A Lexicalized Tree Adjoining Grammar for Thai *

Siripong Potisuk

Department of Electrical and Computer Engineering
The Citadel, the Military College of South Carolina
171 Moultrie Street
Charleston, South Carolina 29409 USA
siripong.potisuk@citadel.edu

Abstract. This paper describes an alternative formalism for Thai syntax parsing based on a lexicalized tree adjoining grammar (LTAG). We first briefly present some formal background concerning LTAG, which is necessary for an understanding of LTAG and its application to Thai. Specifically, we address several issues regarding difficulties in parsing Thai sentences and how to resolve these issues using LTAG. Such difficulties arise for several reasons as follows. For one thing, Thai sentences do not contain delimiters or blanks between words while Thai words lack inflectional and derivational affixes. Moreover, inconsistent ordering relations within and across phrasal categories characterize Thai sentences as well as the fact that they sometimes contain discontinuous sentence constituents in their construction. Finally, we discuss future research on a novel almost-parsing approach based on LTAG for handling compound multi-word extraction in automatic Thai word segmentation.

Keywords: Thai-LTAG, Thai syntax parsing

1 Introduction

Research on the syntactic analysis of Thai sentences by computer has been carried out for several decades. Thai grammars have been developed utilizing various grammar formalisms based on the two theories of syntax, namely dependency grammar and phrase-structure (constituency) grammar or their combination. Vorasucha (1986) was one of the first researchers to use Gazdar's Generalized Phrase-Structure Grammar (GPSG) in his research. Immediate Dominance (ID) rules and Linear Precedence (LP) rules were used. Pornprasertsakil et al. (1990) later employed additional Feature Specification Default (FSD) constraints of GPSG to describe three types of sentence structures including verb and noun phrases. Aroonmanakun (1990) developed a nondeterministic parser called CUPARSE based on a dependency representation of syntax. The parser uses a chart as its central data structure. As part of a project on machine translation of Asian languages, Sornlertlamvanich and Phantachart (1992) used a combination of phrase-structure and dependency grammars. Phrase-structure grammar rules were used to identify locally well-formed phrase patterns, and thus reduce lexical ambiguities, based on the relatively fixed relation of the positions of Thai words and their syntactic roles. Then, syntactic dependency structures among the words were generated based on verb subcategorization information. Finally, the syntactic dependency structure was mapped to a semantic one by utilizing a lexical functional grammar. Wuwongse and Pornprasertsakul (1993) introduced a probabilistic approach using a least exception logic (LEL) model of default reasoning to resolve ambiguities. Potisuk and Harper (1996) proposed an alternative formalism

* The author would like to thank the Citadel Foundation for its financial support in the form of a presentation grant.

called Constraint Dependency Grammar (CDG). CDG parsers rule out ungrammatical sentences by propagating constraints developed based on a dependency-based representation of syntax. CDG is capable of efficiently analyzing free-order languages because order between constituents is not a requirement of the grammatical formalism. Since Thai exhibits significant word order variation, using CFG to describe Thai is cumbersome because numerous rules would be needed to cover all possible configurations of a constituent. Secondly, The CDG approach provides a uniform mechanism of constraint propagation for each knowledge source, i.e., lexical, syntactic, semantic, and pragmatic information, in resolving ambiguities during parsing. The constraints for each knowledge source can be independently developed and applied. A CFG parser, on the other hand, does not provide a good coordinating scheme because it is incapable of selectively invoking different knowledge sources.

In this paper, we propose an alternative formalism based on Lexicalized Tree Adjoining Grammar (LTAG) for Thai syntax parsing. Before describing our implementation of LTAG for Thai, we present a brief introduction to Thai morphology and syntax in the next section and argue why LTAG appears to be an attractive choice for Thai grammar development.

2 An Overview of Standard Thai

Standard Thai, the focus of this research, is the official language of Thailand and used in conducting official business and legal matters. It is also the medium of instruction in government schools throughout the country. Since a formal linguistic analysis of standard Thai is not a primary goal of this research, this overview is by no means intended as an exhaustive linguistic description of standard Thai. In this section, we present an overview of standard Thai concerning its morphology and syntax. We also outline several issues regarding difficulties in parsing Thai sentences. The overview is given to provide a linguistic framework for subsequent description of Thai grammar development.

2.1 Morphology

A continuous stream of text can be broken up by readers into smaller and meaningful parts- a discourse into sentences, sentences into words, and words into morphemes. A morpheme is the smallest linguistic unit which has a meaning or grammatical function. Lexical morphemes have meaning in and of themselves; grammatical morphemes specify the relationship between two lexical morphemes. Free morphemes can stand alone; bound morphemes cannot.

Although Thai is commonly thought to be a monosyllabic language like Chinese, many words are disyllabic and trisyllabic. Most disyllabic words are names of plants and animals. Others are Khmer-related. The following describes the word formation process in Thai.

Affixation

The first process is the affixation. Affixes (i.e., prefixes and suffixes) are bound grammatical morphemes which are attached to other lexical morphemes (i.e., roots) to compose words. There are two types of affixes: inflectional and derivational affixes. Inflectional affixes are grammatical devices used to indicate syntactic or semantic relations between different words, i.e., the concepts of plurality, possessive case, degree of comparison, or tense. Derivational affixes are used to indicate semantic relations within a word, but not syntactic relations outside the word. They also indicate a change in meaning or syntactic category (i.e., part of speech) for the roots to which they are attached (e.g., un-).

With the exception of loanwords, affixes are not part of the morphological structure of Thai. Neither inflectional nor derivational affixes are used to compose words. In other words, Thai is considered an isolating or analytical language in which sentences are constructed from sequences of free morphemes; each word consists of a single morpheme, used by itself with its meaning intact. Semantic and grammatical concepts are expressed through the use of free morphemes rather than the use of a change of form or affix. For example, unlike English which

indicates plurality through a change of form (e.g., I to We) or tense through the use of affix (e.g., -ed for past tense), Thai uses additional words.

Nevertheless, Thai does make limited use of affixation. Most affixes are of Khmer or Indic (Pali and Sanskrit) origin. This process accounts for restricted types of nominalization, semantic extension, intensification, and stylistic marking. Adjectival or stative nominalization are formed with the prefix 'khwaam-'; action-verb nominalizations are formed with the prefix 'kaan-' (meaning activity). Some verbs admit either type of nominalization. Specialized nominalizations to form agent nouns, often occupational, are formed with the prefix 'nak-'. In addition, Khmer-derived nasal infixes, e.g. –am(n)-, are a common form of derivation.

Compounding

Compounding is a process which forms new words from two or more independent words, a mere juxtaposition or a collocation of two or more words with no alteration in the consonants and/or vowels. The component words of the compound can be free morphemes, words derived from affixation, or even words of the compound can be free morphemes, words derived from affixation, or even words formed by compounding. Compounding in Thai, regularly of the type head + modifier, is widespread and usually serves a function similar to affixation. Compounds can be distinguished from syntactic phrases by differences in stress patterns. The stress placement is determined by the morphological derivation of the compounds. For example, the first syllable of a 2-syllable compound is always unstressed. For written language, this distinction is nonexistent.

Reduplication

Reduplication is process of forming new words either by doubling an entire free morpheme (total reduplication) or part of it (partial reduplication). Thai makes extensive use of reduplication. Total reduplication is possible for most adverbs of manner and for adjectivals, and is used to intensify or express command. Collective pluralization of a restricted set of nouns denoting humans is derived through total reduplication. The reduplicative extension with alternation of vowels and sometimes tones and consonants is a style-sensitive means of intensification and semantic elaboration. Also, colloquial intensification of adjectivals involves reduplication, with tone shifted to high on the first reduplicated syllable. For action verbs, reduplication may iconically encode a durative or repetitive aspect like progressive in English.

2.2 Syntax

Syntax describes sentence structure or the principles by which words are organized into a sentence in a language. Syntacticians classify Thai as having a subject-verb-object (SVO) pattern as the most favored word order. Thus, Thai is commonly referred to as an SVO language. However, other common orders frequently appear, especially in colloquial or informal conversation. The following paragraphs address lexical categories and subcategorization, as well as basic word order.

Lexical Categories and Subcategorization

In any given language, a large number of words often exhibit the same properties, which suggests that they can be grouped together into classes called lexical categories. Word categorization is usually based on different criteria (or their combination): morphological, syntactic, or semantic properties. For example, according to Aroonmanakul (1990), lexical categorization of Thai words consists of five major categories based on semantic properties of words: Noun(N), Verb(V), Determiner (DET), Auxiliary (AUX), and Relator (REL). Each major category is further refined or subcategorized using both syntactic and semantic criteria. Subcategorization represents a finer distinction within each major category. A noun is subcategorized into cardinal noun, proper noun, pronoun, common noun, and classifier. A verb

is either an active, stative, existential, or equative verb. Ther are left and right determiners depending on the position of the noun it modifies. Subcategories for auxiliaries include left modal, left irrealis, left aspect, left attribute, right attribute, and right aspect auxiliary. A relator can be a preposition, complementizer, or connector. In addition to these five major categories, there is a special class of words called particles. Their primary function is to end utterances. There are three groups of particles. One group marks a statement and forms yes-no question; the second shows respect toward the addressee; and the third indicates a speaker's mood.

Basic Word Order

As mentioned above, the simplest sentence construction is subject-verb-object. However, there are other constructions, such as simple comparative construction, complementation, embedding, topicalization, yes/no questions, and content (wh-) questions. Inversion of word order does not occur except in the case of topicalization. This topic-initial word order is object-subject-verb (OSV), where O is topical or thematic. Another characteristic feature of Thai is the obligatory use of classfier when quantifier is present with noun. The most usual order is noun+quantifier+classifier. The head noun determines the choice of classfier. This construction becomes discontinuous when new information quantifiers are present, and the noun is separated from the quantifier and the classifier. A more controversial construction is verb serialization where verbs occur in series with or without and explicit intervening conjunction. Verb order may represent the progress of a complex event or a passive construction like English.

2.3 Difficulties in Parsing Thai Sentences

Despite increasing concerted efforts among Thai universities and government agencies, a satisfactory approach to analyzing Thai sentences has not been obtained. Difficulties in parsing Thai sentences arise for several reasons.

First, written Thai sentences do not contain delimiters or blanks between words. Unlike English, Thai words are not flanked by a blank space. Words are concatenated to form a phrase or sentence without explicit word delimiters. This creates a problem for the syntactic analysis of Thai sentences because most parsers operate on words as the smallest syntactic unit in a sentence. To overcome this problem, a word segmentation module must be added to the front end of most Thai parsers. It may seem, on the surface, that the problem has been solved. But, on the contrary, a new problem has been created. Instead of analyzing a single sentence, a parser must now analyze multiple sentence hypotheses comprising a combination of all possible words generated by the word segmentation process. For example, given the following string of Thai characters, ภาพรอออกฉาก, two possible sentence hypotheses are generated given dictionary lookup.

a. ภา — พร — ออ — อก — ฉาก

b. ภาพ — รอ — ออก — ฉาก

Secondly, Thai words lack inflectional and derivational affixes. Since words in Thai do not inflect to indicate their syntactic function, the position of a word in a sentence alone often shows its syntactic function. Hence syntactic relationships are primarily determined by word order, and structural ambiguities often arise. For example, without a subject-verb agreement feature or disambiguating context, there is no way of differentiating a 2-syllable noun-verb sequence from a 2-syllable compound noun in the same sequence of words. The following example illustrates the problem.

Compound: สมชายเป็นคนเจ้าชู้ เขามี **<u>คนรัก</u>** มาก
'Somchai is a flirting kind. He has a lot of girlfriends.'

Sentence: สมชายเป็นดารานิสัยดี เขามี **<u>คนรัก</u>** มาก
'Somchai is not a stuck-up movie star. Many fans love him.'

Thirdly, inconsistent ordering relations within and across phrasal categories characterize Thai sentences. Based on word order typology, the majority of the world's languages usually exhibit consistent ordering relations across phrasal categories. The head (i.e., the central, obligatory member) of the phrase is usually placed either consistently before its modifiers and complements as in the so-called head-initial languages or after its modifiers and complements as in head-final languages. Such classification of a particular language is made according to the majority of its ordering relations within phrases, and thus, the distinction is based on tendencies, not exclusivity. In Thai, a noun, the head of a noun phrase, always precedes its modifying adjectives and determiners. The verb phrase, however, exhibits less consistency. Although a verb, the head of the verb phrase, always precedes its object, its modifying auxiliaries can either precede or follow it. In addition, constituents which optionally occur with the head in both noun and verb phrases, such as determiners and quantifiers, tend to be less consistent in their ordering as well. The following example contains three syntactically correct versions of the same sentence, '*He often invites his friends to have dinner at his house*', each with a different ordering of constituents.

เขาชวนเพื่อน ๆ มากิน <u>กัน</u> **ที่บ้าน** *บ่อย ๆ*
เขาชวนเพื่อน ๆ มากิน <u>กัน</u> *บ่อย ๆ* **ที่บ้าน**
เขาชวนเพื่อน ๆ มากิน **ที่บ้าน** <u>กัน</u> *บ่อย ๆ*

Lastly, Thai sentences sometimes contain discontinuous sentence constituents in their construction. In grammatical analysis, discontinuity refers to the splitting of a construction by the insertion of another grammatical unit. In other words, discontinuity occurs when the elements which make up the constituents are interrupted by elements of another constituent in a sentence. Consider the following example in which the object noun phrase (NP-obj) is interrupted by the auxiliary (aux.).

เพื่อน	ยืม	หนังสือ	ไป	สอง	เล่ม
friend	borrowed	book	(aux.)	two	(classifier)

| NP-subj. | V. | ---------- NP-obj ---------- |

3 A Description of LTAG Parsing Formalism for Thai

In this section, we introduce a different approach to Thai syntactic parsing based on a Lexicalized Tree Adjoining Grammar (LTAG). We will first briefly present some formal background concerning LTAG, which is necessary in understanding our implementation for Thai as well as the rest of the paper.

3.1 An Overview of LTAG Formalism

Lexicalized Tree Adjoining Grammar is a tree-rewriting grammar formalism unlike Context-Free Grammars and Head Grammars which are string-rewriting formalisms. LTAG is based on Tree Adjunct Grammar (TAG) formalism originally proposed by Joshi, Levy, and Takahashi (1975) and later extended to include lexicalization (Schabes, et al., 1988) and unification-based structures (Vijay-Shanker and Joshi, 1991).

The primitive elements of the LTAG formalism are elementary trees. Each elementary tree is associated with at least one lexical item on its frontier since one of the important characteristics of LTAG is that it is lexicalized. The lexical item associated with an elementary tree is called the anchor of the tree. An elementary tree serves as a complex description of the anchor and provides a domain of locality over which the anchor can specify syntactic and semantic (predicate-argument) constraints. That is, each lexical item is anchored to a tree structure that encodes subcategorization information. For example, a transitive verb is anchored to a tree that includes a subject NP and an object NP. Elementary trees are of two types: initial trees and auxiliary trees.

Initial trees are minimal linguistic structures that contain no recursion. They include trees containing the phrasal structures of simple sentences, NPs, etc. In initial trees, all internal nodes are labeled by non-terminals and all leaf nodes are labeled by terminals or non-terminal nodes marked for substitution. The symbol (↓) is used to indicate the substitution site.

Auxiliary trees are used in representing recursive structures which are adjuncts to basic structures, such as adverbials, adjectivals, etc. All internal nodes of an auxiliary tree are labeled by non-terminals whereas all leaf nodes are labeled by terminals or non-terminals nodes marked for substitution, except for one non-terminal node marked as foot node. The foot node is marked by a symbol (*) and must have the same label as the root node of the tree.

The LTAG formalism defines two operations for combining elementary trees: substitution and adjunction operations. The substitution operation inserts an elementary tree into another elementary tree at the substitution node. The root label of the tree being inserted must match the label at the substitution node of the inserted tree. In an adjunction operation, an auxiliary tree is inserted into an initial tree. The label of the root and foot node of the auxiliary tree must match that of the node of the initial tree where the auxiliary tree is to adjoin. The resulting initial tree is split at the adjoined node by the auxiliary tree.

Another characteristic of LTAG is that it is feature-based. A lexical or grammatical feature structure is associated with each node in a tree. This feature structure contains information about how the node interacts with other nodes in the tree. For example, by using the case feature, we can constrain a transitive verb to anchor to a tree that takes a subject NP with nominative case marker and an object NP with accusative case marker. Each node of an elementary tree is associated with two feature structures (FS), the top and bottom. The bottom FS contains information relating to the sub-tree rooted at the node while the top FS contains information relating to the super-tree at that node. Features may get their values from three different sources: 1) the morphological information of the lexical items that anchor the tree: 2) the structure of the tree itself; and 3) the derivation process of unifying features from trees that adjoin or substitute. Parsing of a sentence is successful only when both feature structures in each node of the tree unify. The result of combining the elementary trees is called the derived tree. In addition, there is also a derivation tree which is the tree that records the history of composition of the elementary trees associated with the lexical items in the sentence. This derivation tree can be interpreted as a dependency tree with unlabeled arcs between words of the sentence. This implies that elementary trees combine both phrase-structure information and dependency information in a single representation.

3.2 LTAG Implementation for Thai

According to Joshi et al. (2002), large scale grammars have been constructed by hand for English at University of Pennsylvania, USA (XTAG Group 2001) and French (at the TALANA group, University of Paris 7, France) and somewhat smaller ones for German (at DFKI, Saarbrücken, Germany), Korean and Chinese (at University of Pennsylvania, USA), and Hindi (at CDAC, Pune, India). To the best of our knowledge, LTAG has never been applied to Thai. Our decision to try to apply LTAG to Thai syntax parsing can be attributed to several considerations. For one thing, the structure of Thai (i.e., its grammatical system) has been extensively studied by Thai linguists, and there is enough evidence pointing to the fact that phrase-structure grammar is well suited for describing Thai. Although there are many formalism based on phrase-structure theory, such as CFG, we opted for LTAG because of its generative capabilities and its key properties including lexicalization, extended domain of locality and factoring of recursion from Domain of Dependency. This is very important for Thai because Thai words lack inflectional and derivation affixes as mentioned above. The position of the word in the sentence along with its lexical category (possibly its subcategory) determines its syntactic function and how it is related to the rest of the sentence. This suggests that the formalism should be able to capture dependency information among words within the sentence as well. And, as mentioned at the end of the previous section, elementary trees of LTAG can capture both phrase-structure and dependency information at the same time.

Next, we give an example of our LTAG implementation for Thai. In particular, we provide elementary trees based on LTAG parsing of the following Thai sentence:

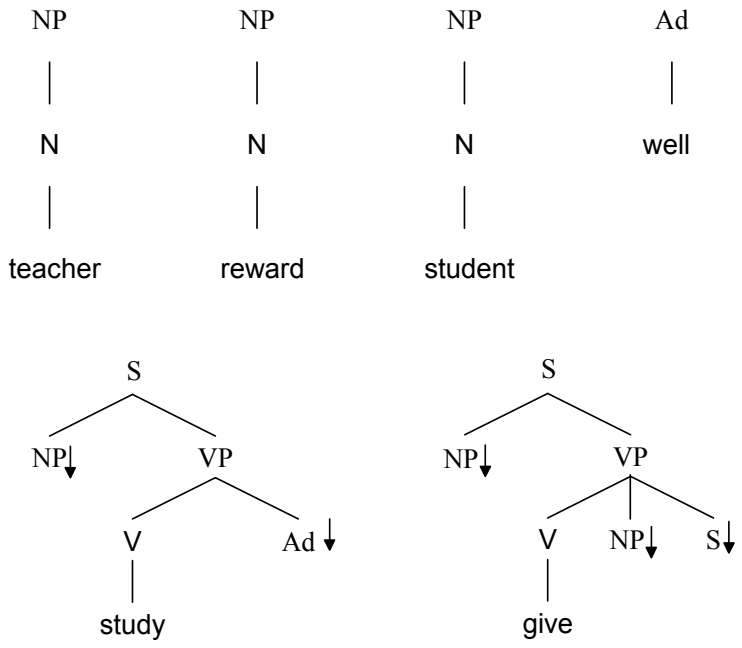
ครูแจกรางวัลนักเรียนเรียนดี (‘The teacher gives rewards to excellent students.’)

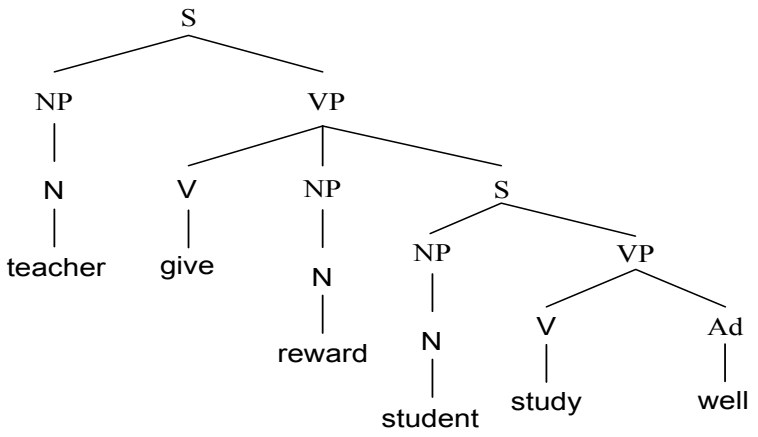
ครู – แจก – รางวัล – นักเรียน – เรียน – ดี

Teacher – give – reward – student – study – well

| NP-subj | V | direct obj | indirect obj (N Clause) |

The elementary tree for each lexical item and the resulting derived tree are shown below.





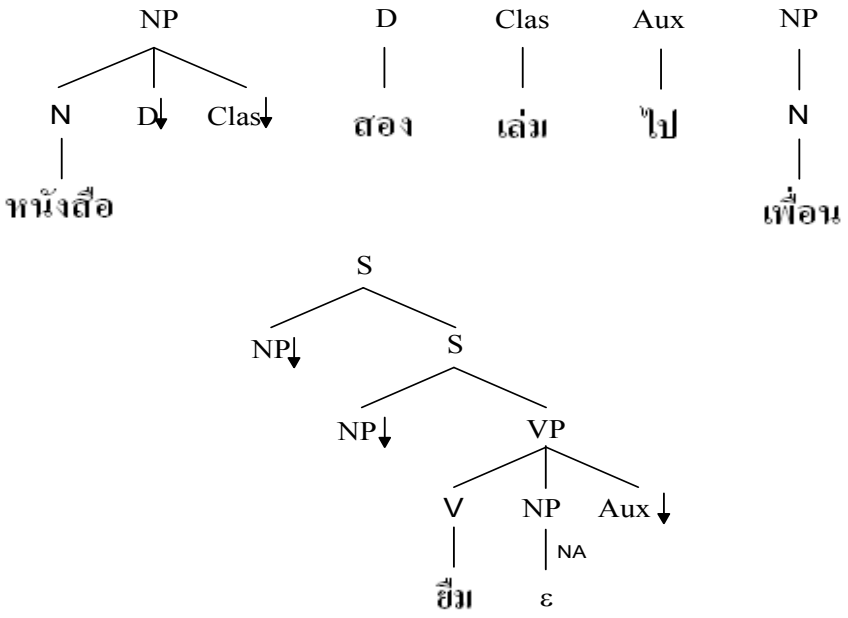
3.3 Handling Discontinuous Constituents in Thai

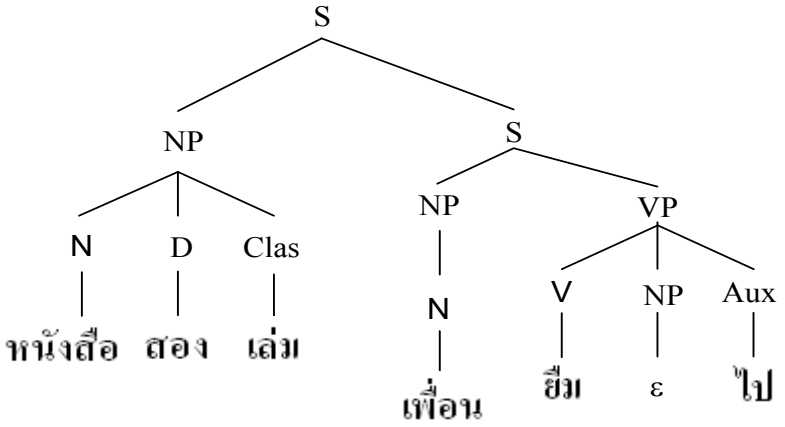
As mentioned in section 2.3 above, one of the difficulties in parsing Thai arises from the problem of Thai sentence construction involving discontinuous constituents. According to Panupong (1989), this kind of sentence construction is nothing more than a special manifestation of Topicalization, where the object noun phrase is topicalized at the beginning of the sentence. As a result, we will use topicalization frame in constructing elementary trees for sentences with discontinuous constituents as shown below. The sentence mentioned in section 2.3 has the following equivalent syntactic representation.

หนังสือสองเล่ม เพื่อน ยืมไป

| ----- NP-obj----- | NP-subj | VP |

The elementary tree for each lexical item and the resulting derived tree are shown below.





4 Discussion and Future Research

From above it can be seen that the elementary trees are designed such that only those elements on which the lexical item imposes constraints appear within a given elementary tree. However, for a large scale grammar, each lexical item is more than likely be associated with as many elementary tree as the number of different syntactic contexts in which the lexical item can appear. As a result, this increases the local ambiguity for a parser in choosing the correct elementary tree for each lexical item. This problem can be exemplified by the compounding process in Thai morphology. For example, consider the following sentences containing two Thai monosyllabic words **ข้อ** and **ต่อ** in different contexts taken from Aroonmanakun (2002).

a. ฟิตเฟล็กซ์เลือกทำธุรกิจเกี่ยวกับ **ข้อต่อ** และสายอ่อนทุกชนิด (compound noun meaning a joint)

b. ที่นายปิ่นกล่าวมาเป็นเพียง**ข้อต่อ**สู้หนึ่งเท่านั้น (compound noun meaning an argument)

c. มอบนโยบาย 5 **ข้อต่อ** ดร.สมบูรณ์ (classifier – preposition 'to')

d. หากคณะกรรมการปล่อยให้ผู้รับเหมาแข็ง**ข้อต่อ**ราคาประมูล (แข็ง**ข้อ** – **ต่อ** → compound verb meaning to defy - preposition 'to')

The compounding process has long been a problem in the design of many Thai word segmentation algorithms. This represents a classic compound multi-word extraction in Thai morphological analysis. This problem arises because Thai words are concatenated to form a phrase or sentence without explicit word delimiters. Many researchers viewed this phenomenon as a word segmentation problem. However, we believe that the problem can be solved at a higher level of analysis, namely the syntactic level. This is because the ambiguities that arise from the compounding process are primarily syntactic ambiguities since each word has an inherit part of speech associated with it.

We believe that LTAG description of Thai syntax can help solve this problem and now are considering a novel almost-parsing approach to compound multi-word extraction in Thai word segmentation. This approach is inspired by the idea of supertagging proposed by Bangalore and Joshi (1999). Due to the scope of this paper, details of the approach cannot be given. However, the method involves the construction of supertags (elementary trees) for different context in which a lexical item can appear. Then, the selection of the appropriate supertag for each lexical item called supertag disambiguation is carried out using statistical distributions of supertag co-occurrences collected from a corpus. Supertag disambiguation results in a representation that is effectively a parse, hence the name 'almost parse'. A final parse can be obtained by combining the individual supertags selected for each lexical item in the sentence. We are in the process of constructing supertags for most, if not all, structures of compound words in Thai.

References

Aroonmanakun, W. 1990. *A Dependency Analysis of Thai Sentences for a Computerized Parsing System*. Master's Thesis, Department of Linguistics, Faculty of Arts, Chulalongkorn University (Thailand).

Aroonmanakun, W. 2002. Collocation and Thai Word Segmentation. *Proceedings of SNLP-Oriental COCOSDA*.

Bangalore, S. and A. K. Joshi. 1999. Supertagging: An Approach to Almost Parsing. *Computational Linguistics*. 25(2): 237-265.

Joshi, A. K., L. Levy, and M. Takahashi. 1975. Tree Adjunct Grammars. *Journal of Computer and System Sciences.*

Panupong, V. 1989. *The Structure of Thai: Grammatical System*. Ramkamhaeng University Press, Bangkok.

Potisuk, S. and M. P. Harper. 1996. CDG: An Alternative Formalism for Parsing Written and Spoken Thai. *Proceedings of the Fourth International Symposium on Languages and Linguistics*, 1177-1196.

Pornprasertsakul, A., V. Wuwongse, and K. Chansaenwilai. 1990. A Thai Generalized Phrase-Structure Grammar and Its Parser. *Proceedings of the International Conference on Computer Processing of Chinese and Oriental Languages,* Hunan, PRC.

Schabes, Y., A. Abeille, and A. K. Joshi. 1988. Parsing Strategies with Lexicalized grammars: Application to Tree Adjoining Grammars. *Proceedings of the 12th International Conference on Computational Linguistics (COLING)*, Budapest, Hungary.

Sornlertlamvanich, V. and W. Phantachart. 1992. Information-based Language Analysis for Thai. *Proceedings of the 3rd International Symposium on Language and Linguistics,* Bangkok, Thailand. 497-511.

Vijay-Shanker, K. and A. K. Joshi. 1991. Unification-Based Tree Adjoining Grammars. *Unification-based Grammars.* MIT Press, Cambridge, Massachusetts.

Vorasucha, V. 1986. Thai Syntax Analysis Based on GPSG. *Regional Symposium on Computer Science and Its Application (with Emphasis on Artificial Intelligence),*KMITL, Bangkok, Thailand.

Wuwongse, V. and A. Pornprasertsakul. 1993. Thai Syntax Parsing. *Proceedings of the Symposium on Natural Language Processing in Thailand,* Chulalongkorn University, Bangkok, Thailand. 446-467.