New UD Treebank for Thai

Siriluck Rattananiyomkul, a doctoral student at Université Paris Nanterre, France

Thai, UD, syntactic treebank, annotation

This project started from attempting to annotate a Thai corpus for my PhD thesis titled *Syntactically Categorizing Word Classes in Thai through the Development of the Thai Dependency Treebank* with the implementation of the UD guidelines, but it was very difficult to annotate it explicitly and consistently. There have been several challenging issues arising while dealing with each annotation layer for Thai.

The first annotation layer of Thai which must be done correctly, explicitly, and consistently to enable other further layers to be done in the same way is word segmentation. This annotation layer is one of the most challenging tasks because most words in written Thai texts are not delimited as like those of Western languages. Punctuation is hardly used in writing as well. The way Thai words are segmented thus leads to the way they are tagged and the way they are syntactically annotated, including the way a sentence is segmented. The major issue of word segmentation in Thai is to decide whether to segment each single word separately or combine it with a preceding or following one. What is decided differentiates POS tagging and meaning for the segmented words. When the meaning of a few words is changed, the sentence meaning can be changed. Likewise, when a few parts of speech are changed, the sentence structure is also changed. All the changes could also impact on a sentence boundary which is another challenge of Thai. The linguistic characteristics of Thai also cause difficulty in analyzing and processing the language linguistically and computationally. Thai is an isolating and tone language with no grammatical markers, so their words are in base and nonfinite form. In addition, Thai words are very fluid, and some of them go through grammaticalization (Iwasaki et al., 2005). Word classes in Thai are then very difficult to be categorized. Compounding is mainly used to coin a new word with specific meaning as well.

All the difficulties of Thai can be resolved by context which plays a major role of enabling better understanding the syntactic and semantic information, so the decisions on the annotations cannot be done freely. This means that at least a few native Thai speakers are required to process all the annotation layers: word and sentence segmentations, POS tagging, and syntactic annotation; each could have different readings and interpretations. The annotations thus tend to be rather inconsistent. To maintain each annotation layer of Thai among the annotators consistently and explicitly, the guidelines should be developed specifically with the linguistic aspects of Thai as well as for each annotation layer. Currently, Thai has only the PUD treebank whose 1000 Thai sentences were tested and annotated with no Thai-specific guidelines and no maintainers. All the mentioned issues have made me decide to develop a new UD treebank for Thai with the Thai-specific guidelines of the annotation layers: word segmentation, sentence segmentation, and POS tagging in order to analyze the syntactic structures of Thai through the UD framework and to present the actual structures of Thai without the analysis influenced by some Indo-European grammar. And since the PUD treebank of Thai has been created without the guidelines, consistency, and maintainers, I decided to willingly correct the inconsistently segmented words and tagged

parts of speech in Thai with the same guidelines to be developed in order that both Thai PUD and new UD treebank of Thai can have the same segments and parts of speech.

The project has been started since last year. I have planned to build a Thai corpus with at least 5,000 sentences to be annotated. To have different syntactic structures of Thai, the corpus is composed of a variety of written Thai texts in different genres, such as political news, short stories, IT articles, lifestyle articles, legal documents, Thai articles written with Royal terms, etc. Each text is firstly segmented with the automatic tool (Thai Tokenization), https://huggingface.co/spaces/pythainlp/newmm_online, developed by Pythainlp to help speed up the process. After that, I manually check and correct each segment as necessary by mainly considering context as well as following the guideline of word segmentation which is developed alongside building this new treebank. The next step is to segment each sentence in the texts by counting on context and some signal words, including the guideline to be created. And the process of Thai part-of-speech categorization is implemented mainly with the syntactic distribution (co-occurrence) which Prasithrathsint (201) discusses as the best criterion fitting in with the natural characteristics of Thai. After all these 3 annotation layers completely done, each sentence is going to be syntactically annotated with the UD guidelines and UD Annotatrix, the browser-only annotation tool.

As mentioned, several challenging and unclear issues have arisen while annotating the written Thai texts with the implementation of the UD guidelines, and they have not yet been resolved:

- Would it be possible to develop the guidelines of word segmentation for Thai which would not rely mainly on the native Thai speakers' context readings, and still be consistent and explicit?
- Categorizing word classes in Thai is very difficult because of the natural characteristics of Thai words – being very fluid. Some word classes in Thai, such as adjectives, are problematic and very challenging that there seem to be no ways to do. Would it be possible to develop some explicit rules of POS categorizing through the syntactic distribution to solve the issue?
- Would there be some consistent and precise ways to set a sentence boundary for a language, like Thai, with a very little of word and sentence delimiting and punctuation using?
- There are several grammatic aspects of Thai which still cannot be annotated with the UD guidelines clearly, such as proper nouns in Thai, nominalizations in Thai, classifiers in Thai, serial verb construction, passive voice, a sentence without a verb, etc.

I decided to apply for the first UniDive training school of 2024 to brainstorm about all the unclear issues with people who have had more experiences, to hope for the answers, solutions, and/or ideas for such issues, and also to learn how to use the tools offered in this Summer School for my project and PhD thesis.

References

Amara Prasithrathsint. 2010. *Parts of Speech in Thai: A Syntactic Analysis*. A.S.P. Publishers, Bangkok.

Shoichi Iwasaki and Preeya Ingkaphirom. 2005. *A Reference Grammar of Thai*. Cambridge University Press, New York.