เป็น</w><w NCMN>กระบวนการ</w></EDU><EDU><w COMPF>ที่</w><w VERB>ขัดแย้ง</w><w NPRO>กันเอง</w><w ADVERB>ตลอด</w><w ADVERB>เวลา</w></EDU><w SPACE>space</w><EDU><w COMPF>ซึ่ง</w><w AUX>สามารถ</w><w VERB>นำ</w><w VERB>ไป</w><w PREP>สู่</w><w PFN>การ</w><w VERB>สร้าง</w><w NCMN>คุณค่า</w><w CCORIN>และ</w><w NCMN>อัตลักษณ์</w></EDU><EDU><w COMPF>ที่</w><w VERB>แตกต่าง</w><w NPRO>กัน</w><w PFAV>อย่าง</w><w VERB>หลากหลาย</w></EDU>

ภาคผนวก ข ตัวอย่างไฟล์ ARFF ซึ่งเป็นรูปแบบไฟล์ที่ใช้ในการฝึกฝนและทดสอบแบบจำลอง

@relation test-data-edu

@attribute POS-P
{NCMN,VERB,SPACE,PREP,PUNC,AUX,PFN,COMPF,NPRP,ADVERB,CCOR,FOREIGN,DET,CCO
RIN,NNUM,CSUB,NPRO,CSBI,NCLS,MNCF,NEG,PFAV ,ADJ,COMPNF,PT,MNCB,MCN,NON }

@attribute POS-B
{NCMN,VERB,SPACE,PREP,PUNC,AUX,PFN,COMPF,NPRP,ADVERB,CCOR,FOREIGN,DET,CCO
RIN,NNUM,CSUB,NPRO,CSBI,NCLS,MNCF,NEG,PFAV ,ADJ,COMPNF,PT,MNCB,MCN,NON }

@attribute POS-A
{NCMN,VERB,SPACE,PREP,PUNC,AUX,PFN,COMPF,NPRP,ADVERB,CCOR,FOREIGN,DET,CCO
RIN,NNUM,CSUB,NPRO,CSBI,NCLS,MNCF,NEG,PFAV ,ADJ,COMPNF,PT,MNCB,MCN,NON }

@attribute DM {Y,N}

@attribute Space {Y,N}

@attribute Punc {Y,N}

@attribute Label {Boundary,NonBoundary}

@data

CCOR,NCMN,SPACE,Y,N,N,Boundary
SPACE,CCOR,NCMN,N,N,N,NonBoundary
NCMN,SPACE,NCMN,N,Y,N,NonBoundary
NCMN,NCMN,COMPF,N,N,N,NonBoundary
COMPF,NCMN,VERB,N,N,N,Boundary
VERB,COMPF,DET,N,N,N,NonBoundary
DET,VERB,SPACE,N,N,N,NonBoundary
SPACE,DET,PREP,N,N,N,NonBoundary
PREP,SPACE,MNCF,N,N,N,Boundary
MNCF,PREP,PUNC,N,N,N,NonBoundary
PUNC,MNCF,SPACE,N,N,N,NonBoundary

SPACE,PUNC,NCLS,N,N,N,NonBoundary
NCLS,SPACE,CSBI,N,N,N,NonBoundary
CSBI,NCLS,NEG,N,N,N,NonBoundary
NEG,CSBI,ADVERB,N,N,N,NonBoundary
ADVERB,NEG,VERB,N,N,N,NonBoundary
VERB,ADVERB,VERB,N,N,N,NonBoundary
VERB,VERB,VERB,N,N,N,NonBoundary
VERB,VERB,VERB,N,N,N,NonBoundary
VERB,VERB,PFN,N,N,N,NonBoundary
PFN,VERB,VERB,N,N,N,NonBoundary
VERB,PFN,NCMN,N,N,N,NonBoundary
NCMN,VERB,VERB,N,N,N,NonBoundary
VERB,NCMN,ADVERB,N,N,N,NonBoundary
ADVERB,VERB,SPACE,N,N,N,NonBoundary
SPACE,ADVERB,CCOR,N,N,N,NonBoundary
CCOR,SPACE,AUX,Y,Y,N,Boundary
AUX,CCOR,AUX,N,N,N,NonBoundary
AUX,AUX,VERB,N,N,N,NonBoundary
VERB,AUX,VERB,N,N,N,NonBoundary
VERB,VERB,VERB,N,N,N,NonBoundary
VERB,VERB,PFAV,N,N,N,NonBoundary
PFAV,VERB,VERB,N,N,N,NonBoundary
VERB,PFAV,NCMN,N,N,N,NonBoundary
NCMN,VERB,SPACE,N,N,N,NonBoundary
SPACE,NCMN,CSUB,N,N,N,NonBoundary
CSUB,SPACE,VERB,Y,Y,N,Boundary
VERB,CSUB,VERB,N,N,N,NonBoundary
VERB,VERB,VERB,N,N,N,NonBoundary

ภาคผนวก ค ตัวอย่างผลลัพธ์ที่แสดงบนหน้าจอของโปรแกรมวีก้า

=== Run information ===

Scheme:weka.classifiers.functions.SMO -C 1.0 -L 0.001 -P 1.0E-12 -N 0 -V -1 -W 1 -K "weka.classifiers.functions.supportVector.PolyKernel -C 250007 -E 2.0"

Relation: train-data-edu

Instances: 68814

Attributes: 8

POS-P

POS-B

POS-A

DM

Space

Punc

Lable

Test mode:user supplied test set: size unknown (reading incrementally)

=== Classifier model (full training set) ===

SMO

Kernel used:

Poly Kernel: K(x,y) = <x,y>^2.0

Classifier for classes: Boundary, NonBoundary

Number of support vectors: 21544

Number of kernel evaluations: 1361254762

Time taken to build model: 18320.79 seconds

=== Predictions on test split ===

inst#, actual, predicted, error, probability distribution

1	2:NonBound	2:NonBound		0 *1
2	2:NonBound	2:NonBound		0 *1
3	2:NonBound	2:NonBound		0 *1
4	2:NonBound	2:NonBound		0 *1
5	2:NonBound	2:NonBound		0 *1
6	2:NonBound	2:NonBound		0 *1
7	1:Boundary	1:Boundary		*1 0
8	2:NonBound	2:NonBound		0 *1
9	2:NonBound	2:NonBound		0 *1
10	2:NonBound	2:NonBound		0 *1
11	2:NonBound	2:NonBound		0 *1
12	2:NonBound	2:NonBound		0 *1
13	2:NonBound	2:NonBound		0 *1
14	2:NonBound	2:NonBound		0 *1
15	2:NonBound	2:NonBound		0 *1
16	2:NonBound	2:NonBound		0 *1
17	2:NonBound	2:NonBound		0 *1
18	2:NonBound	2:NonBound		0 *1
19	1:Boundary	1:Boundary		*1 0
20	2:NonBound	2:NonBound		0 *1
21	2:NonBound	2:NonBound		0 *1
22	2:NonBound	1:Boundary	+	*1 0
23	2:NonBound	2:NonBound		0 *1
24	2:NonBound	2:NonBound		0 *1
25	2:NonBound	2:NonBound		0 *1
26	2:NonBound	2:NonBound		0 *1
27	2:NonBound	2:NonBound		0 *1
28	2:NonBound	2:NonBound		0 *1
29	1:Boundary	1:Boundary		*1 0

=== Evaluation on test set ===

=== Summary ===

Correctly Classified Instances	7398	96.7565 %
Incorrectly Classified Instances	248	3.2435 %
Kappa statistic	0.8251	
Mean absolute error	0.0324	
Root mean squared error	0.1801	
Relative absolute error	16.6831 %	
Root relative squared error	56.869 %	
Total Number of Instances	7646	

=== DETailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
	0.771	0.007	0.93	0.771	0.843	0.882	Boundary
	0.993	0.229	0.971	0.993	0.982	0.882	NonBoundary
Weighted Avg.	0.968	0.204	0.967	0.968	0.966	0.882	

=== Confusion Matrix ===

a	b	<-- classified as
666	198	a = Boundary
50	6732	b = NonBoundary

ประวัติผู้เขียนวิทยานิพนธ์

นางสาวนลินี อินต๊ะซาว เกิดเมื่อวันที่ 19 มกราคม พ.ศ. 2529 ที่จังหวัดลำปาง สำเร็จการศึกษาระดับปริญญาตรีศิลปศาสตรบัณฑิต สาขาภาษาอังกฤษ มหาวิทยาลัยเชียงใหม่ ในปีการศึกษา 2550 จากนั้นทำงานตำแหน่งครูผู้สอน (อัตราจ้าง) หลักสูตรสองภาษาที่โรงเรียนอนุบาลเชียงใหม่ และเข้าศึกษาต่อในหลักสูตรอักษรศาสตรมหาบัณฑิต ภาควิชาภาษาศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย ในปีการศึกษา 2553