Language, Information and Computation(PACLIC 11), 1996, 109-118

Ambiguous (((Par(t)(it))((ion))(s))(in)) Thai Text

Doug Cooper
Center for Research in Computational Linguistics, Bangkok
doug@nwg.nectec.or.th http://seasrc.th.net

Abstract

Despite the importance of segmentation to a variety of software applications, almost nothing is known about the characteristics or distribution of ambiguous *partitions* (eg. *to_pend* vs. *top_end*) in Thai text. By using special-purpose code to investigate a large (~400K word) text corpus, we were able to extract 36,267 such sequences, involving 9,253 distinct examples. Of these, a little more than two-fifths involved genuinely ambiguous partitions. We classify partitioning problems into distinct categories, report on many of their statistical and lexical characteristics, and describe heuristics for choosing the correct partition that do not depend on the availability of a large segmented corpus.

1. Introduction

Thai writing does not use spaces to segment text into words. While open text contains many obvious *separation* points (**bigdog** vs. **big_dog**), and a smaller group of questionable *bind* points that are usually permissible either way (**toolbox** vs. **tool_box**), there is inevitably a residue of ambiguous *partition* points (**to_pend** vs. **top_end**) for which computer segmentation is essentially random. This causes difficulty for many software applications: line-breaking, spell-checking, machine-assisted translation, text-to-speech, optical character recognition, full-text indexing, corpus-based dictionaries, etc.

Yet despite the importance of segmentation, little is known about the characteristics of ambiguous partitions in Thai, or its orthographic cousins Khmer, Lao, and Burmese. Work has been slow and progress poor due to a lack of formal, concrete analysis. We know the problem's gross characteristics, and the general direction of solutions, but there are few theories to guide the way or allow comparison of research results.

This paper describes experiments carried out on a 2 megabyte (roughly 400,000 word) Thai corpus. We collected nearly ten thousand distinct alternative segmentations of at least two words in length. Of these, a little more than two-fifths involved genuinely ambiguous partitions. We classify partitioning problems into distinct categories, report on statistical and lexical characteristics of ambiguous partitions, and describe heuristics for disambiguating that do not depend on the availability of a large segmented corpus.

Our results make several contributions to understanding Thai text segmentation. First, we categorize breakpoints in a way that distinguishes between choices that are and are not semantically significant, and show how to collect them automatically and consistently. Second, we find that ambiguous instances are fairly rare (roughly 5% of word break opportunities), and have a pronounced Zipfian distribution — a relatively small number of circumstances produce a great deal of ambiguity; and show ways of collecting low-frequency items that exhibit the same behavior. Third, we find that contrary to the canonical examples, resolving ambiguity does not usually depend on knowing or understanding the context it occurs in. Finally, we suggest new methods — stop nodes, go collocates, and analysis of hidden 'swing strings' — to aid in automated disambiguation.

2. Prior Work

There is extensive literature on text segmentation for Asian languages; those for Chinese are typical (Wu 1993, Chiang 1992, Chang 1993). Approaches to segmenting Thai text are surveyed in (Vonvgipanond 1993, Sornlertlamvanich 1993, Wuwongse 1993). In

general, dictionary-based maximal matching is followed; the segmentation that contains the fewest words is selected as correct. See (Haas 1942, 1946, 1964, Noss 1964, Luksaneeyanawin 1984, Vongvipanond 1992) for discussion of underlying linguistic issues.

While Chinese and other languages continue to make incremental improvements (eg. Maosong 1995), the literature on more advanced approaches in segmenting Thai is nearly non-existent; a notable exception describes a Viterbi-based approach to using statistical information derived from grammatical tags (Pornprasertsakun 1994), but even with restricted input grammar, results were poor. More recently (Kawtrakul 1995, 1996) combines various statistical and grammar-based methods; these tend to depend on a training corpus, and report testing only on a relatively small (~200 sentences) dataset.

Aside from frequent citation of canonical examples of ambiguous partitions, we could find no English or Thai-language literature that specifically addressed the partitioning problem, or attempted to classify different kinds of ambiguity in any way. Moreover, the large text corpora needed for more sophisticated approaches to the problem are not available; even the text corpus we used is relatively small, and contains a considerable amount of highly specialized text (eg. government documents, textbooks).

3. Methodology

Our 2 megabyte test sample consisted of 42 selections of hand-segmented, grammatically tagged Thai text (LINKS). The original text was split into some 415,844 words over 53,242 lines, leaving 362,602 potential error points. We removed spaces and tags, then replaced English text, numbers, and punctuation (unambiguous breakpoints) with newlines. A dictionary-based method resegmented the text, generating all possible parse trees in the process. We intentionally used a very large word list — over 70,000 entries, including all words from the text sample — to maximize opportunities for ambiguous partitions, and to ensure that every sentence would be segmentable.

Finally, special-purpose software selected outcomes that involved alternative partitions at least two words long. Given the string **topend**, we would select **top_end / to_pend** as an ambiguous partition. However, given **toolbox**, we would not choose **toolbox / tool_box** as alternatives. With a few notable exceptions (น้ำ_ดี nám dii = *good water* vs. น้ำดี námdii = *bile*), these are not open to ambiguous interpretation unless the context is at least three words long (which gives the central word the opportunity of binding either left, right, or not at all). Moreover, the exocentric exceptions should be found in any ordinary dictionary, while very, very large numbers of unambiguous compounds are an inescapable artifact of any large corpus-based word list.

This procedure described above produced some 36,267 candidate sequences, of which 9,253 were distinct (available on-line, along with most of the derived data dis-

Class	Type	Ambiguous	Example
Lexical	*Partition*	*yes*	มากกว่า = มาก กว่า | *มา กก ว่า
Contextual	*Partition*	*yes*	ความจำเป็น = *ความจำ เป็น | ความ จำเป็น
Contextual	*Bind*	*maybe*	ช่างพูด = ช่างพูด |ช่าง พูด (*good talker* or *the artisan said*)
Two-way	*Bind*	*no*	ที่จะต้อง = ที่จะ ต้อง | ที่ จะต้อง
One-way	*Bind*	*no*	ทรัพย์สิน = ทรัพย์ + สิน

Table 1 Kinds of segmentation decisions. Usually, the go/no go choice *Can a newline be inserted here?* is applied in strictly local terms to guide classification; it changes the local meaning of partitions, but not that of binds. Contextual binds, in contrast, are only potentially ambiguous when considered in a larger context; eg. for translation. Because they have both meanings when written either way, it is not clear that these are segmentation decisions at all.

cussed here, at the Southeast Asian Language Data Archives, http://seasrc.th.net/sealda). We investigated three groups in detail: the most frequent 5%, 5% selected at random from the remainder, and 5% taken at random from single-appearance entries.

4. Categorizing Segmentation Decisions

Our first concern was to distinguish between segmentation points that affected subsequent applications of the text, and those that did not. We derived two basic classes from the data: ***partitions*** that did affect sentence semantics, and ***binds***, which did not.

Alternative partitions involve two distinct sequences of words or compounds. They fall into two roughly equal classes: *lexical* partitions involving isolated letters, and *contextual* partitions involving full words or affixes. Excepting intentional pun-like constructions, proper partitions (in context) can always be chosen correctly and consistently. Binds, in contrast, tend to involve alternative ways of considering serial constructions. While the meaning of binds is not ambiguous, it is difficult to label segmentation decisions as correct or not because the alternatives do not affect semantics.

Our basic test for class membership was whether inserting a newline affected local semantics. We intentionally ignored a transitional class of *contextually ambiguous binds*, in which an affix binds to its neighbor, but still permits a newline to be inserted. We ignore these in the present analysis because they are essentially 'phantom' segmentations whose existence depends on the needs of subsequent applications. In summary:

— *Lexically ambiguous partitions* break on sub-word boundaries, and yield alternative sequences of entirely different words.

มากกว่า = มาก กว่า | *มา กก ว่า
*more than = more+than | *come hug (says) that*

— *Contextually ambiguous partitions* break on word or affix boundaries. They typically involve affixes that can bind either left or right, or exocentric compounds.

ความจำเป็น = *ความจำ เป็น | ความ จำเป็น
*need (n) = *memory+to be | nominalizer+must*

— *Two-way binds* occur when a central affix binds to either its left or right neighbor without significantly affecting the meaning of the phrase, eg:

ที่จะต้อง = ที่จะ ต้อง | ที่ จะต้อง
which will have to = which will+must | which+will have to

— *One-way binds*, which were not produced by our selection method, but were inferred from the data, are typically *endocentric compound* constructions whose meaning is easily derived from their constituents; eg. คำซ้อน kʰam sɔ́ɔn, or overlapping words; terms that have similar meanings but different origins or euphonic sounds.

	Most frequent 5%	Random 5% (excluding most frequent)	5% (from singles)	Actual, top 5%	Estimated in text
Sample size	460	460	460		
Lexical	95 = 21%	115 = 25%	113 = 24.5%	4,335	2.5%
Contextual	95 = 21%	97 = 21%	97 = 21%	4,667	2.5%

Table 2. Distribution of ambiguous partitions. The top 5% of the sample accounts for just over half of the actual appearances in the corpus; we estimate frequency in the corpus by doubling the actual counts, then dividing by the number of decision points (~363K).

ทรัพย์สิน = ทรัพย์ + สิน
wealth = wealth (Pali/Sanskrit) + wealth (Thai)

— *Contextually ambiguous binds* were also inferred. For example, verbs like ไป (*go*) and มา (*come*) have common meanings, but are also used as auxiliaries that indicate the manner, duration, intensity, tense, etc. of other verbs. However, in writing and in most applications, a newline may be inserted between each word without changing sentence semantics, or losing the information required to bind them correctly if a later application (like translation) requires it:

กิน ไป แล้ว = กินไปแล้ว | กิน ไปแล้ว
eat+go+already = ate it all up | ate, and then left

Assigning terms to specific categories was not overly difficult. Only three of the lexically ambiguous terms were not easily classed; ie. it was not obvious whether a term was a new word or an exocentric compound. In 36 cases that appeared to be contextually ambiguous, it was not entirely clear whether a phrase involved true contextual ambiguity or merely two-way binding. Nevertheless, in almost all cases it is possible to avoid inserting a space or line-break incorrectly. These are somewhat subtle points, eg: ทางการ ศึกษา means *the way of doing + education*, while ทาง การศึกษา means *the way + of educating*; the latter alternative is more generally correct in writing.

5. Discussion

Distribution The 36,267 candidate sequences demonstrated a strongly Zipfian distribution: 85 distinct forms (<1%) accounted for 25% of the candidates, 50% of the appearances were accounted for by just under 5% of the distinct forms, and well over half of the distinct forms appeared just once each. Actual counts of the partition types are summarized in table 2; figure 3 gives an idea of the implications of the numbers.

Situations in which incorrect partitioning changed the meaning of the sentence were far less common than generally thought. Estimates based on the percentage of forms in the three groups we investigated in detail, and on projection of the actual counts of the most frequent forms, both indicate that lexically and contextually ambiguous partitions probably account for about 5% (± 2.5%) of the total space insertions.

The small number of truly ambiguous partitions indicates that performance of segmentation algorithms must be measured in isolation, focusing specifically on their ability to resolve semantically significant partitions. Minor differences in counting even insig-

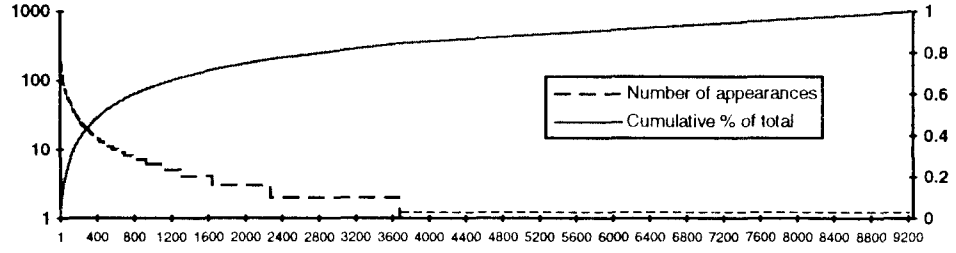

Figure 3 A relatively small number of distinct candidates account for a large number of potentially ambiguous partitions. The most frequent candidate appears 365 times; the top 452 account for half of the total in the text, and about 5,500 appear just once. This strongly Zipfian distribution implies that while it is impossible to anticipate all ambiguous partitions, dealing with the most common is a worthwhile investment of time.

nificant one-way binds can overwhelm count statistics; a major improvement in dealing with true ambiguity may be masked by trivial differences in reporting methods. See (Sproat 1994, Wu 1994, Cooper 1996) for more on the reporting problem.

Their distribution also implies that, contrary to the usual analytical and statistical approaches, brute-force methods can give good results. In this case, Zipf's law makes statistics work for us: dealing with a fraction of forms solves a majority of potential errors. Moreover, given the lack of context dependency (see below) required for disambiguation, we are able to gather precisely the data we need — ambiguous partitions — from unsegmented text, and need not depend on segmented corpora.

One cautionary note involves the very large dictionary we used to spot potential ambiguity. While detection of lexically ambiguous partitions does not appear to be significantly affected by the addition of lexical phrases and compounds to the segmentation word list, arguments can be made that contextually ambiguous partitions are both over- and under-reported; we continue to investigate.

Context Dependency in Disambiguation Our next concern was finding the degree to which correct resolution of ambiguous partitions depends on context. If recognizing the right partition is not context sensitive, then *a*) partitions can be decided wholesale, and *b*) correct outcomes can be saved for future reuse.

In general, this was the case. As noted above, we examined some 15% of the sample term-by-term. Of 323 lexically ambiguous partitions, all could be disambiguated correctly without reference to the surrounding text; in 9 cases neither alternative was right.

Of the contextually ambiguous partitions, less than 20% (51 items) could not be disambiguated in isolation. Frequently, one phrase appeared in various combinations; eg.

Type	Forms	Correct Partition	Incorrect Partition	Meaning
L	365	การ บริหาร	กา รบ ริ หาร	*service* vs. *blackbird fight divide*
C	362	การ ฝึกอบรม	การฝึก อบ รม	*training* vs. *baked umbrella*
C	339	การ ดำเนิน	การดำ เนิน	*conducting* vs. *diving hill*
L	261	มาก กว่า	มา กก ว่า	*more than* vs. *come b'bird says*
C	214	ทาง เศรษฐกิจ	ทางเศร ษฐกิจ	*economic direction* vs. nonsense
L	218	การ เปลี่ยนแปลง	การเปลี่ยน แป ลง	*changing* vs. *changing flat down*
C	188	การ ลงทุน	การลง ทุน	*investing* vs. *go down capital*
C	186	การ พัฒนาข้าราชการ	การพัฒนา ข้า ราช การ	*development* vs. *develop servant ...*
C	176	เจ้า หน้าที่	เจ้าหน้า ที่	*officer* vs. *boss face at*
C	170	ความ จำเป็น	ความจำ เป็น	*need* vs. *memory is*
C	164	ตัว ดำเนินการ	ตัวดำ เนิน การ	*conductor* vs. *black body hill task*
L	159	การ สร้าง	กา รส ร้าง	*constructing* vs. *b'bird taste neglect*
C?	136	การ จัดการ	การจัด การ	*management* vs. *set up task*
L	132	การ กระทำ	การก ระ ทำ	*doing* vs. *case scrape to do*
C	125	การ ปรับปรุง	การปรับ ปรุง	*improving* vs. *changing to season*
L	122	การ กระ	การก ระ	*replacing* (frag) vs. *case sweep*
L	107	แน่ นอน	แน่น อน	*certainly* vs. *solid not*
L	96	การ สอน	กา รส อน	*teaching* vs. *b'bird taste not*
L	89	ทัศน คติ	ทัศ นค ติ	*attitude* vs. *ten seven blame*
L	86	การ ส่งเสริม	การส่ง เส ริม	*supporting* vs. *sending diverge edge*

Table 4 L(exical) and C(contextual) ambiguity. Assignment of categories posed little difficulty, and the correct partition could almost always be selected without reference to the underlying text. Despite their frequency, not all of these examples are common in ordinary text; the first three terms (*service, training, conducting*) obviously reflect specialized subject matter.

จำ+เป็น with various affixes. It is possible the most likely partition actually occurs in all cases; but we did not cross-check the original text in this study. It is also worth noting that contextually ambiguous partitions are less critical for some common segmenting applications (eg. text-to-speech, indexing) in which a word-by-word split is satisfactory.

Maximal Matching This strategy is based on the premise that the segmentation that produces the fewest words is probably correct. In our three test sets, maximal matching was *always* correct when it could be applied; ie. when the alternatives had different numbers of words (see table 5). This was generally the case with lexically ambiguous partitions (about 90% were asymmetrical), but far less so with contextually ambiguous partitions (just over half were asymmetrical). Intuitively, this makes sense, because contextually ambiguous partitions frequently involve the left-right binding of a single word; hence, both sides have the same number of words.

Stop Nodes and Go Collocates We hypothesized that we would find a class of words that appeared frequently within trial segmentations, but would never (or almost never) be correct. This is based on an analysis of letters and words whose orthographic characteristics make them exceptionally prone to causing ambiguous partitions.

For example, ก, the third most frequent consonant in typical Thai text, is by far the most common first letter in dictionary entries (partly because it is one of the few letters that start consonant clusters). Because it is one of eight 'regular' final consonants, it is a common final letter. A typical dictionary word list (18,151 entries) contained 2,064 words beginning with ก (11.4%) and 1,194 words ending with ก (6.6%).

Thus, words that end or begin with ก have a high affinity for joining their neighbors to produce lexically ambiguous partitions (eg. in symmetrical examples of types 1 and 2 below, we found 70 different forms accounting for 453 actual entries). For example, the first case below involves only common words; the second and third rely on words (บอ and กก) that are progressively less so — in fact, they never occur in the text at all.

มากว่า = มาก ว่า | *มา กว่า *(many) + (says) that | *come + (more) than*
บอกว่า = บอก ว่า | *บอ กว่า *says that | *madcap + (more) than*
มากกว่า = มาก กว่า | *มา กก ว่า *many + more than | come + cuddle + (says) that*

By comparing the 7,420 distinct words that appeared among all trial partitions to the actual hand-segmented word list, we were able to identify a set of *stop nodes* that invariably flagged the wrong partition for our data set. Figure 6 shows the most frequent of the 2,383 terms that did appear within trial partitions, but which were never found in the original text. Overall, these terms fell into three classes:

Maximal matching . . .	Sample	Fails	Succeeds	Inapplicable
Lexically ambiguous partitions	323	0	292	31 (10%)
Contextually ambiguous partitions	289	0	162	127 (44%)

Includes both binds and two-way partitions	Forms (9,253)	Total (36,267)
Symmetrical (ตาก ลม / ตา กลม)	5,138 (55%)	23,239 (64%)
Asymmetrical (มาก กว่า / มา กก ว่า)	4,115 (45%)	13,028 (36%)

Table 5 Maximal matching can only be applied if the trial partitions are asymmetrical. In the sample we inspected term-by-term, it always worked *when applicable*. However, nearly half of the contextually ambiguous partitions could not be dealt with this way, and a clear majority of potential partitions in the actual text were symmetrical as well. We suspect that to some extent, this may be an artifact of the very large dictionary we used to generate candidates.

- *Artifacts and combining forms*. These terms are in the dictionary, but essentially never appear as standalone words, eg. **ราช** is defined as 'King' but is invariably used to mean 'royal' in combining forms.
- *Obsolete and learned words*. Often seldom-used historical terms, eg. **อง**, 'prince.'
- *Ordinary words*. Everyday words that simply happened to be absent from our sample, eg. **งา**, 'tusk.'

The first and second groups suggest a new approach to building dictionaries for segmentation: mark such words as 'present,' but do not allow them to be produced as the result of ambiguous partitions. In other words, allow the terms to be recognized as actual words, which of course they are, but reject them whenever they appear as one alternative of an ambiguous partition: they are stop nodes, not stop words.

The third group holds great interest for further work, because we anticipate that any calculated frequency statistics will tend to exclude them, even if they are correct. Table 7 shows a sample of such words that do actually appear in the text sample.

In forthcoming work, we test our ability to spot correct appearances of these low frequency terms by a *forced training* technique that relies on collocates found in dictionary entries. This approach relies on the fact that in Thai, many words have collocates that are predictable, even though finding them in open text may require a prohibitively large segmented corpus. For nouns, these include classifiers; for verbs, these include prepositions, completative verbs (eg. *you look and see*) and other secondary verbs, and for adjectives, these include restricted modifiers and intensifiers. We treat these terms as 'go collocates' — neighbors that indicate a low probability word is likely to be correct.

Word	Meaning	Forms	Total
ระ	scrape	201	943
ราช	royal (pref)	223	878
กระ	replace (pref)	152	809
ริ	start	138	745
เนิน	hill	49	742
อบ	bake	89	686
อง	prince	163	586
ข้า	servant	87	579
ประ	to affix	202	565
การทำ	doing	60	522
การจัด	setting up	84	516
การดำเนิน	conducting	67	515
งา	tusk	124	467
นข	fingernail	116	458
การดำ	diving	25	440
กก	cuddle	51	439
งก	greedy	82	438
ตัวดำ	black body	27	328
การให้	giving	56	322
บุ๋	dented	181	307

Word	Meaning	Total	Forms
ราช	royal (pref)	878	223
ประ	to affix	565	202
ระ	scrape	943	201
บุ๋	dented	307	181
อง	prince	586	163
กระ	replace (pref)	809	152
ริ	start	745	138
นอ	rhino horn	224	130
งา	tusk	467	124
นข	fingernail	458	116
อบ	bake	686	89
ข้า	servant	579	87
การจัด	setting up	516	84
งก	greedy	438	82
ยอ	compliment	135	68
การดำเนิน	conducting	515	67
กร	arm (royal)	222	60
การทำ	doing	522	60
จา	to speak	152	58
กอ	a clump	138	56

Table 6 Frequent trial partitions that do not appear in the text, by total (left) and by number of forms (right). Many, but not all, of these terms *always* mark an incorrect partition. They can be thought of as stop nodes that can be listed in the dictionary as 'present,' but should not be chosen when they appear as one alternative of an ambiguous partition.

Hidden Terms One of the more powerful tools in our arsenal is the ability to focus attention on the ambiguous partitions that appear most frequently. We considered the possibility that some common contextually ambiguous partitions might be 'hidden' by being embedded within longer strings. As a result, they might not be noticed as being exceptionally frequent, or general rules for resolving them might not be applied.

For example, จำ, which means *know* or *remember* as a standalone term, appears in every possible alternative: it may be prefixed or suffixed, show up in endocentric or exocentric compounds, or simply appear coincidentally in the middle of another word.

We found candidates in this class by writing code that selected only symmetric patterns of the following form, restricting the interior string to 2-15 characters:

string1+string2 string3 | *string1 string2+string3*

Our data set contained a total of 4,806 patterns of this form, representing 21,833 actual entries. The central 'swing string' consisted of 807 distinct terms; of these, 707 were ordinary words, 94 were compounds that did not appear in a basic dictionary word list, 4 were transcribed foreign words, and 2 were fragments.

We are interested [illegible] because they contain the difficult cases that [illegible] terms that appear [illegible] phrases, not just the swing strings) introduces terms that are more text-dependent, or which are essentially artifacts of the large dictionary we used for partitioning.

Our analysis of this data is still underway. Nevertheless, we can see in table 9 that a particular phenomenon is responsible for many of the contextually ambiguous partitions

Term	Meaning	Forms	Total	Real		Term	Meaning	Forms	Total	Real
กา	blackbird	200	1541	1		ทา	paint	32	65	1
รม	to smoke	135	1082	[illegible]		มูล	dung, origin	23	63	2
ย่าง	roast	241	580	[illegible]		หลา	yard	13	55	[illegible]
การณ์	case	52	390	2		ลำ	clf. boat, plane	11	46	[illegible]
รส	taste	32	[illegible]			สน	interest	12	31	[illegible]
[illegible]	[illegible]	27	[illegible]			คำ	advance / word	13	31	[illegible]
[illegible]	[illegible]	25	[illegible]	[illegible]		จ	note	16	[illegible]	
แฝ	fire	24	298	2		นม	breast	11	21	1
พล	strength	21	235	1		เผย	to open	13	15	1
มาตร	meter	57	183	2		วา	app'x two yards	11	14	1
ตุ่ม	below	56	160	3		สิ	emphasis part	19	32	2
ชา	tea	34	152	1		โครง	structure	76	227	8
สา	content	49	139	1		ทด	raise up	18	35	2
เทศ	foreign	25	123	1		อม	hold in mouth	18	42	2
ริม	edge	20	120	2		ผ่อน	to	9	16	1
อา	uncle	51	116	3		ลัก	steal	9	38	1
เห	to deviate	47	110	1		ติ	blame	35	171	4
โยชน์	about 10 miles	21	98	1		ป้อง	protect	17	36	2
ดำ	black	44	96	3		ลาย	pattern	17	65	2
งม	grope in water	25	94	1		ไข	grease	16	100	2

Table 7 Relatively rare words that are common as trial partitions. We think that it is very likely that they have predictable collocates (classifiers, auxiliary verbs) that can be found in dictionary entries, and hence do not require large text corpora.

— the intersection of two or more of the high-frequency terms in a single phrase, for example, ความ, จำ, and เป็น. In effect, all the potential alignments of bindings seem to occur — the 22 examples consist of only 25 words.

While all of these terms have very high individual and collocational frequencies, we would argue that some of the lexical phrases are more tightly bound than others. In future work, we investigate whether we can define transitive orderings that can be applied to novel circumstances. For example, we list these in order of increasing likelihood:

จำ → จำหลัก → ความจำ → จำใจ → จำนวน → จำเป็น → ประจำ

6. Conclusions and Further Work

In recent years work on segmentation for Thai has focused on analytical methods involving statistical and grammatical analysis of large, segmented text corpora. But while this may hold promise for the long run, at present we have neither the text corpus, nor the grammatical understanding of Thai, nor sufficient understanding of the segmentation problem itself to make objectively measureable progress.

The analysis presented here argues that segmentation has distinct and separable aspects, and that both performance and our ability to measure performance can be improved by focusing on specific aspects of the problem. In particular, we find that ambiguous partitions — which we feel are the most critical, because they affect the text's meaning — can be isolated and attacked independently, using methods that do not depend on having large text corpora at our disposal.

This paper has been primarily descriptive. We look forward to working with different dictionaries and texts to test both our analysis, and the new methods we propose.

Ordered by # of forms			Ordered by actual counts			Ordered by actual/forms			
Term	Forms	Actual	Term	Forms	Actual	Term	Forms	Actual	Ratio
การ	1011	2871	การ	1011	2871	เศร	1	214	214.0
ที่	142	569	เป็น	121	593	ควบ	1	124	124.0
ตัว	131	501	ที่	142	569	ดำ	4	475	118.7
เป็น	121	593	ตัว	131	501	ดำเนิน	3	311	103.6
ความ	102	327	จะ	46	480	อบ	1	81	81.0
ได้	90	301	ดำ	4	475	หาก	1	74	74.0
งาน	75	424	ทำ	45	465	แข่ง	2	146	73.0
ไป	69	197	งาน	75	424	กรณี	1	64	64.0
ทาง	68	354	ให้	62	423	คล่อง	1	41	41.0
ว่า	64	210	จัด	20	384	ปฏิบัติ	6	244	40.7
ให้	62	423	ทาง	68	354	วาง	3	120	40.0
ผล	51	154	ความ	102	327	ขยาย	3	117	39.0
มา	48	150	ดำเนิน	3	311	ต่าง	2	77	38.5
จะ	46	480	ได้	90	301	คือ	1	38	38.0
ทำ	45	465	จำ	15	244	ประมวล	2	74	37.0
กัน	37	61	ปฏิบัติ	6	244	ขน	2	72	36.0
มี	36	150	หน้า	19	243	แปล	2	71	35.5
ดี	32	102	ลง	17	238	แลก	1	35	35.0
เข้า	30	104	บริหาร	7	220	ตัดสิน	1	33	33.0
ต่อ	29	60	เศร	1	214	นัก	4	131	32.7

Table 8 The 'swing strings' of symmetrical alternatives. Ordering by forms highlights affixes; ordering by counts reflects the subject matter. The count/form ratio is very text-dependent; it suggests where to look for context-independent ambiguity.

	Term	Correct	Incorrect		Term	Correct	Incorrect
1	จำ	ความ จำกัด	ความจำ กัด	3	จำ	รู้จำ ได้ ??	รู้ จำได้ ??
1	จำ	ความ จำใจ	ความจำ ใจ	39	จำ	ประจำ ปี	ประ จำปี
1	จำ	ความ จำเดิม	ความจำ เดิม	5	จำ	ประจำ ใจ	ประ จำใจ
1	จำ	ความ จำนง	ความจำ นง	6	จำ	ถูก จำกัด	ถูกจำ กัด
1	จำ	ถูก จำขัง	ถูกจำ ขัง	1	จำกัด	ข้อ จำกัดความ	ข้อจำกัด ความ
1	จำ	ท่องจำ หลัก	ท่อง จำหลัก	1	จำนวน	เลขจำนวน หนึ่ง ?	เลข จำนวนหนึ่ง?
1	จำ	ประจำ ได้	ประ จำได้	44	จำนวน	เลข จำนวนเต็ม	เลขจำนวน เต็ม
1	จำ	ประจำ ต้อง	ประ จำต้อง	2	จำนวนเต็ม	เลขจำนวนเต็ม บวก ?	เลข จำนวนเต็มบวก ?
1	จำ	ประจำ เป็น	ประ จำเป็น	1	จำเป็น	โดย จำเป็นต้อง	โดยจำเป็น ต้อง
12	จำ	ความจำ หลัก	ความ จำหลัก	4	จำเป็น	ความจำเป็น จะ	ความ จำเป็นจะ
170	จำ	ความ จำเป็น	ความจำ เป็น	6	จำเป็น	ความจำเป็น ต้อง	ความ จำเป็นต้อง

Table 9 Extracting and ordering the hidden 'swing strings' is invaluable for understanding low-probability cases that have similar characteristics, but are not identical. For example, **ประจำ** is always the correct partition; it appears in five different forms. Note that all of these are symmetrical, so maximal matching cannot be applied.

7. Acknowledgements

We gratefully acknowledge Namfon Buntua's assistance in classifying, disambiguating, and translating the ambiguous partitions, and the kind permission of Wantanee Phantachat of the NECTEC *Linguistics and Knowledge Science Lab* for use of their corpus.

8. References

Chang, C.H, and Chen, C.D. 1993. SEG-TAG: A Chinese Word Segmentation and Part-Of-Speech Tagging System. In *Proceedings of Natural Language Processing Pacific Rim Symposium* 319-327, Fukoka, Japan.

Chiang, T.H, Chang, T.S., Lin, M.Y, and Su, K.Y. 1992. Statistical Models for Word Segmentation and Unknown Word Resolution. In *Proceedings of ROCLING V*, 121-146, Taipei, Taiwan.

Cooper, Doug. 1993. Compared to What? Measuring the Performance of Thai Segmentation Algorithms. Technical Report 9, Center for Research in Computational Linguistics, Bangkok.

Haas, Mary. 1964 *Thai-English Student's Dictionary*. Stanford University Press.

Haas, Mary. 1946 Techniques of Intensifying in Thai. *Word*, Vol. 2.

Haas, Mary.1942 Types of Reduplication in Thai. In *Studies in Linguistics*, Vol 1, 1:1-4.

Kawtrakul, Asanee, et al 1996. A Gradual Refinement Model for a Robust Thai Morphological Analyzer. In *COLING-96: 16th International Conference on Computational Linguistics*, Copenhagen, Denmark.

Kawtrakul, Asanee, et al. 1995. A Lexicon Model for Writing Production Assistant System. In *Proceedings of the Symposium on Natural Language Processing in Thailand '95*, Kasetsart University, Thailand..

Luksaneeyanawin, Sudaporn. 1984 Some Semantic Functions of Reduplicatives in Thai. In *Selected Papers from the First Int. Symposium on Language and Linguistics (Pan Asiatic Linguistics)*, Chiang Mai University.

LINKS 1996 Tagged Text Corpus (unpublished). Language and Knowledge Science (LINKS) Laboratory, Bangkok.

Maosong, S., and T'sou, B.K. 1995. Ambiguity Resolution in Chinese Word Segmentation. In *Proceedings of the 10th Pacific Asia Conference on Language, Information, and Computation*, City University of Hong Kong.

Noss, Richard B. *Thai Reference Grammar*. Foreign Service Institute, State Dept., Washington D.C.

Pornprasertsakul, Ampai. 1994 *Thai Syntactic Analysis* PhD Thesis, Asian Institute of Technology,

Sornlertlamvanich, Virach. 1993. *Word Segmentation for Thai in a Machine Translation System*. National Electronics and Computer Technology Center (in Thai).

Sproat, Richard, Chilin Shih, William Gail, and Nancy Chang. 1994. A Stochiastic Word Segmentation Algorithm for a Mandarin Text-to-Speech System. In *Proceedings of the 32nd Annual Conference of the Association for Computational Linguistics*, 66-72, Las Cruces, New Mexico.

Vongvipanond, Peansiri E 1992. Lexicological Significance of Semantic Doublets in Thai. In *Papers on Tai Languages, Linguistics, and Literatures*, Northern Illinois University.

Vongvipanond, Peansiri E. 1993. Linguistic Problems in Computer Processing of the Thai Language. In *Proceedings of the Symposium on Natural Language Processing in Thailand*, Chulalongkorn University.

Wu, Dekai and Fung, Pascal. 1994. Improving Chinese Tokenization with Linguistic Filters on Statistical Lexical Acquisition. *ANLP*.

Wu, Z, Tseng, G. 1993. Chinese Text Segmentation for Text Retrieval: Achievements and Problems. *Journal of the American Society for Information Science*, 532-542.

Wuwongse, Vilas and Pornprasertsakul, Ampai. 1993. Thai Syntax Parsing. In *Proceedings of the Symposium on Natural Language Processing in Thailand*, Chulalongkorn University.