How Do Thais Tell Letters Apart?

Doug Cooper <doug@chula.ac.th>
Center for Research in Computational Linguistics, Bangkok

Introduction

Pity the foreign student who thinks that he has finally mastered the Thai alphabet's loops and turns. After he leaves his ก ไก่ alphabet primer behind, he finds that carefully memorized rules for telling letters apart are nowhere to be found.

Hungry for a meal, he looks for ฐ ร้านอาหาร, but can only find ร้านอาหาร after รานอาหาร. When he finally follows his nose, he searches the menu in vain for **ข้าวผัค** before deciding to try *ข้าวผัค* and ข้าวผัด. Imagine his surprise when two orders of a third dish — *ข้าวผัด* — appear *on the bill instead!*

This paper investigates the reasons for our student's dilemma. We will find that while Thai printing fonts and handwriting vary considerably from the reference letterforms, letters have consistent secondary characteristics easily recognized by fluent Thai speakers. Unfortunately, these characteristics are obscured by traditional reading and writing instruction, and are not taken into account by prototype optical character recognition (OCR) systems.

For example, consider this elementary rule: ค is distinguished from ด by the inward or outward orientation of the letter's head. Although the rule is true, it doesn't help us decide what this letter is: ค. At ordinary text sizes, the head's position in this everyday printing font is ambiguous, and cannot be deciphered by either students or OCR programs.

But if we see ค and ด in various print styles, we can derive secondary characteristics and infer new rules. A new salient feature — the bar's origin, rather than the circle's orientation — emerges to resolve the ambiguity:

คด → คด → คด → คด

ด's bar always starts at the base of the letter, while ค's bar creeps up the left side. In effect, if the bar is too short for the reference alphabet rule to apply, the letter is probably ค, not ด.

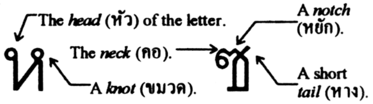
Overall, we will find, first, that certain secondary characteristics are usually retained regardless of style, and second, that inspecting just a few letters is usually enough to let us *predict the entire alphabet's design.* We also find that a variety of foreign influences and stylistic conventions (some of which are introduced simply to make letters distinct) have been incorporated into widely used fonts.

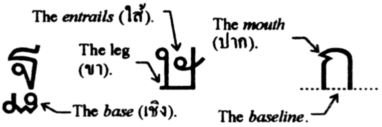
I'll begin by defining terms we'll need to describe Thai letterforms, and summarize traditional ways of describing them. Next, we investigate variations from the reference standard, and see that they may be unpredictable. After a close look at the alphabet, I discuss how fluent readers cope with unfamiliar styles.

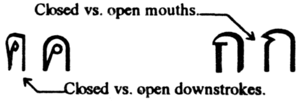
I'll close with specific recommendations for Thai language instruction, and discuss the *implications* for Thai-language OCR and OCR font design. We find, surprisingly, that students would benefit from the methods currently used in programs: getting detailed descriptions of the physical characteristics that distinguish letters. Computers, in turn, would benefit from applying the methods — considering the letter in context —used by fluent Thai speakers.

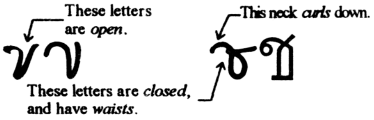
Anatomy of Thai Letterforms

We'll begin with some terminology. The nomenclature of Thai characters is not universal, but we can use these descriptive terms:








These letters are *open*.
This neck *curls* down.
These letters are *closed*, and have *waists*.

The Traditional Approach

Both the Thai and English-language literature traditionally describe characters in terms of the head's orientation. Mary Haas's *The Thai System of Writing* is typical (emphasis hers):

> *All consonants except ก and ธ 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.* (HAAS 1956*)*

Haas also includes many pages of handwritten and 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)* (BROWN 1979:a,b) also introduce the reference style, while dealing extensively with a variety of handwriting styles in two appendices. For example, Appendix 1 of Volume R has many hand-written 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 ข, ช, and ซ can be omitted completely (though ซ must keep its jags), and the difficult narrow parallel lines can be replaced by various kinds of loops.* (BROWN 1979)

A 1991 study by Gandour and Potisuk, *Distinctive Features of Thai Consonant Letters*, proposed a classification system 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.* (GANDOUR 1991)

This study shows what happens when the standard introduction to the reference alphabet is taken at face value. Even within the reference alphabet, we can 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 ฟ ฝ and ผ พ (their entries 9 and 142, table 2) — only the orientation of the heads is assumed to be significant. But 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.* (HIRANVANICHAKORN 1993)

> *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.* (KIMPAN 1993)

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

> *Reasons [for] rejection and misidentification were mainly due to differences in the number of holes ... between input data and models.* (HIRANVANICHAKORN 1993)

> *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.* (KIMPAN 1993)

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; for examples, see the 5" by 7" flip-books like *รวมแบบลายมือ*, which contain page after page of hand-lettered samples.

From this wealth of designs, we can focus on five primary variations 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, and both horizontal and vertical lines are regular and perpendicular. Here are typical letters from Cordia New:

  ก ข จ ท พ อ

- The *craft* style (เขียนตัวศิลปะ). A highly calligraphic, Indian-influenced style, drawn with a broad pen or brush point. Heads are semi-circular at most; if possible, letters have a distinctive horizontal top bar, a style found in modern Devanagari printing fonts (like those used for Sanskrit, Hindi, and Marathi). These letters are from JS Chanok:

  ก ข ง ท ร ว

- The *tail* style (เขียนเล่นหาง). Characteristics include fairly regular pen thickness, and an exaggerated tail that may wrap around the body. These letters are from JS Wansika:

  ง ซ ป ล ว อ

- 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 ค ว น บ ห

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

  ข ง ช พ ย ร

See (DANVIVATHANA 1987) for additional discussion of the origin 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. There are slight variations between instances of the reference style; eg. Cordia New adds a small leg to letters that have a left corner at the baseline, eg. ฆ ฏ บ.

Groups of Letters

Within each alphabet style, we find groups of letters whose forms are so similar that their designs are inextricably bound together.

This is due to Thai's origins. When he defined the modern Thai alphabet sometime before 1283, King Ramkhamhaeng began with the cursive Khmer script (derived in turn from Indian scripts) then modified it to account for the sounds of 13th century Sukhothai Thai as well (see ANUMAN 1968, BROWN 1985, DANVIVATHANA 1987).

As a result, many letters differed from the start only because of added notches and tails, while others have grown closer together over the centuries. Today, many letters are essentially *isomorphs*; identical but for the orientation of a head or knot.

In the dozen or so groups set out below, note that inspecting just one letter is often sufficient to let us anticipate what *all* the members of the group are going to look like. I've laid out consonants first, followed by numbers, various marks, and vowels.

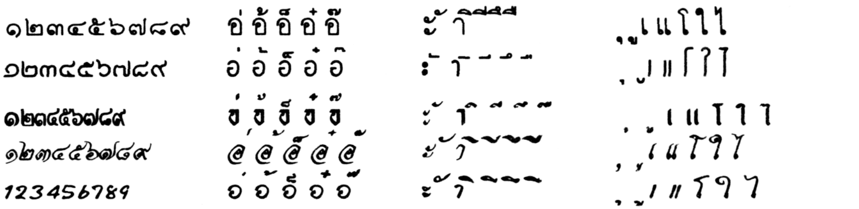
ขชซย บปษนมฆ กภถฤฦฎฏ ณฌญฒ ผฝพฟฬ คศดต จฐฉ งวอฮ ลส รธ ทฑห

ขชซย บปษนมฆ กภถฤฦฎฏ ณฌญฒ ผฝพฟฬ คศดต จฐฉ งวอฮ ลส รธ ทฑห

ขชซย บปษนมฆ กภถฤฦฎฏ ณฌญฒ ผฝพฟฬ คศดต จฐฉ งวอฮ ลส รธ ทฑห

ขชซย บปษนมฆ กภถฤฦฎฏ ณฌญฒ ผฝพฟฬ คศดต จฐฉ งวอฮ ลส รธ ทฑห

ขชซย บปษนมฆ กภถฤฦฎฏ ณฌญฒ ผฝพฟฬ คศดต จฐฉ งวอฮ ลส รธ ทฑห



Styles: reference (Cordia New), modern (JS Thanaporn), craft (JS Chanok), tail (JS Wansika), script (JS Sirium)

Internal Design Differentiation

Partitioning the alphabet into groups highlights the phenomenon of *internal design differentiation* — the introduction of artificial features to compensate for ambiguity. Systematic modifications in style are balanced by an internal pressure that develops inside the alphabet itself.

Internal design differentiation is an important concept in Thai font design. It leads to unpredictable changes in letterforms, and can present insurmountable problems for OCR.

For example, design for compact printing creates a bit of ambiguity in Cordia New (our reference font) — the pair ฃ/ซ is hard to distinguish.

Other reference-style-like fonts compensate by extending the neck *downward* slightly. Even at ordinary sizes, below, the head of the second letter clearly hangs below the head of the first.

ฃ ซ (Cordia New — *difficult to distinguish*)
ฃ ซ (Angsana New)
ฃ ซ (Dillenial UPC)
ฃ ซ (JS Prasoplarp)

That example was easy. In contrast, note the differences between ง and ว 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:

งวอ → งวอ → งวอ

There's no way that we could have predicted just where and what those extra variations would be. In one case a bar replaces the letter's head; in the other, it replaces the letter's tail. Note also that the relative proportions of the ง tail and ว head are reversed. Look at what happens when I mix the fonts:

งว → งว / งว

Consequently, particular features are less important than the requirement that they vary from each other: if one letter's tail is extended, another's head must be abbreviated; if one line is straight, another will curl. And for the TSL student (no less than for the OCR program) it implies that certain letters must be identified in context, or studied as a group.

Three Degrees of Variation

Letterform variations run the gamut from the straightforward and obvious to the unexpected and occasionally indecipherable:

Primary variations: บ becomes บ
Secondary variations: ค becomes ค
Tertiary variations: ท becomes ທ

Primary variations involve a single guideline, like 'delete the circle' or 'extend the tail.' Other rules are prompted by the instrument, real or imaginary, used to draw the letters. For example, in the craft style circular heads are replaced by angled wedges that are more easily drawn with a brush:

บพน becomes บพน

Secondary variations entail bringing the letter's lesser characteristics to the fore. The best example is the progression that leads to the ค/ค variation:

คด → คด → คด → คด

Tertiary variations are unpredictable, and often reach outside the alphabet in search of alternative designs. For example, the letterforms ທ and ຫ are the historical forebears of ท and ห, and are still found in the modern Lao alphabet. Other letterforms come from modern Roman designs. Here are reference, Thai, and Roman letters:

หรลทนง หsanuv ksanuv

Before we look at the alphabet, let's survey the variants more closely.

<u>Primary Variations</u> The mouth is most likely to be simplified or deleted. With one exception, it adds no information to the reader's reading of any letter.[1] The orientation of the head (or head substitute) alone distinguishes ก from ภ and ถ, and the location of the knot decides between ณ and ฌ. For instance:

กถภณฌ **กถภณฌ** กถภณฌ

The head goes next, for the same reason, and the knot soon follows. The traditional craft style is a transitional style; the head is minimized, but not done away with entirely. Most modern styles prune the letter as severely as possible:

ม ม ม ม ม ม ม

Vowels, tone marks, and letter bases or entrails are usually simplified as well:

แนะอ้ แนะอ้ แนะอ้ แนะอ้

Exaggeration is the other chief primary variation. Exaggerated styles tend to be easy to read; eg. the extended tail style, like its English counterpart, is common in wedding invitations and formal announcements:

เขียนเล่นหาง

<u>Secondary Variations</u> Secondary variations frequently originate in handwriting shortcuts. Consider how ค and ด move from standard to script-like to modern fonts:

คด *คด* **คด** คด คด

Because of the direction in which the head starts, ค tends to close the angle at the lower left corner far more quickly than ด does. The head disappears entirely, then the downstroke merges with the upstroke, and we're left with ก.

จ and ฐ follow the same progression. The head and first downstroke of both letters is abbreviated and reattached, and the tail of ฐ is folded along the body of the letter:

จฐ จฐ จฐ **จฐ** จฐ

Secondary variations also appear as minor features of the basic letterforms become more prominent. For instance, with the head in place, the height of the *internal notch* in พ/ผ and ฟ/ฝ is usually ignored.

But when the head is minimized or removed, 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 พ or ฟ:

พผฟฝ **พผฟฝ** พผฟฝ

The ร/ธ pair demonstrate a kind of 'sympathetic' secondary variation. In the reference form, the head and tail of ร are entirely different. Little by little, the original letter loses its distinctive appearence, and acquires the more symmetrical, less technically demanding shape of the English letter S.

รธ รธ *รธ* รธ **รธ**

What's unexpected is that ร drags ธ along for the ride — both letters usually change in tandem.

Finally, some 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:

ญฎฏฐปฝฮ **ญฎฏฐปฝฮ**

<u>Tertiary Variation</u> A few letters are regularly transformed from the reference style into shapes that are not easily predicted or prescribed. We have already seen two letters that derive from historical shapes:

ท ห ท ห

Various modern letterforms into this category as well. Again, ห is the most dramatic example, but ญ and ย can vary greatly from the norm, too:

หญ หญ **หญ** ย**ย**ยย

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

A few highly stylized conventions are also encountered. An extended neck, as in ช/ซ, is a predictable secondary variation. A *cleft* or slightly upward-curling head, in contrast, is a completely artificial convention that replaces the neck, complete with head and notch, in headless styles:

ชซท ชซท ชซท ชซท ชซท

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

กขคฆง **กขคฆง** กขคฆง

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

A Close Look at the Letters

Let's take a methodical look at each group of letters.

First Group: **ขชซย**

The first group presents two issues: telling ช from ซ, and telling ข from the others. While the reference rules focus on the notch and tail, characters taken in isolation often have imperceptible notches (ซ), false tails (ข), and missing heads (ข).

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

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

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

ขช ขช ขช ขช

Examples of enlarged heads/enlongated necks:

ชซ ชซ ชซ ชซ

Finally, here are missing or highly stylized heads:

ชซ ชซ ชซ ชซ ชซ

ชซ ชซ ชซ ชซ ชซ

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 ย, which is why that letter is in this group:

ชซย ชซย **ชซย** **ชซย**

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

Second Group: **บปษนมฆ**

Letters in this group are fairly easy to distinguish once the reader stops looking for heads and circular knots. The transition is:

บนมฆ *บนมฆ* บนมฆ บนมฆ

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

บนมฆ **บนมฆ** บนมฆ

There are also cases in which the design distance between letters is vanishingly small. Worse, the close resemblance between these highly stylized letters and their Roman alphabet counterparts is jarring, and it is difficult for foreign students to avoid seeing the letter 'u,' below:

บนมบ บนมบ **บนมบ**

In the very popular craft style, conflict arises between ข and บ because letters are given curved bottoms to accentuate the flat bar heads. This reverses the pattern of the reference style:

ขบ **ขบ** *versus* กล **กล**

Third Group: **กภถฤฦฎฏ**

Variations in this group have to do with opening, closing, and deleting, the letter's mouth and head.

กถ กถ กถ กถ

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

ฎฏ ฎฏ ฎฏ ฎฏ ฎฏ

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 difference between letters. Here, even though ภ's mouth is removed, ถ's mouth is retained so that it can be distinguished from ค:

กภถค กภถค กภถค

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 ค; note that it's the only one with a rounded upper-left corner:

กภถค กภถค กภถค

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:

กภถค กภถค กภถค กภถค

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

คค คค ภถ ภถ

Fourth Group: ณฌญฒ

This group has a combination of the second and third groups' features. Usually, letters are legible no matter how much they have been altered:

ณ ฌ ญ ฒ ณ ฌ ญ ฒ

Nevertheless, they may go quite far from the reference standard. On the first line, left, below, note the crossbar protruding slightly to indicate a knot. We also see the base of ญ attach to the body of the letter in the second line's examples:

ณฌญฒ ณฌญฒ

ณฌญฒ ณฌญฒ

These letters are also good gauges of the degree to which letters can be simplified. Below, note that ฒ can be reduced far more than either ต or น alone — ฒ has few letters it can be mistaken for.

ตนญฒ ตนญฒ ตนญฒ

Fifth Group: ผฝพฟฬ

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 ฬ, in turn, almost any loop or notch will do.

In practice, there are really two secondary clues to look for: first, ผ 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 พ or ฟ:

พผฟฝฬ พผฟฝฬ พผฟฝฬ

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

พผฟฝฬ พผฟฝฬ พผฟฝฬ

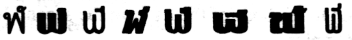
However, the notch's height is the surest indicator. Note that in some of the examples below, the head's orientation really is ambiguous in comparison to the reference standard:

พผฟฝฬ พผฟฝฬ พผฟฝฬ

The exceptions have vanishingly small heads and no notch variation, and are very hard to read:

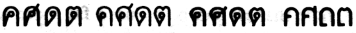
พผฟฝฬ พผฟฝฬ

ฬ 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 variatión. As we have seen before, such variations are not difficult for humans to distinguish, but they can pose problems for OCR. For example:



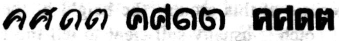
Sixth Group: คศดต

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:

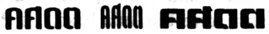


In general, if the downstroke attaches at the baseline, the letter is ด, even if there's no distinguishable head, eg. ด.

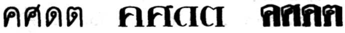
Almost any closing of the downstroke (ie. reattaching) hints at a ค, and almost any opening implies that the letter is ต:



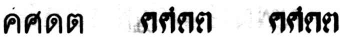
The notch that distinguishes ต from ด disappears rapidly. Almost any flattening or break in the top bar must be recognized as an indicator of ต:



A small foot sometimes, but not always, distinguishes ต and ด from ค and ศ. These designs may be unpredictable, but at least they're consistent:

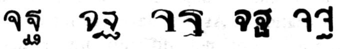


The craft style is a special case. The font samples below show that the downstroke is reattached mid-bar for all letters. A new secondary distinction is introduced: the upstroke turns slightly outward for the ค/ศ pair, and slightly inward for ค/ศ.

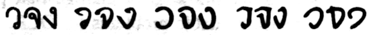


Seventh Group: จฐฉ

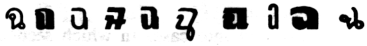
These letters are quite different from each other in the reference style, but are subject to a great deal of variation. The variations themselves are all familiar from other groups; the most dramatic involve shortening and reattaching the downstroke:



Note that the downstroke of จ 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 จ from headless forms of ว. The letter's tail, in turn, tends to stay long, which helps distinguish จ from headless forms of ง. In the last example below, the extended downstroke is prompted by the need to keep ว identifiable:

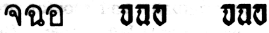


The third letter of this group, ฉ, often ends up with an appearance that is recognizable, but not very attractive:

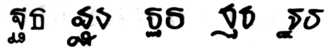


In general, the size of ฉ distinguishes it from จ, while the knot, or knot substitute, distinguishes it from ว and ข.

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



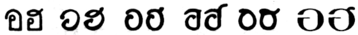
And, were it not for the base, ฐ would have a potential conflict with ธ in a variety of fonts:



Eighth Group: งวอฮ

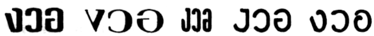
The large number of words these characters appear in make this group the most problematic. In addition, this group has the largest number of internally prompted variations; two letters may be modified in opposite ways in different fonts.

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



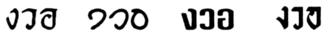
Aside from pointing out that the tail of ฮ may be reduced to a bare dot, we'll ignore ฮ.

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 ว, not ง:

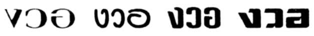


Note that ง frequently resembles a Roman J.

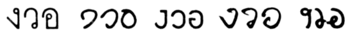
If the head is large and closed, or nearly closed, the letter is อ. Note that this contradicts the student's expectation that ว's head will usually be written larger than อ's:



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

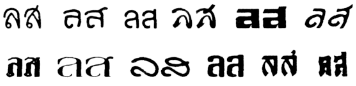


Of course, there are variations that defy these conventions. The letters below are readily identifiable within the group, but usually have to be guessed from context if seen in isolation:



Ninth Group: ลส

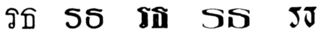
At last we arrive at a group that poses practically no problems. While both letters vary, they are so distinct from the rest of the alphabet that they are almost always recognizable.



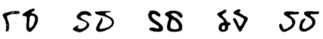
As noted earlier, some variations of ส are interesting because of the obvious debt they owe to Roman alphabet design.

Tenth Group: รธ

This group is also influenced by Roman letter design, and the result is often barely recognizable. Here are the most common variations; note that ธ always manages to retain a slightly longer head:



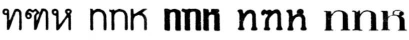
Handwritten styles subject these letters to a great deal of stress, even in fonts or style that are otherwise fairly clear. Note the large, closed loop of ธ in the first three examples, below:



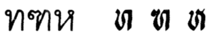
The only letter that comes close is อ, but the final direction of the tail is enough to distinguish the letters easily.

Eleventh Group: ทฑห

The three final consonants are subjected to two drastic, but usually consistent, variations. The first is modernization:



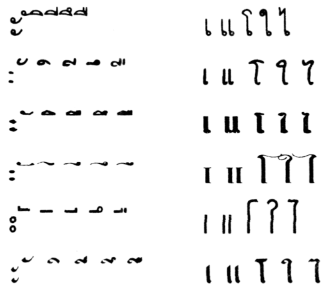
As before, the resemblance to Roman letters is obvious. The second major change is reversion to the historic letterforms, most commonly seen in the craft style. Here are the same three letters:



As noted earlier, this variation must be memorized; it can't be seen as reasonably deriving from any intermediate style.

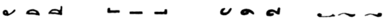
Twelveth Group: Vowels and Tone Marks

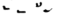
Alterations in vowels and tone marks follow one basic rule: simplify. The key characters tell most of the tale by themselves:



ใ usually has at least a minimal circle, but in the case of โ and ไ, a hint of direction is sometimes all there is.

The over vowels are a little more difficult to read very quickly. In general, อั curls up, while อิ curls down or remains flat. A '~' shape is usually, but not invariably, อี.



Finally, even when heads are greatly simplified, the tone mark ้ retains a hint of a curved head, no matter how much ั has been trimmed; eg.: 

Tactics of Fluent Readers

As we've seen, common Thai fonts include letters that:

- are ambiguously designed — ว,
- rely on alternative historical designs — ห,
- are essentially indistinguishable — พ, and
- are so far from the standard that they are practically arbitrary symbols — ญ.

If Thai basal readers and ก ไก่ practice books only present the reference alphabet, how do fluent readers deal with non-standard styles? Alternative strategies include:

- They understand shortcuts derived from handwriting, and applied to print fonts.
- They recognize unstated stylistic conventions defined by consistent use.
- They learn alternative designs, like the historical craft style, that are not taught explicitly.
- Spoken fluency tells them that the letter is unique in its context — it might not have to be 'read' at all in order to be identified.

A variety of clues are learned through repeated exposure. Some letters have nonstandard, but conventional, representations (as, in the Roman alphabet, 'a' and 'g' take the place of 'ɑ' and 'ɡ'). The curve of the initial stroke of พ and ผ implies the orientation of the head, whether the head is there or not. Similarly, the simple curve ɔ *always* indicates ว, and never ง, despite their similarity.

The alternative alphabets, particularly the craft style, rely on an entirely different set of clues. Evidently, they are learned without ever being presented formally. It is interesting to note that *among* fonts of the craft style there is almost no variation whatsoever — as though the secondary characteristics of this style are unfamiliar even to native Thai readers.

That letters can be disambiguated through context is an easily demonstrated phenomenon:

In *Fuzzy Letters and Thai OCR* (Cooper 1995) I investigated the general question of disambiguating minimally distinguished letter pairs through local (word-length) context. It turns out that Thai's large alphabet, diverse origins, and relatively short word-list make it relatively unlikely that two letters with similar appearance will be ambiguous.

Consider an especially difficult pair like ข/ช. A search of more than 16,000 entries in an on-line copy of the Dictionary of the Royal Academy (ROYAL 2525) located 1,038 words that included the letter ข and 314 words that included the letter ช. 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 ข *or* ช is in a word in order to recognize the word correctly. The pairs (like similar vowels) that suffer in highly stylized alphabets are even less likely to coincide. Add the reader's understanding of the word's meaning in ambiguous cases, and it is no surprise that fluent speakers can read even *scribbled handwriting*.

We can go further, and predict that any letter that can be identified solely by position will tolerate a great deal of variation in appearance. Indeed, this is the case for ห. Because of its unique role in shifting consonants from low to high class, ห can frequently be recognized no matter what it looks like (as in ใหม่ and ใหญ). The range of acceptable styles — from ห to ห — comes as no surprise.

Implications for Teaching Thai

What does this all mean for the introductory Thai student? First and foremost, we must recognize that focusing exclusively on the reference letterforms is as frustrating 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 second-language students approach unfamiliar letterforms from very different vantage points. The fluent reader relies on spoken fluency when he reads. A new print style may be momentarily puzzling, but a fluent speaker can derive its underlying rules and conventions without conscious effort.

But the second-language 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, we have to help the student explicitly. I recommend extending 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 ผ and ฝ, the open waist of ข and lengthened neck of ช, and the closed downstroke of ก.
- Second, show common variations from the reference alphabet as well as 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 ว rely on accepted conventions, and how letters like ฟ derive from historical influences.
- Third, draw the student's attention to the way that each letter fits into a group of similar letterforms, and show 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. We must begin by accepting that any system that is based on recognizing individual letters probably won't work. There are three fundamental reasons for this:

- First, the difference between letters, especially of non-reference styles, won't always overcome ambiguity introduced by the physical printing process. Even in the reference style, letters aren't always identifiable
- Second, variations in letterforms are not always predictable exaggerations or simplifications of existing styles. Specific features of some letterforms can vary to the extent that two letters may be identified when viewed side-by-side, but not when inspected independently.
- Third, computer-based desktop publishing systems make it inevitable that printed text is going to contain an ever-increasing variety of letter styles. The problem is getting worse, not better.

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. Tactics include:

- Using dictionary lookup to reject illegal constructions.
- 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.
- As a last resort, basing a guess on usage statistics for specific letters and words.

The first techniques are an immediate possibility; see (COOPER 1995) The third is an active research area in Thai language analysis; see, for instance, (WUWONGSE 1993, SONLERTLAMVANICH 1992).

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.
- Every letter should have at least two features that distinguish it from other letters.

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

For instance, the letters on the left, below, have two clear differences. If the head goes, 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.

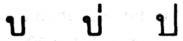
พ ผ   บ ป

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

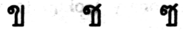
บ   บ่   ป

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 here:



The left and center characters have at least two differences; ฑ 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 students must be recognized in order to tell letters apart.

We can class variations into three categories. Primary variations involve an easily stated, consistent rule for simplifying or exaggerating the basic letterform. 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. Occasionally, we also see artificial changes that result from the need to maintain a reasonable design distinction between letters.

We can 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 Thai-as-a-second-language pedagogy, and Thai OCR. For OCR font design, I showed that 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, students would do better to use the methods now used by computers, and computers would better profit by the approach used by native Thai students.

References

Brown, J. Marvin. *AUA Language Center Thai Course: Reading and Writing Workbook (mostly reading)* American University Alumni Association Language Center, 1979a.

Brown, J. Marvin. *AUA Language Center Thai Course: Reading and Writing Workbook (mostly writing)* American University Alumni Association Language Center, 1979b.

Cooper, Doug. *Fuzzy Letters and Thai Optical Character Recognition*. In *Proceeding, Symposium on Natural Language Processing '95*, Kasetsart University, 1995.

Danvivathana, Nantana. *The Thai Writing System*. Ph.D. thesis, published by Helmut Buske Verlag, Hamburg, 1987.

Gandour, Jack and Potisuk, Siripong. Distinctive Features of Thai Consonant Letters. *Journal of Language and Linguistics* 9:2 (2534), Thammasat University, 1991.

Haas, Mary. *The Thai System of Writing*. Spoken Language Services, Inc./ American Council of Learned Societies, 1956.

Hiranvanichakorn, Pipat and Boonsuwam, Monlada. Recognition of Thai Characters. In *Proceedings of the Symposium on Natural Language Processing in Thailand*, Chulalongkorn University, 1993.

Kimpan, Chom and Walairacht, Somsak. Thai Characters Recognition. In *Proceedings of the Symposium on Natural Language Processing in Thailand*, Chulalongkorn University, 1993.

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.

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

Appendix 1: Collected Characteristics

This appendix summarizes the secondary characteristics discussed in this paper.

First Group: ขชซย

— ข never has a waist; opening the letter is a common alteration.

— ช and ซ, in contrast, frequently nip in just before the tail.

— An elongated neck and/or enlarged head is used to distinguish between ข and ช.

— A missing or slightly downward-curling line head generally indicates ช, while a flat, squiggly, or slightly upward-curling line head shows ซ.

— If the waist pinches from the left side, the letter is ย, otherwise it's one of the others, eg.: ย ข

Second Group: บปษนมฆ

— Heads are almost invariably omitted: บ.

— A small foot or attached bar is used to replace a knot: ม ม.

— If it looks like the Roman letter u in a serif font, it's the Thai letter น.

— If it looks like the Roman letter U in a sans serif font, it's the Thai letter บ.

— A flat bar is sometimes all that distinguishes บ from ษ: บ ษ

— In most fonts, almost any stylized head indicates ฆ rather than ม: ม ม ฆ.

— In the craft style, conflict arises between ข and บ because *a*) heads are minimal, and *b*) they are given curved bottoms to accentuate other letters' flat bar tops. The distinction must be memorized: ขบ ขบ.

Third Group: กภถฤฦฎฏ

— The mouth is frequently done away with, so that ก becomes ก.

— When ฎ and ฏ are simplified, almost any squiggle indicates a ฎ: ฎ ฎ ฏ ฏ

— Even though ก and ภ are simplified into ก and ภ, ถ will retain its mouth if necessary so that it can be distinguished from ด: ถ ด

— Absent a clear indication of a head and mouth, if the upper-left corner is sharp, the letter is probably ถ; if the upper-left corner is rounded, it is more likely to be ด: ถ ด ถ ด

— Craft-style letters don't have mouths; consequently ด is ด, not ก or ถ, and ค is ค, rather than ก or ภ.

Fourth Group: ณฒญฌ

— Heads are frequently omitted, and a small foot or attached bar replaces the knot: ณ ณ.

— The base of ญ is frequently attached to the letter: ญ ญ

— The interior head and downstroke of ฌ are frequently either minimized or done away with: ฒ ฌ. A notch or dip in the top bar may be all that distinguishes these letters from ณ and ญ.

Fifth Group: ผฝพฟฬ

— ผ and ฝ have a low central notch. They almost always retain some hint of a head: ผ ฝ ผ ฝ

— If there is no head of any sort, the letter is almost invariably พ or ฟ: พ ฟ พ ฟ.

— If the first downstoke curves into the letter, it's พ or ฟ; if it curves away from the letter, it is ผ or ฝ: พผ ฟฝ.

— If the notch and downstroke are ambiguous, look for some indication, however slight, of a head: ผพผพ ฝฟฝฟ.

— Any nick or notch in the tail indicates ฬ, eg. ฬ ฬ ฬ ฬ ฬ ฬ.

Sixth Group: คศดต

— If you can't tell whether a letter is ค or ด, it must be ค, eg.: ค ค ค.

— If the downstroke attaches at the baseline, the letter is ด (rather than ถ), even if there's no distinguishable head: ด ด ด.

- — If the downstroke 'closes,' or reattaches along the upstroke, the letter is ค: ค ค ค.
- — Almost any flattening or break in the top bar means the letter is ต, rather than ด: ดต ดต ดต ดต.
- — A small foot is sometimes added to distinguish ด and ต from ค and ศ: ดต ดต.
- — 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 ค/ศ pair, and slightly inward for ด/ต: คศ ดต.

Seventh Group: ขฐฉ

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

Eighth Group: งวอฮ

- — The tail of ฮ often attaches to the letter, and may be cut to a dot or whisker: อฮ อฮ อฮ.
- — If the head is large and closed, or nearly closed, the letter is probably อ, rather than ว: วอ วอ วอ วอ วอ วอ.
- — If the head and tail are symmetrical, the letter is ว, not ง: งว งว งว งว งว.
- — If it looks like a Roman letter J, it's inconclusive — sometimes ง, sometimes ว: งว งว.
- — ง usually has an extended tail, regardless of what happens to the head: ง ง ง ง.

Ninth Group: ลส

- — If it looks like the Roman letter a, it's ล: ล ล ล ล ล ล ล ล.
- — Almost any loop or whisker makes the letter ส, not ล: ลส ลส ลส ลส ลส ลส ลส ลส ลส.

Tenth Group: รธ

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

Eleventh Group: ทฑห

- — If it looks like the Roman letter n, it's really the letter ท: ท ท ท ท.
- — 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: ท ฑ ห.

Twelveth Group: Vowels and Tone Marks

- — If it looks like a Roman I or II, it's really เ or แ.
- — If it looks like a colon, :, it's really ะ.
- — If it looks like a question mark, ?, it's really ไ. If it turns right or left, it's ใ or ไ: ไ ใ ไ.
- — The vowels ั ี ื are frequently simplified to ั ี ี.
- — If it looks like a tilde, ~, i's probably ื.
- — The tone mark ้ retains a hint of a curved head, even if ั has been simplified: ั ี ั้