The Identification of Word Classes in Thai

Stanley Starosta

University of Hawai'i

INTRODUCTION

This article is intended as a contribution to solving the problem of finding a generally acceptable and grammatically justifiable set of syntactic word classes for Thai. It begins with a consideration of the basic question of how syntactic word classes should be established, and then applies the results to several open questions in Thai syntax, contrasting some previous approaches with constrained dependency analyses of the same phenomena in terms of the kinds of conflicting criteria which have been applied in establishing word classes. The two particular areas I will focus on are the questions of (1) which word class to assign "classifiers" to, and (2) whether or not Thai has a distinct class of prepositions.

The grammatical framework within which the following discussion will be framed is an approach to dependency grammar referred to as LEXICASE. Dependency grammar is an ancient approach to grammatical analysis (cf. Covington, 1986) which analyzes sentences in terms of pairwise relations between words, and lexicase is a version of dependency grammar which localizes information about these pairwise links in the lexical entries themselves.

For the purposes of this paper, lexicase dependency grammar can be viewed as having three salient characteristics: it is GENERATIVE (formal and explicit), it is strictly CONSTRAINED, and it is UNIVERSAL. One of the universal constraints it imposes on languages is a strictly limited inventory of word classes: no language (including Thai) may contain a word class which is not identical with, or a subclass of, one of the following eight classes: *V* (verb), *N* (noun), *Adj* (adjective), *Det* (determiner), *Adv* (adverb), *P* (preposition or postposition), *Cnjc* (conjunction), or *Sprt* (sentence particle), though not every language necessarily utilizes all eight classes (cf. Dixon, 1977). This requirement is not some kind of arbitrary edict, but rather an empirical hypothesis: if it is correct, it will make possible greater language-specific and cross-linguistic generalizations, and it can be proven wrong by proposing an alternate set (or no fixed set at all) and showing how this alternative approach makes it possible to capture more and better grammatical generalizations. This is the crucial factor that makes grammars written in different frameworks comparable: all grammars try to capture generalizations, and the grammar that does a better job of capturing generalizations is a better grammar. [1]

---

[1]As an example, Savetamalya (1989) analyzes words like Thai *níi* 'this' as a determiner, while Amara Prasithrathsint (p.c.) considers them to be adjectives. Savetamalya showed that the distribution of *níi* is quite different from the distribution of other adnominal modifiers, so that it would at least have to be considered a special subclass of adjectives. Considerations of cross-linguistic generality then tell us that a class of adnominal modifiers which always occur at the periphery of an NP and which express deictic meanings are better analyzed as determiners than as adjectives.

As a brief illustration of how this criterion may be applied, consider the advantages of proposing a fixed inventory of lexical categories for the analysis of auxiliary verbs in English or Thai (cf. Savetamalya, 1987; Indrambarya, 1994): the above eight-term inventory makes no provision for a distinct grammatical category *AUX, INFL, or I.* It thus excludes the Chomskyan *I(NFL)* analysis, an analysis which reliably results in the loss of numerous syntactic and morphological generalizations, the creation of such otherwise unmotivated transformational rules as "V-movement" and "I-movement,"[2] and an increase in the abstractness of the associated grammatical representations. That is, the constraint forces the linguist to adopt an analysis in which "auxiliaries" are a subclass of the class of verbs, an analysis which turns out to be demonstrably superior to the Chomskyan alternative in capturing morphological and syntactic generalizations (Starosta, 1991).

WORD CLASSES

In writing a syntactic description of a language, one thing of which we may be sure is that the description will have to refer to word classes. This follows from some simple but fundamental considerations.

The first consideration is that every language must have words. Let us start off by defining a FREE FORM as a stretch of speech bounded at both ends by silence. If we follow Leonard Bloomfield (1926, pp. 153–164; 1933 p. 178) in defining a word as a minimal free form, then we can find a set of free forms which are not composed of parts which themselves occur as free forms. Such forms are minimal free forms, that is, WORDS.

We know that these words must be identified in a grammar as grouping into classes from the fact that not every sequence of words is equally acceptable to native speakers of a language. Speakers are fairly consistent in identifying certain sequences as belonging to their language and others as not belonging to their language, and from the point of view of generative grammar, the content of a syntactic description is the internalized knowledge that allows speakers to do this in a fairly consistent and intersubjectively verifiable way. Logically, there are two ways in which that knowledge could be internally represented, an extensional way and an intensional way. The speaker might be assumed to have an internalized list of possible strings of words, which would be an extensional representation, or he might have access to a set of "well-formedness conditions" that a string must meet in order to qualify as belonging to his language. This would be an intensional mechanism. Because the set of strings is potentially infinite, we can rule out the first alternative, and focus our attention on the second.

---

[2]Claims for the necessity of V/I movement made within the Chomskyan framework, as presented for example in Pollock (1989), are circular because they assume as given that there must be a category I to begin with. When this category is eliminated, the generalizations captured by his analysis can be reformulated much more naturally in terms of the dependency grammar notion of FINITE VERB DOMAIN (cf. Savetamalya, 1987). Note that I am not denying the existence of the set of facts that V/I movement is intended to account for, but only the necessity of using powerful transformational rules to account for them.

We have just in effect defined syntax as the study of which words can co-occur in phrases, that is, non-minimal free forms. Once again, we can consider two alternative ways in which syntactic information might be encoded: we might hypothesize that the necessary well-formedness conditions on strings were encoded as a list of specific pairs of words which are allowed to occur with each other. Once again, however, such a list would be a sizable subset of the Cartesian product of the entire vocabulary of a language, a list which seems much too large to be plausibly memorized. Moreover, even if it were memorized, where would it come from? It could hardly have been memorized from experience, since speakers can consistently judge collocations as good or bad even for pairs that are very unlikely to have ever been produced before.

Again, we are faced with the necessity of positing a more abstract ability than simple listing to account for syntactic competence. I can find no alternative to assuming that each word is identified as belonging to one of a fairly small set of classes, and that the well-formedness conditions which are the content of a syntactic description are stated in terms of these classes rather than directly in terms of individual lexical entries. All we need to do then is to group words with identical properties into sets and write our grammatical rules to refer to these sets, and the problem is solved. This paper will consider some strategies for accomplishing this, and draw implications for the resolution of some controversial questions in the analysis of Thai "classifiers" and prepositions.

CONSTRAINTS VERSUS FLEXIBILITY

Throughout this paper, I will be assuming that it is desirable to find a rigorous and consistent set of syntactic criteria for uniquely determining the syntactic "part of speech" of each Thai word, and that it is desirable to fit the parts of speech for Thai into a limited universal set. However, neither of these goals is universally accepted. In the first section of this paper, I would like to consider a proposal made by Eric Schiller which rejects both of these assumptions. I will begin by quoting the first page of his paper, "Parts of speech in Southeast Asian languages: An autolexical view" (Schiller, 1992, p. 777):

> O. Introduction
>
> The syntax of Southeast Asian languages often seems quite difficult when observed from a perspective based on the study of European languages. This complexity is often compounded when one applies a theoretical perspective which forces lexical items into fixed syntactic categories determined by what are claimed to be universal considerations. This paper uses the notion of syntactic polysemy (Schiller, 1989) or syntactic flexibility (Ratliff, to appear) to discuss the nature of word classes in Khmer and a few other Southeast Asian languages. Specifically, I will concentrate on several words which appear in a wide variety of syntactic contexts, not merely nouns and verbs, but also modals, adverbs, prepositions, and classifiers.
>
> 1. "Parts of speech"
>
> By using the autolexical technique of separating syntactic considerations from semantic considerations (Sadock, 1991), and having a distinct inventory of word classes (or categories) at each level, the often confusing problem of determining "parts-of-speech" is made much clearer. Categories which have traditionally been at least somewhat controversial, such as "relator-nouns," "classifiers," and "coverbs," are easier to deal with when syntactic, semantic, and morphological considerations are dealt with separately. These notions have a tendency to be defined in purely language-specific terms, usually by positional factors since morphology is not

> much help in mainland Southeast Asian languages. For pedagogical purposes it is often useful to determine lexical categories simply on the basis of co-occurrence restrictions. However, this approach runs into real problems in the languages which permit widespread deletion, as is the case with most of the isolating languages of Southeast Asia.

I will consider three general points raised by this passage, 1) the question of flexibility versus rigid categories, 2) the question of universals versus language-specific analyses, and 3) the question of polysemy versus homophony.

Flexibility versus Rigidity

Schiller's paper can be seen as a plea for flexibility in the assignment of syntactic categories. Flexibility, however, is the sworn enemy of science. The content of a theory is its constraints, and a "theory" which allows everything, as autolexical grammar does, tells us nothing. To be testable, a theory must be constrained and rigid. If new data cannot be accommodated by such a theory, then the theory has been falsified and must be corrected. If new data are found to fit into the cast iron pigeonholes provided without any alteration being necessary, then the theory has been confirmed, though of course no scientific theory can ever be proven absolutely correct.

Universals versus Language-Specific Analyses

There are at least two questions we might ask in connection with the roles of universals in language analysis: *should* languages be analyzed in terms of universal categories, and *can* languages be analyzed in terms of universal categories? For the structuralists, of course, the answer to the first question was negative; it was felt that to try to fit observations into preexisting categories was circular and compromised the objectivity of the analysis. However, it was a genuine contribution of Chomskyan linguistics to point out that if linguistics was to be a science, it had to be a search for generalizations, for simple explanations that covered a broad range of data, and that a true scientific theory should cover all the phenomena in the domain with a minimum amount of explanatory apparatus.

From these considerations, it follows that a science of language should strive to cover all the phenomena of human language with a single coherent set of principles. That is, a science of linguistics should be a search for universals. That raises the second question: *can* languages be analyzed in terms of universal categories? If not, we have to give up an attempt to construct a science of language as the term "science" is generally understood. With the stakes so high, it is obvious that this course should not be taken without first making a good-faith effort to accomplish this goal. I think Schiller has given up too soon.

It is of course difficult, outside of mathematics, to prove that something is impossible. A claim such as Schiller's that it is impossible to find a universal set of syntactic categories and/or a universal set of criteria to identify such a set will only stand as long as no one has actually produced a universal set of categories and/or promulgated such a set of criteria. However, (1) the lexicase dependency grammar framework *has* proposed a universal set of eight categories (Starosta, 1988, pp. 51–52), and tested them in analyses of parts of more than 70 different languages,