บทความวิชาการ

คลังต้นไม้ภาษาไทย: แนวคิด การสร้าง และการประยุกต์ใช้

THAI TREEBANK: CONCEPTS, CONSTRUCTION, AND APPLICATIONS

ธีระพล ลิ้มศรัทธา
Theerapol Limsatta

สาขาวิชาเทคโนโลยีสารสนเทศ คณะเทคโนโลยีดิจิทัลและนวัตกรรม มหาวิทยาลัยเซาธ์อีสท์บางกอก
Department of Information Technology, Faculty of Digital Technology and Innovation,
Southeast Bangkok University
Author email: theerapol.lim@gmail.com

Received: June 4, 2025
Revised: December 2, 2025
Accepted: December 9, 2025

บทคัดย่อ

การสร้างคลังต้นไม้เป็นทรัพยากรพื้นฐานที่สำคัญในการประมวลผลภาษาธรรมชาติ เพื่อใช้ประโยชน์จากโครงสร้างไวยากรณ์ในรูปแบบต้นไม้ซึ่งช่วยให้การตีความประโยคมีความถูกต้อง คลังต้นไม้สามารถสร้างได้ทั้งแบบไม่มีเครื่องมือช่วยหรือกึ่งอัตโนมัติ และแบ่งเป็นสองกลุ่มหลักคือ คลังต้นไม้แบบโครงสร้างวลี และคลังต้นไม้พึ่งพา ภาษาไทยมีคลังต้นไม้เช่น CG Treebank และคลังต้นไม้ของสุธีที่ใช้ไวยากรณ์พึ่งพาคลังต้นไม้ไวยากรณ์ ถือเป็นคลังข้อมูลที่บรรจุประโยคพร้อมการวิเคราะห์เชิงวากยสัมพันธ์ในรูปแบบโครงสร้างต้นไม้ เพื่อสะท้อนความสัมพันธ์เชิงไวยากรณ์ระหว่างคำหรือวลี องค์ประกอบสำคัญของคลังต้นไม้ ได้แก่ ชุดข้อมูลข้อความต้นฉบับ, การตัดคำที่แม่นยำ, การตัดแบ่งชนิดของคำโดยใช้ระบบการตัดคำที่เหมาะสมกับภาษาไทย, โครงสร้างต้นไม้ไวยากรณ์, คำอธิบายประกอบมาตรฐาน, และรูปแบบข้อมูล การเปรียบเทียบกับภาษาอื่น ๆ เช่น Penn Treebank และ Universal Dependencies แสดงให้เห็นว่าภาษาไทยมีลักษณะเฉพาะ เช่น การไม่มีการเว้นวรรคระหว่างคำ, การละองค์ประกอบในประโยค, และการใช้คำหลายหน้าที่ บทความนี้ใช้แนวคิดทฤษฎีไวยากรณ์ X-bar ที่อธิบายโครงสร้างภายในวลี การสร้างคลังต้นไม้ไวยากรณ์ภาษาไทยมีความท้าทายเนื่องจากลักษณะเฉพาะของภาษา โดยมีการประยุกต์ใช้ X-bar กับไวยากรณ์ภาษาไทยโดยการดัดแปลงให้เข้ากับลักษณะเฉพาะ เช่น การไม่มีส่วนขยายด้านซ้าย (Specifier) ชัดเจน และการจัดการโครงสร้างซ้อน รวมถึงการรองรับการละองค์ประกอบในประโยคโดยแสดงโหนดที่ถูกละการกำหนดมาตรฐานการจัดโครงสร้างที่ดีควรประกอบด้วยคู่มือที่ครอบคลุม ระบบตรวจสอบความสอดคล้อง และตัวอย่างที่หลากหลาย คลังต้นไม้ภาษาไทยมีความสำคัญต่อการพัฒนาเทคโนโลยีภาษาศาสตร์คอมพิวเตอร์อย่างมาก เช่น การพัฒนาระบบวิเคราะห์ไวยากรณ์อัตโนมัติ การปรับปรุงระบบแปลภาษาและการสอนภาษาไทยและภาษาศาสตร์คลังต้นไม้ไม่เพียงเป็นทรัพยากรเชิงเทคนิคแต่ยังเป็นฐานข้อมูลที่มีคุณค่าทางภาษาศาสตร์ วัฒนธรรม และการอนุรักษ์ภาษา เพื่อต่อยอดงานวิจัยและนวัตกรรมด้านภาษาศาสตร์คอมพิวเตอร์ของไทย

คำสำคัญ: ไวยากรณ์ภาษาไทย, คลังต้นไม้, ทฤษฎีเอ็กซ์-บาร์

บทความวิชาการ

Abstract

Treebank construction is a fundamental resource in natural language processing, leveraging grammatical structures in tree format to ensure accurate sentence interpretation. Treebanks can be created manually or semi-automatically and are categorized into phrase structure treebanks and dependency treebanks. For Thai, notable treebanks include CG Treebank using Categorial Grammar and Suthee's treebank using dependency grammar. A treebank is a data repository containing natural language sentences with syntactic analysis in tree structures, reflecting grammatical relationships between words or phrases. Key components include original text data, accurate word segmentation, Part-of-Speech tagging with appropriate Thai tag sets, syntactic tree structures, standard annotation guidelines, and data formats. Comparisons with other languages like English (Penn Treebank) and Universal Dependencies highlight unique Thai characteristics such as absence of word spacing, ellipsis of sentence components, and polysemous word usage. This article describes the X-bar theory, which explains internal phrase structures. Thai treebank construction poses challenges due to the language's specific characteristics. The application of X-bar theory to Thai grammar requires adaptation, including handling the absence of clear specifiers, managing nested structures, and accommodating null nodes for omitted elements. Establishing robust annotation standards involves comprehensive guidelines, standardized constituent types and POS tags, validation tools, and diverse annotated examples. Thai treebanks are crucial for advancing NLP technologies, particularly for automatic parsing systems, machine translation improvement, and Thai language education. Beyond technical utility, Thai treebanks serve as valuable linguistic, cultural, and language preservation databases, fostering further research and innovation in Thai computational linguistics.

Keywords: Thai Grammar, Treebank, X bar Theory

บทนำ

การสร้างคลังต้นไม้เป็นทรัพยากรพื้นฐานที่สำคัญในการทำงานด้านการประมวลผลภาษาธรรมชาติ เช่น การแปล การวิเคราะห์ความรู้สึก การจับประเด็นข้อความ โดยอาศัยประโยชน์ของโครงสร้างไวยากรณ์ในรูปแบบต้นไม้ที่จะทำให้ประโยคมีความหมายถูกต้องในการตีความของประโยค ตัวอย่างเช่น การใช้ประโยชน์ในการตัดคำ “หลวงตามหาบัวไม่อยู่” การตัดคำที่ถูกต้อง ต้องตัดให้ได้ว่า หลวงตา/มหาบัว/ไม่/อยู่ แทนที่จะตัดได้ว่า หลวง/ตาม/หา/บัว/ไม่/อยู่ เพราะโครงสร้างประโยคตามการตัดคำแบบแรกมีเป็นไปได้มากกว่าถึงทั้งสองแบบการตัดคำจะถูกไวยากรณ์ทั้งคู่ เมื่อได้การตัดที่ถูกต้องก็สามารถนำไปใช้ประโยชน์ในด้านอื่น ๆ ต่อไปได้ เช่น นำไปใช้กับการวิเคราะห์ความรู้สึกในด้านบวก หรือลบ เพราะแนวการวิเคราะห์ความรู้สึกมักมีพื้นฐานมาจากใช้ค่าความถี่ของคำ การสร้างคลังต้นไม้อาจสร้างโดยไม่มีเครื่องมือช่วยหรือสร้างแบบกึ่งอัตโนมัติ

โดยใช้โปรแกรมแจกแจงช่วยกำกับร่วมกับนักภาษาศาสตร์ช่วยตรวจสอบความถูกต้องของประโยคอีกครั้ง คลังต้นไม้มีแยกเป็นสองกลุ่มหลักคือ คลังต้นไม้แบบโครงสร้างวลี (Phrase structure) เช่น คลังต้นไม้เพ็นน์ (Penn Treebank) และคลังต้นไม้พึ่งพา (Dependency Treebank) เช่น คลังต้นไม้พึ่งพาปราก (The Prague Dependency Treebank)

ในภาษาไทยมีคลังต้นไม้ชื่อ ซีจี (CG Treebank) [1] ที่ใช้ทฤษฎีไวยากรณ์เคทาโกเรียล (Categorial Grammar) ซึ่งจัดอยู่ในกลุ่มคลังต้นไม้โครงสร้างวลี และคลังต้นไม้ของสุธี [2] ใช้ไวยากรณ์พึ่งพา แม้ทั้งสองประเภทคลังต้นไม้สร้างจากไวยากรณ์ที่ต่างกัน แต่ใช้แนวการสร้างมาจากคลังคำที่มีคำกำกับหมวดหมู่ จับคู่คำเพื่อเชื่อมโยงความสัมพันธ์กันได้ โดยไม่จำเป็นต้องสร้างเหมือนอย่างไวยากรณ์เพิ่มพูน (Generative Grammar) บทความนี้อธิบายแนวคิดทฤษฎีไวยากรณ์ภาษาที่เป็น X-bar ที่อธิบายโครงสร้างภายในวลี (Phrase Structure) การใช้ความน่าจะเป็นเพื่อถ่วงน้ำหนักแต่ละหน่วยคำในโครงสร้างประโยค และดูทุกความเป็นไปว่าโครงสร้างประโยคใดมีความน่าจะเป็นสูงสุด แนวทางการสร้างคลังต้นไม้ และการประยุกต์ใช้ในด้านต่าง ๆ

1. ความหมายและองค์ประกอบของคลังต้นไม้ไวยากรณ์

คลังต้นไม้ไวยากรณ์ (Treebank) เป็นคลังข้อมูลที่บรรจุประโยคจากภาษาธรรมชาติพร้อมการวิเคราะห์เชิงวากยสัมพันธ์ (Syntactic Analysis) ซึ่งแสดงในรูปแบบโครงสร้างต้นไม้ (Tree Structure) เพื่อสะท้อนความสัมพันธ์เชิงไวยากรณ์ระหว่างคำหรือวลีภายในประโยค โครงสร้างดังกล่าวอาจอยู่ในรูปแบบของโครงสร้างแบบวลี (Constituency Structure) หรือโครงสร้างแบบพึ่งพา (Dependency Structure) ทั้งนี้ขึ้นอยู่กับแนวทางการวิเคราะห์ที่ใช้ [3], [4]

1.1 แนวคิดพื้นฐานของคลังต้นไม้ไวยากรณ์

แนวคิดของคลังต้นไม้ คือการจัดระบบข้อมูลภาษาโดยอ้างอิงจากโครงสร้างไวยากรณ์ที่มนุษย์สามารถตีความได้ ซึ่งทำให้สามารถนำไปฝึกแบบจำลองทางภาษาศาสตร์เชิงคำนวณ (Computational Linguistic Models) ได้อย่างมีประสิทธิภาพ องค์ประกอบสำคัญของคลังต้นไม้ ได้แก่ คำที่ถูกตัดและชนิดของคำหรือวลีที่จัดกลุ่มตามหน้าที่ทางไวยากรณ์และความสัมพันธ์เชิงพึ่งพาระหว่างคำ

1.2 องค์ประกอบหลักของคลังต้นไม้

คลังต้นไม้ไวยากรณ์โดยทั่วไปมีองค์ประกอบสำคัญดังนี้:

- ชุดข้อมูลข้อความต้นฉบับ ข้อความหรือประโยคที่ถูกคัดเลือกจากแหล่งต่าง ๆ เช่น ข่าวสาร วรรณกรรม หรือบทสนทนา เพื่อความหลากหลายเชิงรูปแบบและบริบท
- การตัดคำ (Word Segmentation) โดยเฉพาะภาษาไทยซึ่งไม่มีการเว้นวรรคระหว่างคำการตัดคำที่แม่นยำเป็นปัจจัยสำคัญต่อคุณภาพของคลังต้นไม้ [5] ซึ่งอาจต้องอาศัยผู้เชี่ยวชาญตรวจทาน
- การตัดแบ่งชนิดของคำ (Part-of-Speech Tagging) คำแต่ละคำจะถูกกำหนดชนิด เช่น NCMN (คำนามสามัญ), VACT (กริยากระทำ) โดยใช้ระบบแท็กที่เหมาะสมกับภาษาไทย เช่น Thai POS Tagset ของ NECTEC
- โครงสร้างต้นไม้ไวยากรณ์ (Syntactic Tree Structure) โครงสร้างนี้แสดงให้เห็นความสัมพันธ์ระหว่างคำหรือวลีโดยอาจใช้แนวทาง Constituency หรือ Dependency
- คำอธิบายประกอบมาตรฐาน (Annotation Guideline) คลังต้นไม้ ที่ดีจะต้องมีคู่มือหรือแนวปฏิบัติที่สอดคล้องกันในการจัดวางโครงสร้าง เช่น Universal Dependencies Guidelines [4]
- รูปแบบข้อมูล (Data Format) – เช่น Bracketed format หรือ CoNLL-U format ซึ่งรองรับการนำไปใช้งานกับเครื่องมือวิเคราะห์ไวยากรณ์อัตโนมัติ

1.3 การเปรียบเทียบกับคลังต้นไม้ของภาษาอื่น

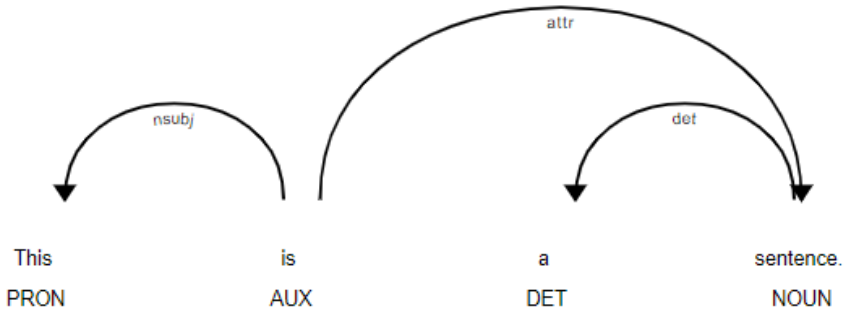
การเปรียบเทียบกับคลังต้นไม้ของภาษาอื่น ๆ แสดงให้เห็นถึงลักษณะเฉพาะของภาษาไทย ตัวอย่างเช่น:

- Penn Treebank (ภาษาอังกฤษ) ใช้โครงสร้างแบบวลี (Constituency-Based) โดยมีการจัดกลุ่มวลีชัดเจน และสามารถประยุกต์ใช้กับโมเดลทางภาษาศาสตร์หลากหลาย [3] มีการอ้างอิงกันมาก เพราะได้สร้างมาตั้งแต่ปี ค.ศ. 1989 และยังคงพัฒนาต่อเนื่อง โดยสมาคมข้อมูลภาษาศาสตร์ (Linguistic Data Consortium) การสร้างใช้ทั้งระบบอัตโนมัติและตรวจและแก้ไข โดยผู้เชี่ยวชาญของขั้นตอนการแบ่งประเภทของคำ (Part of Speech) การแจงประโยคและการเขียนประกอบคำอธิบายเพิ่มเติมสำหรับประโยคคำพูดที่ไม่สมบูรณ์
- Universal Dependencies (ภาษาต่าง ๆ) เป็นมาตรฐานโครงสร้างแบบพึ่งพาที่เน้นการสร้างคลังต้นไม้ข้ามภาษามีแนวทางการกำหนด “หัวคำ” (Head) และ “คำที่พึ่งพา” (Dependent) ที่สอดคล้องกันในหลายภาษา เช่น อังกฤษ ญี่ปุ่น จีน และไทย [4]
- Japanese Kyoto University Text Corpus ใช้คลังต้นไม้พึ่งพา เช่นเดียวกับภาษาไทย แต่ลักษณะการเรียงคำและการใช้คำช่วย (Particles) ทำให้สามารถวิเคราะห์โครงสร้างแบบพึ่งพาได้ชัดเจนกว่าภาษาไทยที่มีการละคำได้บ่อย เช่น การละประธานหรือกรรม [6]

ภาษาไทย มีลักษณะเด่นคือไม่มีการเว้นวรรคระหว่างคำ การละองค์ประกอบในประโยค เช่น ประธานหรือกรรม และการใช้คำหลายหน้าที่ในบริบทต่างกัน จึงทำให้การสร้างคลังต้นไม้ต้องใช้วิธีที่เหมาะสมกับธรรมชาติของภาษา เช่น การมีตัดคำเฉพาะทาง หรือการจัดกลุ่มคำแบบยืดหยุ่นมากขึ้น [7]

```
(S
  (NP-SBJ
    (NP (NNP Pierre) (NNP Vinken))
    (, ,)
    (ADJP (NP (CD 61) (NNS years)) (JJ old))
    (, ,))
  (VP
    (MD will)
    (VP
      (VB join)
      (NP (DT the) (NN board))
      (PP-CLR (IN as) (NP (DT a) (JJ nonexecutive) (NN director)))
      (NP-TMP (NNP Nov.) (CD 29))))
  (. .))
```

รูปที่ 1 ต้นไม้ไวยากรณ์ ของ Penn Treebank



รูปที่ 2 แผนภาพโครงสร้างความสัมพันธ์แบบพึ่งพา

1.4 ความสำคัญในการวิจัยและการประยุกต์ใช้

คลังต้นไม้ เป็นเครื่องมือสำคัญในหลายงานวิจัยและระบบประมวลผลภาษาธรรมชาติ เช่น:

- การฝึกโมเดลการแจงประโยค เช่น คลังต้นไม้ของ CG Treebank ก็ถูกสร้างขึ้นเพื่อเป็นแหล่งข้อมูลภาษาไทยเพื่อนำไปฝึกแจงประโยคตามแบบไวยากรณ์เคทาโกเรียล
- การแปลภาษาด้วยเครื่อง
- การเข้าใจความหมายของประโยค
- การวิเคราะห์โครงสร้างของข้อความในงานวรรณกรรมหรือนิติศาสตร์

คลังต้นไม้จึงไม่เพียงแต่เป็นแหล่งข้อมูล แต่ยังเป็นรากฐานของเทคโนโลยีภาษาธรรมชาติที่มีความแม่นยำสูงในระดับเชิงไวยากรณ์

2. การสร้างคลังต้นไม้ไวยากรณ์ภาษาไทย

การสร้างคลังต้นไม้ไวยากรณ์ภาษาไทยมีความท้าทายหลายประการเนื่องจากภาษาไทยมีลักษณะเฉพาะทางวากยสัมพันธ์ที่ต่างจากภาษายุโรป เช่น การไม่มีการผันคำ การไม่มีการเว้นวรรคระหว่างคำ และการที่โครงสร้างประโยคสามารถละประธาน กริยา หรือกรรมได้โดยไม่ผิดไวยากรณ์ ด้วยเหตุนี้การออกแบบโครงสร้างไวยากรณ์สำหรับคลังต้นไม้ภาษาไทยจึงต้องอาศัยกรอบทฤษฎีทางภาษาศาสตร์ที่ยืดหยุ่นและเหมาะสม หนึ่งในทฤษฎีหลักที่ถูกนำมาใช้ คือ ทฤษฎี X-bar (X-bar Theory) ซึ่งเป็นทฤษฎีโครงสร้างวากยสัมพันธ์ที่อยู่ภายใต้แนวทางไวยากรณ์เพิ่มพูน

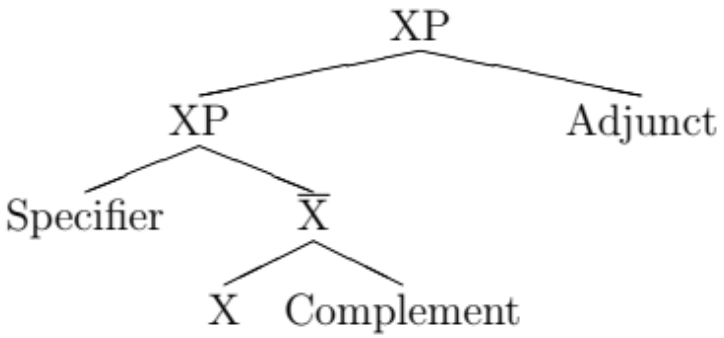
2.1. ทฤษฎี X-bar กับการวิเคราะห์ไวยากรณ์

ทฤษฎี X-bar เป็นกรอบแนวคิดที่อธิบายโครงสร้างภายในวลี (Phrase Structure) โดยเสนอว่าองค์ประกอบไวยากรณ์ทุกประเภท (คำนาม กริยา คุณศัพท์ ฯลฯ) มีโครงสร้างร่วมกันในรูปแบบสามระดับ ได้แก่:

- X (head): ส่วนหัวของวลี
- X (X-bar): ระดับกลางซึ่งประกอบด้วยหัวและส่วนขยายบางส่วน
- XP (Maximal Projection): วลีทั้งหมด ซึ่งอาจรวมส่วนขยายด้านซ้าย (Specifier) และด้านขวา (Complement)

โดย XP ใช้เพื่อลดทอนจำนวนกฎให้น้อยลง โดยให้เหลือเพียงกฎเดียวแต่ครอบคลุมกฎทั้งหมด ในทฤษฎีเอ็กบาร์ (X’) จึงลดเหลือเพียงกฎเดียว คือ X’ ⟶ X (XP)

โดยที่ X แทนหมวดหมู่คำหลัก เป็นตัวแปรสำหรับแทนประเภทของ N, V, Adj, Adv, P ให้ XP แทน NP, VP, AP, PP และ X' แทน N', V', Adj', Adv', P'



รูปที่ 3 โครงสร้างทั่วไปของ XP

ตัวอย่างเช่น ในภาษาอังกฤษ เช่น "The Big Dog" ถูกวิเคราะห์เป็น NP (Noun Phrase) ซึ่งมี "Dog" เป็น Head, "Big" เป็น Modifier และ "The" เป็น Specifier

เมื่อประยุกต์กับภาษาไทย เช่น วลี "สุนัขตัวใหญ่" โครงสร้างตาม X-bar จะเป็นดังนี้:

- XP = NP (Noun Phrase)
- Head = "สุนัข"
- Modifier = "ตัวใหญ่" (ซึ่งอาจเป็น part of NP หรือ AdjP แล้วแต่การตีความ)
- ไม่มี Specifier ชัดเจนในภาษาไทย ซึ่งแสดงให้เห็นข้อแตกต่างเชิงโครงสร้างระหว่างภาษาไทยกับภาษาอังกฤษ

2.2. การประยุกต์ใช้ X-bar กับไวยากรณ์ภาษาไทย

แม้ทฤษฎี X-bar จะถูกพัฒนามาจากการศึกษาภาษาอังกฤษและภาษาอินโด-ยูโรเปียนอื่น ๆ แต่ก็สามารถปรับประยุกต์เพื่อใช้อธิบายภาษาไทยได้ในหลายกรณี โดยต้องอาศัยการดัดแปลงบางประการ:

- Specifier: ภาษาไทยมักไม่ปรากฏ specifier ชัดเจนในวลี เช่น ไม่มีคำนำหน้านามเหมือน "The" หรือ "A" ดังนั้น NP ในภาษาไทยจึงมักประกอบด้วย Head และ Modifier เท่านั้น [5]
- คำลำดับ (Modifiers): ภาษาไทยสามารถมีคำขยายได้หลายรูปแบบ เช่น นามขยาย กริยาขยาย คุณศัพท์ขยาย ซึ่งต้องออกแบบโครงสร้างต้นไม้ให้รองรับความหลากหลายนี้ [5]
- โครงสร้างซ้อน: ภาษาไทยสามารถมีวลีซ้อนหลายชั้น เช่น "หนังสือของนักเรียนของโรงเรียน" ซึ่งต้องการการวิเคราะห์แบบ Recursive ตามแนวทางของ X-bar [8]
- การละองค์ประกอบ: ภาษาไทยสามารถละประธาน หรือกรรมในบางบริบท เช่น "กินข้าวแล้ว" ซึ่งต้องออกแบบโครงสร้างที่สามารถ "คาดหมาย" โหนดที่ถูกละได้ [9]

ด้วยข้อพิจารณาดังกล่าว การนำทฤษฎี X-bar มาใช้กับการสร้าง คลังต้นไม้ภาษาไทย จึงไม่ใช่การลอกแบบตรงจากภาษาอังกฤษ แต่เป็นการใช้แนวคิด "โครงสร้างไวยากรณ์ที่มีระดับชั้น" แล้วปรับให้เข้ากับลักษณะเฉพาะของภาษาไทย

2.3. ตัวอย่างโครงสร้างตาม X-bar ของภาษาไทย

ประโยค: “นักเรียนอ่านหนังสือ” สามารถวิเคราะห์ได้ในโครงสร้างแบบ Constituency โดยอิงทฤษฎี X-bar ดังนี้:

[S [NP นักเรียน] [VP [V อ่าน] [NP หนังสือ]]]

โดยสามารถเพิ่มระดับ X-bar ได้ เช่น

[S [NP [N′ [N นักเรียน]]] [VP [V′ [V อ่าน] [NP[N′ [N หนังสือ]]]]]]

2.4. การถ่วงน้ำหนักด้วยความน่าจะเป็น

โครงสร้างประโยคที่เพิ่มค่าถ่วงน้ำหนักเรียกอีกชื่อว่า ไวยากรณ์ความน่าจะเป็นแบบไม่พึงบริบท (Probability Context Free Grammar) เป็นการตรวจทุกความเป็นได้ของต้นไม้ที่สร้างเป็นประโยค แล้ววัดว่าโครงสร้างประโยคใดมีความน่าจะเป็นสูงสุด

จากรูป 4 เมื่อนำมาคำนวณความน่าจะเป็นของต้นไม้โครงสร้างประโยคแรก และ ต้นไม้โครงสร้างที่สองได้ความน่าจะเป็นของต้นไม้ไวยากรณ์แบบที่สองมากกว่า

$$P(S_1) = 1.0 \times 0.1 \times 0.3 \times 0.7 \times 1.0 \times 1.0 \times 0.18 \times 1.0 \times 0.1 = 0.000378$$

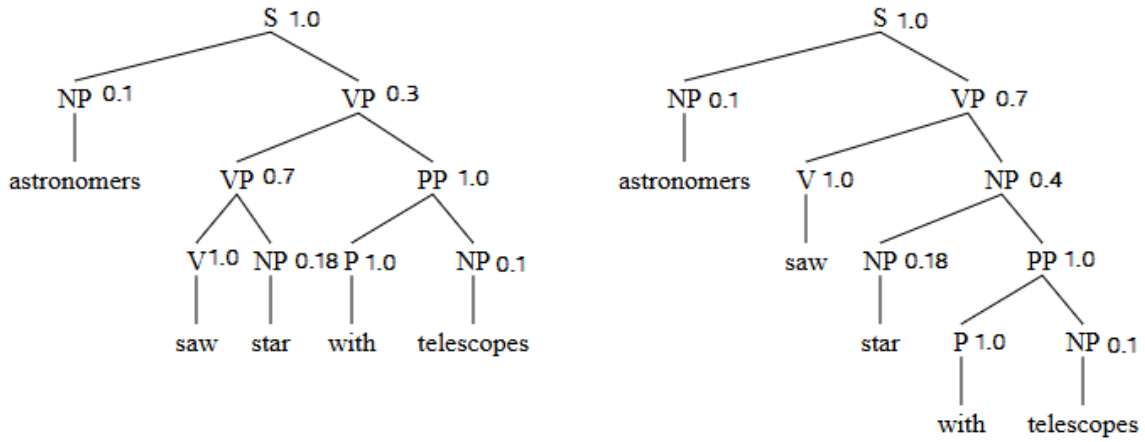
$$P(S_2) = 1.0 \times 0.1 \times 0.7 \times 1.0 \times 0.4 \times 1.8 \times 1.0 \times 1.0 \times 0.1 = 0.000504$$

ดังนั้นโครงสร้างต้นไม้แบบที่สอง () มีโอกาสถูกสร้างได้มากกว่า

(S)

(NP astronomers)

(VP (V saw) (NP (NP star) (PP (P with) (NP telescopes))))) (p=0.000504)



รูปที่ 4 ต้นไม้ไวยากรณ์ ที่สร้างได้สองแบบ (ซ้าย และ ขวา)

2.5. การนำไปใช้งานในคลังต้นไม้

การกำหนดโครงสร้างต้นไม้โดยอิงทฤษฎี X-bar

- โครงสร้างมีความสม่ำเสมอและเป็นระบบ
- ง่ายต่อการวิเคราะห์อัตโนมัติและการสกัดคุณลักษณะทางไวยากรณ์
- สามารถนำไปฝึกแบบจำลองสังเคราะห์ประโยค (Syntactic Generation) ได้ดี
- รองรับการขยาย Treebank ไปยังวลีที่ซับซ้อนมากขึ้น

อย่างไรก็ตาม ยังต้องมีการกำหนดอนุกรมวิธานของประเภทวลีและโหนดให้ชัดเจน เพื่อให้สามารถใช้ในเครื่องมือกับป้ายชื่อได้อย่างเป็นระบบ เช่น ในโครงการ Thai Treebank ของ NECTEC [10] ซึ่งเป็นโครงการที่พัฒนาต่อเนื่องจากคลังคำ Best [11]

3. มาตรฐานและแนวทางการจัดโครงสร้าง

การจัดโครงสร้างไวยากรณ์ในคลังต้นไม้ต้องอาศัยแนวทางที่ชัดเจนและสอดคล้องกัน เพื่อให้การวิเคราะห์ประโยคมีความแม่นยำสม่ำเสมอ และสามารถใช้งานร่วมกับเครื่องมือภาษาศาสตร์คอมพิวเตอร์ต่าง ๆ ได้อย่างมีประสิทธิภาพ โดยเฉพาะอย่างยิ่งในกรณีของภาษาไทย ซึ่งมีลักษณะเฉพาะที่แตกต่างจากภาษาในกลุ่มอินโด-ยูโรเปียน การกำหนดมาตรฐานจึงต้องพิจารณาให้เหมาะสมกับโครงสร้างภาษาไทย ทั้งในระดับคำ วลี และประโยค

3.1. โครงสร้างระดับวลี (Phrase Structure)

ภาษาไทยเป็นภาษาที่มีโครงสร้างแบบวลีเป็นหลัก (Phrase-Based Language) ซึ่งสามารถใช้ได้ทั้งการจัดแบบ Constituency Tree และ Dependency Tree โดยที่การจัดโครงสร้างในระดับวลีควรพิจารณาองค์ประกอบต่อไปนี้:

- วลีคำนาม (NP - Noun Phrase): เช่น “นักเรียนหญิงเก่ง” โดย “นักเรียน” เป็นแกน (Head) ของวลี
- วลีกริยา (VP - Verb Phrase): เช่น “อ่านหนังสือ”
- วลีบุพบท (PP - Prepositional Phrase): เช่น “ในห้องเรียน” ซึ่งเริ่มต้นด้วยคำบุพบท
- วลีกริยาวิเศษณ์ (AdvP - Adverbial Phrase): เช่น “อย่างรวดเร็ว”

แต่ละวลีควรแยกโครงสร้างภายในให้ชัดเจนตามบทบาทในประโยค เช่น ตำแหน่งประธาน (Subject), กรรมตรง (Direct Object), กรรมรอง (Indirect Object), ส่วนขยาย (Modifier) เป็นต้น [8]

3.2. มาตรฐานการแท็กชนิดวลีและโหนด (Constituent Labels)

โหนดต่าง ๆ ในต้นไม้ควรมีการกำหนดชื่อมาตรฐาน เพื่อให้สามารถวิเคราะห์หรือแปลงข้อมูลได้อย่างมีระบบ โดยทั่วไปจะใช้ชื่อเช่น:

- S: ประโยค (Sentence)
- NP: วลีคำนาม
- VP: วลีกริยา
- PP: วลีบุพบท
- ADJP: วลีคุณศัพท์
- ADVP: วลีกริยาวิเศษณ์
- N, V, PRP, AUX, etc.: ชนิดคำที่เป็น โหนดปลายทาง (Terminal Nodes)

ในกรณีภาษาไทย อาจมีการเพิ่มโหนดเฉพาะ เช่น CL (Classifiers - ลักษณนาม), PART (คำช่วย), หรือ NEG (คำปฏิเสธ) ซึ่งจำเป็นสำหรับการตีความไวยากรณ์ในบริบทของภาษาไทย [12]

3.3. การจัดการลำดับคำและการละองค์ประกอบ

ภาษาไทยมีลำดับคำแบบ SVO (ประธาน – กริยา – กรรม) เป็นหลัก แต่สามารถสลับได้ในบางบริบท เช่น การย้ายกรรมขึ้นต้นเพื่อเน้นหรือการละประธานเมื่อทราบจากบริบท เช่น:

- “เขากินข้าวแล้ว” ⟶ “กินข้าวแล้ว”
- “ข้าวเขากินแล้ว” (เพื่อเน้น “ข้าว”)

โครงสร้างต้นไม้ต้องสามารถรองรับการละประธานหรือกรรมโดยแสดงโหนดที่ถูกละ (Null Nodes) หรือโดยใช้แท็กพิเศษ เช่น PRO หรือ Ø เพื่อแสดงองค์ประกอบที่แฝงอยู่ [7]

3.4. ความสอดคล้องกับมาตรฐานสากล

แม้จะเน้นลักษณะเฉพาะของภาษาไทย การสร้างคลังต้นไม้ที่ดีควรคำนึงถึงความสามารถในการใช้งานร่วมกับมาตรฐานสากล เช่น:

- Penn Treebank-style: เน้นการแบ่ง Constituent Structure แบบ Top-down และการแท็ก POS แบบละเอียด [3]
- Universal Dependencies (UD): มาตรฐานระดับการพึ่งพา ที่ใช้ได้กับหลายภาษารวมทั้งภาษาไทย โดยมีการปรับปรุง Schema ให้สอดคล้องกับภาษาที่มีโครงสร้างไม่เป็นเชิงรูป (Non-Inflectional) เช่น ไทย ลาว และเวียดนาม [8]

สำหรับภาษาไทยมีโครงการที่ใช้แนวทาง UD อย่างเป็นระบบ ได้แก่ Thai UD Treebank โดย NECTEC และโครงการ TWT (Thai Word Treebank) ซึ่งจัดโครงสร้างแบบ พึ่งพา พร้อม POS tagging แบบ Universal POS [13]

3.5 แนวทางการวิเคราะห์ไวยากรณ์เฉพาะของภาษาไทย

การจัดการคำหลายบทบาท: คำเดียวกันในภาษาไทยอาจเป็นได้ทั้งคำนาม กริยา หรือคำวิเศษณ์ เช่น “เดิน” ⟶ เดิน (V), ทางเดิน (N), เดินเร็ว (Adv) ซึ่งต้องอาศัยบริบทและแนวทาง Annotation ที่ชัดเจน [9]

คำลักษณนาม: ภาษาไทยมีลักษณนาม ซึ่งไม่พบในภาษาอังกฤษ เช่น “นักเรียน สอง คน” ⟶ “สอง” เป็นตัวเลข “คน” เป็นลักษณนาม ต้องแยกออกจากคำนามหลัก [8]

คำช่วยและอนุภาค: เช่น “นะ”, “สิ”, “ซิ”, “ล่ะ”, “ละ”, ซึ่งมีความสำคัญในระดับวัจนปฏิบัติ (Pragmatics) และควรมีแท็กเฉพาะ เช่น PART หรือ INTJ เพื่อแสดงหน้าที่

ตัวอย่างโครงสร้างต้นไม้ (แบบ Constituency)

```
[S [NP [N นักเรียน]] [VP
        [V กิน]
                [NP
                        [NUM สอง]
                        [CL คน]
                        [N ข้าว]]] ]
```

3.6 ข้อเสนอแนะ

การกำหนดมาตรฐานที่ดีควรประกอบด้วย:

- คู่มือการจัดโครงสร้างที่ครอบคลุมกรณีหลากหลาย
- ตารางคำย่อของ Constituent Types และ POS tags
- ระบบการตรวจสอบความสอดคล้องของโครงสร้าง (Tree Validation Tools)
- การจัดทำตัวอย่าง เพิ่มคำอธิบาย (Annotated) ที่มีความหลากหลายและครอบคลุมลักษณะภาษาไทย

4. การประยุกต์ใช้

คลังต้นไม้ไวยากรณ์ภาษาไทยมีบทบาทสำคัญในการพัฒนาเทคโนโลยีทางภาษาศาสตร์คอมพิวเตอร์ เนื่องจากเป็นแหล่งข้อมูลเชิงโครงสร้าง (Syntactic Structure) ที่ได้รับการตรวจสอบและจัดระเบียบแล้วอย่างเป็นระบบ คลังต้นไม้ที่มีคุณภาพจะช่วยยกระดับความแม่นยำของระบบประมวลผลภาษาธรรมชาติ โดยเฉพาะอย่างยิ่งกับภาษาไทยซึ่งมีความซับซ้อนในเชิงวากยสัมพันธ์อย่างมาก

4.1. การพัฒนาระบบวิเคราะห์ไวยากรณ์อัตโนมัติ

หนึ่งในประโยชน์หลักของคลังต้นไม้คือการใช้เป็นข้อมูลฝึก (Training Data) สำหรับสร้างแบบจำลองการวิเคราะห์โครงสร้างประโยคอัตโนมัติไม่ว่าจะเป็นแบบ Constituency Parsing หรือ Dependency Parsing ซึ่งเป็นรากฐานของการแปลภาษาการทำความเข้าใจความหมายของประโยค และระบบถามตอบในภาษาไทย [5]

ตัวอย่างเช่น Thai Dependency Treebank ที่พัฒนาโดย NECTEC ได้ถูกนำไปฝึกระบบที่ใช้โมเดลแบบการเรียนรู้ด้วยเครื่องจักร (Machine Learning) และ การเรียนรู้เชิงลึก (Deep Learning) เช่น BERT-Based Parsers สำหรับภาษาไทย ซึ่งสามารถลดข้อผิดพลาดในการวิเคราะห์โครงสร้างไวยากรณ์ลงได้อย่างมีนัยสำคัญ [13]

4.2. การปรับปรุงระบบแปลภาษา

การแปลภาษาอัตโนมัติ (เช่น Thai–English หรือ Thai–Chinese) จะมีประสิทธิภาพมากขึ้นเมื่อระบบสามารถเข้าใจโครงสร้างวลีและบทบาทของคำในประโยคอย่างถูกต้อง คลังต้นไม้ช่วยให้แบบจำลองทราบว่าคำใดเป็นประธาน กริยา หรือกรรม ซึ่งมีผลต่อการจัดลำดับคำและการรักษาความหมายที่ถูกต้องในภาษาปลายทาง [14]
คลังต้นไม้ภาษาไทยที่เชื่อมโยงกับ คลังต้นไม้ ของภาษาอื่น (Parallel Treebanks) ยังสามารถใช้ในงาน Cross-Lingual Alignment และการระบบฝึกการแปลแบบ Supervised ได้อย่างมีประสิทธิภาพ

4.3. การสกัดคุณลักษณะทางวากยสัมพันธ์ (Syntactic Feature Extraction)

คลังต้นไม้ช่วยในการสกัดคุณลักษณะเชิงไวยากรณ์ เช่น รูปแบบวลี ความลึกของต้นไม้ ตำแหน่งของประธานหรือกรรม ซึ่งสามารถนำไปใช้ในหลายงาน เช่น:

- การตรวจสอบความถูกต้องของประโยค (Grammar Checking)
- การประเมินความซับซ้อนของข้อความ (Text Complexity)
- การสร้างประโยคในระบบสนทนา (Natural Language Generation)

โดยเฉพาะกับภาษาไทยซึ่งมักมีการละองค์ประกอบ เช่น ประธาน หรือกรรม คลังต้นไม้ที่เพิ่มคำอธิบายอย่างละเอียดจะช่วยระบุส่วนที่แฝงอยู่ และเป็นฐานข้อมูลสำหรับระบบทำความเข้าใจภาษา (Natural Language Understanding) ที่ลึกซึ้งยิ่งขึ้น

4.4. การวิจัยภาษาศาสตร์เชิงคำนวณ (Computational Linguistics)

Treebank ภาษาไทยยังเป็นแหล่งข้อมูลสำคัญสำหรับการศึกษาคุณลักษณะทางไวยากรณ์ เช่น:

- ความแตกต่างเชิงวากยสัมพันธ์ระหว่างถ้อยความระดับปากกับระดับทางการ
- การศึกษาโครงสร้างที่ไม่ตรงตามไวยากรณ์ (Ill-Formed Syntax)
- การวิเคราะห์โครงสร้างซ้อนและความไม่ชัดเจนในการตีความ (Ambiguity)

คลังต้นไม้ที่มีการเพิ่มคำอธิบาย อย่างพิถีพิถัน เช่น คลังต้นไม้ภาษาไทยของโครงการ NECTEC หรือ Thai UD Treebank ช่วยให้นักวิจัยสามารถตรวจสอบข้อสันนิษฐานทางภาษาศาสตร์ได้จากข้อมูลจริงขนาดใหญ่

4.5. การสอนภาษาไทยและภาษาศาสตร์

ในด้านการศึกษาคลังต้นไม้ไวยากรณ์ภาษาไทยสามารถนำไปใช้เป็นเครื่องมือการสอนสำหรับ:

- การสอนหลักไวยากรณ์ไทย
- การสอนภาษาศาสตร์เชิงวิเคราะห์ การวิเคราะห์เพื่อการสอนบนพื้นฐานของคลังคำ (Comprehensive Of Corpus-Based Instruction) ซึ่งเป็นวิธีที่นำภาษาที่ใช้งานจริงมาใช้ในการสอน ซึ่งพบว่ามีส่วนช่วยยกประสิทธิภาพประสิทธิภาพการเรียนภาษาได้สูงขึ้น [15]
- การเปรียบเทียบโครงสร้างภาษาต่าง ๆ

ผู้เรียนสามารถมองเห็นภาพรวมของโครงสร้างประโยค และเข้าใจบทบาทของคำแต่ละชนิดได้จากต้นไม้ไวยากรณ์ ซึ่งดีกว่าการอธิบายด้วยข้อความล้วน นอกจากนี้ยังช่วยให้เข้าใจการใช้ภาษาที่ถูกต้องตามบริบท และลดการสับสนในการใช้คำที่มีหลายบทบาทในภาษาไทย

5. บทสรุป

คลังต้นไม้ไวยากรณ์ภาษาไทยเป็นทรัพยากรภาษาศาสตร์ที่มีความสำคัญอย่างยิ่งต่อการวิจัยและพัฒนาเทคโนโลยีภาษาธรรมชาติในบริบทของภาษาไทย บทความนี้ได้เสนอภาพรวมของแนวคิดพื้นฐานเกี่ยวกับคลังต้นไม้ โดยชี้ให้เห็นถึงความหมาย องค์ประกอบ โครงสร้าง และมาตรฐานที่จำเป็นต้องใช้ในการสร้างคลังต้นไม้ที่มีความสอดคล้องและสามารถนำไปใช้ในทางปฏิบัติได้

การสร้างคลังต้นไม้ไวยากรณ์สำหรับภาษาไทยนั้นมีความท้าทาย เนื่องจากลักษณะเฉพาะทางไวยากรณ์ของภาษาไทย เช่น การละประธาน การไม่มีการผันคำตามกาล หรือพหูพจน์ และการใช้ลักษณนาม ซึ่งต้องอาศัยทฤษฎีวากยสัมพันธ์ เช่น ทฤษฎี X-bar ในการจัดโครงสร้างวลี และการกำหนดมาตรฐานที่เหมาะสมกับบริบทของภาษาไทยโดยเฉพาะ

คลังต้นไม้ ที่ได้รับการจัดทำอย่างเป็นระบบมีบทบาทสำคัญต่อการพัฒนาระบบวิเคราะห์ไวยากรณ์อัตโนมัติ ระบบแปลภาษา การทำความเข้าใจภาษา การสกัดคุณลักษณะทางวากยสัมพันธ์ ตลอดจนการศึกษาและการสอนภาษาไทย โดยเฉพาะในยุคที่การประมวลผลภาษาด้วยปัญญาประดิษฐ์ คลังต้นไม้ต้องการข้อมูลเชิงโครงสร้างที่มีคุณภาพสูง

ท้ายที่สุดคลังต้นไม้ไวยากรณ์ภาษาไทยไม่เพียงแต่เป็นทรัพยากรเชิงเทคนิค หากยังเป็นฐานข้อมูลที่มีคุณค่าในเชิงภาษาศาสตร์ วัฒนธรรม และการอนุรักษ์ภาษา ซึ่งสามารถนำไปต่อยอดในงานวิจัยและนวัตกรรมด้านภาษาศาสตร์คอมพิวเตอร์ของไทยได้หลากหลาย

เอกสารอ้างอิง

[1] T. Ruangrajitpakorn, K. Trakultaweekoon, and T. Supnithi, "A syntactic resource for Thai: CG treebank," in *Proc. of the 7th Workshop on Asian Language Resources,* pp. 96-101, 2009. (in Thai)

[2] S. Sudprasert, "A Dependency Tree Annotation Manual for Thai Language (Version 1.4)," 2008. [Online]. Available: http://github.com/crishoj/thcg. [Accessed: May 30, 2024]. (in Thai)

[3] M. P. Marcus, B. Santorini, and M. A. Marcinkiewicz, "Building a large annotated corpus of English: The Penn Treebank," *Computational Linguistics*, vol. 19, no. 2, pp. 313–330, 1993.

[4] J. Nivre, M.-C. de Marneffe, F. Ginter, Y. Goldberg, J. Hajič, C. D. Manning, et al., "Universal Dependencies v1: A multilingual treebank collection," in *Proc. of the 10th International Conference on Language Resources and Evaluation (LREC)*, 2016.

[5] K. Kosawat, M. Boriboon, T. Charoenporn, and V. Sornlertlamvanich, "The Thai National Corpus (TNC): Corpus-based linguistic resources for Thai language processing," in *Proc. of the 7th Workshop on Asian Language Resources*, 2009.

[6] H. Isahara, C. Kruengkrai, and S. Shirai, "Thai Treebank and applications," in *Proc. 6th Workshop on Asian Language Resources (ALR)*, Hyderabad, India, pp. 65–72, 2008.

[7] P. Boonkwan, N. Thanachart, and T. Charoenporn, "Thai Dependency Treebank: Annotation guideline and corpus," in *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pp. 1234-1240, 2016.

[8] T. Aroonmanakun, "Issues in tagging and parsing the Thai language," in *Proc. 17th Pacific Asia Conf. on Language, Information and Computation (PACLIC 17)*, Sentosa, Singapore, pp. 219–226, 2003.

[9] C. Wirote and V. Sornlertlamvanich, "Thai grammar extraction using statistical and rule-based approach," in *Proc. 4th Int. Conf. on Language Resources and Evaluation (LREC'04)*, Lisbon, Portugal, 2004.

[10] NECTEC, "Thai Treebank Project: Guidelines and Corpus Development," 2010. [Online]. Available: http://www.thaicorpora.net. [Accessed: Jun. 9, 2023]. (in Thai)

[11] K. M. K. Boriboon, K. Kriengket, P. Chootrakool, S. Phaholphinyo, S. Purodakananda, T. Thanakulwarapas, and K. Kosawat, "Best corpus development and analysis," in *Proc. 2009 Int. Conf. on Asian Language Processing*, pp. 322–327, 2009.

[12] J. Nivre, M.-C. de Marneffe, F. Ginter, J. Hajič, C. D. Manning, S. Pyysalo, S. Schuster, F. Tyers, and D. Zeman, "Universal Dependencies v2: An evergrowing multilingual treebank collection," *arXiv preprint* arXiv:2004.10643, 2020.

[13] S. Sornlertlamvanich, K. Charoenporn, and T. Aroonmanakun, "A deep syntactic parsing approach for Thai using Universal Dependencies," in *Proc. 34th Pacific Asia Conf. on Language, Information and Computation (PACLIC 34)*, 2020.

[14] A. Piamsa-Nga, "Improving Thai-English machine translation via syntactic reordering based on treebank," *Kasetsart Journal of Social Sciences*, vol. 39, no. 2, pp. 235–244, 2018. (in Thai)

[15] D. Li, N. Noordin, L. Ismail, and D. Cao, "A systematic review of corpus-based instruction in EFL classroom," *Heliyon*, vol. 11, no. 2, pp. 1–14, 2025.