

How Do Thais Tell Letters Apart?

Doug Cooper
<doug@nwg.nectec.ac.th>

Abstract

Many Thai printing fonts, like most Thai handwriting, vary considerably from 'standard' Thai letterforms. Nevertheless, native Thai speakers can easily read text that baffles both TSL (Thai as a Second Language) students and OCR (optical character recognition) systems. This paper demonstrates that Thai letters have secondary characteristics that are recognized by fluent Thai speakers, but which are obscured by traditional TSL reading and writing instruction, and are ignored by prototype OCR systems. I show how these secondary characteristics are consistently applied in font design and handwriting, and suggest ways of bringing them to the attention of both students and computers.

Introduction

Pity the poor TSL student who feels that he has finally mastered the loops and turns of the Thai alphabet. As soon as he ventures away from his ก ฦๅ alphabet primer, he finds that the rules for recognizing letters he has learned so carefully are nowhere to be seen.

Hoping for a meal, he hunts for a ร้า๓๑๒๒, but can only find one ร้า๓๑๒๒ after another. After searching the menu in vain for ข้า๓๑๒๒ he settles for ข้า๓๑๒๒. Imagine his surprise, then, when a third dish — ข้า๓๑๒๒ — appears on the bill instead!

Thus, even though most Thai letters are invariably described as consisting of a circular *head* followed by a continuous stroke, a little investigation shows that this description only applies to the *reference style* — letters written in a careful hand, or printed using traditional typewriters or typesetting fonts. Variations abound, and it is clear that fluent Thai readers identify them as readily as fluent English readers identify variants like a/a or g/g.

Letterform variations in Thai presents two special problems for TSL students and OCR systems. First, the changes are pervasive; a style change alters practically every letter in the alphabet. Second, the changes often undermine

the rules of the reference forms. As a result, the more effectively the TSL student masters the standard letterforms, the more thoroughly he is confused; the more thoroughly an OCR system is programmed to identify standard letterforms, the more easily it is tripped up.

For example, consider this elementary rule:

ฦๅ is separated from ฦๅ by the inward or outward orientation of the letter's head.

The rule is obviously true, but it does not suffice to determine what this letter — ฦๅ — is. At ordinary text sizes, the head's position in this common printing font is ambiguous. Our hypothetical TSL student looks for a magnifying glass, and the OCR program spins its wheels..

But there is a more productive approach: look at ฦๅ and ฦๅ in various styles to tease out secondary characteristics and infer new rules:

ฦๅ ฦๅ → ฦๅ ฦๅ → ฦๅ ฦๅ → ฦๅ ฦๅ

The ambiguity is resolved by changing the salient feature: look at the bar's origin instead of checking the circle's orientation. ฦๅ's bar always starts at the base of the letter, while ฦๅ's bar tends to creep up the left side. In effect, if the bar is too short for the reference alphabet's rule to apply, the letter is probably ฦๅ, and not ฦๅ.

This paper looks at the ways in which Thai print styles vary, and formalizes the implicit rules by which letters are distinguished from one another. Our investigation relies on two sorts of comparisons:

This research was supported by the Center for Research in Computational Linguistics
Author's address: 1617 Ratchaprarop Tower Mansion
99 Soi Bunprarop, Ratchaprarop Road
Makasan, Bangkok, Thailand 10400
(662) 246-9311 (-28) ext 1617 Fax (662) 246-9329, 5468

- Look for the features that maintain a minimal *design distance* between similar letters in a given style.
- Track these letters across different fonts, and see what happens as designs become more highly stylized in their departure from the reference standard.

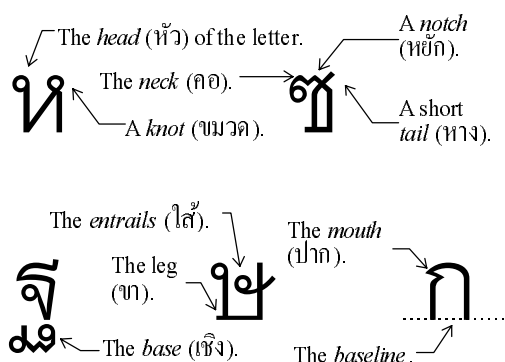
I'll begin by defining the terms we'll need to describe Thai letterforms, and summarize the traditional approaches to describing the letters. Then we'll look at basic print styles; our analysis will be made easier by breaking the alphabet into groups of letters that tend to be built along similar lines.

Next, we'll investigate different kinds of variations from the reference standard, and see that not all of them are predictable. After a close look at the alphabet, I'll address the fundamental question: how do fluent readers manage to learn unfamiliar styles?

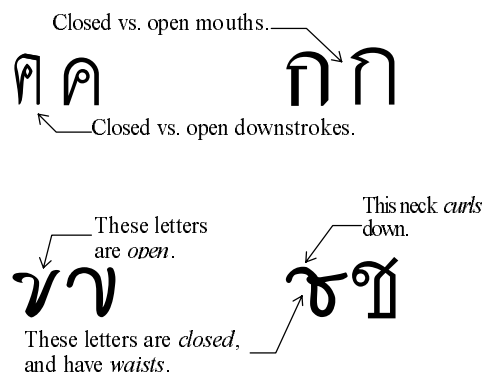
I'll close with specific recommendations for TSL instruction, and discuss the implications for Thai-language OCR. We will find, somewhat surprisingly, that TSL students would benefit from the methods currently applied to computer programs: in effect, being given detailed descriptions of the physical characteristics that distinguish letters. Computers, in turn, would benefit from applying the methods — considering the letter's context — commonly used by fluent Thai speakers.

Anatomy of Thai Letterforms

Let's begin by defining the distinctive features of Thai letters. The nomenclature of Thai characters isn't universal, but we can come up with some reasonably descriptive terminology:



We will also need some non-standard comparative terms for describing the differences between letters, and for referring to design innovations:



The Traditional Approach

Traditionally, letters are described in terms of the letter's starting point, and the head's orientation. The classic introduction for TSL students, Mary Haas's *The Thai System of Writing*, puts it this way (emphasis hers):

All consonants except ก and ข are started with the production of their characteristic little CIRCLE ... It is very important to note whether the circle is to the RIGHT or to the LEFT of its connecting line. [HAAS56]

Of course, with her usual thoroughness, Haas also includes four pages of handwritten material, and a half-dozen pages of printed samples, that demonstrate variations from the norm.

J. Marvin Brown's two-volume series *AUA Thai Course (mostly reading)* and *AUA Thai Course (mostly writing)* [BROWN79a, BROWN79b] also introduce the reference style, while dealing extensively with a variety of handwriting styles in two appendices.

For example, Volume W includes an appendix that goes through the entire alphabet and supplies four sample hands for each character. Appendix 1 of Volume R has many, many handwritten samples, along with a letter-by-letter commentary on handwriting styles, eg.:

The loops themselves are frequently omitted ... The difficult inner and outer loops of ก, ข, and ฃ can be omitted completely (though ฃ must keep its jags), and the difficult narrow parallel lines can be replaced by various kinds of loops. [BROWN79b]

Brown's discussion of how vowels and tone-marks are attached to various letters is unmatched, and clearly explains handwriting that would otherwise be incomprehensible.

More recently, a 1991 study by Gandour and Potisuk, *Distinctive Features of Thai Consonant Letters*, proposed an extensive classification system for the distinctive features of consonants as part of an investigation of spelling errors made by a Thai speaker. For the researchers' purposes, 17 features were required to distinguish between letters. They noted that:

As many as seven of the features deal with various attributes of loops: 4 with the beginning loop, 2 with the body loop, and 1 with the tail loop. [GANDOUR91]

This study is particularly interesting because it demonstrates the systematic oversimplification that results from taking the standard introduction to the reference alphabet at face value. Even within the reference alphabet, it is easy to find letters that incorporate distinctions not accounted for by their orthographic feature set.

For example, the authors find a visual 'feature difference' of just 1 for the pairs ก ฃ and ฃ ก (their entries 9 and 142, table 2) — only the orientation of the heads is assumed to be significant. Yet, the difference in the height attained by the central strokes is just as pronounced in their article's typeface as it is here. Indeed, we will see that in many fonts, letters are distinguished solely on this basis.

Computerized approaches to optical character recognition (OCR) for Thai have also focused on the reference alphabet and head. The 1993 *Symposium on Natural Language Processing in Thailand* includes two articles on Thai OCR:

Many characters have small holes called the heads of characters, and the drawing of the characters begins by tracing these heads. [HIRANVANICHAKORN93]

There is always present a small circle portion which is called the head of the character ... Internal [and] external heads [are] the two styles of heads of Thai characters. [KIMPAN93]

Both teams go on to point out that Thai OCR systems encounter particular difficulty when the head varies:

The reasons [for] rejection and mis-identification were mainly due to differences in the number of holes ... between input data and models. [HIRANVAN-ICHAKORN93]

A recognition rate of 98.20% for testing data has been obtained. The ill-classified characters occurred if the head of the character is broken. [KIMPAN93]

As we will see, the head is generally the first feature to go. In a real sense, then, Thai OCR confronts the same problems, and has the same success, as TSL students in recognizing rapid handwriting and nonstandard letterforms.

Five Basic Letter Styles

There are literally dozens of Thai letter styles. The interested reader is referred to 5" by 7" flip-books like *รวมแบบลายมือ*, which contain page after page of hand-lettered samples. Many of these are now appearing as outline fonts that can be used with computer typesetting programs as well.

From this wealth of designs, we can focus on five primary variations that the reader is likely to encounter:

- The *classic* style (ตัวไทยเดิม = classic style, or เขียนตัวบรรจง = write letters precisely) dates from the time of King Narai (ca. 1680). Line weight has little or no variation, letters have complete circular heads where appropriate, and horizontal and vertical lines are regular and perpendicular. Here are typical letters from Cordia New:

ก ข จ ท พ อ

- The *craft* style (เขียนตัวศิลป์). A highly calligraphic, Indian-influenced style, drawn with a broad pen or brush point. Heads are no more than semi-circular, and wherever possible, letters are given a distinctive hori-

zontal top bar, a style that is retained in modern Devanagari printing fonts (used for Sanskrit, Hindi, and Marathi, among others) as well. These typical letters are from JS Chanok:

ก ข ง ต ฎ ฐ

- The *tail* style (เขียนเส้นหาง). Characteristics include fairly regular pen thickness, and an exaggerated tail that wraps around the body of some letters. These typical letters are from JS Wansika:

ง ข ฃ ฅ ๕ ๖

- The *modern* style (เขียนสมัย). Usually drawn with a single pen thickness, letters have no heads, and are simplified as much as possible. These letters are from JS Thanaporn:

ก ก จ ฃ ฅ ฆ

- Various *script* styles (เขียนหวัด = scribble). Characterized by a rapid, flowing line with heads minimized, corners often rounded, and some letters (particularly ก, ฃ, and ฅ) opened up. This sample is from JS Sirium:

ก ข ฃ ฅ ฆ ๕

See *The Thai Writing System* [DANVIVA-THANA87] for considerable discussion of the history of Thai scripts.

By definition, I assume that the *standard reference style* is synonymous with the classic style, and has the letterforms that appear in ก ไก่ practice books, letter charts, and Thai basal readers. I'll use the font Cordia New for examples. Note that there are slight, but common, variations between instances of the standard reference style. For instance, Cordia New adds a small leg to letters that have a left corner at the baseline, eg. ก ฃ ฅ.

Internal Design Distances

It might not be apparent at first, but part of the beauty of the reference style is that the underlying similarity between letters is not always obvious. Within a group letters may be close, but they don't really begin to blur together until they

are drawn in a more simplified or stylized manner.

Indeed, the only common exception to this rule is the pair ก/ฃ, which is not very well differentiated in the font I'm using as a reference sample. Even under the best of circumstance, though, a tiny notch in a short neck is almost indiscernable.

What we find, though, is that other designs for the reference alphabet deal with the problem by extending the neck *downward* slightly. Even in the small sizes used for the examples below, you can see that the head of the second letter hangs a bit below the head of the first.

ก ฃ (Angsana New)

ก ฃ (Dillennial UPC)

ก ฃ (JS Prasoplar)

The introduction of a secondary feature of this sort is prompted by the need to maintain a minimum *internal design distance* of some sort. Without it, people couldn't tell letters apart.

The concept of internal design distance is worth noting because it lets us make a prediction:

If a stylistic variation makes letters ambiguous, letterforms in the ambiguous group will be modified until a reasonable design distance between letters is achieved again.

Regardless of the nature of a design variation, it will be balanced by an internal pressure that develops within the alphabet itself — a pressure that forces letters to be different from one another.

What may be difficult to predict is how that pressure will manifest itself. On occasion, even a clear picture of the sort of variations that are being introduced won't be enough to tell us what the final letterforms are going to look like.

Instead, it will be necessary to see how certain characters are drawn relative to each other. However, once we've seen one member of the set, we may have enough information to make predictions about the others.

For example, note the differences between ก and ฃ in the center and right-hand examples below. In both cases, the new style gets rid of the original letter's circular head. But since this change alone might make the letters ambiguous, additional variations turn up to maintain a reasonable design difference between the two letters:

Primary variations: ບ becomes ບ

Secondary variations: ປ becomes ປ

Tertiary variations: ທ becomes ທ

Primary variations involve a single guideline, like ‘delete the circle’ or ‘extend the tail.’ It is easy to see how the new letterform varies from the reference letter, and to predict what other letters are going to look like.

Some primary variations are prompted by the instrument, real or imaginary, being used to draw the letters. For example, in the craft style circular heads are usually replaced by angled wedges that are more easily drawn with a brush:

ປພຟ becomes ປພຟ

Secondary variations usually entail bringing the letter’s lesser characteristics to the fore. We have already seen the progression that leads to the ດ/ຄ variation:

ດດ → ດດ → ດດ → ດດ

Tertiary variations cannot be predicted by any reasonably stated set of design rules. They reach outside the alphabet in search of alternative letterforms. For example, the letterforms ທ and ທ are the historical forebears of ທ and ທ, and can still be found in the modern Lao alphabet in essentially identical form.

Other letterforms seem to be inspired by designs that originate in modern Roman alphabets. Here are reference, Roman, and Thai letters:

หกร รส ลาล ทกท นูน งว

Before we take a methodical look at the alphabet, let’s survey the variants more closely.

Primary Variations

The mouth is the single feature most likely to be simplified or deleted. With one exception, its presence or absence adds no information to the reader’s understanding of any letter.¹ It is the orientation of the head (or head substitute) alone that distinguishes ປ from ປ and ປ, and the location of the knot that decides between ປ and ປ. For instance:

¹ The exception, unexpectedly, is prompted by the need to distinguish between ປ and ດ in simplified fonts. Note how the mouth is retained here: ດ ດ.

กถกถถถ กถกถถ กถกถถ

A second common primary variation lies at the letter’s head. The traditional craft style can be seen as a transitional style; the head is minimized, but not done away with entirely. Most modern styles go the whole way, and minimize the letterform as severely as possible:

ม ม ม ม ม ม ม

In general, if the head has been done away with, the knot, if there is one, is drastically simplified as well.

Simplification as a form of primary variation is particularly evident in vowels, tone marks, and letter bases or entrails. For example, consider these minimal vowels and tone marks:

แะอ๋ ๓ะอ๋ ๓ะอ๋ ๓ะอ๋

Exaggeration is another common form of primary variation, but such fonts tend to be easy to read. For instance, the extended tail style, like its English counterpart, is frequently used for wedding invitations and formal announcements:

เสียดเสียดเสียด

We can find designs in which regular exaggerations are consistently applied

กขคฆง กขคฆง กขคฆง

Secondary Variations

Secondary variations usually have a more complex history. Frequently, we must turn to the shortcuts that appear in handwriting to understand how these modifications were arrived at.

For example, consider the appearance of ດ and ດ as they move from the standard to script-like to modern fonts:

ດດ ດດ ດດ ດດ ດດ

Already we can see that, simply because of the direction in which the head starts, ດ tends to close the angle at the lower left corner far more quickly than ດ does. The head disappears entirely, then the downstroke merges with the upstroke, and we’re left with ດ.²

² Brown puts it this way: ‘The different directions of the loops in ດ and ດ transform in different ways

You can see exactly the same sort of modification in จ and จู. The head and first downstroke of both letters is abbreviated and reattached, and the tail of จู is folded along the body of the letter:

จ จู จ จู จ

Exaggerating (and occasionally introducing) secondary features is common. I've already mentioned the extended neck and lowered head in some ฦ/ฦ designs. This principle is applied to greater effect in distinguishing between ฦ/ฦ and ฦ/ฦ. As you can see, the internal notch is much lower for ฦ and ฦ, presumably because the writer leaves a bit of extra room for the internal head.

However, when the head is minimized or removed entirely, the notch height alone is usually sufficient to tell the letters apart. As a rule, if the notch is the full letter height *and there's no head*, the letter is ฦ or ฦ:

ฦ ฦ ฦ ฦ ฦ ฦ

Two exceptions prove the rule; they are practically unreadable at ordinary text sizes:

ฦ ฦ ฦ ฦ ฦ ฦ

Another dramatic secondary variation is seen in the ฦ/ฦ pair. In the reference form, the head and tail of ฦ are entirely different. Then, little by little, the original letter loses its distinctive appearance, and acquires the more symmetrical, less technically demanding shape of the English letter S.

ฦ ฦ ฦ ฦ ฦ ฦ

In effect, ฦ drags ฦ along for the ride; the latter is changed in tandem with the former.

Certain special-purpose styles introduce variations in order to meet specific printing needs. For example, fonts like Kobori Allcaps are intended to be used for newspaper headlines. As such, they minimize any over- or under-structure, either by shortening it, or by folding it into the character itself. For example:

ฦ ฦ ฦ ฦ ฦ ฦ

the overall shapes of their letters and then frequently disappear.' [BROWN79a] p. 82.

Tertiary Variation

Finally, a few letters are regularly transformed from the reference style into shapes that are not easily predicted or prescribed. For example, we have already seen two letters that derive from historical shapes:

ฦ ฦ ฦ ฦ

I would be inclined to put some modern letterforms into this category as well. Again, ฦ is the most dramatic example, but ฦ and ฦ can vary greatly from the norm, too:

ฦ ฦ ฦ ฦ ฦ ฦ

A few highly stylized conventions are also encountered. For example, in our discussion of internal design distances I showed how the neck of ฦ might be extended. A *cleft* or slightly upward-curling head, in contrast, stands in for a neck complete with head and notch in styles that do away with the neck entirely:

ฦ ฦ ฦ ฦ ฦ ฦ

Finally, it is possible to find artistic fonts that are intentionally designed to be deciphered, rather than read. For example:

ฦ ฦ ฦ ฦ ฦ ฦ

Such fonts are interesting for the insight they provide into the designer's mind, especially in exposing his perception of the internal design distance between letters. However, we do not commonly encounter them in print.

A Close Look at the Letters

Let's turn our attention to a methodical look at the letter groups themselves.

First Group: ฦ ฦ ฦ

The first group can be difficult to distinguish, even within the reference alphabet. There are really two issues: telling ฦ from ฦ, and telling ฦ from the others.

The reference rules focus on the notch and the tail. But when characters are taken in isolation, we find imperceptible notches (ฦ), false tails (ฦ), and missing heads (U).

In practice, there are three secondary characteristics to look for:

- The character's *waist*. 𑜋 never has a waist; indeed opening the letter is a common alteration. In contrast, 𑜊 and 𑜌 frequently nip in just before the tail.
- An elongated neck and/or enlarged head is used to distinguish between 𑜊 and 𑜌.
- A missing or slightly downward-curling line head generally indicates 𑜊, while a flat, squiggly, or slightly upward-curling line head shows 𑜌.

The last distinction can be very fine indeed, and it's often necessary to see both characters together to tell what's going on.

Here are some open and closed waists. Note that the slightly closed waist on the third example distinguishes 𑜋 from 𑜌 in this style:

𑜊𑜊 𑜊𑜌 𑜌𑜌 𑜌𑜊

Examples of enlarged heads/enlongated necks:

𑜊𑜊 𑜊𑜊 𑜊𑜊 𑜊𑜊

Finally, here are some missing or highly stylized heads:

𑜌𑜌 𑜌𑜌 𑜌𑜌 𑜌𑜌 𑜌𑜌

𑜌𑜌 𑜌𑜌 𑜌𑜌 𑜌𑜌 𑜌𑜌

In some cases, it is all but impossible to tell the letters apart without seeing them in context.

There is also an occasional conflict from 𑜌, which is why that letter is in this group:

𑜊𑜊𑜌 𑜌𑜌𑜌 𑜌𑜌𑜌 𑜌𑜌𑜌

The secondary rule is subtle, but consistent: if the waist pinches from the left side, the letter is 𑜌, otherwise it's one of the others.

Second Group: 𑜌𑜌𑜌𑜌

The letters in this group are fairly easy to distinguish once the reader stops looking for heads and circular knots. We can trace the transition fairly easily:

𑜌𑜌𑜌𑜌 𑜌𑜌𑜌𑜌 𑜌𑜌𑜌𑜌 𑜌𑜌𑜌𑜌

Nevertheless, the reader must be alert to inconsistently applied changes. In these examples, knots are modified in different ways, depending on the letter:

𑜌𑜌𑜌𑜌 𑜌𑜌𑜌𑜌 𑜌𑜌𑜌𑜌

And there are always cases in which the design distance between letters is so small that sharp eyes are required. Note also that the close resemblance between these highly stylized letters and their Roman alphabet counterparts is rather jarring. It is difficult for TSL students to prevent themselves from seeing the letter 'u,' below:

𑜌𑜌𑜌𑜌 𑜌𑜌𑜌𑜌 𑜌𑜌𑜌𑜌

In the very popular craft style, conflict can arise between 𑜌 and 𑜌. It happens because letters are given curved bottoms to accentuate the flat bar heads, eg:

𑜌𑜌 𑜌𑜌 *versus* 𑜌𑜌 𑜌𑜌

This reverses the pattern of the reference style.

Third Group: 𑜌𑜌𑜌𑜌𑜌𑜌

The primary variations in this group have to do with opening, closing, and doing away with the letter's mouth and head.

𑜌𑜌 𑜌𑜌 𑜌𑜌 𑜌𑜌

Regardless of mouth or head variations, the bases of 𑜌 and 𑜌 are often greatly simplified.

𑜌𑜌 𑜌𑜌 𑜌𑜌 𑜌𑜌 𑜌𑜌

These typify the kind of change that poses nearly insurmountable problems for OCR. It is not that a computer cannot detect a squiggle as well as a human reader can. Rather, the human is better at knowing if very slight variations, as in the three examples on the right above, are intentional or not.

We also run into cases in which secondary characteristics have been introduced in order to maintain the required design distance between letters. Here, note that 𑜌's mouth is done away with, but 𑜌's mouth is retained so that it can be distinguished from 𑜌:

𑜌𑜌𑜌𑜌 𑜌𑜌𑜌𑜌 𑜌𑜌𑜌𑜌

If there are no other distinguishing features, letters that originally had mouths are usually given a sharp, upper-left corner. Below, the fourth letter in each group is 𑜌; note that it's the only one with a rounded upper-left corner:

กกด กกด กกด

Again, deciding what is sharp can be much easier for humans than for computers. The two sets on the left, below, are slightly sharper than the ones on the right, but all four are very difficult for machines to interpret:

กกด กกด กกด กกด

Finally, as a rule craft-style letters do not have mouths. Features that might easily be interpreted as highly stylized mouths are actually reattached downstrokes:

คค คค กค กค

Fourth Group: ฅฅฅ

This group has a combination of the second and third groups' features. In most cases, the letters are legible regardless of the degree to which they have been altered:

ฅ ฅ ฅ ฅ ฅ ฅ ฅ ฅ

Nevertheless, they can vary quite far from the reference standard. On the first line, left, below, note that the crossbar protrudes slightly to indicate the presence of a knot. We also see the base of ฅ attach to the body of the letter in the second line's examples:

ฅฅฅ ฅฅฅ
ฅฅฅ ฅฅฅ

These letters are also good gauges of the degree to which letters can be simplified. Below, note that ฅ can be reduced more than either ค or ฅ alone — the changes are really quite drastic!

คคค ฅคค ฅคค

Fifth Group: ฝฝฝ

We have already seen the main secondary characteristic of this group: the height of the central notch is usually sufficient to tell the letters apart. For ฝ, in turn, almost any loop or notch will do.

In practice, there are really two secondary clues to look for: first, ฝ and ฝ have a low central notch, and second, they practically always retain some hint of a head. If there is no head, the letter is almost invariably ฝ or ฝ:

ฝฝฝ ฝฝฝ ฝฝฝ

The curve of the first descending line also carries a slight hint — into the body for ฝ, ฝ, or ฝ, and away from the body for ฝ or ฝ:

ฝฝฝ ฝฝฝ ฝฝฝ

However, the notch's height is the surest indicator. Note that in some of the examples below, the head's orientation really is ambiguous in comparison to the reference standard:

ฝฝฝ ฝฝฝ ฝฝฝ

Once again, I'll point out that the exceptions, which have vanishingly small heads, and no notch variation, are very difficult to read:

ฝฝฝ ฝฝฝ ฝฝฝ

ฝ is not a common letter. Because it appears in so few common words, it is easy to identify, and is subjected to an extreme degree of variation. As we have seen before, such variations are not difficult for humans to distinguish, but they can pose problems for OCR. For example:

ฝ ฝ ฝ ฝ ฝ ฝ ฝ ฝ

Sixth Group: คคค

Letters in this group vary in three steps: first the head goes, then the downstroke is shortened, and finally the downstroke is reattached further along the side of the upstroke:

คคค คคค คคค คคค

In general, if the downstroke attaches at the baseline, the letter is ค, even if there's no distinguishable head, eg. ค.

Almost any closing of the downstroke (ie. reattaching) hints at a ค, and almost any opening implies that the letter is ค:

คคค คคค คคค

As you can see, the notch that distinguishes ค from ค disappears rapidly. Almost any flattening or break in the top bar must be recognized as an indicator of ค:

คคค คคค คคค

A small foot is sometimes, but not always, used to distinguish ค and ค from ค and ค.

These designs may be unpredictable, but at least they're consistent:

ក្រីក្រ ក្រីក្រ ក្រីក្រ

The craft style is a special case. As you can see in the font samples below, the downstroke is reattached mid-bar for all letters. A new secondary distinction is introduced: the upstroke turns slightly outward for the ក/ក្រ pair, and slightly inward for ក្រ/ក្រ.

ក្រីក្រ ក្រីក្រ ក្រីក្រ

Seventh Group: ឡឡ

These letters are quite different from each other in the reference style, but are subject to a great deal of variation. The variations themselves should all be familiar from other groups; the most dramatic involve shortening and reattaching the downstroke:

ឡឡ ឡឡ ឡឡ ឡឡ ឡឡ

Note that the downstroke of ឡ curves into the body of the letter slightly when the head is missing. This secondary feature usually leads to a sharp corner at the baseline, and distinguishes the letter from headless forms of ឡ. The letter's tail, in turn, tends to stay long, which helps distinguish ឡ from headless forms of ឡ. In the last example below, note the extended downstroke prompted by the need to keep ឡ identifiable:

ឡឡ ឡឡ ឡឡ ឡឡ ឡឡ

The third letter of this group, ឡ, often ends up with an appearance that is recognizable, but not very attractive:

ឡ ឡ ឡ ឡ ឡ ឡ ឡ ឡ ឡ ឡ

In general, the size of ឡ distinguishes it from ឡ, while the knot, or knot substitute, distinguishes it from ឡ and ឡ.

As in the previous group, the craft style poses special problems. Here are two slightly different versions of the craft style. Note that inward or outward curve is seen only at the very beginning of the downstroke:

ឡឡ ឡឡ ឡឡ

And, were it not for the base, ឡ would have a potential conflict with ឡ in a variety of fonts:

ឡឡ ឡឡ ឡឡ ឡឡ ឡឡ

Eighth Group: ឡឡ

The large number of words these characters appear in make this group the most problematic. In addition, this group has the largest number of internally prompted variations; two letters may be modified in opposite ways in different fonts.

Let's begin with a quick look at ឡ and ឡ. Note that no matter how wild the variation, these two are clearly distinguished:

ឡឡ ឡឡ ឡឡ ឡឡ ឡឡ

Aside from pointing out that, as usual, the tail of ឡ can be reduced to an unrecognizable dot, we'll ignore ឡ for a while.

The three remaining letters are difficult because they all lose their heads and tails, but at different rates. Nevertheless, we can rely on a few secondary characteristics. To begin with, if the head and tail are symmetrical, the letter is ឡ, not ឡ:

ឡឡ ឡឡ ឡឡ ឡឡ ឡឡ

Note that ឡ frequently resembles a Roman J.

If the head is large and closed, or nearly closed, the letter is ឡ. Note that this contradicts the TSL student's expectation that ឡ's head is usually written larger than ឡ's:

ឡឡ ឡឡ ឡឡ ឡឡ

An extended tail generally marks ឡ. If the tail is very long, it can be straight or slightly open (below left); if it's shorter, it's usually slightly closed:

ឡឡ ឡឡ ឡឡ ឡឡ

Of course, there are variations that defy these conventions. As you can see, they're readily identifiable within the group, but usually have to be guessed from context if seen in isolation:

ឡឡ ឡឡ ឡឡ ឡឡ ឡឡ

Ninth Group: ឡឡ

At last we arrive at a group that poses practically no problems. Although both these letters vary, they are so distinct from the rest of the

ନିଶି ତଳ ଗତ ନିଶି **ବତ ବତ**

ଶୁଣ ତାହା ଏହି ଗୀତ ଶୁଣି ଶୁଣି

Tenth Group: 55

၇၈ ၈၈ ၉၈ ၈၈ ၇၈

FD SD SD SD SD

Eleventh Group: ททท

ທຳທ ກກ ກກກ ກກກ
ກກກ

၆၇၆

Twelveth Group: Vowels and Tone Marks

٢ ٣ ٤ ١ ٢ ٣ ٢ ٣ ٤ ١ ٢ ٣

How *Do* Fluent Readers Learn?

It seems obvious that fluent Thai readers rely on their general knowledge of legal Thai constructions. The phenomena of *closure* comes into play; the unfamiliar character is automatically filled in by the reader's expectation of what he expects to see. Perhaps the best-known demonstration of this effect is:

THE UNIVERSITY OF CHICAGO

Page 11

main, along with an unstated understanding that the text will probably be relevant to the content of our discussion.

In more formal terms, I would propose that people draw on two types of knowledge to interpret unfamiliar letterforms:

- *Syntactic* knowledge of legal constructions; in particular, the knowledge that one possibility for the letter results in a correctly spelled word, while the other is nonsense.
- *Semantic* knowledge of the word's meaning; understanding the reader can use to guess at what the word should be, or to decide between two equally legal possibilities.

Faith in these phenomena is the basis of various forms of *cloze* testing. The student is given a text passage from which every *n*th word has been deleted, yet is expected to replace the word correctly. Modifications of the cloze test — the *C-Test* and the *X-Test*, which remove the second and first half, respectively, of every *second* word — come even closer to duplicating the fundamental problem of OCR. (See [KAN-CHANA94] for X- and C-test examples.)

Nor is it necessary for letters to be completely unambiguous. Consider an especially difficult pair like ๗/๗. A search of the more than 16,000 entries in an on-line copy of the Dictionary of the Royal Academy [ROYAL82] located 1,038 words that included the letter ๗ and 314 words that included the letter ๗. However, only 100 words were identical except for these letters; ie. only 100 words were potentially ambiguous.

Another way to put this is to say that well over 90% of the time, a fluent Thai speaker only needs to know that either ๗ or ๗ is in a word in order to recognize the word correctly. Add the reader's semantic understanding of the word in ambiguous cases, and it is no surprise that fluent speakers can read even scribbled handwriting.

It is also likely that fluent readers don't even attempt to identify all the letters in advance. Rather, the reader 'chunks' words into groups of letters that are initially identified by their overall appearance, including vowels and tone marks. If necessary, individual letters can be inspected after the fact; again, the fluent reader can instantly reject most possible alternatives because the resulting words would not make sense.

If this is the case, incidentally, we would expect that a letter that can be identified unambiguously merely by position would tolerate a

great deal of variation in appearance. Indeed, this is the case for ๗. Because of its unique role in shifting consonants from the low to the high class, ๗ frequently appears in contexts like ๗๗ and ๗๗, where it can be identified no matter what it looks like. As a result, its range of acceptable styles — from ๗ to ๗ — should come as no surprise.

Finally, certain styles have become accepted conventions as the result of regular use. For example, the curve alone of the initial stroke of ๗ and ๗ implies the orientation of the head — whether the head is there or not. Similarly, the simple curve ๗ *always* indicates ๗, and never ๗, despite their similarity. This sort of abbreviation is extremely confusing for TSL students, and can't really be understood until the student spends some time practicing writing.

The point of this digression is that fluent speakers need not explicitly learn the secondary distinctions between letterforms that this paper has described. Instead, because they are frequently exposed to new forms in unambiguous contexts, they are continually trained to look for the tiny differences that help maintain the internal design distance between letters — they learn what tricks are used, and which letters tend to vary in lockstep. As a result, they can instantly recognize variations they might not even be able to consciously describe.

Implications for TSL Methodology

What does this all mean for the introductory TSL student? First and foremost, we must recognize that focusing exclusively on the reference letterforms is going to be as frustrating to the student as learning only formal styles of speech would be. Signs and menus, notes and advertisements — he wants to read and understand them all, regardless of how they are written.

Second, we must accept that fluent readers and TSL students approach reading unfamiliar letterforms from entirely different vantage points. The fluent reader relies on spoken fluency when he deciphers unfamiliar letterforms. A new print style may be momentarily puzzling, but a fluent speaker can derive the style's underlying rules and conventions without conscious effort.

By definition, though, the TSL student is not a fluent speaker. He may not be able to decipher many letters, and certainly cannot easily infer the rules that underly a slightly exotic font's

variations from the reference standard, nor detect the secondary characteristics that remain in common.

As a result, it is necessary to help the TSL student explicitly. The range of variation in Thai print styles is certainly less pronounced than the range of lower- and upper-case letters in the Roman alphabet, and if first graders can handle the latter, I would think that adult TSL students can manage the former.

In conclusion, I think the TSL teacher should extend the standard introduction to the Thai alphabet in three ways:

- First, point out the secondary characteristics that are not usually explicitly mentioned. These include the shortened centers of ก and ก, the open waist of ข and lengthened neck of ข, and the closed downstroke of ก.
- Second, show common variations from the reference alphabet along with the standard forms. In particular, show how the craft and modern styles do away with circular heads, how script styles hint at the head's original orientation, how letters like อ rely on accepted conventions, and how letters like ข derive from historical influences.
- Third, draw the student's attention to ways that internal design distance is made consistent and maintained by showing how each individual letter fits into a group of similar letterforms, and how one letter's appearance can influence, and help predict, the form of others.

Appendix 1 contains a summary of the secondary characteristics discussed in this paper.

Implications for Thai OCR

We also find implications, both encouraging and discouraging, for Thai-language OCR. Again, it seems to me that we must begin by accepting that any system that is based on recognizing the reference style alone is going to fall short of the goal of reading Thai electronically. There are three fundamental reasons for this:

- First, the design distance between letters, especially of non-reference styles, is not always sufficient to overcome ambiguity introduced by the physical printing process. Even with the reference style, the internal

design distance is not always sufficient to distinguish letters clearly.

- Second, as a result of the availability of computer-based desktop publishing systems, it seems inevitable that printed text is going to contain an ever-increasing variety of letter styles.
- Third, specific features of some letterforms may vary to the extent that two letters may be identified when viewed side-by-side, but not when inspected independently.

On the other hand, computers are unlike TSL students in one essential way — the computer is able to mimic some aspects of perfect spoken fluency. As a result, an OCR program can, in many instances, unambiguously interpret words even when it cannot decipher individual letters. Three techniques are called for:

- *Syntactic disambiguation*; using dictionary lookup to reject illegal constructions, and narrow the solution set.
- *Statistical guesswork*; using knowledge based on actual usage of specific letters and words.
- *Semantic selection*; selecting one potentially legal candidate over another by looking at its function as a part of speech, or its meaning as one-half of a doublet.

The first techniques are an immediate possibility; see [COOPER95a]. The third is an active research area in Thai language analysis; see, for instance, [WUWONGSE93, SONLERT-LAMVANICH92].

Thai OCR Font Design

If a Thai OCR font is built, its design should consciously draw on the secondary characteristics this paper has described. There are two key principles to follow:

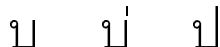
- An OCR font is not meant for computers alone; humans should find it as readable as an ordinary book font.
- The font should have an internal design distance of at least two features. In other words, every letter should have at least two features that distinguish it from all other letters.

We can achieve these goals by incorporating secondary characteristics *without* doing away with each letter's primary characteristics.

For example, the letters on the left, below, have two clear differences. If the head vanishes, or the interior is smudged, we can still tell the letters apart. In contrast, the pair on the right has just one distinction: the length of the tail.

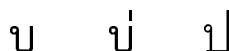


The problem this causes is obvious if I throw a tone mark into the picture:



Absent other features, it is very difficult for a computer to tell if the center character has a tone mark or merely a broken tail.

Where should the extra feature come from? Easy — get rid of one letter's foot (and slightly round the lower left corner as well) and we have added an inconspicuous secondary feature:



We can apply the same redistribution of features heres:



The left and center characters have at least two differences; ๒ has a closed waist as well as a tail. But the center and right-hand characters have just one easy-to-distinguish feature: the notched neck. Take away the center letter's foot, and you've restored balance: each letter has at least two distinctive marks.

Conclusion

This paper has looked at some of the ways in which common Thai printing fonts and handwriting styles vary from the reference standard. We have seen that in many cases, secondary characteristics that are not usually brought to the attention of TSL students must be recognized in order to tell letters apart.

For convenience, we can class variations into three categories. Primary variations usually involve applying an easily stated, consistent rule for simplifying or exaggerating the basic letter-form. Secondary variations are more dramatic and can usually be traced through a sequence of small changes. Tertiary variations often involve outside influences, such as the historic forms of

letters. In addition, there are changes that result from the need to maintain a reasonable design distance between letters.

It is possible to go through the alphabet letter-by-letter, and explicitly describe the secondary characteristics that distinguish letters from each other. However, we soon see that the variety and overall inconsistency of change makes it unlikely that OCR software will ever be able to recognize non-standard fonts without error on the basis of letter shapes alone.

I concluded with a variety of recommendations for TSL pedagogy, and Thai OCR. For OCR font design, I showed how it is possible to design fonts that are readable, but still maintain an internal design distance of at least two features between all letters.

Overall, I suggested that students can be taught to distinguish between letters on the basis of secondary characteristics, but that OCR software would do better to emulate fluent Thai speakers, and attempt to distinguish ambiguous letters by context. In effect, TSL students would do better to use the methods now used by computers, and computers would better profit by the approach used by Thai students.

Appendix 1: Collected Characteristics

This appendix summarizes the secondary characteristics discussed in this paper.

First Group: ขขฃ

- ข never has a waist; opening the letter is a common alteration:
- ข and ค, in contrast, frequently nip in just before the tail.
- An elongated neck and/or enlarged head is used to distinguish between ข and ค.
- A missing or slightly downward-curling line head generally indicates ข, while a flat, squiggly, or slightly upward-curling line head shows ค.
- If the waist pinches from the left side, the letter is ค, otherwise it's one of the others, eg.: ข ค

Second Group: บปผมหม

- Heads are almost invariably omitted: บ.
- A small foot or attached bar is used to replace a knot: บ บ.
- If it looks like the Roman letter U in a serif font, it's the Thai letter บ.
- If it looks like the Roman letter U in a sans serif font, it's the Thai letter ป.
- A flat bar is sometimes all that distinguishes ป from ฝ: บ ฝ
- In most fonts, almost any stylized head indicates ผ rather than ม: บ ผ ม.
- In the craft style, conflict arises between ข and ป because a) heads are minimal, and b) they are given curved bottoms to accentuate other letters' flat bar tops. The distinction must be memorized: ขป ม.

Third Group: กฃคฃกฃค

- The mouth is frequently done away with, so that ก becomes ก.
- When ก and ค are simplified, almost any squiggle indicates a ก: ก ก ก
- Even though ก and ค are simplified into ก and ก, ก will retain its mouth if necessary so that it can be distinguished from ค: ก ก
- Absent a clear indication of a head and mouth, if the upper-left corner is sharp, the letter is probably ก; if the upper-left corner is rounded, it is more likely to be ค: ก ค ก ค
- Craft-style letters don't have mouths; consequently ก is ค, not ก or ก, and ก is ค, rather than ก or ก.

Fourth Group: ณณญณ

- Heads are frequently omitted, and a small foot or attached bar replaces the knot: ณ ณ.
- The base of ณ is frequently attached to the letter: ณ ณ
- The interior head and downstroke of ณ are frequently either minimized or done away with: ณ ณ. A notch or dip in the top bar may be all that distinguishes these letters from ณ and ณ.

Fifth Group: ผฝพฟพ

- ผ and ฝ have a low central notch, and they practically always retain some hint of a head: ผ ผ ผ ผ
- If there is no head of any sort, the letter is almost invariably พ or ฟ: ผ ผ ผ ผ.
- If the first downstroke curves into the letter, it's พ or ฟ; if it curves away from the letter, it is ผ or ฝ: ผ ผ ผ ผ.

- If the notch and downstroke are ambiguous, look for some indication, however slight, of a head: **wwwwww** **wwwwww**.
- Any nick or notch in the tail indicates **𑀮**, eg. **𑀮 𑀮 𑀮 𑀮 𑀮 𑀮**.

Sixth Group: 𑀭𑀮𑀯

- If you can't tell whether a letter is **𑀭** or **𑀮**, it must be **𑀭**, eg.: **𑀭 𑀭 𑀭**.
- If the downstroke attaches at the baseline, the letter is **𑀭** (rather than **𑀮**), even if there's no distinguishable head: **𑀭 𑀭 𑀭**.
- If the downstroke 'closes,' or reattaches along the side of the upstroke, the letter is **𑀮**: **𑀮 𑀮 𑀮**.
- Almost any flattening or break in the top bar means the letter is **𑀮**, rather than **𑀭**: **𑀮𑀮 𑀮𑀮 𑀮𑀮**.
- A small foot is sometimes added to distinguish **𑀭** and **𑀮** from **𑀭** and **𑀮**: **𑀭𑀮 𑀭𑀮**.
- In the craft style, the downstroke attaches mid-bar for all letters. A new secondary distinction is introduced: the upstroke turns slightly outward for the **𑀭/𑀮** pair, and slightly inward for **𑀭/𑀮**: **𑀭𑀮 𑀭𑀮**.

Seventh Group: 𑀯𑀰𑀱

- If heads have been done away with, a slightly closed downstroke, or a sharp corner at the bottom, means that the letter is probably **𑀯**, rather than **𑀰**: **𑀯𑀯 𑀯𑀯 𑀯𑀯 𑀯𑀯**.
- A closed downstroke and generally longer tail distinguish **𑀯** from headless forms of **𑀰**: **𑀯𑀯 𑀯𑀯 𑀯𑀯**.
- The knot, or knot substitute, distinguishes **𑀯** from **𑀰**: **𑀯𑀯 𑀯𑀯 𑀯𑀯**.
- If a narrow letter has a base, it's **𑀯**, no matter what it looks like: **𑀯𑀯 𑀯𑀯 𑀯𑀯 𑀯𑀯**.
- In the craft style, the head of **𑀯** curves out slightly, the head of **𑀰** curves in slightly, and **𑀯** has a knot: **𑀯 𑀯 𑀯**. The letter **𑀰** does not conflict because it stays open: **𑀰**.

Eighth Group: 𑀱𑀲𑀳

- The tail of **𑀱** frequently attaches to the letter, and may be reduced to a dot or whisker: **𑀱𑀱 𑀱𑀱 𑀱𑀱**.
- If the head is large and closed, or nearly closed, the letter is probably **𑀱**, rather than **𑀲**: **𑀱𑀱 𑀱𑀱 𑀱𑀱**.
- If the head and tail are symmetrical, the letter is **𑀲**, not **𑀱**: **𑀱𑀱 𑀱𑀱 𑀱𑀱 𑀱𑀱 𑀱𑀱**.
- If it looks like a Roman letter J, it's inconclusive — sometimes **𑀱**, sometimes **𑀲**: **𑀱𑀱 𑀱𑀱**.
- **𑀱** usually has an extended tail, regardless of what happens to the head: **𑀱 𑀱 𑀱 𑀱**.

Ninth Group: 𑀳𑀴

- If it looks like the Roman letter **a**, it's **𑀳**: **𑀳 𑀳 𑀳 𑀳 𑀳 𑀳 𑀳**.
- Almost any loop or whisker makes the letter **𑀳**, not **𑀴**: **𑀳𑀳 𑀳𑀳 𑀳𑀳 𑀳𑀳 𑀳𑀳 𑀳𑀳 𑀳𑀳**.

Tenth Group: 𑀴𑀵

- If it looks like the Roman letter **S**, it's really **𑀴**: **𑀴 𑀴 𑀴 𑀴 𑀴 𑀴**.
- **𑀴** retains a longer downstroke, even when the letter is greatly modified: **𑀴𑀴 𑀴𑀴 𑀴𑀴 𑀴𑀴 𑀴𑀴 𑀴𑀴**.

Eleventh Group: 𑀶𑀷𑀸

- If it looks like the Roman letter **n**, it's really the letter **𑀶**: **𑀶 𑀶 𑀶 𑀶**.

- If it looks like the Roman letter K, it's really the letter ห: ห ห ห ห.
- The craft style uses historic letterforms that must be memorized: ฦ ฦ ฦ.

Twelveth Group: Vowels and Tone Marks

- If it looks like a Roman l or ll, it's really ๑ or ๒.
- If it looks like a colon, :, it's really ๓.
- If it looks like a question mark, ?, it's really ๔. If it turns right or left, it's ๕ or ๖: ๗ ๘ ๙.
- The vowels ๑ ๒ ๓ are frequently simplified to ๔ ๕ ๖.
- If it looks like a tilde, ~, it's probably ๗.
- The tone mark ๘ retains a hint of a curved head, even if ๘ has been simplified: ๘ - ๘.

References

- [BROWN79a] Brown, J. Marvin. 'AUA Language Center Thai Course: Reading and Writing Workbook (mostly reading)' American University Alumni Association Language Center, 1986.
- [BROWN79b] Brown, J. Marvin. 'AUA Language Center Thai Course: Reading and Writing Workbook (mostly writing)' American University Alumni Association Language Center, 1986.
- [DANVIVATHANA87] Danvivathana, Nantana. 'The Thai Writing System.' Ph.D. thesis, published by Helmut Buske Verlag, Hamburg, 1987.
- [GANDOUR91] Gandour, Jack and Potisuk, Siripong. 'Distinctive Features of Thai Consonant Letters.' *Journal of Language and Linguistics* 9:2 (2534), Thammasat University, 1991.
- [HAAS56] Haas, Mary. 'The Thai System of Writing.' Spoken Language Services, Inc./American Council of Learned Societies, 1956.
- [HIRANVANICHAKORN93] Hiranvanichakorn, Pipat and Boonsuwam, Monlada. 'Recognition of Thai Characters.' In 'Proceedings of the Symposium on Natural Language Processing in Thailand,' Chulalongkorn University, 1993.
- [KIMPAN93] Kimpan, Chom and Walairacht, Somsak. 'Thai Characters Recognition.' In 'Proceedings of the Symposium on Natural Language Processing in Thailand,' Chulalongkorn University, 1993.
- [SORNLERTLAMVANICH92] Sornlertlamvanich, Virach, and Phantachat, Wantanee. 'Information-based Language Analysis for Thai.' In 'Pan-Asiatic Linguistics: Proceedings of the Third International Symposium on Language and Linguistics.' Chulalongkorn University Printing House 1992.
- [WUWONGSE93] Wuwongse, Vilas and Pornprasertsakul, Ampai. 'Thai Syntax Parsing.' In 'Proceedings of the Symposium on Natural Language Processing in Thailand,' Chulalongkorn University, 1993.