TaLAPi – A Thai Linguistically Annotated Corpus for Language Processing

AiTi Aw, Sharifah Mahani Aljunied, Nattadaporn Lertcheva, Sasiwimon Kalunsima
Institute for Infocomm Research
1 Fusionopolis Way #21-01
Connexis (South Tower)
Singapore 138632
E-mail: {aaiti,smaljunied,lertchevan,kalunsimas}@i2r.a-star.edu.sg

Abstract

This paper discusses a Thai corpus, TaLAPi, fully annotated with word segmentation (WS), part-of-speech (POS) and named entity (NE) information with the aim to provide a high-quality and sufficiently large corpus for real-life implementation of Thai language processing tools. The corpus contains 2,720 articles (1,043,471words) from the entertainment and lifestyle (NE&L) domain and 5,489 articles (3,181,487 words) in the news (NEWS) domain, with a total of 35 POS tags and 10 named entity categories. In particular, we present an approach to segment and tag foreign and loan words expressed in transliterated or original form in Thai text corpora. We see this as an area for study as adapted and un-adapted foreign language sequences have not been well addressed in the literature and this poses a challenge to the annotation process due to the increasing use and adoption of foreign words in the Thai language nowadays. To reduce the ambiguities in POS tagging and to provide rich information for facilitating Thai syntactic analysis, we adapted the POS tags used in ORCHID and propose a framework to tag Thai text and also addresses the tagging of loan and foreign words based on the proposed segmentation strategy. TaLAPi also includes a detailed guideline for tagging the 10 named entity categories.

Keywords: Word Segmentation, Natural Language Processing, Thai Language Resources

1. Introduction

There are only a few Thai language text resources open for public access currently. The ORCHID corpus (Sornlertlamvanich, 1997) was the first Thai POS-tagged corpus developed in collaboration between Communications Research Laboratory (CRL) of Japan and National Electronics and Computer Technology Center (NECTEC) of Thailand. It contains about 2 MB (or about 400,000 words) of the proceedings of an NECTEC annual conference. The BEST corpus (Kosawat, 2009) is a large Thai text corpus developed under the BEST (Benchmark for Enhancing the Standard of Thai language processing) project. It contains text from 4 different genres, namely, academic articles, encyclopaedic text, novel(s), and news. The texts in the corpus were manually segmented into words by linguists and comprise about 5 million words. This corpus is freely available for research and academic purposes.

Different strategies were adopted in the construction of the two corpora. The ORCHID corpus regards word segmentation as a task combined with POS tagging. Thus, wherever an element in a sequence is taggable with a different POS tag from its adjacent elements, that element is considered as a segmentable unit. On the other hand, the BEST corpus tends to be segmented based on smallest units, as long as the sequence of words are 'composite'. There are thus differences between the ORCHID and BEST corpora in terms of segmentation. For example |หนังสืออ้างอิง| (*references*) remains a single unit in the ORCHID corpus versus the segmented |หนังสือ| (*book*) and |อ้างอิง| (*refer*) in the BEST corpus. In addition, in the BEST corpus words are not POS-tagged and named entities are not segmented.

In these two corpora, no specific mention has been made to address loan and foreign word sequences in relation to word segmentation task. We view this as an area that requires a systematic approach because loan and foreign words normally comprise more than one minimal unit and annotators are often faced with the decision whether to segment or not to segment the sequence. This difficulty is particularly apparent in named entity phrases. In the BEST corpus (Boriboon et.al, 2009), segmentation for non-native sequences follow donor language segmentation rules with the following proposed segmentations[1]. Non-native sequences are not explicitly addressed in the ORCHID corpus.

i. *Unadapted Foreign Phrase*: |Le| |français| |pour| |les| |élèves| |thaïs| (|*French*| |*for*| |*students*| |*Thai*|)
ii. *Transliterated English Phrases*: |มาย| |แฟร์| |เลดี| (|*My*| |*Fair*| |*Lady*|); |อเมริกัน|ฟุตบอล| (|*American*|*football*|)
iii. *Adapted Phrase* (adapted the Thai language structure): |ทีม|ฟุตบอล| (|*team*|*football*|)

TaLAPi is a comprehensive corpus with complete word segmentation (WS), part-of-speech (POS) and named entity (NE) information for the development of Thai word segmentation, POS and named entity annotation tools. In this corpus, we propose slightly different but more comprehensive WS and POS annotation schemes to address the different usages of loan and foreign words in real-world data. The same scheme also serves the annotation of Thai native words quite well, with the aim of providing more useful information for the further processing of words and their linguistic-syntactic

[1] "|" indicates segmentation boundary

contexts.

This paper reviews prior work on Thai segmentation with reference to foreign and loan word treatment. We also discuss the types of multiple unit foreignisms and explain our segmentation approach based on structural-grammatical integrity. We introduce the linguistic tags of this corpus followed by the special considerations for named-entity tagging. We then conclude the paper with the corpus evaluation.

2. Related Work

Discussions of Thai language processing (Haruechaiyasak et.al, 2008; Supnithi et.al, 2004) often focused on the difficulty of word segmentation in terms of the absence of orthographic indicators as word boundaries. Aroonmanakun (2007) pointed out the inadequacy of segmenting on the basis of lexeme as a unit, which is likely to result in too many larger-than-word units. This led to the proposal of a concept of minimal integrity unit (the smallest word which has its own meaning) being the target of Thai word segmentation tasks, leaving the larger multi-word detection (such as name-entity) to a separate task. There was also the idea of uninterruptability (Chaicharoen, 2002) as a good characteristic for Thai wordhood whenever more than one minimal unit is detected, while distinctions between context-dependent and context-independent segmentation (Supnithi et.al, 2004) were also highlighted.

Works on loanwords in Thai have included studies on the main donor languages, namely Sanskrit, Pali, Khmer, Chinese and more recently English. These include studies on phonological changes (Kang, 2010) and tonal rule adaptations (Potisuk, 1999). The description of loanwords in Sujaduk (2006) classified Thai loanwords into the followings, but excluded the appearance of words written as foreign script.

i. ***Transliterations****: e.g.* |แบงก์| (*bank*) where the whole word แบงก์ is transliterated

ii. ***Loan blends*** *: e.g.* |ภาษา|ลาติน| (|*language*|*latin*|) where ภาษา in native Thai is mixed with ลาติน, which is transliterated.

iii. ***loan translations****: e.g.* |บาท|วิถี| *(*|foot|path|) where both Thai words are translated from English

iv. ***loan shifts****: e.g.* |มลายู| *(*|*malayo*|) where the word is shifted from the original sound and spelling to a similar existing word in another language

3. The Structural-Grammatical Approach for Thai Word Segmentation

The 4.2 million-word TaLAPi corpus comprises 5,489 news articles (a total of 3,179,318 words with an average of 579 words per article), and 2,720 articles (a total of 1,043,471 words with an average of 383 words per article) in the entertainment and lifestyle (NE&L) category. We adopted the minimal integrity consideration as the baseline for segmentation, but post-modified the minimal units for compositionality and structural integrity for all elements. This is to obtain a segmented corpus that essentially reflects semantic accuracy by making non-composite elements into one unit, and identifying foreign elements based on their level of adaptation in Thai texts.

3.1 Foreign and Loan Word Sequences in Thai

In this task, we defined multiple-unit foreign and loan words as those sequences which are regarded in the donor or recipient language as containing more than one minimal identifiable unit or word. We kept our study to noun and noun phrases, as they have been found to have higher borrowability than other word classes (Hock, 1991; Morimoto, 1999) and classified the sequences into 4 types.

Type A (Completely Unadapted**):** appears in non-Thai script (also referred to as foreign words)
Type B (Phonologically Adapted): transliterations only, with no ordering changes or additional Thai words (also referred to as loan word transliterations in Sujaduk (2006))
Type C (Phonologically Adapted with Structural Adaptation): transliterated with some adaptations of Thai grammar
Type D (Others): *Long sequences displaying various combinations of the above.*

A study on Thai journalistic writing (Kapper, 1992) noted that Thai borrowings from English appear at about an average of 3.43 words per news article. In our task, observations revealed that the above 4 types of multiple units appear 25 times among every random 100 sentences. In other words, there is one foreign word sequence in every 4 sentences. This is much higher than the observations made in 1990s and constitutes evidence of foreign language adaptation resulting from language contact commonly found in today’s language use. The NE&L domain shows an even higher frequency of 33 for every 100 sentence. Among all four categories, Type B is the most common (68%) multiple foreign unit (from English) seen in the corpus.

3.2 Segmentation Methodology

Here, we extended the work done in the BEST corpus to provide a more extensive framework to segment loan and foreign words for addressing two main issues. They are (1) the unclarity of incoming foreign sequences’ minimal units, and (2) the compositionality of words within these sequences. As mentioned above, the segmentation approach is first considering minimal integrity unit and then followed by compositionality assessment.

Segmentation for Type A is straight forward as it refers to all untransliterated foreign word sequences. Common examples of these from English are titles of movies and books, product names, and technical terms. These appear in Thai texts exactly as they do in the original donor language. Examples include ‘Give Me All Your Luvin’, and ‘Near Field Communication’. These foreign words appearing in non-native Thai scripts would not be segmented from each other.

Sequences of Type B include transliterated non-Thai names, company names, common nouns and many others. Apart from compositionality and true compounding, segmentation of Type B would also depend on their level of conformity to Thai NP structure. For English human names, we proposed segmentation using space as the delimiter as in |ไมเคิล| |แจ็คสัน| (|Michael| |Jackson|). However, for other named entities which do not conform to typical Thai head-first NP structure, we proposed the whole transliterated unit to be treated as one, as in |คอล เซนเตอร| (|call center|) and |ช็อปปิ้งมอลล์| (|shopping mall|), as they reflect an English head-last NP structure.

Type C refers to loan words which are transliterated and have undergone some kind of structural adaptation. For example, เกมออนไลน์ (*online game*), appears at first glance to be a transliterated sequence from English without additional adaptation. However, we can say with a degree of certainty that the original English sequence is ‘online game’ and not ‘game online’. We proposed to segment sequences of Type C, which are compositional, into their units, i.e. |เกม|ออนไลน์| (|game|online|). Another example is คลิปวิดีโอ (*video clip*), displaying a head-first Thai-style structure of English word ‘video clip’. This would be further segmented as |คลิป|วิดีโอ| (|clip|video|).

Thai noun phrases have been described as having clear, ‘well-behaved’ head-first structure (Diller, 1993), with qualifiers of the noun phrase following the head. It is therefore relatively easy to know if a transliterated or adapted foreign language phrase violates this structure as described in the above examples. However, in real data, the relationship between borrowed elements may not be so easily seen, particularly in named entities. Type D refers to foreign sequences in longer NPs, particularly in named entity phrases containing foreign word elements with combinations of Type A-C appearing with native Thai words in the presence of spaces in the entity spans.

The principle of Type D segmentation is (1) not to segment the named entity which does not follow the typical Thai head-first structure even if there are spaces or native Thai word in the name entity; (2) segment according to space if the entity containing spaces between element and words relation within the element is unclear (this includes person names); (3) treat non-Thai script as one unit within the name entity span. Below are some examples which show unclear relationship between its elements within the entities. It can be seen that some phrases clearly contain entity indicators in native Thai (e.g. ‘hotel’, ‘company’), while some less obviously (e.g. ‘entertainment’, ‘international’, ‘motor’), while others without overt entity indicators (e.g. ‘Tesco Lotus’).

i. |***โรงแรม***|แมนดาริน| |โอเรียนเต็ล| (|Hotel|Mandarin| |Oriental|)
   - ***โรงแรม***(hotel) in native Thai

ii. |เทสโก้| |โลตัส| (|Tesco| | Lotus|)
   - relationship within element is unclear

iii. |บริษัท| |ฟอร์ด| |มอเตอร์| |จำกัด| (|Company| |Ford| |Motor| |Limited)
   - Thai head-first structure

iv. |แคนทารี บีช เขาหลัก โฮเทล วิลล่า แอนด์ สวีทส์| (|Kantary beach Kaoluk Hotel Villa and Suite|)
   - transliterated without Thai head-first structure

v. |โตโยต้า| |Camry| (|Toyota| |Camry|)
   - non-Thai script as separate unit

3.3 Benefits to Thai Language Processing

We believe the proposed structural-grammatical approach provides a more holistic approach for Thai word segmentation and benefit the development of Thai language processing systems in various ways. First, it discriminates between foreign sequences according to their level of integration into the Thai language. Second, by treating those un-integrated, transliterated sequences which have different composition from Thai nouns as one unit, such as |เครดิตการ์ด| (|credit card|), it helps in the analysis of Thai structure. It can also help to reduce ambiguities in the analysis since post-modified qualifiers are more likely to be modifying the whole unsegmented unit, rather than any individual element. This treatment can also be useful for improving machine translation accuracy of longer phrases.

In addition, by forcing sequences of Type B displaying structurally violating units to be available only as multiword units, it helps to generate better quality dictionary entries and reduce the possibilities presented to automatic segmentation. While at the same time, keeping sequences that fit into Thai structure in their minimal integrity units encourage annotation consistency based on lexical unit-hood.

4. TaLAPi POS Tags

To reduce the ambiguities in POS tagging and better reflect the usage of loan word and foreign words in Thai text, we adapted the POS tags in ORCHID and proposed a framework to tag all words in corpus based on the above segmentation strategy.

4.1 Differences in ORCHID and TaLAPi POS Tags

The design of TaLAPi tags aims to support the syntactic analysis of Thai language for language applications such as named entity extraction and machine translation. The ORCHID corpus utilizes 14 POS’s with 47 subcategories. Driven by the view of enhancing Thai analysis, we found some of these subcategories can be shared while some desired categories are unavailable. Thus the ORCHID POS tags were revised to 12 categories with 35 subcategories.

4.1.1 Nouns and Pronouns (NN, NR, PPER, PINT, PDEM)

In the category of noun, we proposed 2 subcategories of common noun (NN) and proper noun (NR). ORCHID has 6 noun subcategories - proper noun, cardinal number, ordinal number, label noun, common noun and title noun. We found the syntactic role of number to be clear in real data and provided a separate tag for number in TaLAPi. The PPER, PINT and PDEM constitute the 3 (personal, interrogative, and demonstrative) pronoun subcategories used in TaLAPi, roughly corresponding to ORCHID's pronoun set, with the exception of the reflexive pronoun.

4.1.2 Reflexive Words (REFX)

We proposed REFX as a new category and not under the pronoun category to annotate words like กัน (*together*), ซึ่งกันและกัน (*each other*), กันและกัน (*each other*), ต่อกัน (*each other*), as we felt their function to be rather different from the other pronouns, often appearing to modify verbs reflexively or reciprocally.

4.1.3 Determiners (DPER, DINT, DDEM, PDT)

TaLAPi's determiner set of 4 subcategories was simplified from ORCHID's subset of 9, with the latter focusing on the representation of definiteness. Our subcategories are more indicative of the word-type component of determiner words, and include possessive pronouns used as determiners (DPER). Here is an example of a DPER, เธอ (*your*), attached to a head noun:

> |ฉัน/PPER|ชอบ/VV|กระเป๋า/NN|เธอ/DPER|
> (|I|like|bag|your|).

We also recognize the need to reflect the relative position of determiners in Thai noun phrases. Thus, the pre-determiner (PDT) subclass was introduced in TaLAPi, in addition to the demonstrative (DDEM) and interrogative (DINT) words functioning as determiners in NPs.

4.1.4 Noun-modifying Adjectives and Attributive Verbs (JJA, JJV)

Adjectives and verbs in Thai appear in predicate positions in clauses as well as within NPs, post-modifying their heads. To capture the latter, which have an important role in the analysis of Thai noun phrase chunking especially, we introduced JJA and JJV to indicate adjectives and verbs functioning as noun modifiers respectively. This is to contrast these words when they function as predicates in clauses as discussed in 4.1.5.

4.1.5 Verb and Adjectives as Predicates, and Auxiliaries (VV, VA, AUX)

ORCHID has three subcategories under verb. They are active verbs, stative verbs and attributive verbs. Our classification of verbs applies only to verb and adjective words that appear as predicates in Thai clauses. This allows us to distinguish the predicative and modifying function of verbs and adjectives to facilitate further structural analysis. In addition, we collapse the semantic distinctions of 'action' and 'state' verbs in ORCHID, the main verb categorization in ORHCID, into one class VV in TaLAPi. We also treat ORCHID's attributive verb (VATT), when used as predicate, as VA (adjective predicate in clauses) in TaLAPi. Moreover, we proposed one AUX category (AUX), while ORCHID has 5.

4.1.6 Adverb and Negative Word (ADV, NEG)

Same as ORCHID, TaLAPi has classes for negator words (NEG) and adverbs (ADV). However, ORCHID has 4 subdivisions of adverbs, while TaLAPi has none. Sentential and normal adverbs are not distinguished in TaLAPi, while adverbs with prefixes and iterative markers (ORCHID's ADVP and ADVI respectively) are treated as separately segmented elements and require no new categories.

4.1.7 Classifiers (CL)

TaLAPi has only one classifier (CL) class as compared to 5 in ORCHID.

4.1.8 Numbers (OD, CD)

As mentioned earlier, numbers have a clear syntactic role in Thai text. We introduced two tags to annotate words that indicate quantity (CD), and numbers indicating non-quantity (OD) as in dates, itemized lists, phone numbers, etc. This is different from ORCHID which has 4 subcategories for number.

4.1.9 Prefixes (FXN, FXAV, FXAJ, FXG)

ORCHID subcategorizes prefixes into nominal prefix and adverbial prefix which, in our view, does not cover all the prefixes' usage in Thai. Thus, we proposed two more prefix subcategories. They are group prefix (FXG) and adjectival prefix (FXAJ). Both are productive and can change the category of element that they are attached to. FXG is the prefix used to identify the group of noun, possibly changing a singular noun to plural. For example, the root word meaning 'betray' is changed into 'betrayers' when attached with the FXG พวก (phuak) in the phrase |พวก|ทรยศ| (|phuak|betray|). Likewise, FXAJ น่า (na) changes the root verb following it to derive the meaning of '*cute*' in |น่า|รัก| (|na|love|).

4.1.10 Linking Words (P, COMP, CNJ)

The treatments of linking words – words that coordinate, conjoin and complete – differ slightly across the 2 corpus. Whilst both have one category each for prepositions (without subcategorization), ORCHID regards conjunctions as having 3 subtypes, but TaLAPi recognizes conjunctions and coordinators as one (CNJ) category. Another difference is the category of complementizers (COMP) - that link nouns or verbs to their complement clauses or complement words – is not present in ORCHID. The COMP words in TaLAPi are treated as subordinating conjunctions and relative pronouns in ORCHID. As these COMP words function to complete preceding elements with following ones, we feel the need to have a separate category. They do not behave like other pronouns, and not quite like conjunctions which are restricted to clausal conjunctions and true coordinators only.

4.1.11 Foreign Words (FWN, FWV, FWA, FWX)

Another set of category that does not exist in ORCHID which is used in TaLAPi is the class 'foreign words'. With 4 subcategories, these refer to elements in the corpus which are not in Thai script and are not names or proper nouns (see section 4.2 below). An example is the English words in the sequence '*rule by law*' in the following context: |นิติกลวิธี/NN| |หรือ/CNJ| |rule by law/FWN |. We found such non-Thai items rather frequently and used in different ways in Thai sentences. As such, it is beneficial to keep track of the POS of these foreign words, with different subcategories as in noun (FWN), verb (FWV), adjective (FWA) and other POS (FWX).

4.1.12 Others (PAR, PU, IJ, X)

Both ORCHID and TaLAPi have one category each for punctuation (PU) and interjection (IJ), without subcategorization. We mapped the two classes of endings in ORCHID to the single particle class (PAR) in TaLAPi. Finally, we included a category (X) for words when none of the available tags are appropriate.

All of the above tags are sufficient for us to construct the TaLAPi corpus. Table 1 shows the distribution of POS tags in TaLAPi corpus.

4.2 Tagging of Loan Words and Foreign Words

The tagging of loan words and foreign words is made straight forward by our proposed segmentation strategy. The tagging follows the syntactic behavior of the words in Thai sentences. Words that appear in Thai sentences which are written in non-Thai script will be treated as foreign words. However, not all foreign words are tagged with foreign word tags. As mentioned in 4.1.11, foreign word tags are only used to tag regular non-proper nouns, verbs, adjectives and other category non-Thai script words. For example, |my/FWX|โลก(world)/NN| where '*my*' cannot be tagged as FWN, FWV or FWA, the tag FWX is used. Proper nouns that remain in the English script, such as [*GOOG*], *Apple*, *IBM*, would be tagged as NR. This is because some of these proper nouns have been adapted and used as part of the Thai language. The tagging of other loan words which are in Thai script, typically reflecting transliteration, would be tagged according to the behavior of the words in the Thai sentence, following the same principle as native Thai words.

4.3 POS Distribution in News and NE&L corpora

Table 1 compares the distribution of POS in NE&L and NEWS domain. As can be seen, NN has the highest number of use in both domains. Though we expect the news domain to have more named entities than NE&L, the percentages of NR in both domains are actually quite comparable. The reason for this is that phrases that make up named entities, like organization or division names, are comprised of individually segmented words which are not proper nouns, but in fact common nouns. For example, in such NE spans, words like 'ministry'(กระทรวง) and 'university' (มหาวิทยาลัย) are NN tagged, not NR.

POS	NE&L		News	
NN (non-proper noun)	250562	24.01%	859531	26.93%
NR (proper noun)	58904	5.65%	183096	5.58%
PPER (personal pronoun)	15250	1.46%	21681	0.71%
PINT (interrogative pronoun)	936	0.09%	5346	0.13%
PDEM (demonstrative pronoun)	4252	0.41%	5556	0.25%
REFX (reflexive word)	7725	0.74%	14869	0.46%
DPER (possessive to PPER)	1154	0.11%	861	0.03%
DINT (noun determiner)	1301	0.12%	1230	0.04%
DDEM (determiner)	12396	1.19%	39302	1.26%
PDT (quantifier determiner)	13784	1.32%	28963	0.84%
JJA (noun-modifying adjective)	39122	3.75%	67170	1.76%
JJV (noun-modifying verb)	14303	1.37%	67406	1.72%
VV (verb in predicate)	224048	21.47%	704963	22.88%
VA (adjective in predicate)	12656	1.21%	12590	0.30%
AUX (auxiliary verb)	58508	5.61%	162243	5.11%
ADV (adverb)	40094	3.84%	91428	2.94%
NEG (negative word)	12526	1.20%	36884	1.19%
CL (classifier)	19919	1.91%	60927	1.97%
OD (non-quantifying number)	7100	0.68%	37273	1.16%
CD (cardinal number)	11842	1.13%	51160	1.59%
FXN (noun-type prefix)	30932	2.96%	118023	3.66%
FXAV (adverb type prefix)	2509	0.24%	11245	0.37%
FXAJ (adjective type prefix)	1104	0.11%	1298	0.06%
FXG (group type prefix)	1378	0.13%	6981	0.27%
P (preposition)	64019	6.14%	196339	6.36%
COMP (complementizer)	43673	4.19%	166060	5.44%
CNJ (coordinator and clause conjunction)	45496	4.36%	144214	4.36%
FWN (noun in non-Thai script)	1165	0.11%	2963	0.11%
FWV (verb in non-Thai script)	37	0.00%	44	0.00%
FWA (adjective in non-Thai script)	44	0.00%	79	0.00%
FWX (other in non-Thai script)	639	0.06%	91	0.00%
PAR (particle)	5221	0.50%	8670	0.25%
PU (punctuation)	39964	3.83%	72677	2.27%
IJ (interjection)	837	0.08%	162	0.01%
X (others)	71	0.01%	162	0.01%
Total number of words	**1,043,471**		**3,181,487**	

Table 1: Distribution of POS Tags in TaLAPi corpus.

We can also observe distinct writing styles between the two domains by studying their POS distributions. The language used in NEWS has a higher level of formality and conciseness, and this is reflected by the fewer use of personal pronouns (PPER) and particles (PAR). Likewise, in the NEWS text, fewer numbers of VA, JJA, and ADV are found, which typically provide more detailed, evaluative, and sometimes elaborate descriptions to the topics. NE&L has more text covering conversations between an interviewer and an interviewee; thus, there are higher occurrences of PPER and PAR compared to NEWS.

As can be seen above, the distribution of the types of POS across the 2 text domains serves as a useful basis for the study of domain style analysis, not solely for the structural analysis of Thai to build language processing tools.

5. Thai Named Entity Annotation

Phanarangsan et al. (2006) proposed a simple named entity guideline on Thai, focusing on person name, title name, organization name and location name. As far as we know, there are no other guidelines available publicly for Thai named entity tagging. We performed a comprehensive study on our corpus and proposed a guideline to tag 5 categories of Thai named entity with 10 subcategories. This aims to cover the majority of the named entities found in the two text genres.

5.1 Principles of Named Entity Annotation

In our guideline, special attention has been made to the annotation of compound expressions, possession structure, abbreviations, common words used as proper noun, and proper names included as part of another named entity or another untagged noun phase. We found these areas have the most ambiguities among the annotators and provided detailed instructions to avoid inconsistencies during annotation. The general principles are:

i. Individual noun phrases in a compound expression will be tagged separately.
   - [|จังหวัด|ตาก|(|*province*|*Tak*|)]$_{\text{LOC}}$ และ(*and*)[|จังหวัด|แม่ฮ่องสอน|(|*province*|*Mae Hong Son*|)] $_{\text{LOC}}$

ii. Compound expressions with names separated by space will be tagged separately.
   - [|มหาวิทยาลัย|กรุงเทพ|(|*University*|*Bangkok*|)]$_{\text{ORG}}$ | | [|ประเทศ|ไทย|(|*country*|*Thai*|)] $_{\text{ORG}}$

iii. Multi-names expression in which there is elision of the head of one conjunct, should be tagged as single expression.
   - [|เกาหลี|เหนือ|และ|ใต้|(|*Korea*|*North*|*and*|*South*|)] $_{\text{ORG}}$

iv. Unified noun phrases in the compound expression will be tagged as a whole.
   - [|กระทรวง|เกษตร|และ|สหกรณ์|(|*Ministry*|*Agriculture*|*and*|*Co-operation*|)]$_{\text{ORG}}$

v. Possessor name or its possessory name will be tagged as part of the named entity only if they are part of the whole proper name.
   - [|สำนัก|ข่าว|กรอง|แห่ง|ชาติ|(|*agency*|*intelligence*|*of*|*national*|)]$_{\text{ORG}}$

vi. Common words used as names of named entity will be tagged, e.g. *Mueang* is a common word means city.
   - [|อำเภอ|เมือง|(|*district*|*Mueang*|)]$_{\text{LOC}}$

vii. Proper name included in another untagged noun phrase or to name another entity will be tagged.
   - |เทศกาล|ภาพยนตร์|(|*festival*|*film*|)[|เมือง|คานส์|(|*city*|*Cannes*|)]$_{\text{LOC}}$

viii. Entities transliterated or written in non-Thai script will be tagged.
   - [|นาตาลี| |เกลโบวา|(|*Natalie*| |*Glebova*|)]$_{\text{PER}}$

     [|มิสยูนิเวิร์ส|(|*Miss Universe*|)]$_{\text{DES}}$

5.2 Thai Named-Entity Tags

The five categories of Thai named-entity tags are described below. The tags were designed to keep semantic units as a whole as far as possible to facilitate tasks in machine translation.

5.2.1 Abbreviation (ABB)

We introduced ABB to tag short-forms of TTL, DES, PER, ORG, LOC and BRN, e.g. [|น.ส.|(*Miss*)]$_{\text{ABB\_TTL}}$.

5.2.2 Title and Designation (TTL, DES)

Phanarangsan et al. (2006) categorized titles, roles and appositives as title entities, whereas we categorized titles, roles and appositives into two subcategories based on whether that word is used as title or designator. It is because the role between the two entities is distinct. TTL is used for kinship relation and social status whereas DES for designators like job titles, profession and political positions, awarded titles, and professional and academic titles, e.g. [|นาย|(|*Mister*|)]$_{\text{TTL}}$[|อรุณ|(|*Aroon*)]]$_{\text{PER}}$, [|ศาสตราจารย์|(|*Professor*|)]$_{\text{DES}}$[|อรุณ|(|*Aroon*|)] $_{\text{PER}}$.

5.2.3 Proper Noun (PER, LOC, ORG, BRN)

Person, organization and location names are the main NE types under MUC-7(Chinchor, 1998) and Phanarangsan et al. (2006) also mentioned these three named entities. In our guideline, we added "brand" as one entity type because brand is an important category under NE&L domain. The following describes how the four NEs are being used.

- PER is the label for Thai or foreign human name.
- LOC is tagged when political boundaries, geographical names and construction names are mentioned in the text.
- ORG is for tagging names of government organization, names of non-government organization, office, union and coalition of governments.
- BRN is the label for names of brands, products, trademarks and television channels.

5.2.4 Quantities (MEA, NUM)

Phanarangsan et al. (2006) did not mention number expressions in their guideline. We included them in TaLAPi as they are common entities found in our corpus. MEA is used when the quantity is in percentages, standard units, ratios or phrases indicating capacity of things.

NUM indicates quantity of things, excluding cases tagged as MEA, e.g. |เงิน|(|*money*|)| |[|5| |บาท|(|*baht*|)]$_{\text{MEA}}$, |กระดาษ|(|*paper*|)| |[|5|]$_{\text{NUM}}$ | |แผ่น|(|*pieces*|).

5.2.5 Date and Time (DTM)

This category is to tag expressions indicating time and time period, e.g. [|17| |เมษายน|(|*April*|)]$_{\text{DTM}}$.

All other segmented words which are not categorized under the above 5 categories will be given the tag "Other" (O). Each entity suggested in TaLAPi corpus has its own importance and we believe the corpus is a good Thai resource for Thai NER tasks. Table 2 shows the distribution of NE tags in the TaLAPi corpus.

NE	**NE&L**		**News**	
ABB	1798	2.60%	48027	14.19%
TTL	3698	5.35%	27219	8.04%
DES	2299	3.33%	12370	3.65%
PER	22390	32.41%	64259	18.98%
LOC	10656	15.43%	42254	12.48%
ORG	7182	10.40%	54656	16.15%
BRN	4111	5.95%	942	0.28%
MEA	4824	6.98%	43612	12.88%
NUM	6754	9.78%	21949	6.48%
DTM	5363	7.76%	23219	6.86%
Total Number of NE	**69,075**		**338,507**	
NE Freq in the Corpus	**10.78%**		**18.33%**	

Table 2: Distribution of NE Tags in TaLAPi corpus.

5.3 NE Distribution in News and NE&L corpora

From Table 2, we can make a few observations. Entity ABB, TTL, ORG, and MEA have higher occurrences in the NEWS domain as compared to the NE&L domain. The differences in occurrences among PER, LOC, and ORG in NEWS domain are also not as much as in the NE&L domain, which sees a much higher frequency in PER than LOC or ORG. These can be explained as articles in NEWS domain usually have more factual information where person, place, organization and even quantity are important sources of information to be referred to in the content. Compared to the NEWS domain, we see a much higher occurrence for entity type BRN in NE&L domain. This can be explained as product names and brands are usually mentioned in articles for entertainment, technology and leisure, especially promotional and marketing write-ups. The findings clearly reflect the domain difference in data distributions as shown in the following examples.

NEWS domain
|ยอด(*total*)|ขาย(*sale*)|รถ(*car*)|**[**โตโยต้า(*Toyota*)**]**$_{\textbf{ABB\_ORG}}$|ใน(*in*) |ตลาด(*market*)|รวม(*total*)|ช่วง(*period*)|**[**9| |เดือน|(*months*)**]** $_{\textbf{MEA}}$| อยู่(*be*)|ที่(*at*)|**[**1.99| |แสน(*hundred thousands*)**]**$_{\textbf{NUM}}$ |คัน|, …

NE&L Domain
|ลูกค้า(*customer*)|ที่(*who*)|จอง(*book*)|**[** บีเอ็มฯ(*BMW*)**]**$_{\textbf{ABB\_ORG}}$| **[**X1**]**$_{\textbf{BRN}}$ |และ(*and*) | **[** ซีรีส์(*series*)|3**]**$_{\textbf{BRN}}$ |เลือก(*choose*)|รับ(*receive*) |ชุด(*set*) |กอล์ฟ(*golf*) | Set 1| หรือ(*or*) | Set 2|.

6. Annotation Process and Corpus Evaluation

The annotation of the TaLAPi corpus was done by 11 annotators on the NE&L domain and 10 annotators on the NEWS domain over a period of 18 months. Each annotator was assigned a specific task based on their specialisation on WS, POS or NE. The annotations were randomly validated by two Thai linguists who would provide feedback to the team on the discrepancies found. The team would then re-check the errors and correct the annotations if necessary. Annotators who produced more consistent and reliable results were given more tasks during the re-annotation stages.

Annotation evaluations were performed manually. We randomly selected 21 articles with 2,090 words done by 4 different annotators from the NE&L domain. The manual evaluation shows an accuracy of 97.58% on WS, 92.58% on POS and 98.79% on NE. A separate evaluation was performed on 10 randomly selected articles with 5,281 words on the NEWS domain. The manual evaluation shows an accuracy of 97.58% on WS, 90.02% on POS, and 96.65% on NE. We also performed an inter-annotator agreement check between our two evaluators using 5 articles with 9,264 words. We achieve an agreement of 99.85%, 99.03% and 99.46% for WS, POS and NE respectively.

7. Conclusions

This paper discusses the different types of English foreign and loan word sequences in noun phrases discovered during our annotation of TaLAPi corpora. We introduced the concept of structural-grammatical integrity in Thai word segmentation and discuss its benefits for improving the development of language processing tools. We also present our POS and NE tags and discuss their distributions in two text genres.

In the near future, we would like to compare our annotation accuracy with an automatic tool trained from the TaLAPi corpus. We would also like to perform open tests on automatic tools trained on TaLAPi and ORCHID corpus. In addition, we will continue to further improve the annotation accuracy on all areas. The TaLAPi corpus will also be used to study transfer learning, so as to boost cross-domain named entity recognition performance.

8. References

Aroonmanakun, W. (2007). Thoughts on Word and Sentence Segmentation in Thai, In *Proc. of the Seventh Symposium on Natural Language Processing*, 85-90.

Chaicharoen, N. (2002). Computerized Integrated Word

Segmentation and Part-of-Speech Tagging of Thai, MA Thesis, Dept. of Linguistics, Chulalongkorn University.

Chinchor, N. (1998). MUC-7 Named Entity Task Definition Dry Run Version, Version 3.517 September 1997. In *Proc. of the Seventh Message Understanding Conference (MUC-7)*. Fairfax, Virginia: Morgan Kaufmann Publishers, Inc. URL: ftp://online.muc.saic.com/NE/training/guidelines/NE.task.def.3.5.ps.

Diller, A. (1993). "Diglossic Grammaticality in Thai", in William Foley (ed.), *The Role of Theory in Language Description*, 393-420.

Haruechaiyasak, C., Kongyoung, S., & Dailey, M. (2008). A Comparative Study on Thai Word Segmentation Approaches. In *Proc. of Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology*, 1:125-8.

Hock, H.H. (1991). Principles of Historical Linguistics, (2nd ed.). Mouton De Gruyter.

Kang, Y. (2010). The emergence of phonological adaptation form phonetic adaptation: English loanwords in Korean, Phonology, 27(2):225-53.

Kosawat, K., Boriboon, M., Chootrakool, P., Chotimongkol, A., Klaithin, S., Kongyoung, S., et al. (2009). "BEST 2009 : Thai Word Segmentation Software Contest", In *Proc. of SNLP'2009*, Bangkok, Thailand, October, 2009.

Morimoto, Y. (1999). Loan Words and Their Im-plications for the Categorial Status of Verbal Nouns. in S. Chang, L. Liaw, and J. Ruppenhofer (eds.) In *Proc. of the 25th Annual Meeting of the Berkeley Linguistics Society*, 371-82.

Phanarangsan, K., Arnold, S., Mandel, M., & Walker, C. (2006) . Simple Named Entity Guidelines Version 6.4-Thai. (n.p).

Potisuk, S., Harper, M.P., & Gandour, J. (1999). Classification of Thai tone sequences in syllable-segmented speech using the analysis-by-synthesis method, *IEEE Trans. on Speech and Audio Processing*, 7(1):95-102.

Sornlertlamvanich, V., Takahashi, N., & Isahara, H. (1997). Building A Thai Part-of-Speech Tagged Corpus (ORCHID), *The Journal Of the Acoustical Society of Japan* (E) 20(3):140-89.

Sujaduk, W. (2006). A Study of English Loan Words in Thai Newspapers, BA Thesis, English Language Department, Rangsit University, Thailand.

Supnithi, T., Kosawat, K., Boriboon, M., & Sornlertlamvanich, V. (2004). Language Sense and Ambiguity in Thai, In *Proc. of the 8th Pacific Rim International Conference on Artificial Intelligence (PRICAI2004) Workshop in Language Sense and Computer*.