CDG: AN ALTERNATIVE FORMALISM FOR PARSING WRITTEN AND SPOKEN THAI

Siripong Potisuk

Department of Electrical Engineering, Academic Division, Chulachomklao Royal Military Academy, Nakorn-nayok 26000, THAILAND

and

Mary P. Harper

School of Electrical and Computer Engineering, Purdue University, West Lafayette, IN 47907, USA

1. Introduction

The impetus for this research arose during an investigation of the language modelling aspects of an automatic speech understanding system of Thai. A good language model not only improves the accuracy of low-level acoustic models of speech, but also reduces task perplexity (the average number of choices at any decision point) by making better use of high-level knowledge sources including prosodic, syntactic, semantic, and pragmatic knowledge sources. A language model often consists of a grammar written using some formalism which is applied to a sentence by utilizing some sort of parsing algorithm. For example, a set of context-free grammar (CFG) production rules can be used to parse sentences in the language defined by that grammar. CFG is a phrase-structure representation of syntax (Chomsky, 1963). Another example is constraint dependency grammar (CDG). CDG parsers rule out ungrammatical sentences by propagating constraints. Constraints are developed based on a dependency-based representation of syntax. Some parsing algorithms combine phrase-structure and dependency grammars. In this paper, a CDG parsing approach is adopted. In the next section, we present a contrastive description of the two major approaches to representing the syntax of natural languages in order to motivate our choice of dependency grammar for parsing Thai sentences.

1.1 Dependency *vs.* phrase-structure grammar

In the theory of the syntax of natural languages, there are currently two major methods of representing the syntactic structure of natural sentences: dependency grammar and phrase-structure (constituency) grammar. There has been no third approach developed although combinations of the two major methods above have been used, e.g., lexical-functional grammar, case grammar, relational grammar, word grammar, etc. (Mel' čuk, 1988).

As a formal syntactic representation, dependencies have been studied and explored for centuries by traditional syntacticians of European, Classical, and Slavic languages. Lucien Tesnière (1959) was credited as the first syntactician who formalized and laid the groundwork for subsequent investigations of the theory. Unfortunately, the dependency formalism has not gained great popularity among today's syntacticians, who generally favor constituency grammar. Constituency or phrase-structure grammar was formulated in North America in the early 1930's by Leonard Bloomfield (1933), primarily for describing the syntax of English. The theory was seriously advanced by Noam Chomsky, and his transformational-generative approach has been accepted throughout the world. Phrase-structure syntax gradually forced dependency syntax into relative obscurity. Nevertheless, there have been some attempts to defend the use of dependency syntax, and several linguists have contributed to this cause. For example, Mel' čuk (1988) presented an argument for the case of dependency formalism and bravely claimed that "dependencies are much better suited to the description of syntactic structure (of whatever nature) than constituency is." His contrastive description of the two methods is summarized as follows.

A dependency grammar describes the syntactic structure of a sentence by using a dependency tree (D-tree) to establish dependencies among words in terms of head and dependents. A D-tree shows a relational characteristic of the syntactic representation in the form of hierarchical links between items, i.e., which items are related to which other items and in which way. On the other hand, a phrase-structure grammar uses a phrase-structure tree (PS-tree) to describe the groupings of words into the so-called *constituents* at different levels of sentence construction. A PS-tree shows which items go together with other items to form tight units of a higher order, a distributional characteristic of a grouping within a larger grouping.

A *tree* is a network consisting of nodes which are linked in a tree-like structure (i.e., with a stem and branches). In a syntactic tree, a node which represents a word or lexical item, the smallest syntactic unit, is called a terminal node; a node which represents an abstract syntactic grouping or phrase, such as noun phrase (NP), verb phrase (VP), prepositional phrase (PP), etc., is called a non-terminal node. A D-tree contains only terminal nodes; no abstract representation of syntactic groupings is used. On the contrary, a PS-tree contains both terminal and non-terminal nodes; most nodes are, however, non-terminal. This hierarchical representation in terms of

terminals and non-terminals in a PS-tree leads to the notion of syntactic class membership of an item (i.e., categorization as belonging to an NP, VP, etc.). Syntactic class membership is a way of labelling syntactic roles in a PS-tree because a PS-tree does not and cannot specify the types of syntactic links existing between two items in a natural and explicit way. On the other hand, class membership is not specified in a D-tree. Instead, a D-tree puts a particular emphasis on specifying in detail the type of any syntactic relation between two related items. Such syntactic relations are, for example, predicative, determinative, coordinative relations, etc. In addition, in terms of the ordering of nodes, nodes must be ordered linearly in a PS-tree. In a D-tree, however, nodes are not necessarily in a linear order.

From the above, one can draw the following conclusion. The PS-representation is suitable for languages like English which have a rigid word order and a near absence of syntactically driven morphology. On the other hand, the D-representation is suitable for languages like Latin or Russian which feature an incredibly flexible (but far from arbitrary) word order and very rich systems of morphological markings. In these languages, word arrangements and inflectional affixes are obviously contingent upon relations between words rather than upon constituencies.

1.2 The difficulty of parsing Thai.

Research on the syntactic analysis of Thai sentences by computer has been carried out for over a decade. Thai grammars have been developed utilizing various grammar formalisms based on the above two theories of syntax or their combination. However, the most popular formalism has been the phrase-structure representation of syntax. Vorasucha (1986) was one of the first researchers to use Gazdar's Generalized Phrase-Structure Grammar (GPSG) in his research. Syntactic rules were written in ID and LP rule formats. Pornprasertsakul, et al. (1990) later employed additional FSD constraints of GSPG to describe three types of sentence structures including verb and noun phrases. Aroonmanakul (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 lexical functional grammar. Wuwongse and Pornprasertsakul (1993) introduced a probabilistic approach using a *least exception logic* (LEL) model of default reasoning to resolve

ambiguities. 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 the following reasons.

First, written Thai sentences do not contain dilimiters 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 dilimiters. 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 algorithm. For example, given the following string of Thai characters (Luksaneeyanawin, 1993), ภาพรออกฉาก, two possible sentence hypotheses are generated given dictionary lookup. Note that the string พรอ is not a legitimate Thai word based on a Thai-English dictionary (McFarland, 1960).

a. ภาพ รอ ออก ฉาก

b. ภา พร ออ ก 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 shows its syntactic function. Hence, syntactic relationships are primarily determined by word order, and structural ambiguity often arises. 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 comprising the same sequence of words. The following example illustrates the problem.

Compound: เขาเป็นคนเจ้าชู้ มี คนรัก มาก

'He is a flirting kind. He has a lot of girlfriends.'

Sentence: เขาเป็นดาราณิสัยดี มี คนรัก มาก

'He 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 world's languages usually exhibit consistent ordering relations across phrasal categories. The