# BUILDING A THAI PART-OF-SPEECH TAGGED CORPUS (ORCHID)

Virach Sornlertlamvanich
National Electronics and Computer Technology Center,
Ministry of Science Technology and Environment, Thailand.
Department of Computer Science, Tokyo Institute of Technology, Japan.
virach@cs.titech.ac.jp
Naoto Takahashi
Machine Understanding Division, Electrotechnical Laboratory,
Ministry of International Trade and Industry, Japan.
ntakahas@etl.go.jp
Hitoshi Isahara
Intelligent Processing Section, Communications Research Laboratory,
Ministry of Posts and Telecommunications, Japan.
isahara@crl.go.jp

**Keywords:** POS-tagged corpus, Thai language, probabilistic trigram tagging, part of speech, linguistic resource

**Subject:** Submitted to the "Special Issue on Speech Database/Assessment for Oriental Languages"

**Type of submission:** Paper

**Author's mailing address:**

Virach Sornlertlamvanich
Machine Understanding Division,
Electrotechnical Laboratory,
Agency of Industrial Science and Technology,
Ministry of International Trade and Industry
1-1-4 Umezono, Ibaraki 305-8568, Japan

**No. of pages:** 20
**No. of figures:** 2
**No. of tables:** 5

**Running title:** Building a Thai Part-Of-Speech Tagged Corpus

**Paper No.:** E-98-39

# BUILDING A THAI PART-OF-SPEECH TAGGED CORPUS (ORCHID)

**Abstract**

ORCHID (Open linguistic Resources CHanelled toward InterDisciplinary research) is an initiative project aimed at building linguistic resources to support research in, but not limited to, natural language processing. Based on the concept of an open architecture design, the resources must be fully compatible with similar resources, and software tools must also be made available. This paper presents one result of the project, the construction of a Thai part-of-speech (POS) tagged corpus, which is a preliminary stage in the construction of a Thai speech corpus. The POS-tagged corpus is the result of collaborative research between the Communications Research Laboratory (CRL) in Japan and the National Electronics and Computer Technology Center (NECTEC) in Thailand, with technical support from the Electrotechnical Laboratory (ETL) in Japan. In this paper, we propose a new tagset, based on the results of a prior multilingual machine translation project. The corpus is annotated on three levels: the paragraph, sentence, and word levels. Text information is maintained in the form of the *text information lines* and the *number lines*, which are both utilized in data retrieval. Both word segmentation and POS tagging were carried out by way of a probabilistic trigram model. Rules for syllable demarkation were additionally used to reduce the number of candidates in computing tagging probabilities. Some typical problems in POS assignment are also formalized to resolve ambiguity.

# 1. INTRODUCTION

Natural language processing (NLP) represents a key technology in any highly computerized community. The "Research and Development Cooperation Project on a Machine Translation System for Japan and Neighboring Countries", or so-called Multilingual Machine Translation Project (MMT project) (Komurasaki , 1995), was one attempt to diffuse such NLP technology on an international level, commencing in 1987. It continued until 1992, and was proceeded by a two year follow-up program. The project was conducted between five Asian countries, namely Thailand, Japan, China, Indonesia, and Malaysia.

What were the more significant results of the MMT Project for subsequent NLP research? The MMT project was successful in developing a prototype multilingual machine translation system, some software tools to support NLP research, and linguistic data such as dictionaries, corpora, and grammars. Among these achievements, the linguistic data, and especially the corpora, are re-usable for other research purposes and also for related projects. However, the single most significant result of the MMT project would be the stimulation of NLP-related research in the Asian region.

Our new project, ORCHID, was initiated in 1996 as a successor to the original MMT project, to continue collaboration in NLP research among these countries; initially, collaborative research is being carried out solely between Thailand and Japan. The Thai and Japanese writing systems are similar in that they have their own peculiar character sets and no delimiters between words, although the languages themselves are completely different in terms of grammar rules and other linguistic phenomena. We believe that we can benefit from combining efforts to solve analogous problems, through bilateral collaboration.

ORCHID is also focused at technological and personnel interchange between Thailand and Japan. As a first step in this interchange, LINKS (Linguistic and Knowledge Science Laboratory) of the National Electronics and Computer Technology Center (NECTEC) in Thailand and KARC (Kansai Advanced Research Center) of the Communications Research Laboratory (CRL), operated under the auspices of the Japanese Ministry of Posts and Telecommunications, in collaboration with the Electrotechnical Laboratory (ETL), are currently jointly developing the ORCHID tagged corpus for Thai (Sornlertlamvanich et al., 1997; Charoenporn et al., 1997; Sornlertlamvanich et al., 1998). The corpus is being tagged with LINKS's original part-of-speech (POS) tagset, which is an improved version of the tagset used in the MMT project (Muraki et al.,

1989). The ORCHID corpus contains about 2MB (or about 400K words) of the proceedings of a NECTEC annual conference. CRL's contribution has been to apply its research on automatic POS tagging using neural networks to the Thai POS tagging task, and look at the automatic extraction of linguistic knowledge from the tagged corpora. NECTEC is focusing its research on natural language processing for Thai and is preparing linguistic resources for the development of a machine translation system and other NLP applications. ETL has involved itself in the development of a multilingual editor, called Mule, which supports the Thai language.

The POS set adopted as the tagset for the ORCHID corpus is a carefully revised version of the 45 element Thai POS set used in the MMT project, and contains 47 POSs for as great a coverage of real-world texts as possible.

The remainder of this paper is structured as follows: Section 2 describes the process used in designing the ORCHID text corpus with details of its structure and the construction procedure. Section 3 discusses the corpus tagset, and Section 4 discusses some problematic tagging issues and sets up guidelines for determining the appropriate tag.

## 2. MARKING-UP THE TEXT CORPUS

In the ORCHID corpus, text is marked up with our originally designed marker schema aimed at maintaining all necessary information. The markers are not yet committed to any standard mark-up language, such as SGML; such mark-up languages involve considerable overhead to produce marker schema compatibility, excessive for our needs in POS tagging. However, we plan to extend our mark-up strategy to meet SGML standards when the data in our corpus gets to a certain size.

### 2.1. Structure of the Text Corpus

Markers are classified into 2 types: 1) *text information lines*—lines beginning with the character '%', and 2) *number lines*—lines beginning with the character '#'. Neither type forms a part of the original text, and therefore the given special characters are utilized to delimit such lines from the text. As the text is processed line-wise, it is necessary to keep annotational information within a single line for each of the two marker types.

*Text information lines* are used to store text information, as shown in Table 1. Text information is given in both Thai and English. If either of these is absent in the original text, the text in the given language is translated into the other language for accessibility between the two languages.

Most Thai texts indicate the year of publication in the form B.E. (Buddhist Era). In this case, the year is converted into the corresponding year A.D. (Anno Domini), to avoid confusion.

Lines beginning with the character '%' followed directly by a registered token string, as given in Table 1, are also identifiable as comment lines.

Table 1: The Mark-up Schema for Text Information Lines

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

*Number lines* are used to indicate the sequence of lines in the text. There are 2 sub-types of number lines, as shown in Table 2, used either to index the paragraph number or the line number within the containing paragraph. These numbers are automatically generated when other mark-up processes are completed and maintained for consistency during editing.

Table 2: The Mark-up Schema for Number Lines

| Mark-up | Description |
|---|---|
| #P[number] | Paragraph number of the text. |
| | The number in brackets shows the sequence number within the text. |
| #[number] | Sentence number within the paragraph. |
| | The number in brackets the sequence number within the paragraph. |

Besides the line mark-up, there are three other marker types, as described in Table 3. Since there is no explicit work break character in ordinary Thai texts, a line can be terminated at: 1) a space character, 2) a

suitable break within a word (as governed by syllable construction restrictions), or 3) the end of a word. As a result, a newline character can be read as either a space character or a suitable word break. To explicitly mark up the text, we add the '\\' marker after space at the end of lines terminated by a space character; otherwise we add the '\\' marker right after the last word of the line. The '//' marker is used to mark the end of a sentence. The '/' marker followed by a POS tag is used to mark the appropriate POS for the immediately preceding word. Table 3 is a summary of special characters used in mark-up.

Table 3: Special Characters used in Mark-up

| Mark-up | Description |
|---|---|
| \\ | Line break symbol. |
| // | Sentence break symbol. |
| /[POS] | POS tag for the immediately preceding word. |

All non-alphanumeric special characters are replaced by internally defined strings enclosed by "<" and ">", as listed in Table 4. This is to avoid any ambiguity that may occur with symbols in the text.

Figure 1 shows an example of marked-up text.

Table 4: Defined Strings to Represent Special Characters

| Special character | Defined string | Special character | Defined string |
|---|---|---|---|
| | <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

:

#4

ดังนั้นกำลังของเลเซอร์ต่อหน่วยความยาวจึงสามารถเพิ่มขึ้นได้ค่อนข้างมาก//

ดังนั้น/JSBR

กำลัง/NCMN

ของ/RPRE

เลเซอร์/NCMN

ต่อ/RPRE

หน่วย/NCMN

ความ/FIXN

ยาว/VATT

จึง/XVBM

สามารถ/XVAM

เพิ่ม/VATT

ขึ้น/XVAE

ได้/XVAE

ค่อนข้าง/ADVN

มาก/ADVN

//

#5

ในการวิจัยครั้งนี้เราได้ลองศึกษาการเกิดดิสชาร์จจากลักษณะของรูปทรงของแคโถดที่ใช้ต่างๆ กัน\\

พบว่าการใช้แคโถดเป็นรูปทรงกระบอกกลวงทำให้เกิดกระแสในการดิสชาร์จ//

ใน/RPRE

การ/FIXN

วิจัย/VACT

:

การ/FIXN

ดิสชาร์จ/VACT

//

:

Figure 1: A Sample of the Thai POS-Tagged Corpus

## 2.2. Construction Procedure for the ORCHID Corpus

Thanks to the widespread use of computers in text publishing tasks, numbers of electronic Thai documents are rapidly increasing. At the same time, however, it has become increasingly clear that there are many language features specific to Thai text which remain to be fully supported. Using limited resources, the ORCHID corpus is being constructed based on the procedure shown in Figure 2. Most text has been input via a keyboard, as Thai OCR is still at the development stage and requires high print quality. As such, all processes other than POS tagging are being carried out manually with limited software support.

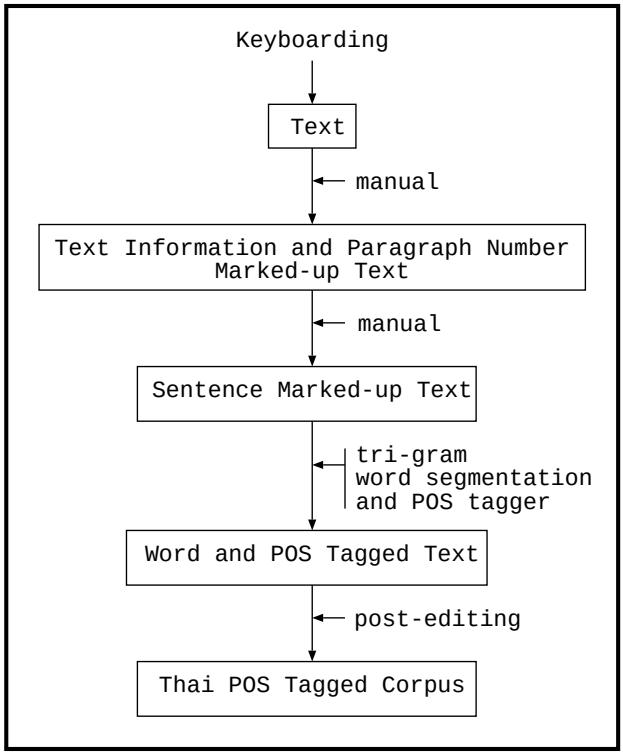


Figure 2: Procedure in Building the ORCHID corpus

**Word Segmentation and POS Tagging** We redefine the problem of word segmentation (Sornlertlamvanich , 1993; Sornlertlamvanich , 1998) in terms of POS tagging. The combination of the most probable sequence of POSs and individual word-level POS assignments determines the most probable combination of word segmentation and POS assignment. Therefore, the tagging task can be restated as finding the most probable sequence of component words and the corresponding POS sequence. Word and POS sequence

probabilities are computed with a trigram model (Church , 1988; Cutting et al., 1992; Nagata , 1994) as shown in Equation 1, where $T$ is a sequence of POSs $\{t_1, \ldots, t_n\}$ and $W$ is the associated sequence of words $\{w_1, \ldots, w_n\}$. We introduce the Viterbi algorithm (Viterbi , 1967) for computing the most probable sequence of POSs and then rank the resultant word sequences according to their probability. To reduce the number of candidates in computing these probabilities, we apply a Thai spelling rule set (Sornlertlamvanich et al., 1996) (made up of constraints on character combinations) which helps in pruning off illegal word segmentations of the input string.

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

## 3. WORD CLASS

As described above, we have developed our word classes (parts-of-speech) to classify words according to their syntactic roles, and implemented them in a dictionary used in a machine translation system (Muraki et al., 1989). The parts-of-speech are comprised of 13 categories, which can be subcategorized into 45 subcategories. They were used both in the analysis and generation modules of the original machine translation system. We revised the original part-of-speech schema through analysis of real-world text data. As a result, we refined some parts-of-speech to clarify ambiguities, and produced a new set of 14 categories with 47 subcategories, as shown in Table 5. Significant changes are the subcategorization of classifiers (CLAS) and prefixes (FIXP). We subcategorized the original classifier designation into 5 subcategories and prefixes into 2 subcategories.

In the Thai language, as well as in some other Asian languages such as Japanese and Chinese, classifiers find significant use in quantitative noun phrases (Sornlertlamvanich et al., 1994; Bond et al., 1996). From our study, we realized that classifiers do not only help in expressing quantitative noun phrases, but also play a very important role in forming many types of phrases, including relative pronoun phrases, noun phrases and adverb phrases (see (Sornlertlamvanich et al., 1994) for a detailed discussion). Based on this observation, we subcategorized classifiers into more detailed sub-groups to help in disambiguating phrasal structure.

Another modification was carried out for prefixes, in an attempt to support the construction of noun

phrases and adverb phrases. These two phrases types are ambiguous because of the absence of word inflection in the Thai language.

*(1) ka:n0/FIXN ?@:k1-kam0-lang0-ka:y0/VACT lx4/JCRG ka:n0/FIXN phak4-ph@:n1/VACT thi:2/PREL phi:ang0-ph@:0/VSTA pen0/VSTA sing1/NCMN cam0-pen0/VSTA sam5-rap1/RPRE ma?4-nut4/NCMN thuk4/DDBQ khon0/CNIT*

[Lit.: Sufficient exercise and rest are essential things for all people.]

*(2) ka:n0/FIXN ?@:k1-kam0-lang0-ka:y0/VACT lx?4/JCRG phak4-ph@:n1/VACT thi:2/PREL phi:ang0-ph@:0/VSTA pen0/VSTA sing1/NCMN cam0-pen0/VSTA sam5-rap1/RPRE ma?4-nut4/NCMN thuk4/DDBQ khon0/CNIT*

[Lit.: Sufficient exercise and rest are essential things for all people.]

Both of the above two sentences are grammatical, and they are equivalent in meaning, though the underlined FIXN in (1) is absent in the case of (2). Here, we can define "*ka:n0-phak4-ph@:n1*" as either a single-word noun meaning "taking a rest", or a compound noun composed of "*ka:n0*" (a nominal prefix) and "*phak4-ph@:n1*" (to rest). If we define it as a single-word noun, there will be a problem in describing "*phak4-ph@:n1*" (to rest) as a verb paralleling the noun "*ka:n0-?@:k1-kam0-lang0-ka:y0*" (exercising) in the case of (2). Therefore, we introduce FIXN and FIXV for nominal prefixes and adverbial prefixes respectively, and propose decomposing nominalized nouns into a prefix and a noun, and similarly for prefixed adverbs. As a result, we can explain the grammaticality of sentence (2) above.

We used the 47 subcategories as the POS tagset for the ORCHID corpus. Table 5 lists the entire tagset with examples.

## 4. PROBLEMATIC TAGGING CASES

The Thai language has no inflection and most compound words are created from the concatenation of two or more smaller word units. Moreover, we found that difficulties in tagging occur because of the fixed lexical form, even when the word is used in different positions or roles in a sentence. We thus classified some

Table 5: The Thai Part-of-Speech 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, เสือ 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, อ๊อ |
| 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 | (, ), “, ,, ; |

problematic tagging cases to act as guidelines in determining the correct tagging type in cases of potential ambiguity.

### 4.1. Verbs vs. Prepositions

Many prepositions have the same lexical forms as verbs, making these two types difficult to distinguish between at times. The following are guidelines to aid in making this distinction.

- Prepositions cannot be negated, but verbs can.
- Preposition status can be tested by moving the prepositional phrase around within the same sentential context. Prepositions always accompany the proceeding noun under movement, but verbs do not.

For example,

*(3) m@:5 tha:0 ya:0 ta:m0 tu:a0 khon0-khai2*

[Lit.: A doctor applies ointment along a patient's body.]

- * *m@:5 tha:0 ya:0 mai2 ta:m0 tu:a0 khon0-khai2*

  [Lit.: *A doctor applies ointment NOT along a patient's body.]

- * *tu:a0 khon0-khai2 m@:5 tha:0 ya:0 ta:m0* ∅

  [Lit.: * A patient's body, a doctor applies ointment along ∅.]

Based on the above, "*ta:m0*" (along) is a preposition.

*(4) kra1sx:5 ?e:0 ca?1 ta:m0 kra1sx:5 bi:0 pai0*

[Lit.: The current A will follow the current B.]

- *kra1sx:5 ?e:0 ca?1 mai2 ta:m0 kra1sx:5 bi:0 pai0*

  [Lit.: The current A will NOT follow the current B.]

- * *kra1sx:5 bi:0 kra1sx:5 ?e:0 ca?1 ta:m0 pai* ∅

  [Lit.: *The current B, the current A will follow ∅.]

This suggests that, "*ta:m0*" (to follow) is a verb.

## 4.2. Adverbs vs. Prepositions

In many languages, adverbs have less stringent locational constraints than prepositions. This also applies to the Thai language. Despite this, there are no rigorous rules to distinguish between the two category types, except for the noticeable case of prepositions accompanying a proceeding noun, where preposition status can be diagnosed with the criterion from Subsection 4.1. For example:

*(5) sa:n5 cha?4-nit4 si:0 thu:k1 sa?1-kat1 dai2 trong0 l@:t1 thi:2-s@:ng5*

. **trong0 (right at) is a preposition**

[Lit.: The material C is extracted right at the second test tube.]

*(6) kra1-sx:5 ni:4 wing2 trong0 su:1 khu:a2 bu:ak1* **trong0 (straight) is an adverb**

[Lit.: This current flows straight to the plus polar.]

## 4.3. Verbs vs. Verbal Classifiers

Verbal classifiers (CVBL) are classifiers which derive from verbs or have the same lexical form as verbs. In almost all cases, classifiers are used in very rigid patterns of usage, as discussed in (Sornlertlamvanich et al., 1994). Therefore, when a word ambiguous between a verb and verbal classifier is used in a pattern specific to a classifier, it must be a classifier; otherwise it is a verb. For example:

*(7) kha:w2-sa:n5 k@:p1 yai1 thu:k1 nam0 ma:0 chai4 thot4-l@:ng0*

. **k@:p1 (handful) is a classifier**

[Lit.: A big handful rice is brought for the experiment.]

*(8) dek1 kam0-lang0 k@:p1 kha:w2-sa:n5* **k@:p1 (to scoop) is a verb**

[Lit.: A child is scooping the rice.]

## 4.4. Verbs vs. Auxiliaries

Many verbs and auxiliaries have the same lexical form. In the Thai language, there are two main auxiliary types, classified according to their position relative to the matrix verb. The negation criterion from above

cannot be applied in this case because it is possible to negate both verbs and auxiliaries. Therefore, it is suggested that words ambiguous between a verb and auxiliary be tagged as a verb if there are no other candidates for the matrix verb position. This is based on the assumption that the sentence structure has priority over local phrasal structure. For example:

*(9) ?a:0-ca:n0 dai2 thun0 sa?1-nap1-sa?1-nun5 ca:k1 kra?1-su:ang0* **dai2 (to receive) is a verb**

[Lit.: The professor receives a supporting fund from the ministry.]

*(10) phu:2-ru:am2-wi?4-cai0 dai2 tat1-sin5 ca?1 dam0-noen0-ka:n0 t@:1* **dai2 (-ed) is an auxiliary**

[Lit.: The co-researcher decided to continue the process.]

*(11) khaw5 tham0 ka:n0 thot4-l@:ng0 dai2* **dai2 (can) is an auxiliary**

[Lit.: He can do the experiment.]

### 4.5. Verbs vs. Adverbs

There is also scope for confusion between verbs and adverbs in the case that they have the same lexical form. For example:

*(12) khaw5/PPRS doen0/VACT trong0/ADVN pai0/XVAE ro:ng0-ri:an0/NCMN*

[Lit.: He walks straight to school.]

*(13) khaw5/PPRS doen0/VACT trong0/ADVN*

[Lit.: He walks straight.]

*(14) khaw5/PPRS trong0/VACT pai0/XVAE ro:ng0-ri:an0/NCMN*

[Lit.: He (goes) directly to school.]

"*trong0*" ([go] directly; directly/straight) can be either a verb (VACT) or an adverb (ADVN). There is no problem in (14), as there is no other verb in the sentence, and "*trong0*" ([go] directly) must be a verb to form a sentence. In (12) and (13), as there is a verb "*doen0*" (walk), "*trong0*" (straight) can be more

readily interpreted as a modifier to the verb. Consequently, it would be better to interpret (12) as "He walks straight to school" by taking "*trong0*" (straight) to be an adverb, rather than "He walks and (goes) directly to school" with "*trong0*" (direct) as a verb.

### 4.6. Nominalization

Words in Thai can be nominalized by adding the prefix "*ka:n0*" or "*khwa:m0*" (FIXN) before the root. However, it is often difficult to judge whether it is a noun or a noun phrase that has been nominalized. Thus, we propose considering the nominalized noun or noun phrase as the combination of a prefix and a noun or noun phrase. In this way, we can consistently select between nouns and noun phrases. For example:

*(15)* *[ka:n0/FIXN ?@:k1-kam0-lang0-ka:y0/VACT]$_{NP}$ pen0/VSTA sing1/NCMN thi:2/PREL di:0/VATT*

[Lit.: [Exercising]$_{\mathrm{NP}}$ is a good thing.]

*(16)* *[ka:n0/FIXN kha:4-kha:y5/VACT pha:y0-nai0/RPRE pra1-the:t2/NCMN]$_{NP}$ dai2-kam0-rai0/VSTA koe:n0-kha:t2/ADVN*

[Lit.: [Domestic trading]$_{\mathrm{NP}}$ gains profit more than expected.]

*(17)* *[ka:n0/FIXN wi?4-cai0-lx?4-phat4-tha?4-na:0/VACT choe:ng0/FIXN khun0-na?4-pha:p2/NCMN]$_{NP}$ ca?1/XVBM tham0-hai2/VACT dai2/VACT phon5/NCMN thi:2/PREL thuk1-t@:ng2/VATT*

[Lit.: [Researching and developing qualitatively]$_{\mathrm{NP}}$ will yield a correct result.]

*(18)* *[ka:n0/FIXN ?@:k1-bx:p1/VACT lx?4/JCRG sa:ng2/VACT ba:n2/NCMN]$_{NP}$ chai4-we:0-la:0/VACT na:n0/ADVN*

[Lit.: [Designing and building a house]$_{\mathrm{NP}}$ take a long time.]

*(19)* *[ka:n0/FIXN sa:ng2/VACT ba:n2/NCMN lx?4/JCRG tok1-tx:ng1/VACT]$_{NP}$ chai4-we:0-la:0/VACT na:n0/ADVN*

[Lit.: [Building a house and decorating]$_{\mathrm{NP}}$ take a long time.]

*(20)* *[ka:n0/FIXN wi?4-khr@?4/VACT tha:ng0/FXIN ka:n0-phx:t2/NCMN]$_{NP}$ dai2-phon5/VSTA di:0/ADVN*

[Lit.: [Analyzing medically]$_{\mathrm{NP}}$ yields a good result.]

### 4.7. Nouns vs. Classifiers

Confusion can arise when a common noun and its classifier have the same lexical form, as it is possible that the noun and classifier occur in similar patterns of usage. In this case, we check the type of determiners (DDAC, DDAN, DCNM, or DONM) around the words in question in determining the appropriate POS. The following are templates used in this kind of judgement.

a) Noun Classifier DDAC

b) Noun DDAN

c) Noun DCNM Classifier

d) Classifier DONM

e) **X** DDAC

[**X** in the above template is a classifier if it takes the form of a classifier; otherwise it is a noun.]

For example:

*(21)* *kra?1-da:n0/CNIT ni:4/DDAC suay5/VATT*

[Lit.: This board is beautiful.]

["*kra?1-da:n0*" (board) is a classifier because it can be either a noun or a classifier]

*(22)* *kra?1-da:t1/NCMN ni:4/DDAC suay5/VATT*

[Lit.: This paper is beautiful.]

["*kra?1-da:t1*" (paper) is a noun because it can only be a noun.]

*(23)* *?an0/CNIT ni:4/DDAC suay5/VATT*

[Lit.: This thing is beautiful.]

["*?an0*" (thing) is a classifier because it can only be a classifier.]

## 4.8. Common Nouns (NCMN) vs. Proper Nouns (NPRP)

The NCMN noun marker denotes common nouns, while the NPRP noun marker indicates a particular person, place, organization, institute or painting, or a unique thing. NPRP is considered to be a label and items marked as such are not interpreted literally. In English proper nouns are capitalized, but in Thai there is no distinction in the lexical form of NCMN and NPRP instances. We thus established the following guidelines for tagging a given noun as NPRP.

a) Names of products, for example *win0-do:2 95* (Windows 95), *kho:0-lo:0-na:2* (Corona), *kho:k4* (Coke)

b) Abbreviations, for example *c@:0-s@:4-r@:j3* (CS-100), *nek3-thek1* (NECTEC)

c) Names of people, groups of people, and companies

d) Geographical names, such as names of regions, continents, countries, provinces, etc.

e) Astronomical names, for example *phra3-?a:0-thit3* (the sun), *tha:ng0-cha:ng3-phva:k1* (the Milky Way), *da:w0-?ang0-kha:n0* (Mars)

f) Chemical names, for example *pro:0-ti:n0* (protein), *?@:k3-si3-ce:n0* (oxygen)

g) Scientific names

h) Names of artificial places

i) Names of languages, races, religions, etc.

NPRP items can co-occur with NCMN items as in the following examples.

*(24) rot4/NCMN to:0-yo:0-ta:2/NPRP*

[Lit.: Toyota/NPRP car/NCMN.]

*(25) pro:0-krx:m0/NCMN win0-do:2-95/NPRP*

[Lit.: Windows95/NPRP program/NCMN.]

*(26) bo:0-ri?4-sat1/NCMN ch@:0-ka:n0-cha:ng2/NPRP cam0-kat1/NCMN*

[Lit.: Ch. Construction/NPRP Co.,Ltd./NCMN.]

## 4.9. DCNM, DONM, NLBL & ADV in Ordinal and Quantitative Expressions

DCNM and DONM are distinguished from each other by way of the following test frames:

a) NCMN **X** Classifier

b) NCMN Classifier **X**

If a cardinal number (a numeric figure or a word) occurs between a noun and a classifier, it is tagged as DCNM. If an ordinal number (a word or a numeric figure preceded by "*thi:2*" (-th)) occurs after a classifier, it is tagged as DONM. For example:

*(27) ba:n2/NCMN nung1/DCNM lang5/CNIT*

[Lit.: (House/NCMN) one/DCNM house/CNIT.]

*(28) ba:n2/NCMn lang5/CNIT thi:2-nung1/DONM*

[Lit.: (House/NCMN) first/DONM house/CNIT.]

It is worthy of note that classifiers occurring between nouns and ordinal numbers (DONM) can sometimes be omitted when they have the same lexical form as the noun. Otherwise, ordinal numbers can be tagged as DONM in the following ordinal expressions: *nung1* (one), *dia:w0* (single), *rx:k2* (first), *sut1-tha:j3* (last), *na:2* (front), *kla:ng0* (middle) and *lang4* (last). For example,

*(29) khon0/NCMN (khon0/CNIT) thi:2-nung1/DONM*

[Lit.: the first/DONM person/CNIT.]

*(30) khon0/NCMN rx:k2/DONM*

[Lit.: the first/DONM person/CNIT.]

*(31) ba:n2/NCMN lang5/CNIT sut1-tha:y4/DONM*

[Lit.: the last/DONM house/CNIT.]

However, ordinal expressions can also function as adverbs in modifying a verb. All of the following cases of ordinal expressions are tagged as adverbs.

*(32) khaw5/PPRS s@:p1/VACT dai2/XVAE thi:2-nung1/ADVN*

[Lit.: He/PPRS passes/VACT (the exam) with the first rank/ADVN.]

*(33) khaw5/PPRS ma:0/VACT khon0-rx:k2/ADVN*

[Lit.: He/PPRS comes/VACT first/ADVN.]

### 4.10. Classifier Expressions

Besides the general use of classifiers in the construction of quantitative expressions, relative pronouns, demonstrative nouns, and so on (Sornlertlamvanich et al., 1994), classifiers can be used to construct certain types of verb and noun modifiers (adverb and adjective phrases). Thus, a classifier preceding a verb or noun forms an adverb or adjective phrase, respectively. The following are examples of this construction type.

*(34) ka:n0/FIXN wi4-cai0/VACY choe:ng0/CTYP khun0-na?4-pha:p2/NCMN*

[Lit.: Qualitatively research.]

*(35) ?up1-pa?1-k@:n0/NCMN tha:ng0/CTYP ka:n0-phx:t2/NCMN*

[Lit.: Medical instrument.]

*(36) phon5-pha?1-lit1/NCMN da:n2/CTYP ka:n0-ka?1-se:t1/NCMN*

[Lit.: Agricultural products.]

## 5. CONCLUSIONS

The ORCHID project represents the first attempt to build a Thai POS-tagged corpus. However, the project is not limited to the Thai language and POS-tagged corpora. We are planing to apply the developed technologies to build corpora for other languages which closely resemble the Thai language, and also to

build speech corpora, and corpora tagged with more detailed information such as syntactic tree structure and semantic information. Based on the creation of this first corpus, we hope to study and gain more information about the Thai language beyond the actual corpus construction.

This paper presented a revised version of the Thai part-of-speech schema used in the MMT multilingual machine translation system, and its application to a wider range of real-world Thai text than was the case under the MMT project. While the POS schema is not yet complete, it is able to cover the full scope of text presently at hand, and proved itself to have wider coverage for POS assignment than its forerunner. The verification of the applicability of the POS set through analysis of real-world text, was a crucial sub-aim of building the ORCHID corpus.

The ORCHID corpus is now available for academic and research use. It can be viewed and downloaded from `http://www.links.nectec.or.th/ORCHID/`.

## 6. ACKNOWLEDGEMENTS

We would like to thank Ms. Thatsanee Charoenporn, a researcher at LINKS, for her enormously useful comments on revision of the original part-of-speech schema, and Ms. Virongrong Tesprasit and her team for help in preparing the corpus.

## REFERENCES

Bond, F. and Ogura, K. and Ikehara, S. 1996. Classifiers in Japanese-to-English Machine Translation *Proceedings of COLING'96*, Vol.1, pages 125–130.

Charoenporn, T. and 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.

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

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

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.

Muraki, K. and Sornlertlamvanich, V. and 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.

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.

Sornlertlamvanich, V. 1993. Word Segmentation for Thai in Machine Translation System. *Machine Translation*, NECTEC, pages 556–561, (in Thai).

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

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

Sornlertlamvanich, V. and Charoenporn, T. and Isahara, H. 1997. ORCHID: Thai Part-Of-Speech Tagged Corpus. *Orchid*, TR-NECTEC-1997-001, NECTEC, pages 5–19.

Sornlertlamvanich, V. and Takahashi, N. and Isahara, H. 1998. Thai Part-Of-Speech Tagged Corpus: ORCHID. *Proceedings of Oriental COCOSDA Workshop*, pages 131–138.

Sornlertlamvanich, V. 1998. *Probabilistic Language Modeling for Generalized LR Parsing.* Doctoral dissertation, TR98-0005, Tokyo Institute of Technology, Tokyo, Japan.

Viterbi, A. J. 1967. Error Bounds for Convolutional Codes and an Asymptotically Optimum Decoding Algorithm. *Proceedings of IEEE Transactions on Information Theory*, pages 260–269.