How Do Thais Tell Letters Apart?

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

Introduction

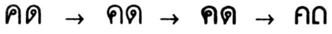
Pity the foreign student who thinks that he has finally mastered the Thai alphabet's loops and turns. After he leaves his ก ไก่ alphabet primer behind, he finds that carefully memorized rules for telling letters apart are nowhere to be found.

Hungry for a meal, he looks for ร้านอาหาร, but can only find ร้านอาหาร after ร้านอาหาร. When he finally follows his nose, he searches the menu in vain for ข้าวผัด before deciding to try *ข้าวผัด* 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:



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.

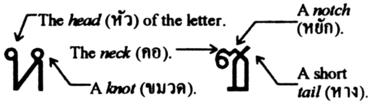
Overall, we will find, first, that certain secondary characteristics are usually retained regardless of style, and second, that inspecting just a few letters is usually enough to let us *predict the entire alphabet's design.* We also find that a variety of foreign influences and stylistic conventions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

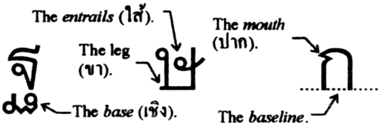
I'll begin by defining terms we'll need to describe Thai letterforms, and summarize traditional ways of describing them. Next, we investigate variations from the reference standard, and see that they may be unpredictable. After a close look at the alphabet, I discuss how fluent readers cope with unfamiliar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

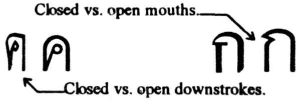
I'll close with specific recommendations for Thai language instruction, and discuss the *implications* for Thai-language OCR and OCR font design. We find, surprisingly, that students would benefit from the methods currently used in programs: getting detailed descriptions of the physical characteristics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

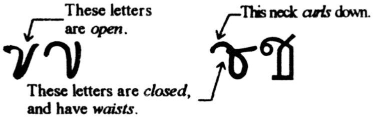
Anatomy of Thai Letterforms

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, ฃ, 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 ข 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 จ 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 ษ 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.

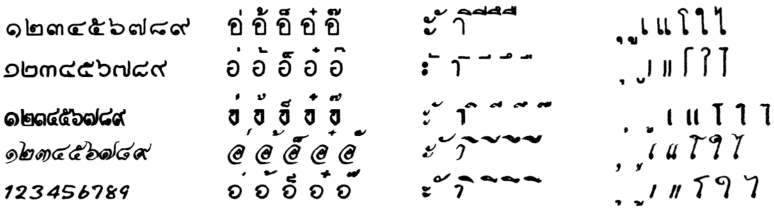
ขชซย บปษนมฆ กภถฤฦฎฏ ณฌญฒ ผฝพฟฬ คศดต จฐฉ งวอฮ ลส รธ ทฑห

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 ฆ/ฑ 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.

ฆ ฑ (Cordia New — *difficult to distinguish*)
ฆ ฑ (Angsana New)
ฆ ฑ (Dillenial UPC)
ฆ ฑ (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 U

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: