Thai UD Treebank

Siriluck Rattananiyomkul

MoDyCo – CNRS UMR 7114, Université Paris Nanterre

Date: 8-12 July 2024

Abstract

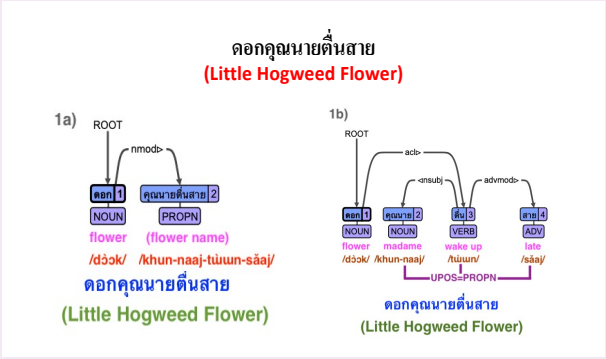
Processing Thai language linguistically and computationally has been very challenging because of the linguistic characteristics and writing system of Thai. While developing a Thai dependency treebank, such as UD Thai-PUD, the number of attempts to resolve such challenges has not been quite successfully done. One of the main reasons is that there are no Thai-specific guidelines of the relevant annotation layers. Through the Thai analysis influenced by some other languages, annotating Thai corpora for a Thai treebank might be inconsistent, and might not present the actual structures of Thai. The purpose of the project therefore is to develop a new Thai UD treebank with the Thai-specific guidelines of the relevant annotation layers: word segmentation, sentence segmentation, and POS tagging to enable consistent annotation, to analyze the syntactic structures of Thai through the UD as well as SUD frameworks and to present the actual structures of Thai without the analysis influenced by some Indo-European grammar.

Thai Language

- SVO and isolating language with 4 tone markers but 5 tone sounds
- No grammatic markers
- Postmodifiers mostly used to express additional information
- Cardinal numbers and quantifiers placed before a modified noun
- Thai words – very fluid and going through grammaticalization
- Compounding mostly used to coin new words
- A very long list of classifiers used to classify nouns
- Context – a crucial clue of syntactic and semantic information
- Language use and style – relying on the hierarchical society and cultural aspects

Word Segmentation

- Due to the way of how to write a Thai text, it is difficult to decide whether to segment each word separately or combine them as one single word.



Sentence Segmentation

- Due to the way of how to write a Thai text, it is difficult to find sentence boundaries even with some whitespace between phrases and clauses used in a text.

2a)

ชายเลี้ยงวัวลุกไปล้างมือในลำห้วยแล้วกลับมานั่งที่เก่า ล้วงหาขวดเหล้าในย่ามที่เหลือเพียงก้นขวดออกมากระดกดื่ม แล้วลงมือแกะห่อข้าว ปั้นข้าวก้อนหนึ่งวางบนใบไม้ที่เด็ดไว้เมื่อไปล้างมือ ใช้ปลายนิ้วจกน้ำพริกปลากระป๋องวางลงบนข้าว นำไปวางไว้บนรากมะตาดที่โผล่พ้นดินเพราะน้ำเซาะในฤดูน้ำหลากให้เจ้าที่ เขากลับมานั่งอีกครั้ง ลงมือจกน้ำพริกขยำกับข้าว จกเข้าปากแล้วกินผักถูดตาม

Plans for Thai UD Treebank

- Collecting around 5,000 Thai sentences from different genres of written Thai texts to build a Thai corpus for a Thai UD treebank
- Manually and/or automatically annotating the corpus with the UD framework, including the SUD framework
- Also contributing Thai annotations to the treebank of UD Thai-TDT; correcting the Thai corpus of UD Thai-PUD by following the guidelines to be developed

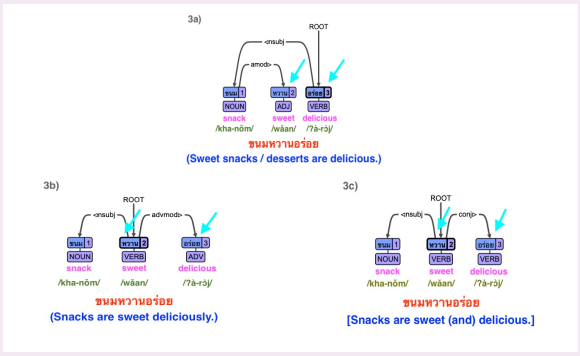
Sample of Written Thai Text

- Most words not delimited
- No capital letters in Thai; a very little of Thai and standard punctuation used
- Whitespace used between clauses and phrases; double whitespace used to end a sentence

ดอกคุณนายตื่นสายริมทางเหลือแต่เครือเถาเหี่ยวแห้งระเลื้อยดิน ถ้าปล่อยทิ้งไว้คงจะแห้งตายไปเอง แต่ที่ร้านข้าวมันไก่ สองตายายจะลงแปลงปลูกใหม่ทุกครั้ง เราสองคนไม่ได้แวะไปนาน ต่อมรับรสตื่นเต้นยินดีรอท่า เราขับมอเตอร์ไซค์ฝ่าเปลวแดดไปด้วยความหวัง แต่พอไปถึงได้แต่ตะลึงงัน ไม่มีร้านข้าวมันไก่ ไม่มีแปลงดอกคุณนายตื่นสาย ลานจอดรถยาวถึงริมถนน ซูเปอร์มาร์เก็ตตั้งตระหง่านแทนที่ อย่างกับมายิดที่ ทุกอย่างใหม่หมดจดไม่เหลือเค้าเดิม ไม่รู้ไปอยู่ไหน ย้ายไปนานเท่าไหร่แล้ว หรือจะเจ็บป่วยไม่สบายหรือจะ... ขออย่าให้เป็นอย่างนั้น ความกลัวมักพาความคิดเตลิดไกล ชายคนนั้นคงไม่ได้ยืนมือรับถุงห่อข้าวมันไก่ครึ่งจานพร้อมน้ำซุปใครไก่ถุงใหญ่อีกแล้ว ท้องฉันร้องไครกครากตัดพ้อด้วยความผิดหวัง ความกลัว กังวลในสิ่งไม่อาจรู้เจือในน้ำลายคล้ายขม มันหนืดเหนียวอยู่เต็มปากเต็มลำคอ หน้ามืดตาลายจนเห็นประเทศรูปขวานกลายเป็นกระเพาะใหญ่ยักษ์ที่กำลังเบียดบด ดูดกลืน ขย้อนเราออกมาเคี้ยวแล้วกลืนซ้ำ เคี้ยวกลืน ซ้ำแล้วซ้ำเล่าไม่หยุดหย่อน ความหิวซุกตัวอยู่ในกระเพาะของคนที่กำลังเดินผ่าน คนที่ไม่รู้จัก คนที่ไม่ถูกมองเห็น นมผง บะหมี่สำเร็จรูป ขนมปัง ปลากระป๋อง ข้าวสาร กลายเป็นหลักฐานของกลางวางแผ่ตรงหน้าผู้ต้องหาลักขโมย เจ้าของไหหน้านั้นจะเป็นใครก็ได้ โชว์หราในหน้าข่าวรายวัน ฉันนึกถึงความลำบากยากแค้นที่ยายเคยเล่า สมัยนั้นคนเราลำบากยากจนกันมาก... แต่มันก็ยังไม่เท่าสมัยนี้เลยนะยาย ถ้าหวาน ประชาชนอย่างเราก็จะอมไว้ลิ้มรสซ่านลิ้นให้นานที่สุด แต่ถ้าขม ไม่ว่าจะสักแค่ไหนเราก็จะฝืนกลืนมันลงไปเหมือนที่ผ่านๆ มา

POS Tagging

- Due to Thai words being very fluid, it is difficult to categorize Thai words into each part of speech.



Promising Procedures and Results

- Instead of following the conventional sequence of annotation, starting firstly with sentence segmentation by depending mainly on the context of a text to gather what is discussed in the same topic
- And later looking for linguistic clues, such as topic changing, sentence subjects, proper nouns, personal pronouns, conjunctions, discourse markers, final particles, etc., to find a sentence boundary
- Segmenting the words of each segmented sentence manually and/or automatically with the Thai Tokenization tool, https://huggingface.co/spaces/pythainlp/newmm_online by PyThainlp
- POS tagging with UPOS by analyzing the true structures of Thai with the framework of the syntactic distribution
- Syntactically annotating the tagged words and segmented sentences with the UD framework, including the SUD one
- To develop the Thai-specific guidelines out of these annotation layers to perform better consistent annotations; and hopefully to introduce the Thai-specific guidelines to any Thai annotations and/or studies