Thai Part-Of-Speech Tagged Corpus: ORCHID

Virach Sornlertlamvanich[1,2], Naoto Takahashi[3], Hitoshi Isahara[4]
[1] Linguistics and Knowledge Science Laboratory, National Electronics and Computer Technology Center
Ministry of Science Technology and Environment, Thailand.
[2] Department of Computer Science, Tokyo Institute of Technology, Japan.
[3] Electrotechnical Laboratory, Japan.
[4] Intelligent Processing Section, Communications Research Laboratory
Ministry of Posts and Telecommunications, Japan.
E-mail: virach@cs.titech.ac.jp, ntakahas@etl.go.jp, isahara@crl.go.jp

Abstract

This paper presents a procedure in building a Thai part-of-speech (POS) tagged corpus, called ORCHID corpus. It is a collaboration project between Communications Research Laboratory (CRL) of Japan and National Electronics and Computer Technology Center (NECTEC) of Thailand, supported by Electrotechnical Laboratory (ETL) of Japan. We propose a new tagset based on the previous research on Thai parts-of-speech for using in a multi-lingual machine translation project. We mark the corpus in three levels:- paragraph, sentence and word levels. The corpus keeps text information in *text information line* and *number line,* which are necessary in retrieving process. We applied a probabilistic tri-gram model for simultaneously word segmenting and POS tagging. Rules for syllable construction are additionally used to reduce the number of candidates for computing the probabilities. The problems in POS assignment are formalized to resolve the ambiguities occurring in case of similar used of some POSs.

1. Introduction

Natural language processing is a key technology for highly computerized communities. To tackle this important technology, we started our collaboration project on NLP research from 1996. "Research and Development Cooperation Project on a Machine Translation System for Japan and Neighboring Countries", this so-called Multi-lingual Machine Translation Project (MMT project) [4] began at the 1987 Japanese fiscal year and continued through the 1992 fiscal year, followed by a two year follow up program. The project consisted of five countries, i.e., Japan, Thailand, China, Indonesia and Malaysia. Our collaboration project is a small successor of this project, to continue collaborations on NLP among these countries. We first started the collaboration between Japan and Thailand. This is because both Thai and Japanese have their own peculiar sets of characters and there are no delimiters between words in both languages. However, the languages themselves, e.g., grammar rules and other linguistic phenomena, are completely different. We believe we can obtain very interesting insights from this kind of bilingual collaboration.

What is the most significant result of the MMT Project for future research on NLP? The MMT project ended, developing its prototype of the multilingual machine translation systems, tools for NLP and linguistic data, such as dictionaries, corpora and grammars. Among these results, linguistic data, especially corpora, are re-usable by other researchers and/or in other projects. However, the most significant result of the MMT project was the stimulation of NLP related research in the region. Therefore, we decided to develop our tagged Thai corpus as a starting point of our collaboration project. Also, we focus on technological and personnel interchange between Japan and Thailand. LINKS (Linguistic and Knowledge Science Laboratory) of the National Electronics and Computer Technology Center (NECTEC) of Thailand and KARC (Kansai Advanced Research Center) of the Communications Research Laboratory (CRL) of The Japanese Ministry of Posts and Telecommunications, in collaboration with Electrotechnical Laboratory (ETL), are developing a tagged corpus for Thai, named ORCHID corpus [1]. The corpus is tagged with LINKS's original part-of-speech (POS) tagset, which is the improved version of the tagset in the MMT System [5]. The ORCHID corpus contains about 2MB (or about 400K words) of the proceedings of the NECTEC annual conference. CRL is conducting research on automatic POS tagging technology using a neural networks approach and automatic extraction of linguistic knowledge from tagged corpora. NECTEC is focusing its research on natural language processing for Thai and preparing linguistic resources for developing either a whole system of machine translation or applications of natural language processing. ETL takes part in developing a multilingual editor, Mule.

In this paper, we propose a procedure in building Thai POS tagged corpus with the design of data structure and the POS. We apply our POS developed from the set which had been used in developing the multilingual machine translation system in the co-operation project with Japan and other four Asian countries [4], as the tagset used in tagging words in the text. The original 45 POSs are carefully revised to be able to cover all roles of words used in the real text. As

a result, a new set of 47 POSs is defined as the tagset. Section 2 described the process in designing the text corpus with the details of structure and procedure in building the ORCHID corpus. Section 3 discusses the word class for using as the tagset in the corpus and Section 4 discusses some problematic tagging and sets up a guideline for making decision in giving a tag.

2. Marking the Text Corpus

The text is marked up with the original designed markers aiming at keeping all necessary information. The markers are not committed to any standard mark-up language, such as SGML, because the standard of such mark-up language causes an excessive procedure to handle the marks-up, though we need only a mark-up for POS tagging. However, we plan to extend our mark-up strategy to meet the SGML standard when the information in our corpus comes up to an extent.

2.1 Structure of the Text Corpus

The markers are classified into 2 classes of *text information line*, a line beginning with a "%" character, and *number line*, a line beginning with a "#" character. Both classes are not referred to as parts of the text therefore, the special characters are introduced to differentiate the lines from the original lines of the text. Texts are processed in line based manner then it is necessary to keep the information within a line for each line class.

The *text information line*, a line beginning with a "%" character, is used for storing a text information as shown in Table 1. The text information is given in both Thai and English. If one of each is absent from the original text, it is translated for the accessibility from both languages. Most of Thai texts show the published year in B.E. (Buddhist Era), they are converted into the year of A.D. to avoid the confusion in referencing. The line beginning with a "%" character followed by a string beside the registered tokens in Table 1 is recognised as a *comment line* for user's use as an additional information.

Table 1 Mark-up for Text Infromation Line

Mark-up	Description
%TTitle:	Title of a document, in Thai.
%ETitle:	Title of a document, in English.
%TAuthor:	Author's name, in Thai.
%EAuthor:	Author's name, in English.
%TInbook:	Title of a book where a document exists, in Thai.
%EInbook:	Title of a book where a document exists, in English.
%TPublisher:	Publisher of a book, in Thai.
%EPublisher:	Publisher of a book, in English.
%Page:	Page number or page range of a document.
%Year:	Published year (A.D.).
%File:	File number of a document. A long document may be separated into a number of files.

The *number line*, a line beginning with a "#" character, is used to sequence the lines in the text. There are 2 types in this line class as shown in Table 2. A number is used to index the line number within a paragraph. The number is automatically generated when other mark-up process is completed and it is always kept consistent when the editing has been done.

Table 2 Mark-up for Number Line

Mark-up	Description
#P[*number*]	Paragraph number of a text. The number in the bracket is shown in a sequence within a text.
#[*number*]	Sentence number of a paragraph. The number in the bracket is shown in a sequence within a paragraph.

Besides the line mark-up, there are another four special characters introduced as shown in Table 3. Because there is no explicit word breaking character used in the real Thai text, a paragraph is usually wrapped at either a space character or a suitable break (the breakable position according to the syllable construction restriction) within a word or at the end of a word. In general, a newline character can be read as whether there is a space character or it is a suitable break. To indicate the break explicitly, we introduce a marker, namely "\\" to break a paragraph into a number of lines avoid the ambiguity when the break is exactly at the space character. The "//" is used to mark the end of a sentence, the "/" followed by a POS is used to mark the appropriate POS tag to the preceding word.

Table 3 Special Characters for Mark-up

Mark-up	Description
\\	Line break symbol.
//	Sentence break symbol.
/[POS]	Tag mark for an appropriate POS of a word.

All special characters other than alphanumeric characters are replaced by internal defined strings enclosed by a pair of "<" and ">", as listed in

Table 4. Figure 1 shows a sample of the marked text, or the POS tagged text.

Table 4 Defined Strings for Special Characters

Special characters	Defined strings	Special characters	Defined strings
	<space>	/	<slash>
!	<exclamation>	:	<colon>
"	<quotation>	;	<semi_colon>
#	<number>	<	<less_than>
$	<dollar>	=	<equal>
%	<percent>	>	<greater_than>
&	<ampersand>	?	<question_mark>
'	<apostrophe>	@	<at_mark>
(	<left_parenthesis>	[	<left_square_bracket>
)	<right_parenthesis>	]	<right_square_bracket>
*	<asterisk>	^	<circumflex_accent>
+	<plus>	_	<low_line>
,	<comma>	{	<left_curly_bracket>
-	<minus>	}	<right_curly_bracket>
.	<full_stop>	~	<tilde>

```
%TTitle: คาร์บอนไดออกไซด์เลเซอร์กำลังสูงแบบไหลเวียนตามแนวแกน
%ETitle: High-Power Compact Axial Flow CO2 Laser
%TAuthor: ผศ.พิพัฒน์ โชคสุวัฒนสกุล
%EAuthor: [Asst. Prof. Pipat Choksuwatanasakul]
%TInbook: การประชุมทางวิชาการ ครั้งที่ 6 โครงการวิจัยและพัฒนาอิเล็กทรอนิกส์และคอมพิวเตอร์ ปีงบประมาณ 2536
%EInbook: The 6th NECTEC Annual Conference
%TPublisher: ศูนย์เทคโนโลยีอิเล็กทรอนิกส์และคอมพิวเตอร์แห่งชาติ กระทรวงวิทยาศาสตร์ เทคโนโลยีและสิ่งแวดล้อม
%EPublisher: National Electronics and Computer Center, Ministry of Science Technology and Environment
                    :
#P5
                    :
#5
ในการวิจัยครั้งนี้เราได้ลองศึกษาการเกิดดิสชาร์จจากลักษณะของรูปทรงของแคโถดที่ใช้ต่างๆ กัน\\
พบว่าการใช้แคโถดเป็นรูปทรงกระบอกกลวงทำให้เกิดกระแสในการดิสชาร์จ//
ใน/RPRE
การ/FIXN
วิจัย/VACT
                    :
//
                    :
```

Figure 1 A Sample of Thai POS Tagged Corpus

2.2 Procedure of Building the ORCHID

Though electronic files of Thai texts are increasing presently because of the widely use of computer in the text publishing, the language depending features of Thai text expose various kinds of topics to be solved. Under the limited resources, the ORCHID corpus is constructed regarding to the procedure shown in Figure 2. Most of the texts are inputted by keyboarding, because a Thai OCR is still in a developing stage and depends mostly on the quality of the printed texts. Therefore, the processes other than the POS tagging process are done manually with some software supports.

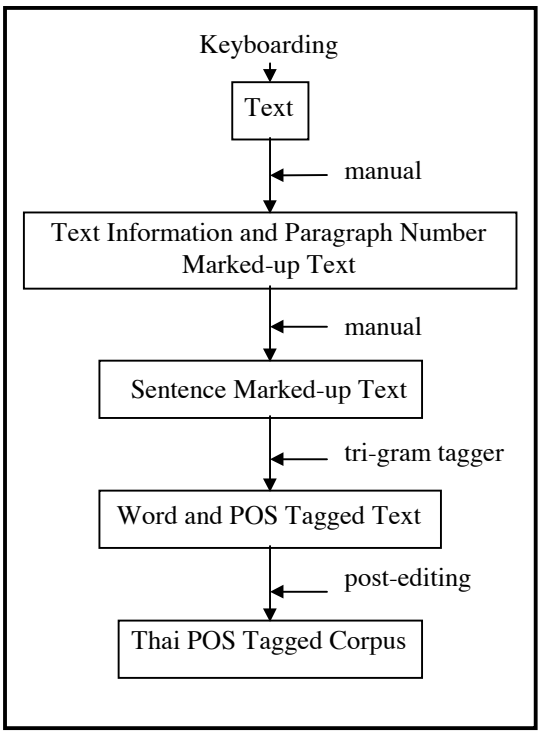
2.2.1 Word Segmentation and POS Tagging

We redefine the problem of word segmentation as the problem of POS tagging. The most probable sequence

of POSs and the POS assigned to a word determines the result of the most probable word segmentation and POS assignment. Probabilities in predicting a consecutive word or POS are computed with a tri-gram model [2,3,6] as shown in Equation 1, where, T is a sequence of POSs $\{t_1, \ldots, t_n\}$and W is a sequence of words $\{w_1, \ldots, w_n\}$. We introduce the Viterbi algorithm for computing for the most probable sequence of POSs and then rank the resultant word sequences according to their probabilities. To reduce the number of candidates for computing for the probabilities, we consult Thai spelling rule set [8] (constraint on combination of characters) which helps in pruning illegal segmenting of the input string.

$$P(W,T) = \prod_{i=1}^{n} P(t_i \mid t_{i-1}, t_{i-2}) \cdot P(w_i \mid t_i) \ldots\ldots (1)$$

Figure 2 Procedure in Constructing ORCHID



3. Word Class

We firstly developed our word class or parts-of-speech to classify words according to their syntactic roles and implemented in a dictionary used in a machine translation system [5]. The parts-of-speech contains 13 categories, which are subcategorized into 45 subcategories. They are essentially used in both analysis and generation modules of a machine translation system. We revised the original parts-of-speech by observing the real text data. As a result, we redefined some parts-of-speech to clarify the ambiguous parts and set up a new set of 14 categories with 47 subcategories, as shown in Table 5. The significant changes are the subcategories for classifier (CLAS) and prefix (FIXP). We subcategorized the original CLAS into 5 subcategories and FIXP into 2 subcategories. Classifier plays an important role in constructing phrases in Thai language, see [7] for detail discussion. Therefore, we subcategorized the CLAS to help in disambiguating phrasal construction.

Another modification is done on the FIXP, which is our attempt to support the construction of noun phrase and adverb phrases which is ambiguous because of the absence of word inflection in changing the syntactic roles of the words.

(1) การ/FIXN ออกกำลังกาย/VACT และ/JCRG <u>การ/FIXN</u> พักผ่อน/VACT ที่/PREL เพียงพอ/VSTA เป็น/VSTA สิ่ง/NCMN จำเป็น/VSTA สำหรับ/RPRE มนุษย์/NCMN ทุก/DDBQ คน/CNIT

(2) การ/FIXN ออกกำลังกาย/VACT และ/JCRG ∅ พักผ่อน/VACT ที่/PREL เพียงพอ/VSTA เป็น/VSTA สิ่ง/NCMN จำเป็น/VSTA สำหรับ/RPRE มนุษย์/NCMN ทุก/DDBQ คน/CNIT

The sentence (2) is still valid and has the equivalent meaning to the sentence (1) though the underlined FIXN is absent. From the above sentences, we may define "การพักผ่อน" either as a single-word noun to mean "taking a rest" or as a two-words noun composed of "การ" (a nominal prefix) and "พักผ่อน" (to rest). If we define it as a single-word noun, there will be a problem in assigning "พักผ่อน" as a verb paralleling with the noun "การออกกำลังกาย" in case of (2). As a result, we introduce FIXN and FIXV for nominal prefix and adverbial prefix consequently, and propose separating a nominalized noun into a prefix and a noun, as well as a prefixed adverb.

We used the 47 subcategories as the tagset for POS tagging in the ORCHID corpus. Table 5 shows the tagset and the examples.

Table 5 Thai Part-of-Speech as the Tagset for ORCHID

No.	POS	Description	Example
1	NPRP	Proper noun	วินโดวส์ 95, โคโรน่า, โค้ก, พระอาทิตย์
2	NCNM	Cardinal number	หนึ่ง, สอง, สาม, 1, 2, 3
3	NONM	Ordinal number	ที่หนึ่ง, ที่สอง, ที่สาม, ที่1, ที่2, ที่3
4	NLBL	Label noun	1, 2, 3, 4, ก, ข, a, b

5	NCMN	Common noun	หนังสือ, อาหาร, อาคาร, คน
6	NTTL	Title noun	ดร., พลเอก
7	PPRS	Personal pronoun	คุณ, เขา, ฉัน
8	PDMN	Demonstrative pronoun	นี่, นั่น, ที่นั่น, ที่นี่
9	PNTR	Interrogative pronoun	ใคร, อะไร, อย่างไร
10	PREL	Relative pronoun	ที่, ซึ่ง, อัน, ผู้
11	VACT	Active verb	ทำงาน, ร้องเพลง, กิน
12	VSTA	Stative verb	เห็น, รู้, คือ
13	VATT	Attributive verb	อ้วน, ดี, สวย
14	XVBM	Pre-verb auxiliary, before negator “ไม่”	เกิด, เกือบ, กำลัง
15	XVAM	Pre-verb auxiliary, after negator “ไม่”	ค่อย, น่า, ได้
16	XVMM	Pre-verb, before or after negator “ไม่”	ควร, เคย, ต้อง
17	XVBB	Pre-verb auxiliary, in imperative mood	กรุณา, จง, เชิญ, อย่า, ห้าม
18	XVAE	Post-verb auxiliary	ไป, มา, ขึ้น
19	DDAN	Definite determiner, after noun without classifier in between	นี่, นั่น, โน่น, ทั้งหมด
20	DDAC	Definite determiner, allowing classifier in between	นี้, นั้น, โน้น, นู้น
21	DDBQ	Definite determiner, between noun and classifier or preceding quantitative expression	ทั้ง, อีก, เพียง
22	DDAQ	Definite determiner, following quantitative expression	พอดี, ถ้วน
23	DIAC	Indefinite determiner, following noun; allowing classifier in between	ไหน, อื่น, ต่างๆ
24	DIBQ	Indefinite determiner, between noun and classifier or preceding quantitative expression	บาง, ประมาณ, เกือบ
25	DIAQ	Indefinite determiner, following quantitative expression	กว่า, เศษ
26	DCNM	Determiner, cardinal number expression	หนึ่งคน, เสื้อ 2 ตัว
27	DONM	Determiner, ordinal number expression	ที่หนึ่ง, ที่สอง, ที่สุดท้าย
28	ADVN	Adverb with normal form	เก่ง, เร็ว, ช้า, สม่ำเสมอ
29	ADVI	Adverb with iterative form	เร็วๆ, เสมอๆ, ช้าๆ
30	ADVP	Adverb with prefixed form	โดยเร็ว
31	ADVS	Sentential adverb	โดยปกติ, ธรรมดา
32	CNIT	Unit classifier	ตัว, คน, เล่ม
33	CLTV	Collective classifier	คู่, กลุ่ม, ฝูง, เชิง, ทาง, ด้าน, แบบ, รุ่น
34	CMTR	Measurement classifier	กิโลกรัม, แก้ว, ชั่วโมง
35	CFQC	Frequency classifier	ครั้ง, เที่ยว
36	CVBL	Verbal classifier	ม้วน, มัด
37	JCRG	Coordinating conjunction	และ, หรือ, แต่
38	JCMP	Comparative conjunction	กว่า, เหมือนกับ, เท่ากับ
39	JSBR	Subordinating conjunction	เพราะว่า, เนื่องจาก, ที่, แม้ว่า, ถ้า
40	RPRE	Preposition	จาก, ละ, ของ, ใต้, บน
41	INT	Interjection	โอ้ย, โอ้, เออ, เอ๋, อ๋อ
42	FIXN	Nominal prefix	การทำงาน, ความสนุกสนาน
43	FIXV	Adverbial prefix	อย่างเร็ว
44	EAFF	Ending for affirmative sentence	จ๊ะ, จ้ะ, ค่ะ, ครับ, นะ, น่า, เถอะ
45	EITT	Ending for interrogative sentence	หรือ, เหรอ, ไหม, มั้ย

46	NEG	Negator	ไม่, มิได้, ไม่ได้, มิ
47	PUNC	Punctuation	(,), “, ,, ;

4. Problematic Tagging

Thai has no inflection and most of the compound words are created from simply combining two or more small word units. We found that the difficulty in tagging occurs because of the unchanging of lexical form though the word is used in different positions or roles in a sentence. We classify some problematic tagging as a guideline for making decision.

4.1 Verb and Preposition

There are a lot of prepositions having the same lexical form as verbs and sometimes hardly making distinction between them. Followings are the additional guides for making the distinction.

a) A preposition cannot be negated, but a verb can.
b) A preposition can be tested by moving the prepositional phrase. A preposition always goes together with the following noun, but a verb does not.

For example,

(3) หมอทายาตามตัวคนไข้

* หมอทายาไม่ตามตัวคนไข้

* ตัวคนไข้หมอทายาตาม ∅

Therefore, “ตาม” is a preposition.

(4) กระแส A จะตามกระแส B ไป

กระแส A จะไม่ตามกระแส B ไป

* กระแส B กระแส A จะตามไป ∅

Therefore, “ตาม” is a verb.

4.2 Adverb and Preposition

In general, adverbs can be placed more freely in a sentence rather than prepositions. There is no any strict rule for discriminating the two categories. But with some noticeable use of a preposition with the following noun, it is recommended to consider for a preposition at first, as to the criteria in 4.1. For example,

(5) สารชนิด C ถูกสกัดได้ตรงหลอดที่ 2 ตรง = Preposition

(6) กระแสนี้วิ่งตรงสู่ขั้วบวก ตรง = Adverb

4.3 Verb and Verbal Classifier

The classifiers, which are classified into the verbal classifier (CVBL) are the classifiers derived from verbs or having the same lexical form as verbs. Classifiers are used in the very rigid patterns as discussed in [7]. Most of the classifiers can be determined by checking with the possible patterns which verbs cannot be conformed to. For example,

(7) ข้าวสารกอบใหญ่ถูกนำมาใช้ทดลอง กอบ = Classifier

(8) เด็กกำลังกอบข้าวสาร กอบ = Verb

4.4 Verb and Auxiliary

Verbs and auxiliaries can have the same lexical form in many cases. In Thai language, there are mainly two types of auxiliary classified by their positions relating to the verb of a sentence. The negation criteria cannot be applied in this case because it is possible to negate both verbs and auxiliaries. Therefore, it is recommended to tag as a verb if there is no other candidate for being the main verb of the sentence. For example,

(9) อาจารย์ได้ทุนสนับสนุนจากกระทรวงฯ ได้ = Verb

(10) ผู้ร่วมวิจัยได้ตัดสินใจจะดำเนินการต่อ ได้ = Auxiliary

(11) เขาทำการทดลองได้ ได้ = Auxiliary

4.5 Verb and Adverb

It is confusable when verbs and adverbs have the same lexical form. For example,

(12) เขา/PPRS เดิน/VACT ตรง/ADVN ไป/XVAE โรงเรียน/NCMN

(13) เขา/PPRS เดิน/VACT ตรง/ADVN

(14) เขา/PPRS ตรง/VACT ไป/XVAE โรงเรียน/NCMN

“ตรง” can be either a verb (VACT) or an adverb (ADVN). There is no problem in (14) because there is no other verb in the sentence then “ตรง” must be a verb to make a sentence. In (12) and (13), there is a verb “เดิน”, then “ตรง” can better be interpreted as a modifier to the verb to make the meaning concisely. Consequently, it is more concise to interpret (12) as “He walks straight to school” by considering “ตรง” as an adverb rather than “He walks and directs to school” by considering “ตรง” as a verb.

4.6 Nominalization

A word in Thai can be nominalized by adding a prefix “การ” or “ความ” (FIXN) before a root word. But it is often difficult to differentiate the case whether it is a nominalized noun or a nominalized noun phrase. We, thus, propose a method to consider the nominalized noun or noun phrase as a composition of a prefix with a following noun or noun phrase. As a result, the decompositional consideration of the nominalized noun or noun phrase produces a consistent solution in the noun and noun phrase interpretation. For example,

(15) [การ/FIXN ออกกำลังกาย/VACT] เป็น/VSTA สิ่ง/NCMN ที่/PREL ดี/VATT

(16) [การ/FIXN ค้าขาย/VACT ภายใน/RPRE ประเทศ/NCMN] ได้ กำไร/VSTA เกินคาด/ADVN

(17) [การ/FIXN วิจัยและพัฒนา/VACT เชิง/FIXN คุณภาพ/NCMN] จะ/XVBM ทำให้/VACT ได้/VACT ผล/NCMN ที่/PREL ถูกต้อง/VATT

(18) [การ/FIXN ออกแบบ/VACT และ/JCRG สร้าง/VACT บ้าน/NCMN] ใช้เวลา/VCAT นาน/ADVN

(19) [การ/FIXN สร้าง/VACT บ้าน/NCMN และ/JCRG ตกแต่ง/VACT] ใช้เวลา/VACT นาน/ADVN

(20) [การ/FIXN วิเคราะห์/VACT ทาง/FIXN การแพทย์/NCMN] ได้ผล/VSTA ดี/ADVN

4.7 Noun and Classifier

In case of a common noun and its classifier form having the same lexical form, we can easily get confused because a noun and a classifier possibly occur in similar patterns. In this case, we use the following testing templates to distinguish a noun from its classifier by considering with some types of determiners (DDAC, DDAN, DCNM and DONM).

a) Noun Classifier DDAC
b) Noun DDAN
c) Noun DCNM Classifier
d) Classifier DONM
e) X DDAC

[X is a classifier if it has a form of classifier else it is a noun.]

For example,

(21) กระดาน/CNIT นี้/DDAC สวย/VATT

[กระดาน is a classifier because it can be either a noun or a classifier.]

(22) กระดาษ/NCMN นี้/DDAC สวย/VATT

[กระดาษ is a noun because it can only be a noun.]

(23) อัน/CNIT นี้/DDAC สวย/VATT

[อัน is a classifier because it can only be a classifier.]

4.8 Common Noun (NCMN) and Proper Noun (NPRP)

NCMN is a class of entity but not an individual while NPRP is a class of noun indicating particular persons, places, organizations, institutes, paintings or unique things, and usually not to be referred to by their meanings. There is no distinction in their syntactic forms between NCMN and NPRP, such as beginning with a capital letter for a proper noun as in English. We then, add the following guidelines for tagging a noun as NPRP.

a) Names of products, for example, วินโดวส์ 95 (Windows 95), โคโรน่า (Corona), โค้ก (Coke)
b) Abbreviation names, for example, จส.100, เน็ค เท็ค (NECTEC)
c) Names of persons, groups of persons, companies
d) Geographical names, such as names of regions, continents, countries, provinces, etc.
e) Astronomical names, for example, พระอาทิตย์ (the sun), ทางช้างเผือก (Milky way), ดาวอังคาร (Mars)
f) Chemical names, for example, โปรตีน (protein), ออกซิเจน (oxygen)
g) Scientific names
h) Names of artificial places
i) Names of languages, races, religions, etc.

NPRP can occur with NCMN as in the following examples.

(24) รถ/NCMN โตโยต้า/NPRP

(25) โปรแกรม/NCMN วินโดวส์95/NPRP

(26) บริษัท/NCMN ช.การช่าง/NPRP จำกัด/NCMN

4.9 DCNM, DONM, NLBL, ADV in Ordinal and Quantitative Expression

DCNM and DONM are classified by the following test frames :-

a) NCMN X Classifier
b) NCMN Classifier X

If a cardinal number (a figure or a word) occurs between a noun and a classifier, it is assigned as DCNM. If an ordinal number (a word or a figure preceding with "ที่") occurs after a classifier, it is assigned as DONM. For example,

(27) บ้าน/NCMN 1/DCNM หลัง/CNIT

(28) บ้าน/NCMN หลัง/CNIT ที่ 1/DONM

It is notable that sometimes a classifier between a noun and an ordinal number (DONM) can be omitted if it has the same lexical form as its noun. Besides an ordinal number can be assigned as DONM, there is a set of ordinal expression which is assigned as DONM. The ordinal expressions can be หนึ่ง (one), เดียว (single), แรก (first), สุดท้าย (last), หน้า (front), กลาง (middle) and หลัง (last). For example,

(29) คน/NCMN (คน/CNIT) ที่ 1/DONM

(30) คน/NCMN แรก/DONM

(31) บ้าน/NCMN หลัง/CNIT สุดท้าย/DONM

However, an ordinal expression can function as an adverb since it modifies a verb. Ordinal expressions in the following cases are all assigned as adverbs of sentences. For example,

(32) เขา/PPRS สอบ/VACT ได้/XVAE ที่ 1/ADVN

(33) เขา/PPRS มา/VACT คนแรก/ADVN

4.10 Classifier Expression

Besides the general use of classifier in the construction of quantitative expression, relative pronoun, demonstrative noun, etc. [7], we introduce a classifier to construct some types of verb or noun modifiers (adverb or adjective phrases). A classifier preceding a verb or a noun forms an adverb or adjective phrase consequently. Followings are some examples of the construction. For example,

(34) การ/FIXN วิจัย/VACT เชิง/CTYP คุณภาพ/NCMN

(35) อุปกรณ์/NCMN ทาง/CTYP การแพทย์/NCMN

(36) ผลผลิต/NCMN ด้าน/CTYP การเกษตร/NCMN

5. Conclusions

The ORCHID corpus is the first project to build Thai POS tagged corpus. It is not limited to the Thai language and POS tagging. We plan to extend our technology to other similar languages which share the similar language features and include other information to the corpus such as syntactic tree bracketing, semantic information, etc. Based on the first created corpus, we hope that we can study and gain more information about the Thai language with some corpus-based studies.

This paper revised the first version of Thai part-of-speech used in developing a multi-lingual machine translation system and applied it to a wider range of the real Thai text. It is not finalized but it can somehow cover all parts of the text we have in hand at present. It is proved to have the widest coverage for POS assignment. The POS confirmed by the real text is another crucial target of the building of the ORCHID corpus.

We are preparing the ORCHID corpus to be available for academic use. See http://www-karc/ips/index.html or http://www.links.nectec.or.th/ for forthcoming details.

6. References

[1] Charoenporn, T., Sornlertlamvanich, V. and Isahara, H. 1997. Building A Large Thai Text Corpus - Part-Of-Speech Tagged Corpus: ORCHID -, Proceedings of NLPRS'97, pages 509-512.

[2] Church, K. W. 1988. A Stochastic Parts Program and Noun Phrase Parser for Unrestricted Text, Proceedings of ANLP-88, pages 136-143.

[3] Cutting, D., Kupiec, J., Pedersen, J. and Sibun, P. 1992. A Practical Part-of-Speech Tagger, Proceedings of ANLP-92, pages 133-140.

[4] Komurasaki, M. 1995. Profile of International R&D Cooperation Project on Multi-lingual Machine Translation (MMT) System. Proceedings of the Symposium on Multi-lingual Machine Translation for Asian Languages, Thailand MMT'95, NECTEC, pages 10-21.

[5] Muraki, K., Sornlertlamvanich, V., Miyabe, T. and Tangdumrongvong, C. 1989. Thai Dictionary for Multi-lingual Machine Translation System, Proceedings of the Regional Workshop on Computer Processing of Asian Language (CPAL), AIT.

[6] Nagata, M. 1994. A Stochastic Japanese Morphological Analyzer Using a Forward-DP Backward-A* N-Best Search Algorithm, Proceedings of COLING'94, pages 201-207.

[7] Sornlertlamvanich, V., Phantachat, W. and Meknavin, S. 1994. Classifier Assignment by Corpus-based Approach, Proceedings of COLING'94, Vol.1, pages 556-561.

[8] Sornlertlamvanich, V. and Tanaka H. 1996. The Automatic Extraction of Open Compounds from Text Corpora, Proceedings of COLING'96, Vol.2, pages 1143-1146.