

# Antikythera Publications

Relational Database Design

<http://www.AntikytheraPubs.com>

[foberle@AntikytheraPubs.com](mailto:foberle@AntikytheraPubs.com)

## DATABASE DESIGN NOTE SERIES

### Exploring Han Script: Globe-Trotting 汉字 and its Kanji Cousin Multi-Script Database Series #7

Prepared by: Frank Oberle

So far in this series we've ignored Hànzì, still in use after 3,000 years by well over a billion people to write a variety of Sinitic and other languages and dialects – and possibly the oldest writing system in continuous use. While much of what was presented in earlier design notes about management of non-Latin Scripts in corporate data remains relevant, Hànzì differs from Arabic, Cyrillic, Devanagari, Hangul, Hebrew, Latin and Thai in a number of ways, primarily because, while those are phonetic – written with alphabets, abjads, abugidas or even syllabaries<sup>α</sup> – Hànzì is primarily logographic.

Our intent, similar to that of earlier notes, is to introduce data stewards to written Hànzì, primarily Chinese, at a very basic level without needing to learn that language. Hànzì is a very complex Script however; depending on the data's origins, there will be differences in usage, standards, conventions, and even conflicting exceptions. But this overview will give you the vocabulary you'll need to undertake whatever further research is needed to handle any issues that may arise.

Prerequisites, such as familiarity with Unicode, multi-script data entry, and the like are covered in previous notes.

Revised August 2021 for public distribution

<sup>α</sup> See Multi-Script Database Series #1, "Exploring Alphabets"



数据库管理

データベース管理

Copyright © 2021 by the Author and Antikythera Publications

Permission is granted to distribute unaltered copies of this document, so long as this is not done for commercial purposes.



## Database Design Note Series on maintaining Multi-Language/Multi-Script Databases

( All available for download from [www.AntikytheraPubs.com/i18n.htm](http://www.AntikytheraPubs.com/i18n.htm) )

1. Exploring Alphabets
2. Exploring Complex Text Layout
3. Exploring UTF-8
4. Evaluating Fonts for use in Multi-Lingual Documents
5. Exploring Bidirectional (BIDI) Text Entry
6. Exploring Arabic Script Behavior
- 7. Exploring Han Script Entry – Chinese**
8. Keyboard Layouts – Hello World
9. Evaluating Bidirectional Text Handling Behavior in Applications

## Table of Contents 篇目

Introduction.....	5
Hànzi's Flavors – 汉字 (simplified) and 漢字 (traditional).....	5
Han Script – The Taxonomy of a Chinese “Character”.....	6
Primitives, Strokes, Radicals, Characters, and Words.....	6
Strokes (笔画).....	6
Exploring Stroke Combinations.....	7
From Strokes (笔画) to Radicals (部首).....	8
Radical Types and Single Character Combinations.....	9
Sorting Single Character Combinations.....	9
Radicals in Multi-Character Combinations.....	10
Using a Chinese Dictionary.....	10
The Ten Most Common Radicals.....	10
Han Script – from Brushes to Computer Keyboards.....	11
Typing Pīnyīn with a Latin Keyboard – a phonetic path to Hànzì.....	11
Typing Hànzì with a Pīnyīn Keyboard – the requisite “Hello, World” Example.....	12
Determining the correct Pīnyīn – Introducing Ruby Fonts.....	15
Typing Hànzì with Alternate (non-phonetic) Methods.....	16
Typing with Strokes – Introductory Example.....	16
Strokes (笔画) Reference Chart.....	17
Stroke Primitives.....	17
Grouping Strokes by Similarity.....	18
Revisiting “Hello World” with the Stroke Entry Method.....	18
Choosing Simplified or Traditional Hànzì.....	19
A Difficult “Simplified” Character in Strokes.....	19
A More Difficult “Traditional” Character in Strokes.....	19
Augmenting Compose Key Definitions to Support Typing Pīnyīn Diacritics.....	20
Hànzi Characters Travel Abroad – Birth of Japanese Writing.....	20
Japanese Kana Scripts.....	21
Hiragana Script “Alphabet”.....	21
Katakana Script “Alphabet”.....	21
“Hello World” – Comparing Chinese Hànzì to Japanese Hiragana+Kanji.....	22
“Database Management” – Comparing Chinese Hànzì to Japanese Katakana+Kanji.....	22

篇

目



piān  
篇

mù  
目

中文

目  
次



mù  
目  
cì  
次

日本語



## Introduction

In previous papers in this series, we used the United Nations' *Universal Declaration of Human Rights* as a common text<sup>1</sup> to illustrate some of the differences across the variety of writing systems that might appear in a modern database. On the right is the Mandarin Chinese version of the two opening sentences of that document.

Each cell<sup>2</sup> is occupied by a single glyph. These glyphs, of which there are thousands,<sup>3</sup> are called 汉字 (hànzì) – a combination of Hàn (汉) meaning “Chinese” (Han people), and 字 (zì), usually translated as “character” or “word.”

### Logograms vs. Characters, Words, and Syllables

The word “character” however is quite misleading when comparing this Script to others we’ve explored so far. Like other ancient writing systems such as cuneiform and hieroglyphics, each hànzì is actually more akin to a logogram (also called a logograph) than a “character” or “word” in the English sense.

The phrase “Chinese character,” however, is so commonly used as a translation for hànzì, that we’ll reluctantly adopt that term in this document, but continue to promote “Chinese writing” as a more useful/accurate translation of 汉字.

Although each hànzì does represent a spoken syllable, Chinese writing is not a syllabary in the same sense as Korean hangul<sup>4</sup>, which are formed from letters (jamo). Likewise, hànzì are not necessarily “words.” Some single syllable hànzì may be words, but many “words” are formed from more than one syllable.

Logograms, written symbols representing a single morpheme or sound, are not historical relics; even in the west, we use many of them on a daily basis. The symbol 6, for example, is understood and interpreted the same way around the globe. Unlike alphabetic characters such as A, however, 6 has a meaning but no inherent phonetic value and, indeed, is pronounced quite differently depending on the speaker’s language. Symbols like %, +, and &, as well as road sign symbols, shorthand, and emoji fall into that same category.

The term “Chinese” as used in this document does not, by the way, refer to a single *spoken* language, but is a collective term for a variety of often mutually unintelligible dialects<sup>5</sup>. For our purposes, we can informally think of Hànzì as being similar to the aforementioned 6 glyph – a commonly understood set of *written* symbols. There is one important caveat to be aware of however.

人人生而自由,在尊严和权利上一律平等。他们赋  
有理性与良心,并应以兄弟关系的精神互相对待。

*Mandarin Chinese (Pǔtōnghuà 普通话) translation of the UN's UDHR Opening*

### Hànzì's Flavors – 汉字 (simplified) and 漢字 (traditional)

In the 1950s, the PRC/Chinese government defined “simplified” replacements for some complex hànzì glyphs to help increase literacy, but traditional Chinese (fántǐ zhōngwén 繁体中文) is still in use both there and elsewhere; either type might be found in the data of a global corporation.

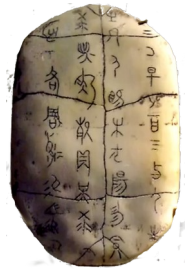
Comparing simplified Chinese (jiǎntǐ zhōngwén 简体中文) to traditional, however, illustrates that “simplified” doesn’t necessarily equate to “simple.” Compare, for example, the respective versions of “love” (ài): 爱 and 愛.

Data from a given geographic area or entity will, of course, have mostly one form or the other. Be aware that the choice of forms can be a sensitive subject.

Other early logographic writing systems such as Cuneiform – used at one time by at least ten disparate spoken languages, and Hieroglyphics – used across Egypt’s vast empire, both dropped out of use thousands of years ago, making hànzì possibly the world’s oldest writing system still in continuous use.



*Mesopotamian cuneiform (left), and early Hànzì (right), are examples of ancient writing systems, but only Hànzì has been in continuous use since it first appeared. Hànzì was (and sometimes still is) written from top-to-bottom, though now mostly appears left-to-right.*



1 See a comprehensive collection of translations of this text in most of the world’s languages at: <https://omniglot.com/udhr/index.htm>

2 The idea of a virtual “character cell” was introduced in the section *Character Codes and Character Cells* on page 8 of the first handout in this series, *DDN-1 Exploring Alphabets*.

3 These are spread across eleven different code planes in the Unicode standard, the two largest of which, u4e00 and u20000, contain over 20,000 and 40,000 glyphs respectively

4 Korean Hangul Script is discussed beginning on page 30 of the second handout in this series, *DDN-2 Exploring Complex Text Layout*.

5 Languages include Mandarin (standard Chinese), Cantonese (Yue), Wu (Shanghainese), Min (Minbei/Fuzhou), Minnan/Hokkien/Fukien, Xiang, Gan, Hakka, and many more.

# Han Script – The Taxonomy of a Chinese “Character”

## Primitives, Strokes, Radicals, Characters, and Words

With the information presented in this document, an ability to write Hànzì won't be required for IT personnel to be able to handle normal maintenance or problem diagnosis, but acquaintance with the basics can often help when determining the most suitable source for assistance if problem escalation becomes necessary.

Explaining hànzì composition in taxonomic terms, while not common in tutorials intended for language study, seems appropriate for programmers and database designers who are familiar with the terms class, hierarchy, and the like. 部首 (bùshǒu), the word for Radical, is in fact composed of the hànzì 部, meaning “unit or section,” and 首, meaning “first” – i.e. “section heading” – so why not “key?”

Using such an analogous perspective, the class hierarchy from lowest to highest would be: Primitives, Strokes, Radicals, Hànzì (our “characters”), and Words. Strokes and Radicals are often referred to collectively as Hànzì's Components.

## Strokes (笔画)

Hànzì glyphs are constructed, not by aggregating miniature letters as we saw with Korean, but from one or more of the brush strokes defined in Unicode Plane U31C0<sup>6</sup> and displayed on the right. The six highlighted cells are discussed in *Typing Hànzì with Alternate (non-phonetic) Methods* on page 16. Less frequently used Strokes are shown in red. Strokes fall into five general classes; the class number for each follows the Unicode value below each stroke.

Brush strokes are not lines (i.e. straight paths between two points); a “stroke” might consist of as many as four connected segments that are painted without lifting the brush. “一” is written with one stroke, but “罐” requires twenty-three.

One characteristic of Strokes that may be surprising to westerners is that both the order and direction in which the strokes are written is far more rigidly defined than it is with Latin alphabetic characters. Stroke order in particular is quite important in some forms of data entry, sorting, and dictionary look-up, though its importance is fading somewhat as computers become “smarter.”

It will also be evident that distinguishing one stroke from another may require careful attention: compare 丩, 丩 and 丩 (31c4, 31d7 and 31df), ㇇, ㇇ and ㇇ (31cb, 31cc and 31e1), 乚 and 乚 (31c8 and 31cd), or 丿 and 丿 (31c0 and 31d3). As with any Script, the font design in use may also be a factor.

A (brush) **Stroke** is the smallest atomic element of a **Hànzì** “character” and, depending on the reference you choose, there are as few as five or as many as thirty-six, but most sources suggest the number in use today is thirty-two.

A **Radical** (部首/bùshǒu) consists of one or more **Strokes**, and the same **Stroke** may appear more than once in a single **Radical**. Once again, depending on the source used, the number of **Radicals** given will vary, but the quantity traditionally is considered to be 214. As with **Hànzì**, there are, naturally, both traditional and simplified forms of some **Radicals**.

Each **Hànzì** consists of a **Radical**, and possibly one or more additional Strokes.

A **Word** consists of one or more cells, each containing one **Hànzì**.

一 U31c0:1	丿 U31c1:5	㇇ U31c2:5	㇇ U31c3:5	㇇ U31c4:5	㇇ U31c5:5	㇇ U31c6:5	㇇ U31c7:5	㇇ U31c8:5
㇇ u31c9:5	㇇ u31ca:5	㇇ u31cb:5	㇇ u31cc:5	㇇ u31cd:5	㇇ u31ce:5	㇇ U31cf:4	一 U31d0:1	丨 U31d1:2
丿 U31d2:3	㇇ U31d3:3	㇇ U31d4:4	㇇ U31d5:5	㇇ U31d6:1	㇇ u31d7:5	㇇ u31d8:5	㇇ U31d9:2	㇇ U31da:2
㇇ u31db:5	㇇ u31dc:5	㇇ U31dd:4	㇇ u31de:5	㇇ u31df:5	㇇ U31e0:5	㇇ u31e1:5	㇇ U31e2:4	㇇ U31e3:

Thirty-six possible bīhuà (笔画 – pen paintings) used to draw Chinese “characters”

Before proceeding to an explanation of Radicals, it will be helpful to take a closer look at how Strokes can morph as they are combined to form Radicals and Hànzì.

Proper names and other details for each stroke, including the use of stroke classes, are given in the *Strokes (笔画) Reference Chart* on page 17. A visual aid for helping to distinguish similar strokes is given in *Grouping Strokes by Similarity* on page 18.

Tip for bored readers: Strokes & Radicals are neither political parties nor  $\sqrt{\text{math}}$  symbols.

6 This standard is titled “CJK Strokes” – the common CJK acronym means Chinese-Japanese-Korean; sometimes you’ll see a V added to mean Vietnamese. Hànzì, borrowed from the original Chinese long ago, were used at one time by these other language families, but are seldom seen today in Korea or Vietnam. They are still encountered in Japanese, however (see page 21). The canonical order given here for strokes (丿, 一, 丨, etc.) can and does vary across Sinitic languages and dialects.

## Exploring Stroke Combinations

As if the similarities between some strokes isn't enough of a challenge, identical strokes used in the same hànzi might be written differently; some are inevitably shrunk or stretched to fit the group into a composite whole.

For this table, refer to the LEGEND in the lower right; the designations **Pīnyīn** and **Kangxi** will be introduced later.

For now, just compare the hànzi in each of the pairs and note the differences. The simplest hànzi shown here (A1 and B1) have two strokes each, while F2 and D3 each have eight.

Compare the wǔ and gàn hànzi in A3 and D2, where adding a single stroke alters the meaning. Then compare the dissimilar hànzi 未 and 为 in C1 and E1 where the sound is identical.

Just as the horizontal strokes in D2 may appear in different lengths, so may vertical ones.

The 士 (shì) in B2, by the way, can also mean warrior or knight. Just as we see in the western tales of the Knights of the Round Table, the most elite medieval warriors in the east also came from the educated classes.

The cattle/cow hànzi in B3 (牛) will be seen again when **Radicals** are introduced.

As with alphabetic scripts, similar hànzi may also be used by typosquatters in homograph attacks – the spoofing of URLs on the web (e.g. www.blinduser.com or blinduser.com instead of blinduser.com<sup>7</sup>). The mechanism is just somewhat different. Though not an issue normally handled by database managers, a quick overview may be useful in cases where malicious entries have infected a corporate data store, and I/T or DBA assistance is required.

1		2		3		
A	刀 Dāo #131 Knife 18 U5200	These two hànzi are both drawn with the 丿 and 丨 strokes, but use different proportions.	土 Tǔ #34 Earth 32 u571F	These hànzi are drawn with two 一 strokes of different widths and one 丨 stroke, but the wider 一 is at the bottom of 土 and the center of 士.	午 Wǔ #687 Noon - u5348	Again, there are two 一 strokes of different widths as well as a 丿 and a 丨 stroke. Note where the horizontal strokes sit on the 丨: 午 牛
B	力 Lì #11 Force 7 u529B	A side by side view helps distinguish them.	士 Shì #35 Scholar 33 u58EB		Niú #340 牛 Cattle, cow 93 u725B	
刀力						
C	未 Wèi #274 Not u672A	The five strokes in these two hànzi are identical. Only the positioning of the dissimilar width 一 strokes determines the meaning.	千 Qiān #354 Thousand u5343	The strokes 丿 and 一 at the top of each hànzi look similar.	住 Zhù #76 Live - u4F4F	Here a single 丿 stroke at the top left changes the hànzi. The other two strokes on the left are moved downward.
D	末 Mò End(ing) - u672B		干 Gàn #223 Dry - u5E72	Also note the relative lengths of the two 一 strokes in the 干 hànzi. Compare both to 午 in A3 above.	往 Wǎng #683 To(ward) - u5F80	Again, note the widths of the three 一 strokes.
E	为 Wèi #576 For - u4E3A	A homonym of the 未 character in C1 above. Note how the 丿 stroke appears in the middle of 为 but on the right of the 办 hànzi; the same strokes can be arranged differently.	坏 Huài #651 (be) Bad - u574F	The addition of a short single 一 stroke at the upper left causes a change in angle of the 丿 stroke, moving its ending point below the 丿 at the left.	LEGEND	H: the hànzi glyph. Pīnyīn: Romanized phonetic rendering. #00: HSK <sup>8</sup> id number. English: one possible translation. 00: the KangXi radical sequence number.
F	办 Bàn #648 (to) Do - u4EBA		环 Huán Ring - u73AF		H Pīnyīn #00 English	
					00 Unicode	

While this overview must of necessity be simplified, it should provide enough subject matter background and vocabulary for data managers to productively assist with troubleshooting and support any remediation efforts.

- URL is short for Universal Resource Locator. To be really “universal” it must be composed of nothing but seven-bit ascii characters, and the highest level domains must be a valid ISO 3166-1 alpha-2 two character country code.
- To permit most of the world to use their own languages and writing systems in web addresses, a format known as punycode is used to convert non-Latin characters to an ascii byte-stream.<sup>9</sup> A Thai URL [www.พรังค์โอเบลลี่.th](http://www.พรังค์โอเบลลี่.th), for example, would be converted to [www.xn-42cf2drup8cb7e8efq3k.th](http://www.xn-42cf2drup8cb7e8efq3k.th). Depending on the browser in use, either or both forms of the URL might be displayed somewhere on the header or within the page.
- A few countries have introduced representations of the high level domains in their local script as well, which has opened up many more opportunities for spoofing.
- One simple test is to detect whether all characters in a single portion of the URL are in the same Unicode Plane.

<sup>7</sup> The digit 1 is used instead of a small L in the first URL; in the second, a Turkish “dotless” small ı (U+0131) is used instead of the Latin/English dotted small i.

<sup>8</sup> 汉语水平考试 (Hànyǔ Shuǐpíng Kǎoshì, aka HSK – is the PRC’s Chinese/Mandarin proficiency test); Taiwan has a similar Mandarin proficiency test known as the TOCFL.

<sup>9</sup> A variety of on-line converters (e.g. <https://www.punyocoder.com/>) are available should you need to experiment with these.

## From Strokes (笔画) to Radicals (部首)

Since Radicals are formed from one or more Strokes, and Hànzì “characters” are formed from Radicals – and perhaps additional strokes – it follows that characters are ultimately formed from Strokes, so the next obvious question is why Radicals exist as an intermediate class.

As an I/T practitioner, you may have already wondered how words written with non-alphabetic characters such as Hànzì are looked up in a dictionary or, more relevant to a data center, how they can be usefully sorted.

At first glance, since 32 strokes isn’t much more than the 26 letters of the English alphabet, it might appear reasonable to sort by strokes; remember, though, that strokes form “characters,” not words. A little thought will show that such an approach would quickly become unwieldy. Radicals are therefore used as what in I/T terms could be called an intermediate class; Radicals are composite groupings of strokes that define (among other things) a subclass of characters. One Radical of each Hànzì can be viewed as the primary key<sup>10</sup> used to perform the highest level sorting of Chinese characters/words.

While this might imply that there is a standard number and order for Chinese Radicals, that unfortunately is not the case. There are 214 in the traditional KāngXī (康熙) set, but some 20<sup>th</sup> century dictionaries began using an updated set containing 227 Radicals with, of course, a modified order. In 2009, the Beijing/PRC government introduced a 201-radical system (Table of Han Character Radicals, 汉字部首表) as a national standard for simplified Chinese. It is therefore important to know which standards are of concern in your own data store(s) and which are supported within your DBMS.<sup>11</sup>

	2E8	2E9	2EA	2EB	2EC	2ED	2EE	2EF
0	丿 2E80	尢 2E90	民 2EA0	纟 2EB0	卅 2EC0	车 2ED0	彳 2EE0	龙 2EF0
Rows removed for clarity								
E	兀 2E8E	步 2E9E	夂 2EAE	卅 2EBE	彳 2ECE	食 2EDE	齿 2EEE	
F	允 2E8F	母 2E9F	纟 2EAF	卅 2EBF	卩 2ECF	食 2EDF	竜 2EEF	

*Extracted from Unicode Publication  
“CJK Radicals Supplement 2E80–2EFF”*

The figure on the left is extracted from one page of the Unicode “CJK Radicals Supplement”<sup>12</sup> that lists 115 of the 214 traditional 康熙字典 (KangXi) Radicals.

In this extract, the radical 丿 (u+2e80) has two Strokes, while 齿 (u+2eee) has eight.

The Radicals are ordered in the Unicode planes catalog by their number of Strokes.

Words are further sorted within Radical subsets by the number of strokes required to complete them<sup>13</sup>. The following section *Radical Types and Single Character Combinations* also serves as a simple example of Hànzì sorted by their Radicals and stroke counts. In that section the Radical 女 forms the “key” for both the 奶 and 娘 Hànzì (as well as others), but since 奶 has two “extra” strokes and 娘 has seven, 奶 will appear first in a dictionary and a “standard” sort. See the illustration below.

The data stored on disk represents Hànzì code points – although the individual Strokes each have Unicode values for standalone use, corporate data is rather unlikely to contain raw strokes, but they will still be sorted appropriately if it does.

### EXAMPLE

The Hànzì Radical 女 is written with the first three strokes in the table on the right. The composite Hànzì one-syllable word 奶 containing that Radical is formed by adding two additional strokes. More examples are shown in *Typing with Strokes – Introductory Example* on page 16.

STROKE 1 丶	STROKE 2 ㇇	STROKE 3 一	STROKE 4 ㇇	STROKE 5 丿
㇇	㇇	女	奶	奶

10 This does not purport to be a proper linguistic definition, but rather an analogy using yet another familiar I/T term. Linguists use descriptions like “semantic indicators” or “shared orthographic components” (e.g. prefixes, suffixes, roots). Another analogy is to view Radicals as a library of common functions that are “included” to compile the Hànzì characters.

11 Documentation supplied by your DBMS vendor will provide an important clue to its capabilities. [https://docs.oracle.com/cd/B12037\\_01/server.101/b10749/ch5lings.htm](https://docs.oracle.com/cd/B12037_01/server.101/b10749/ch5lings.htm) and <https://www.oracle.com/technetwork/database/database-technologies/globalization/twp-appdev-linguistic-sorting-10gr2-132064.pdf> are typical examples for the Oracle RDBMS.

12 See <http://www.unicode.org/charts/> for the most up-to-date Unicode character code charts. All can be freely downloaded as pdf documents if required.

13 ... that is, in addition to the strokes that make up the Radical itself.



## Radical Types and Single Character Combinations

More than 85% of hànzi characters are picto-phonetic, meaning they contain both a semantic element as well as a phonetic element. Most often, but not always, a semantic element (similar to a morpheme) gives a hint as to the hànzi's meaning, i.e. the general class of things to which it belongs (or originally belonged) and under which it can be located in sorted lists such as dictionaries.<sup>14</sup> A phonetic element generally suggests the pronunciation, but due to the passage of time that pronunciation may no longer be current.

The semantic, or key, radical, while most often found on the left or top of a cell, might be placed anywhere; to make things more interesting, semantic radicals might mean different things depending on their position: When placed on the left, 月 (moon/month; radical 74) means body; the hànzi 胖 (pàng) for example combines 月 (body) with 半 (half) to suggest body fat. When placed on the right 月 refers to the passage of time; 其 combines with 月 to form the composite 期 (qī) suggesting a period of time or an “issue” (a particular instance of a periodical, e.g. a magazine).

Any radical can be either semantic or phonetic, and its position within the character doesn't necessarily indicate which; additional strokes may be added to define the hànzi more specifically.

Since radicals often share a character cell with additional strokes, they are often distorted to some degree (compare B with C, where 女 is shrunk to the left side of the cell. In I, 女 appears almost as a subscripted element at the cell's bottom center, while in L it occupies most of the lower portion of the cell. Finally, in example M, 女 is compressed into the bottom right corner. Sometimes the distortion can render the glyph much less recognizable. Though not illustrated here, the basic hànzi for the person class (人 rén) appears as 亻 is the radical form of 人 rén (person) though it doesn't always mean “person” when used in composite characters.

So while English dictionaries have twenty-six “sub-headings,” one for each letter of the alphabet, Chinese dictionaries will have two-hundred-fourteen (assuming use of the traditional kāngxī radicals) “sub-headings.”

As mentioned on page 8, hànzi are then further sorted according to the number of strokes that are added to the key radical to form the character.

The table on the right illustrates how the three-stroke Radical 女 – originally a pictogram of a girl curtsying – combines with other radicals to form composite hànzi that fall into the subset of “female.” In all but example L, 女 is the semantic element and key radical; in L, both the roof radical (kāngxī #40, new #45) and the 女 radical are semantic elements.

SORTING SINGLE CHARACTER COMBINATIONS

	Hànzi	Hànzi Components	Comments	Unicode Value	# of Strokes
A	女	𠃉 + 丿 + 一 strokes z+p+h	Three strokes in the order shown form the traditional kāngxī (康熙) radical number 38 (of 214) or the early twentieth century radical number 73 (of 227).	u2f25	3
B	女	女 woman, female	The word “ <b>female</b> ” is a single character formed from a single radical with no additional components. Note that its Unicode value is different from that of its radical.	u5973 38.0 nǚ	3
C	奶	女 + 乃 female + is (to be)	Literally “ <b>a female is</b> ” suggests meanings like “milk,” “breast” or “lady.” This is sorted in section 38.2 since it is radical 38, with 2 additional strokes added to 女's 3.	u5976 38.2 nǎi	5
D	好	女 + 子 female + child	“ <b>good</b> ” – both 女 and 子 are semantic components, the idea being that a woman loving a child exemplifies “good.” A homonym for 好 means to like or be fond of.	u597d 38.3 hǎo	6
E	妈	女 + 马 female + horse	“ <b>Mom</b> ” (informal) or “nurse”: the word for “horse” (马) is pronounced “mǎ” suggesting a female called “Ma.” As in much of the rest of the world, babies all know 妈妈.	u5988 38.3 mā	6
F	妹	女 + 未 female + “not yet”	“ <b>younger sister</b> ” – while it may be tempting to view 未 as another semantic component, it is phonetic.	U59b9 38.5 mèi	8
G	始	女 + 台 female + platform	“ <b>begin, start</b> ” – 台, meaning “station,” “platform,” “stage” and similar words, is the phonetic component	u59cb 38.5 shǐ	8
H	姐	女 + 且 female + further	“ <b>older sister</b> ” – like 未 in F above, 且 is simply a phonetic component, and reading it as semantic, while tempting, is incorrect, though it could be a mnemonic.	u59d0 38.5 jiě	8
I	威	戍 + 女 military + female	“ <b>prestige, power</b> ” – in this hànzi the primary “female” radical is strokes 4-6 inside the word xū (army/military). 戍, which is another semantic component of this word.	u5a01 38.6 wēi	9
J	娘	女 + 良 female + good	“ <b>Mother</b> ” (formal): traditional form 孃 is pronounced differently – 良 is a phonetic component, notwithstanding its meaning of “good, virtuous, or respectable.”	u5a18 38.7 niáng	10
K	婚	女 + 昏 female + nightfall	“ <b>marriage</b> ” – 昏 means “dusk, nightfall, twilight, dark; to faint, to lose consciousness,” but it serves only as a phonetic component.	u5a5a 38.8 hūn	11
L	安	宀 + 女 roof + female	“ <b>calm, peaceful</b> ” – the secondary “female” radical, still semantic, is below 宀. “Female under a roof” – might be called “domestic tranquility” if we were so inclined.	u5b89 40.3 ān	6
M	矮	矢 + 未 + 女 dart+grain+female	“ <b>short, not tall</b> ” – three semantic components that are “shorter” relative to other elements: dart < spear; grain < tree; female < male.	u77ee 111.8 ǎi/aǐ	13

14 See “Using a Chinese Dictionary” on page 11 for more detail.

## Radicals in Multi-Character Combinations

Just as Radicals are combined with other strokes or radicals to form single “characters”, those hànzi join with others to form multi-syllable words. ㄉ

In *Exploring Stroke Combinations* on page 7, the 牛 (cow/cattle) appears in table cell B3. In DDN-1 we saw how the middle-eastern bovine drawing 𠂇 rotated first to 𠂈 and eventually to the Greek Α (u+0391) and, later, the equivalent Latin A (u+0041) and Cyrillic А (u+0410) characters. In the far east, similar bovine representations such as the very early 𠂉 Oracle Bone Script morphed through the Han seal script 𠂊 and others, eventually becoming the current (牛 u+725b).

On the previous page, a variety of single hànzi combinations using the 女 Radical appeared. The list of multi-character examples below includes one of those (奶).

牛 niú the now familiar (hopefully) hànzi for “cow”  
 牛奶 niú+nǎi cow+milk = Milk (note the female Radical 女 in 奶)  
 牛肉 niú+ròu cow+meat = Beef, as an ingredient  
 牛饭 niú+fàn cow+rice = Beef (with) Rice, as a menu item or meal  
 午饭 wǔ+fàn noon+rice = Lunch (compare the 牛 and 午).

## Using a Chinese Dictionary

1. Begin by finding the identifying Radical or Key – usually on the hànzi’s left or top.
2. Go to the beginning page for that section, shown on the dictionary’s index.

3. Count the number of strokes added to the key radical; go to the beginning of the dictionary’s subsection for that number of added strokes.

Some dictionaries provide a Pīnyīn index to their entries – a useful alternative.

## The Ten Most Common Radicals

The ten most common Radicals – based on their appearance in 7,141 (34%) of the 20,992 Hànzi characters in the Unicode CJK Unified Ideographs block – are shown in the table on the right. The columns represent the following:

- A. The Radical index number in the set of 214 traditional Kangxi Radicals.
- B. The Radical index number in the set of 227 “modern” Radicals.
- C. The simplified Hànzi glyph used with Mandarin and other languages.
- D. Pīnyīn phonetic spelling of the Hànzi character.
- E. Unicode code points for the stand-alone Radical (upper) and the identical single-radical Hànzi character (lower).
- F. The number of strokes in the radical and its Hànzi twin.
- G. An English translation of the single-radical Hànzi character and original, *though not necessarily current*, meaning of the radical when it is used as a semantic element in single hànzi character combinations
- H. The number of hànzi characters containing this Radical that appear in the traditional Kangxi Dictionary of 49,030 characters.
- I. The number of entries in the Unicode “CJK Unified Ideographs” and the Unicode range of ideographs in which the glyph is found.
- J. Hànzi character number in the 汉语水平考试 (Chinese HSK test)<sup>15</sup>

	A	B	C	D	E	F	G	H	I	J
1	140		艸	cǎo	U+2F8B U+8278	6	“grass”	1,902	981 U+8278-864C	
2	85	125	水	shuǐ	U+2F54 U+6C34	4	“water”	1,595	1,079 U+6C34-706A	473
3	75	94	木	mù	U+2F4A U+6728	4	“tree”	1,369	1,016 U+6728-6B1F	80
4	64	55	手	shǒu	U+2F3F U+624B	4	“hand”	1,203	740 U+624B-652E	27, 28
5	30	58	口	kǒu	U+2F1D U+53E3	3	“mouth”	1,146	756 U+53E3-56D6	53
6	61	41	心	xīn	U+2F3C U+5FC3	4	“heart”	1,115	581 U+5FC3-6207	19, 87
7	142	174	虫	chóng	U+2F8D U+866B	6	“insect” “bug” “worm”	1,067	469 U+866B-883F	731
8	118	178	竹	zhú	U+2F75 U+7AF9	6	“bamboo”	953	378 U+7AF9-7C72	65
9	149	185	言	yán	U+2F94 U+8A00	7	“speech” “words”	861	567 U+8A00-8C36	57
10	120	77	糸	sī	U+2F77 U+7CF8	6	“silk”	823	574 U+7CF8-7F35	48

15 Taiwan has a similar standardized test of Mandarin proficiency levels for non-native speakers that is usually known as the TOCFL (Test Of Chinese as a Foreign Language).

## Han Script – from Brushes to Computer Keyboards

Since the invention of the alphabetic typewriter in the nineteenth century, many schemes have been used to permit typing the thousands of Hànzì “characters.” With the advent of computers in the mid-twentieth century, such schemes began to become more affordable and practical. The historical path is out of scope here, but for technologists, it’s as fascinating as the exploits of Babbage and Lovelace<sup>16</sup>. The modern solution to printing Hànzì without needing enormous keyboards and prohibitive amounts of training was to develop a method of phonetically representing Sinitic sounds, ideally with available “standard” keyboards.

Latin Script, familiar to many, and used for the alphabets of languages with such disparate needs as English, Español, Français, Čeština, and Türk, was recommended as the basis for a phonetic input standard for Mandarin to be called **Hànyǔ Pīnyīn** (拼音, literally “Han spell-sound”).

### Typing Pīnyīn with a Latin Keyboard – a phonetic path to Hànzì

Unlike the ability to type Hànzì, the ability to type Pīnyīn is less likely to be of use to data professionals, but is included here in case that need arises. Typing pīnyīn is essentially the same as typing English with the addition of some diacritics. Except when typing the underscore character, the shift key is seldom necessary, though capital letters are sometimes used at the beginning of sentences or proper names to make westerners feel comfortable. Capital letters are always optional, however.

The family of “Chinese” languages are tonal; an identical word with different tones often have different meanings. Mandarin – the focus of this paper – uses five tones, plus an optional neutral tone indicator. though different languages may have more or fewer. Cantonese, for example, has six. In pīnyīn, tones are indicated by adding accent marks/diacritics to the vowels in each syllable – easily done using standard Compose<sup>17</sup> key sequences. Alternatively, some earlier pīnyīn texts use superscripted tone numbers – e.g. “a<sup>1</sup>” may appear instead of “ā” or “a<sup>3</sup>” instead of “ǎ”, etc.

In the original specification for pīnyīn, an apostrophe was used to separate “words”, but this is no longer common except to avoid confusion.

Pīnyīn was a refinement of the early twentieth century Zhuyin/Bopomofo system (still used in some areas), but refined for modern/simplified Chinese. In February 1956, the State Council of the People’s Republic of China issued the “Directives for the Promotion of Putonghua (普通话 Mandarin Chinese)”, followed in 1958 by National Standard GB/T 16159–2012 - “Basic Rules of Chinese Phonetic Orthography” (汉语拼音正词法基本规则), later adopted internationally as ISO 7098:2015.

Pīnyīn has become the most popular method for typing Chinese in the 21<sup>st</sup> century, and variants are available for all common operating systems and devices. Remember, though, that pīnyīn is *not* the “english pronunciation” of Mandarin, and treating it as such may eventually embarrass you.

An alternative stroke-based method of typing Hànzì is presented on page 16.

Tone	Indicator	Glyph	Typing Sequence	Examples
Tone 1 Flat; high level	Macron u02c0	—	[Compose], [a], [_]	ā ē ī ō ū ū
Tone 2 Rising; High Rising	Acute Accent u00b4	´	[Compose], [a], [´]	á é í ó ú ú
Tone 3 Falling-Rising; low	Caron u030c	ˇ	[Compose], [a], [v] <sup>18</sup>	ǎ ě ĭ ǒ ǔ ǔ
Tone 4 Falling; High Falling	Grave Accent u0060	`	[Compose], [a], [´]	à è ì ò ù ù
Tone 5 Neutral; Flat	None		[a]	a e i o u ü

*Diacritics used to indicate Syllable Tones in Mandarin Pīnyīn*

16 For a quick overview: <https://www.bl.uk/history-of-writing/articles/the-double-pigeon-chinese-typewriter>; slightly more detailed coverage: <https://medium.com/@magalhini/the-fascinating-story-of-the-chinese-typewriter-29d32ff09790>. For a full history, see the book “The Chinese Typewriter: A History” by Tom Mullaney or the author’s hour and a half lecture at <https://www.youtube.com/watch?v=KSEoHLnIXYk>.

17 If you aren’t familiar with your operating system’s Compose key functionality, you should be. See page 36 in DDN-2 “Exploring Complex Text Layout” for an introduction.

18 The Compose key sequence for adding a caron (u+030c) is not commonly defined in many operating systems, but must be defined by the user. See the section *Adding and Using Compose Key Sequences* on page . Note that the breve character ˘ (u+0306) is sometimes seen as a substitute for the caron to accommodate certain ill-formed fonts.

## Typing Hànzì with a Pinyin Keyboard – the requisite “Hello, World” Example

As I/T practitioners, we can make use of the traditional “Hello, World” to illustrate how Hànyǔ Pīnyīn can be used to type both simplified (汉) and traditional (漢) Chinese (字) characters, words and phrases using a western style QWERTY keyboard.

In Mandarin, “Hello, World” translates to 你好, 世界. Its pīnyīn representation is “nǐ hǎo, shì jiè.”<sup>19</sup> The intent of modern pīnyīn input methods is to “guess” which Chinese word or phrase is meant as we type. Even more conveniently, they make it unnecessary to include tone markings; in some cases, these aren’t even permitted.

In this paper we assume the use of a specific input method known as “intelligent pīnyīn.” Different pīnyīn engines, although similar, may work slightly differently.<sup>20</sup>

The first step is to activate or switch to the Chinese input method chosen for your particular installation; how this is done varies but usually involves using a “hot key” to switch back and forth between your native keyboard and a Han version.

Using pīnyīn, we might type 你好, 世界 using one of the following sequences:

- n, i, h, a, o, s, h, i, j, i, e, space: resulting in: 你好世界
- n, i, h, a, o, l, \, s, h, i, j, i, e, space: resulting in: 你好、世界
- n, i, h, a, o, ', space, \, s, h, i, j, i, e, space: resulting in: 你好、世界

So why is the digit “1” used, and why the need for a space? In order to help make the usage and capabilities of pīnyīn more clear, we’ll begin by typing the two characters/syllables of the second word “World” (世界) by themselves.

Only then will we enter the pīnyīn for “Hello” and “world” together. By then, you’ll have a good idea how to produce hànzì by typing unaccented pīnyīn.

Later, we’ll discuss how to determine the correct pīnyīn in the first place.

*The first word of the Chinese greeting, nǐ (你), is actually a form of the word “you.” The following word haǒ (好) can mean either “it is good” or “is it good?” – 你好 is therefore more or less analogous to “Are you good?”*

A. Once the pīnyīn engine is activated, press the first letter **[S]** and a selection menu appears – usually near the cursor – looking something like that shown to the right. Some engines have the option to display the menu vertically. Keep in mind that pīnyīn does not care about capitalization, which is only displayed to cater to western sensibilities. There are several things to note about these menus:



*Menu display after typing the first letter s of the Pīnyīn sequence shìjiè.*

1. The number of choices presented – here 5 – is usually configurable, but how this is accomplished may vary across different IMEs. Each choice is given a “menu” number that is used as one means for selecting a particular hànzì.
2. If there are more choices than can be displayed, the menu’s arrows or the PgUp and PgDn keys can be used to present the next selection of characters. Generally, the keyboard arrow keys do **not** serve this purpose, since those continue to be used to move the cursor within the text being entered.
3. As an I/T practitioner, you’ve undoubtedly recognized the similarity to the code completion capabilities of your favorite editor. The bad news is that, just as no two code editors handle code completion in exactly the same manner, no two pīnyīn IMEs work exactly the same way.

4. Don’t confuse the menu option numbers with Pīnyīn tone numbers.
5. The shì syllable we are trying to type is not one of the five most statistically likely so, not surprisingly, when **[S]** is typed the display doesn’t show the 世 character that we want. Although we could click on one of the graphic arrow keys to present a new group of choices, that will quickly become quite tedious – a better option is to continue typing to give the IME more data.
6. The hànzì choices shown here, by the way, may not match what is shown on your own display, because the order in which the choices are presented varies over time; most IME engines “learn” your particular usage patterns and will adapt to them – moving your more frequently used glyphs to the beginning.
7. The third choice is indeed an emoji – not surprising since smart phones are just as common among Chinese-speaking teenagers as western ones. Luckily most pīnyīn engines have an option to make emoji display optional.

B. We continue typing shì by pressing the next character **[H]**.

1. Again, the statistics are not in our favor, and 世 isn’t readily available.



*Menu display after typing the second letter h of the Pīnyīn sequence shìjiè.*

<sup>19</sup> Nǐ hǎo is the common greeting; dà jiā hǎo is a similar collective “hello,” while nín hǎo is a more formal greeting used, for instance, when first meeting someone. Note the tones!

<sup>20</sup> The most commonly used pīnyīn method in mainland China seems to be the Sogou method, which doesn’t seem to be readily available in the west. This will likely change before long, since its features include customizable user interfaces, the ability to download industry-specific probability lists, and other niceties. The fundamentals, however, are the same.

C. We next type the final **[1]** of the first syllable shì and 时 is promoted from 4 to 2.

1. The choices have changed, but shì (世) is still doesn't appear as an option.

1. 是	2. 时	3. 使	4. 市	5. 石	←	→
------	------	------	------	------	---	---

Menu display after typing the third letter i of the Pinyin sequence shìjiè.

D. After typing the **[J]** that begins shìjiè's second syllable, we get lucky.

1. Statistically, shìjiè (世界) is the most common word beginning with shìj.

1. 世界	2. 世纪	3. 时间	4. 🌐	5. 事件	←	→
-------	-------	-------	------	-------	---	---

Menu display after typing the fourth letter j of the Pinyin sequence shìjiè.

E1. When the desired hànzi characters are displayed as a menu option, that word can be chosen immediately in either of two ways.

1. If the desired output is in the first position, the space bar can be pressed to commit<sup>21</sup> 世界 to disk. This may seem quicker to type, but is *only* useful if what we're attempting to type is presented as the first menu option.

Since first place isn't a given, though, this is probably not the most efficient habit to acquire if rapid typing is the end goal.

1. 世界	2. 世纪	3. 时间	4. 🌐	5. 事件	←	→
-------	-------	-------	------	-------	---	---

Menu display after typing the fourth letter j of the Pinyin sequence shìjiè.

2. Alternatively, typing the menu option number key **[1]** will have the same result of committing 世界, making use of menu numbers a better habit.

But: something interesting happens if we do neither and simply continue typing the remaining two pinyin characters ("i" and "e").

E2. Once the **[I]** has been typed, the menu display, as expected, changes again.

1. The word for "world" (shìjiè 世界) that we're looking for no longer appears in the menu as one of the five most likely choices. This might not appear promising, but remain patient and trust your IME.

2. What had been the second menu entry 世纪 (shìjì "century") is now the first. If you're curious, take note of the difference in tone markings on the second i of their respective pinyin representations, shìjìè and shìjì respectively.

1. 世纪	2. 实际	3. 事迹	4. 市级	5. 十几	←	→
-------	-------	-------	-------	-------	---	---

Menu display after typing the fifth letter i of the Pinyin sequence shìjiè.

3. While possibly disconcerting at first, consider that – ignoring tone markings – the string we entered (shiji) is an exact match for shìjì. Of course it wins!  
4. At this point, the choice is either to use the backspace key and return to step E1 or forge ahead with the pinyin entry while disregarding the menu display.

F. Typing the final **[E]** returns 世界 to the first position on the menu.

1. Assuming that "world" (世界) is part of a longer phrase or sentence, the space that would normally be typed next will also serve as the "commit."

2. Note the appearance of the emoji for "world."<sup>22</sup>

1. 世界	2. 是届	3. 🌐	4. 时节	5. 视界	←	→
-------	-------	------	-------	-------	---	---

Menu display after typing the sixth letter e of the Pinyin sequence shìjiè.

是届 is written as shìjiè as well – pinyin with identical tones, i.e. a homonym.

G. Now, as promised, we'll begin again with the "normal" approach to typing the entire "Hello, World" (你好, 世界) phrase and, in the process, get a better sense of how pinyin entry responds. Enter "nǐ hǎo, shì jiè" again by typing the **[N]**.

1. Although 年 (option 1) may be the statistically most common hànzi whose pinyin begins with "n," it isn't what we want. But look more closely. The desired 你 character for nǐ happens to be the second choice in the list.

1. 年	2. 你	3. 🌐	4. 内	5. 能	←	→
------	------	------	------	------	---	---

Menu display after typing the first letter n of the Pinyin sequence nǐ hǎo, shìjiè.

2. We could simply type the numeric character **[2]** to complete the syllable and then proceed to enter the **[H]** that begins the second syllable hǎo. Although this might seem more efficient, we'll ignore the menu choices and continue typing the entire "nǐ hǎo, shì jiè" phrase as if the 你 had escaped our notice.

21 Apologies for appropriating the term "commit," another word that will be perfectly understandable to IT personnel who deal with data.

22 Most Input Method Editors have an option to prevent emoji from being presented; this is to accommodate the few remaining adults using computers.

H. The menu options change again when you type the **I** key. The number of possibilities is of course reduced (as would be expected) to elements beginning with ni. There are several things to note about this display:

1. The hànzi we want, 你, has now become the statistically most likely choice. We *could* either press a space, the **I** or **H** keys, or simply continue typing.

1. 你	2. 尼	3. 呢	4. 泥	5. 妮	←	→
------	------	------	------	------	---	---

Menu display after typing the second letter i of the Pīnyīn sequence nǐ hǎo, shìjiè.

2. The selected 你 would replace the 年 on the display and its three byte UTF-8 representation e4bda0 would be “committed” to disk. **Don’t do this though!**

I. Rather than worrying about intermediate commits, it is more efficient to simply continuing to type the entire phrase and let the computer use its statistics to figure out what we want, so we type the initial **H** of the second word hǎo.

1. The nǐ hǎo sequence (你好) happens to be menu option 3, but we’re not going to be looking carefully until we’ve finished with the entire phrase.

1. 霓虹	2. 拟合	3. 你好	4. 泥孩	5. 你还	←	→
-------	-------	-------	-------	-------	---	---

Menu display after typing the third letter h of the Pīnyīn sequence nǐ hǎo, shìjiè.

2. Notice that only options 3 and 5 have the initial 你 hànzi character we’re looking for. Although option 3 is actually what we’re trying to type, ignore it.

J. We continue by immediately typing the fourth pīnyīn letter **A**.

1. The word we want, 你好, listed third in the previous step, no longer appears as a choice. Nor does the previous first choice 霓虹. Furthermore, the only choice beginning with 你 is not one of the ones shown in step I above.

1. 尼哈	2. 霓虹	3. 你虾	4. 呢哈	5. 拟哈	←	→
-------	-------	-------	-------	-------	---	---

Menu display after typing the fourth letter a of the Pīnyīn sequence nǐ hǎo, shìjiè.

As mentioned earlier, your choices may differ – e.g. you may see 泥哈 or 妮哈.

K. Once we’ve typed the **O**, the word we want, 你好, is now the first menu option.

1. If we wish to type **V** to enter the hànzi comma+space (、) punctuation, we cannot do that in many IME engines without first committing the 你好 or the comma will simply replace the word. But it isn’t necessary to do that.

1. 你好	2. 你号	3. 妮号	4. 呢好	5. 泥好	←	→
-------	-------	-------	-------	-------	---	---

After typing the fifth letter o of the Pīnyīn sequence nǐ hǎo.

2. Continue typing the entire phrase, ignoring the comma as well as the space.

L. Begin typing the second pair of hànzi - shìjiè (世界), meaning “world.”

1. Not surprisingly, the display when the initial **S** is typed doesn’t add the 世 that we want. Just as in step A, though, the 三 is added to 你好 as the most likely. But other choices from step B don’t appear at all.
2. The 🌐 “world” emoji, while a possible interpretation of the standalone word we typed in step A, is far less likely candidate to follow “Hello.”
3. What’s more interesting is that, in options 4 and 5, the IME has changed its mind as to what we might have intended for the first two characters 你好.<sup>23</sup>

1. 你好三	2. 你好所	3. 你好似	4. 尼好三	5. 呢好三	←	→
--------	--------	--------	--------	--------	---	---

After typing the seventh letter s of the Pīnyīn sequence nǐ hǎo, shìjiè.

4. Consider this new selection in terms of pīnyīn spelling rather than meaning:
5. We’ve so far typed “nǐ hǎo s”; the pīnyīn representations for the 尼好三 and 呢好三 in options 4 and 5 are “ní hǎo sān” and “ne hǎo sān” respectively.
6. Without the tone markings, and even with no knowledge of Mandarin at all, the appearance of these choices should make more sense.

M. After typing the **H**, only menu option 3 remains the same as in step L, options 2 and 3 from the corresponding hànzi in step B have disappeared, though step B’s option 4 has been deemed the second most statistically likely to follow 你好.

In this and subsequent steps, the pīnyīn for each option is shown to the right.

1. 你好是	2. 你好时	3. 你好似	4. 尼好是	5. ...	←	→
--------	--------	--------	--------	--------	---	---

After typing the eighth letter h of the Pīnyīn sequence nǐ hǎo, shìjiè.

nǐ hǎo shì    nǐ hǎo shí    nǐ hǎo sì    ní hǎo shì    Alternate possibilities

23 At first glance, this might be surprising, but it shouldn’t be. Regardless of the language used, the equivalents of “hello” and “world” are common enough but, unless you are analyzing a student programmer’s homework, the odds of encountering “world” immediately following “hello” are likely not that good.

N. Typing the third character **[I]** of shì presents the same first two options as step C, but in reverse order and, again, 你好似 and 尼好是 remain the third and fourth most likely guesses as to what we are trying to type based on statistics gathered to date on our machine.

1. 你好时 2. 你好是 3. 你好似 4. 尼好是 5. ... ← →

*After typing the ninth letter i of the Pinyin sequence nǐ hǎo, shì jiè.*

nǐ hǎo shí    nǐ hǎo shì    nǐ hǎo sì    nǐ hǎo shì    Alternate possibilities

O. Once the tenth character **[J]** – beginning the final syllable – has been typed, the first option is extended, but 世界/world still hasn't risen to a likely interpretation.

1. The third option from step D, 时间, appears second here, though not as a word to follow 你好, but 你号.
2. Option 4, 尼好时机, by the way, can be translated as “good timing.”

1. 你好时就 2. 你好时间 3. 你号时间 4. 尼好时机 5. ... ← →

*After typing the tenth letter j of the Pinyin sequence nǐ hǎo, shì jiè.*

nǐ hǎo shí jiù    nǐ hǎo shí jiān    nǐ hào shí jiān    nǐ hǎo shí jī    Alternate possibilities

P. Even typing the second-last character **[I]** doesn't suggest to the engine that we're attempting to type “world” as the object of our 你好 greeting. Observe that:

1. Only three options beginning with Hello are presented.
2. The second option from the previous step has been demoted to third here.

1. 你好世纪 2. 你好时即 3. 你好时间 4. 你号世纪 5. ... ← →

*After typing the eleventh letter i of the Pinyin sequence nǐ hǎo, shì jiè.*

nǐ hǎo shì jì    nǐ hǎo shí jí    nǐ hǎo shí jiān    nǐ hào shì jì    Alternate possibilities

Q. Finally, once the twelfth and, for our purposes, final character **[E]** has been typed, the IME engine decides that, however unlikely, we really are intent on greeting the world, and 你好世界 is displayed as the first option.

1. Still, “Hello, time” (你好时届) might also be possible so it appears second.
2. And in case it believes we're not finished and intend to complete the thought: “You look like \_\_” (你好似届\_\_), that option is fourth on the list.

1. 你好世界 2. 你好时届 3. 你号世界 4. 你好似届 5. ... ← →

*After typing the twelfth letter e of the Pinyin sequence nǐ hǎo, shì jiè.*

3. Now we can press the space bar or the **[1]** key to complete typing 你好世界.  
Summary: Give the pinyin engine enough data and it will very soon be able to provide the correct hànzi conversion. Compare this to a decryption process – it's far easier to decode a complete sentence than an individual word!

### Determining the correct Pinyin – Introducing Ruby Fonts

In order to solve issues with Chinese data, it will sometimes be necessary to follow it from initial entry to its representation on disk, or from that UTF-8 representation to its printed form.

Suppose we have some Chinese text 人人生而自由 (right) and wish to determine the pinyin required to type it. We can either:

- Copy the hànzi and paste it into the Google translate web page at <https://www.google.com/search?q=translate+chinese+to+english>
- OR, more easily if there is a lengthy example to analyze:
- Locate, download and install a suitable ruby-text font.<sup>24</sup>
- Copy the hànzi into a word processor using the ruby font.

rén rén shēng ér zì yóu zài zūn yán hé quán lì shàng yī lǜ píng děng tā men fù  
人人生而自由, 在尊严和权利上一律平等。他们赋

yǒu lǐ xìng hé liáng xīn bìng yīng yǐ xiōng dì guān xì dì jīng shén hù xiāng duì dài  
有理性 and 良心, 并应以兄弟关系的精神互相对待。

Mandarin Chinese (Pǔtōnghuà 普通话) translation of the UDHR Opening  
Compare this ruby font rendering to the identical text shown on page 3.

Ruby (or rubi) text consists of pronunciation aids appearing adjacent to non-alphabetic writing; see the web site [https://en.wikipedia.org/wiki/Ruby\\_character](https://en.wikipedia.org/wiki/Ruby_character) for more information. This example uses pinyin ruby since it is Mandarin; ruby text for other Sinitic languages would of course be different.

24 This has nothing to do with the Ruby programming language or Pokémon and, in fact, predates both by decades. We've had luck with an open source pinyin font named for its acronym FZKTPY (fāng zhèng kǎi tí pīn yīn zì kù-1), from <https://www.fonts.net.cn/font-32489380389.html> (the Běndì Xiàzài “local download” button is [本地下载]).

## Typing Hànzì with Alternate (non-phonetic) Methods

Due to the dramatic rise over the past decades in global availability of personal technology – e.g. computers, smart phones, and other handheld devices – Pīnyīn is now the most commonly used method of “writing” Chinese languages. For I/T practitioners, therefore, familiarity with the phonetic Pīnyīn method for entering Chinese data to perform testing is sufficient, and this section can be ignored.

If, however, your department has any responsibility for supporting less common input methods<sup>25</sup>, this section will provide a useful overview to help understand and address any issues that may arise with such methods.

Various non-phonetic approaches for entering hànzì are still in use. These include:

- ❑ Handwriting detection and recognition on touchscreens or track pads.
- ❑ Keyboard definitions that add a wide variety of similar hánzì to each key on a standard keyboard, e.g. Wubi-98<sup>26</sup>; variants of this may occur on both physical and virtual (on-screen) keyboards. Neither method is covered in this paper.
- ❑ Keyboard (including older cell phone number pads) definitions that add a very small number of strokes – typically less than six – that each represent a class of strokes<sup>27</sup>; statistical analysis of dictionary entries and a user’s past typing habits permits this Wǔbìhuà method to “guess” the desired word or character. This type of method is still in use and has both benefits as well as drawbacks:

**Benefits:** There are still benefits to using Stroke based data entry, including:

- ❑ If the typist already knows how to write hànzì by hand, stroke entry requires no additional training in pīnyīn, and may often result in faster typing speeds.
- ❑ The typist does not need to know the pronunciation which, as mentioned earlier, varies across languages. The “spelling” is entirely based on appearance.
- ❑ Therefore: stroke-based entry can be used for languages other than Mandarin.
- ❑ For the many hànzì with more than five strokes, some methods permit entering only the 1<sup>st</sup> to 4<sup>th</sup> and final stroke to narrow the choices presented for selection.
- ❑ Modern pīnyīn input method editors often support stroke entry as well, so there is no need to install any additional software to experiment with this method.
- ❑ Some stroke entry methods permit the use of a sixth wild card to aid the search. On some Apple products, for example, the letter O and 6 are wild cards.

**Drawbacks:** Of course, there are also drawbacks to stroke based input methods:

- ❑ The typist must know (or be able to guess) the correct stroke order. Without the correct order, this would be akin to looking up a seriously misspelled English word in a dictionary.
- ❑ Hànzì may have more than twenty strokes, complicating typing with strokes.

### TYPING WITH STROKES – INTRODUCTORY EXAMPLE

Before going into detail on using stroke based entry, the short example on the right will illustrate how it works using a few hànzì (女, 好, and 妈) introduced in “Radical Types and Single Character Combinations” on page 9 as examples B, D, and E. This will also confirm that your particular Input Method Editor supports stroke-based entry.

After activating the pīnyīn IME, begin by typing “u” to enter stroke entry mode. The “woman” character 女 can be written by typing the z, p and h keys,<sup>28</sup> which represent the <, /, and 一 strokes. As with normal pīnyīn entry, 女 may not be the first menu choice shown.

The other two examples 好 (“good”) and 妈 (“Mom”) are also used to illustrate use of “stroke classes” as surrogates for the actual strokes.

After the Stroke Reference Chart that follows, we’ll revisit typing “Hello, World” using this method, explaining stroke classes and how to choose the h, n, p, s and z “class keys” shown in the table on the right.

	女: woman, female			女 and 子 combine to form 好: good						马 and. 妈 combine to form 妈: Mom					
Stroke Number	1	2	3	1	2	3	4	5	6	1	2	3	4	5	6
Actual Stroke	<	/	一	<	/	一	冫	丩	一	<	/	一	冫	丩	一
U-Class Stroke	乙	丶	一	乙	丶	一	乙	丨	一	乙	丶	一	乙	乙	一
U-Class “Key”	z	p	h	z	p	h	z	s	h	z	p	h	z	z	h
Cumulative	乙	乃	女	乙	乃	女	奶	奶	好	乙	乃	女	奶	妈	妈
Result	女 nǚ			好 hǎo						妈 mā					
	Example B, page 9			Example D, page 9						Example E, page 9					

Using stroke entry to produce hànzì, 好 (example D) would be entered by typing u, z, p, h, z, s,h, and Enter in sequence.

25 Or if you simply want to become more acquainted with reading and writing any of the many Chinese dialects used around the globe.

26 See [https://en.wikipedia.org/wiki/Wubi\\_method](https://en.wikipedia.org/wiki/Wubi_method) or the wikipedia article at [https://en.wikipedia.org/wiki/Wubi\\_method](https://en.wikipedia.org/wiki/Wubi_method).

27 The comprehensive “Strokes (笔画) Reference Chart” on page 17 introduces the stroke primitives, and groups the strokes by class.

28 There are only five such keys (the other two are s and n), but we’ll hold off on explaining how to determine which keys to use until page 17.



## Strokes (笔画) Reference Chart

The thirty-six strokes introduced in Strokes (笔画) on page 6 are shown on the right – this time grouped into the five general classes<sup>29</sup> used for typing Hànzì using strokes. “Acro” in the legend is short for “acronym,” the Unicode CJK abbreviation for the pīnyīn name.

Each stroke used to build a Chinese character is either itself a primary stroke, or made up of a primary stroke with one of the five stroke primitives appended.

### STROKE PRIMITIVES

- 扁 biǎn A short straight vertical or horizontal line section, e.g. 5.3
- 钩 gōu A slight left or right hook at the end, e.g. 1.2
- 弯 wān A single clockwise line with a soft bend and a left hook at the end, e.g. 5.1
- 斜 xié An oblique counter-clockwise curved line that begins with a straight section, e.g. 5.2
- 折 zhè A multi-segment line with one or more right-angle folds or turns going down or to the right, e.g. 5.4 – 5.15

To type any of the eleven strokes from Classes 1 through 4 (the first column) on a standard keyboard, enter either the abbreviation for that stroke’s class primitive (h, s, p, or n); to enter any of the twenty-four strokes in class 5, use the z key. For numeric keypads, the stroke is typed using the number key.

Take care to distinguish the tí stroke 1.1, used at the end of strokes 2.1 (┘) and 5.10 (ㄣ), from the gōu primitive (the hook) used to complete thirteen other strokes (e.g. 1.2).

The strokes ㄣ, ㄥ, and ㄨ (5.1-5.3) are the only ones that begin with a primitive rather than one of the first four strokes.

1	一	31d0	横
		H	héng ⇒
1.1	㇇	31c0	提
		T	tí ↗
1.2 <sup>30</sup>	㇇	31d6	横钩
		HG	héng gōu

2	丨	31d1	竖
		S	shù ↓
2.1	┘	31d9	竖提
		ST	shù tí
2.2	㇇	31dA	竖钩
		SG	shù gōu

3	㇇	31d2	撇
		P	piě ↙
3.1	㇇	31d3	竖撇
		SP	shù piě

4	㇇	31cF	捺
		N <sup>31</sup>	nà ↘
4.1	丶	31d4	点
		D	diǎn ↘
4.2	㇇	31dD	提捺
		TN	tí nà

5	乙	31e0	横斜弯钩
		HXWG	héng xié wān gōu
5.1	㇇	31c1	弯钩
		WG	wān gōu
5.2	㇇	31c2	斜钩
		XG	xié gōu
5.3	㇇	31c3	扁斜钩
		HXG	biǎn xié gōu
5.4	㇇	31c4	竖弯
		SW	shù wān
5.5	㇇	31c5	横折折
		HZZ	héng zhé zhé
5.6	㇇	31c6	横折钩
		HZG	héng zhé gōu
5.7	㇇	31c7	横撇
		HP	héng piě
5.8	㇇	31c8	横折弯钩
		HZWG	héng zhé wān gōu
5.9	㇇	31c9	竖折弯钩
		SZWG	shù zhé wān gōu
5.10	㇇	31cA	横折提
		HZT	héng zhé tí
5.11	㇇	31cB	横折折撇
		HZZP	héng zhé zhé piě

LEGEND	Hex	Stroke Name (Hánzì)
	Acro.	Stroke Name (Pīnyīn)

5.12	㇇	31cC	横撇弯钩
		HPWG	héng piě wān gōu
5.13	㇇	31cD	横折弯
		HZW	héng zhé wān
5.14	㇇	31cE	横折折折
		HZZZ	héng zhé zhé zhé
5.15	㇇	31d5	横折
		HZ	hèng zhé
5.16	㇇	31d7	竖折
		SZ	shù zhé
5.17	㇇	31d8	竖弯左
		SWZ	shù wān zuǒ
5.18	㇇	31dB	撇点
		PD	piě diǎn
5.19	㇇	31dC	撇折
		PZ	piě zhé
5.20	㇇	31dE	竖折折
		SZZ	shù zhé zhé
5.21	㇇	31dF	竖弯钩
		SWG	shù wān gōu
5.22	㇇	31e1	横折折折钩
		HZZZG	héng zhé zhé zhé gōu
5.23	㇇	31e2	撇钩
		PG	piě gōu

N/A	○	31e3	圈
		Q	quān

29 The ○ character (quān) doesn’t fall into any of these classes so, due to its infrequent use, is ignored here. It is listed only for completeness.

30 There is nothing remotely “official” about the decimal numbers in this chart; they are simply used here to distinguish among the class members

31 Class 4 strokes are typed with an [n] (nà ㇇) when using Intelligent Pīnyīn, but with a [d] (diǎn 丶) when using Shōgǒu Pīnyīn. If using the latter, exchange 4 with 4.1.

## GROUPING STROKES BY SIMILARITY

Grouping strokes by similarity will help distinguish one from another more easily. The sixteen strokes on the immediate right are generally single lines ranging from horizontal through various angles. The “u-mode” key to enter each stroke is shown above the character along with the Unicode Hex value used to display it independently.

U-CODE FOR STROKE:

2-h	2-h	2-h	3-p	3-p	5-z	5-z	2-s	2-s	2-s	5-z	5-z	4-n/d	4-n/d	4-n/d	5-z
—	ㄟ	→	ノ	㇇	㇏	㇐	㇑	㇒	㇓	㇔	㇕	㇖	㇗	㇘	㇙
31d0	31c0	31d6	31d2	31d3	31e2	31c1	31da	31d1	31d9	31ca	31c2	31dd	31cf	31d4	31c3

U-CODE FOR STROKE:

5-z	z	z	z	z	z	z	z	z	z	z
乙	<	ㄥ	㇏	㇏	㇏	㇏	㇏	㇏	㇏	㇏
31e0	31db	31dc	31c4	31d7	31df	31d8	31c7	31c6	31d5	31c5

The 19 strokes on the right and far upper right are all multi-segment lines in class 5:

㇏	㇏	㇏	㇏	㇏
31cd	31ce	31cb	31e1	31cc
㇏	㇏	㇏		
㇏	㇏	㇏		
31c8	31de	31c9		

## Revisiting “Hello World” with the Stroke Entry Method

Recall from the section *Typing Hànzì with a Pīnyīn Keyboard – the requisite “Hello, World” Example* on page 12 that “Hello World” – 你好世界 (nǐhǎo shìjiè) consists of two words of two hànzì characters each. Using stroke entry is similar to using pīnyīn in that you will be presented by a list of possible matches which is updated with each stroke entered. Unlike pīnyīn, each hànzì must be entered independently, either by pressing the space bar or selecting the number of the matching character in the list.

Begin by typing the first six stroke class keys for 你, which, as shown in the table below, are p, s, p, z, s, p. At this point, because the desired hànzì is shown first in the character menu, typing either a space or the number 1 will “commit” 你. Alternatively, of course, typing the final n followed by a space will also work.

Stroke Number	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	21	22	23	24	25	26	27	28	29	
Actual Stroke	ノ	丨	ノ				ㄨ	SPACE	<	ㄥ	一	㇏	丨	一	SPACE	一	丨	丨	一	㇏	SPACE	丨	㇏	一	丨	一					丨
U-Class Stroke	㇇	丨	㇇	乙	丨	㇇	ㄨ		乙	㇇	一	乙	丨	一		一	丨	丨	一	乙		丨	乙	一	丨	一	㇇	ㄨ	㇇	丨	
U-Class “Key”	p	s	p	z	s	p	n <sup>33</sup>		z	p	h	z	s	h		h	s	s	h	z		s	z	h	s	h	p	n <sup>33</sup>	p	s	
DTMF Number Key	3	2	3	5	2	3	4		5	3	1	5	2	1		1	2	2	1	5		2	5	1	2	1	3	4	3	2	
Cumulative Result	㇇	㇇	㇇	㇏	㇏	你	你		乙	乃	女	妳	妳	好		一	丁	卅	卅	世		丨	冂	口	中	田	足	足	界	界	
Result	你 nǐ								好 hǎo								世 shì						界 jiè (and a final SPACE)								

The remaining three hànzì are typed in the same manner.<sup>32</sup> In this particular example, only the final 界 is detected before all its strokes have been entered; even so, it requires entry of eight of the nine strokes before that happens.

At this point, it’s helpful to compare the number of key presses required to type “Hello, World” using pīnyīn and stroke entry methods. Including the final “commit” space, the pīnyīn entry method required 13 strokes, while the stroke entry method needed 30 if typed straight through, though it only required 28 if strokes 7 and 29 were skipped.

For all but professionals and those with calligraphic backgrounds and a fluency in a Chinese language, therefore, this method is seldom used, though it is still widely supported by most IMEs to support those wishing to learn to write hànzì.

32 To permit more space, the final space after 界 is not shown, but is still required for the “commit.”

33 Recall that, in systems using Shogou Pinyin, the “d” key is used for the ㄨ stroke rather than the “n” key shown here.

## Choosing Simplified or Traditional Hànzì

In the section *Logograms vs. Characters, Words, and Syllables* on page 5, we mentioned that there were two versions of some hànzì glyphs – traditional and simplified. So far in this document we’ve assumed the use of simplified characters. With the examples given so far, this made no difference because none of the hànzì have simplified forms. If we wish to type the word “difficult” however, we would need to decide whether to type the simplified version 难 or its traditional predecessor 難. Larger versions of these hànzì are shown in the examples below for easier comparison. To type the simplified 难, we would type    .

For this design note, the Chinese *iBus Intelligent Pīnyīn* Input Method Editor was configured to use the more modern simplified hànzì by default where those are available, and uses the sequence ++ to temporarily toggle to using traditional characters when required. There is no on-screen indicator as to which is active, though this is usually evident from the output. Like most IMEs, intelligent pīnyīn is always reset to its default whenever Chinese input is activated. Typing the traditional 難, therefore, requires         .



### A DIFFICULT “SIMPLIFIED” CHARACTER IN STROKES

The single character word “difficult” – introduced above – was one of those simplified in the 1950s. This set of examples illustrates how to type both the simplified 难 and traditional 難 versions using the stroke entry method.

The table on the right provides the strokes for the simplified hànzì in their proper order. If the input window is watched carefully, the correct glyph is presented by the IME after only five keystrokes, so either       <sup>34</sup> or

Stroke Number	1	2	3	4	5	6	7	8	9	10
Actual Stroke	フ	丶	ノ	丨	丶	一	一	一	丨	一
“U code” Class Stroke	乙	丶	ノ	丨	丶	一	一	一	丨	一
“U code” Key	z	n	p	s	n	h	h	h	s	h
Cumulative	乙	𠂇	以	难	难	难	难	难	难	难
Result	难									

All five stroke “classes” are used in this single hànzì.

Some Difficult Characters	
 nán Simplified U96BE	 nán Traditional U96E3
10 strokes	19 strokes

### A MORE DIFFICULT “TRADITIONAL” CHARACTER IN STROKES

The ‘u mode’ in the intelligent pīnyīn IME requires nineteen strokes to identify the traditional form (and one more to select it); compare this to the seven that are needed when using pīnyīn.

In mainland China, where Sogou Pīnyīn and the fcitx IME are the standard, and where fluency in Chinese is far more common than anywhere else in the world, the aforementioned ability to simply type the first four and final strokes as a “shortcut” makes stroke entry more practical for those who have been “writing” since childhood.

Stroke Number	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
Actual Stroke	一	丨	丨	一	丨	冂	一	一	一	丿	丶	丿	丨	ノ	一	一	一	丨	一
“U code” Class Stroke	一	丨	丨	一	丨	乙	一	一	一	ノ	丶	ノ	丨	丶	一	一	一	丨	一
“U code” Key	h	s	s	h	s	z	h	h	h	p	n	p	s	n	h	h	h	s	h
Cumulative	和	丁	卅	卅	苜	苜	昔	昔	萱	歎	歎	歎	難	難	難	難	難	難	難
Result	難																		

Whether simplified or traditional, only one of the “wild card” strokes (乙) are used.

34 If the purpose of pressing “1” isn’t clear, or why using the space bar gives the same result, return to the more detailed explanation of how to use the IME candidate menus given in *Typing Hànzì with a Pīnyīn Keyboard – the requisite “Hello, World” Example* on page 12.

## Augmenting Compose Key Definitions to Support Typing Pinyin Diacritics

To add compose sequences in most Linux distributions for the caron (u030c – see page 11) used to indicate pinyin tone 3, copy the commands from the block on the right in sequence and run them from the command line. These will add the combinations required to an existing user’s ~/.XCompose file, or create the file if it doesn’t already exist.

If you don’t require the flexibility of typing the combinations in either order – the general convention for most compose sequences – remove alternate lines as appropriate.

These sequences will only take effect after X is restarted, either on the next log in or reboot.

```
Compose=~/.XCompose # Define location of user .XCompose definitions
echo " # Hanyu Pinyin Compose Sequences" >> $Compose
echo \
```

## Hànzì Characters Travel Abroad – Birth of Japanese Writing

Fifteen hundred years ago, Hànzì logograms were already in wide use for writing the broad family of languages on what is now the Chinese mainland when they began to be adopted by an even wider variety of other Asian languages such as Japanese, Korean, and Vietnamese. Though phonetic alphabets or syllabaries have since replaced them in most languages, hànzì are still in use as one of Japan’s three writing systems, where they are known as 漢字, written with the same glyphs, but pronounced Kanji.

Almost immediately after their Japanese adoption, two phonetic “shorthands,” now known as kanamoji (仮名文字) began to be introduced: **Katakana** (カタカナ), an angular form used by religious scribes as well as Japanese government officials, and **Hiragana** (ひらがな), a more-rounded form used by poets and educated females. By the end of the Heian period (800-1200 ce), both Kana forms – having settled into forty-six character syllabic “alphabets” – were commonly intermingled with the adopted Kanji.

In 1548, the first pure alphabetic system for writing Japanese – called Rōmaji (ローマ字, “Roman letters”) was introduced by a Japanese named Yajiro who had become a Jesuit priest; he used the Roman/Latin alphabet to assist Portugese missionaries to preach to the Japanese. Other romanization schemes followed but, until the late twentieth century, were used almost exclusively by non-Japanese, e.g. diplomats, merchant-traders or students.

Over the centuries, Hiragana became the primary Japanese writing system, with Katakana reserved for words adopted from other languages: scientific and technical terms, as well as proper names of non-Japanese people and places are generally written with Katakana. The English term “backup mirror” for instance (what we now refer to as a “rear view mirror” in a vehicle) is written in Katakana script as バックミラー (pronounced “bakkumira”), while the French croissant was added to Japanese as the Katakana word クロワッサン (pronounced “kurowassan”).

In 1946, the government’s Tōyō Kanji List introduced Shinjitai (新字体, “new character form”), simplifying some of the more complex Kanji – much as the Chinese would do a decade later, though the Japanese effort was less extensive.

The introduction of computing devices brought Rōmaji to widespread use in Japan (as Pīnyīn did in China), though with three different writing systems to accommodate, typing Japanese on a QWERTY keyboard is a bit more complex. Some useful examples are presented in DDN-8 *IME Keyboard Layout Charts for selected Languages* (see page 2).

Typing any form of Japanese with Rōmaji is similar to typing Mandarin with Pīnyīn, with a few notable exceptions. Whereas hànzì can be “committed” by pressing either the space bar or Enter key, the space bar in Rōmaji will convert whatever Kana text has not yet been committed into Kanji glyphs if that is possible. It is still necessary, however, to commit the Kanji (using Enter) before proceeding with any further entry of either Kana form.

Another difference is that, if you’re unfamiliar with Japanese writing systems, several Kana may appear to have both capital and small forms, though in positions that seem reversed. On the E key, for example, い is shown *above* い on a Hiragana layout, while the corresponding い appears *above* い on the Katakana map. Though these are sometimes called “capital” and “small” for convenience, the difference between い and い is that the shifted variant simply indicates a weaker and shorter pronunciation.

Ruby text, introduced on page 15 as an aid to reading hànzì by adding pīnyīn above the text, exists for Japanese as well, and is known as furigana (振り仮名 or ふりがな). It is not as useful for I/T personnel however, since it is aimed at providing Hiragana or Katakana representations of more complex Kanji for native Japanese rather than Latin representations for westerners.

# Japanese Kana Scripts

## Hiragana Script “Alphabet”

3040	3041	3042	3043	3044	3045
あ a	あ A	い i	い I	う u	
3046	3047	3048	3049	304a	304b
う u	え e	え E	お o	お O	か Ka
304c	304d	304e	304f	3050	3051
が Ga	き Ki	き Gi	く Ku	ぐ Gu	け Ke
3052	3053	3054	3055	3056	3057
げ Ge	こ Ko	こ Go	さ Sa	ざ Za	し Si
3058	3059	305a	305b	305c	305d
じ Ji	す Su	ず Zu	せ Se	ぜ Ze	そ So
305e	305f	3060	3061	3062	3063
ぞ Zo	た Ta	だ Da	ち Ti	ち Di	っ tu
3064	3065	3066	3067	3068	3069
っ Tu	づ Du	て Te	で De	と To	ど Do
306a	306b	306c	306d	306e	306f
な Na	に Ni	ぬ Nu	ね Ne	の No	は Ha
3070	3071	3072	3073	3074	3075
ば Ba	ぱ Pa	ひ Hi	び Bi	び Pi	ふ Hu
3076	3077	3078	3079	307a	307b
ぶ Bu	ぷ Pu	へ He	べ Be	ぺ Pe	ほ Ho
307c	307d	307e	307f	3080	3081
ぼ Bo	ぽ Po	ま Ma	み Mi	む Mu	め Me
3082	3083	3084	3085	3086	3087
も Mo	ゃ ya	や YA	ゆ yu	ゆ YU	よ yo
3088	3089	308a	308b	308c	308d
よ YO	ら Ra	り Ri	る Ru	れ Re	ろ Ro
308e	308f	3090	3091	3092	3093
わ wa	わ Wa	ゐ Wi	ゑ We	を Wo	ん N
3094	3095	3096	3097	3098	3099
づ Vu	か ka	け ke			っ yori
309a	309b	309c	309d	309e	309f
っ yori	っ	っ	っ	っ	っ

Hiragana and Katakana, collectively referred to by the generic term Kanamoji, meaning Kana Writing – are both syllabaries (syllabic “alphabets”) of 47 “characters,” each of which represents a sound.

Each “character” is shown with its Unicode hexadecimal value as well as its Rōmaji representation below that; the distinction between “capital” and “small” Rōmaji is not always observed however.

Most of these Hiragana and Katakana glyphs with identical Rōmaji represent very similar sounds and many look quite similar to one another – reflecting their common origin as shorthand versions of adopted Hànzì glyphs from the Chinese mainland.

As described above, Hiragana is primarily used to write native Japanese words, while Katakana is used to write words “borrowed” from other languages.

The opening of the United Nations’ *Universal Declaration of Human Rights*<sup>35</sup> is a typical mixture of Hiragana and Kanji:

すべての人間は、生まれながらにして自由であり、かつ、尊厳と権利とについて平等である。人間は、理性と良心を授けられてあり、互いに同胞の精神をもって行動しなければならない。<sup>36</sup>

Some comparisons of Japanese and Chinese writing – including the usual “Hello World” – are shown next for comparison (sketched below)

## Katakana Script “Alphabet”

30a0	30a1	30a2	30a3	30a4	30a5
=	ア a	ア A	イ i	イ I	ウ u
30a6	30a7	30a8	30a9	30aa	30ab
ウ u	エ e	エ E	オ o	オ O	カ Ka
30ac	30ad	30ae	30af	30b0	30b1
ガ Ga	キ Ki	キ Gi	ク Ku	グ Gu	ケ Ke
30b2	30b3	30b4	30b5	30b6	30b7
ゲ Ge	コ Ko	ゴ Go	サ Sa	ザ Za	シ Si
30b8	30b9	30ba	30bb	30bc	30bd
ジ Ji	ス Su	ズ Zu	セ Se	ゼ Ze	ソ So
30be	30bf	30c0	30c1	30c2	30c3
ゾ Zo	タ Ta	ダ Da	チ Ti	ヂ Di	ツ tu
30c4	30c5	30c6	30c7	30c8	30c9
ツ Tu	ヅ Du	テ Te	デ De	ト To	ド Do
30ca	30cb	30cc	30cd	30ce	30cf
ナ Na	ニ Ni	ヌ Nu	ネ Ne	ノ No	ハ Ha
30d0	30d1	30d2	30d3	30d4	30d5
バ Ba	パ Pa	ヒ Hi	ビ Bi	ピ Pi	フ Hu
30d6	30d7	30d8	30d9	30da	30db
ブ Bu	プ Pu	ヘ He	ベ Be	ペ Pe	ホ Ho
30dc	30dd	30de	30df	30e0	30e1
ボ Bo	ポ Po	マ Ma	ミ Mi	ム Mu	メ Me
30e2	30e3	30e4	30e5	30e6	30e7
モ Mo	ャ ya	ヤ YA	ユ yu	ユ YU	ヨ yo
30e8	30e9	30ea	30eb	30ec	30ed
ヨ YO	ラ Ra	リ Ri	ル Ru	レ Re	ロ Ro
30ee	30ef	30f0	30f1	30f2	30f3
ワ wa	ワ Wa	ヰ Wi	ヱ We	ヲ Wo	ン N
30f4	30f5	30f6	30f7	30f8	30f9
ヴ Vu	カ ka	ケ ke	ヴ Va	ヰ Vi	ヱ Ve
30fa	30fb	30fc	30fd	30fe	30ff
ヅ Vo	・	ー	、	ゞ	ㇿ

35 See a comprehensive collection of translations of this text in most of the world’s languages at: <https://omniglot.com/udhr/index.htm>

36 Your mission, should you choose to accept it, is to identify which glyphs in this Japanese text are Kanji and which are Kana.

## “Hello World” – Comparing Chinese Hànzì to Japanese Hiragana+Kanji

Here, both Chinese (你好世界) and Japanese (こんにちは世界) versions of “Hello, World” are displayed.

Note that the glyphs, Unicode values and on-disk UTF-8 representations for Hànzì and Kanji translations of “world” are identical. Also observe the similarity of the first syllable of the Hiragana spelling of “world” (せ, which isn’t used because “world” in this context is considered a “scientific” term) to the first symbol of its Hànzì antecedent (世).

See *Typing Hànzì with a Pīnyīn Keyboard – the requisite “Hello, World” Example* on page 12 and *Revisiting “Hello World” with the Stroke Entry Method* on page 18 for more details on producing this text in Mandarin with Hànzì.

	Hello		World	
Written Chinese (Simplified)	你	好	世	界
Written Pīnyīn	nǐ	hǎo	shì	jiè
Unicode Point	u4F60	u597D	u4E16	u754C
UTF-8 Hex	E4-BD-A0	E5-A5-BD	E4-B8-96	E7-95-8C

	Hello					World	
Written Japanese (Mixed script)	こ	ん	に	ち	は	世	界
Hiragana Only	こ	ん	に	ち	は	(せ)	(かい)
Written Romaji	ko	n'	ni	ti/chi	ha	se	ka - i
Unicode Point	u3053	u3093	u306B	u3061	u306F	u4E16	u754C
UTF-8 Hex	E3-81-93	E3-82-93	E3-81-AB	E3-81-A1	E3-81-AF	E4-B8-96	E7-95-8C

## “Database Management” – Comparing Chinese Hànzì to Japanese Katakana+Kanji

On the cover page of this note, both Chinese Hànzì (数据库管理) and Japanese Katakana (データベース管理) spellings/translations of our profession “Database Management” are displayed.

The Japanese word for “database” (dē-ta-bē-su) is clearly “borrowed” from the English equivalent, so is therefore written in Katakana rather than Hiragana.

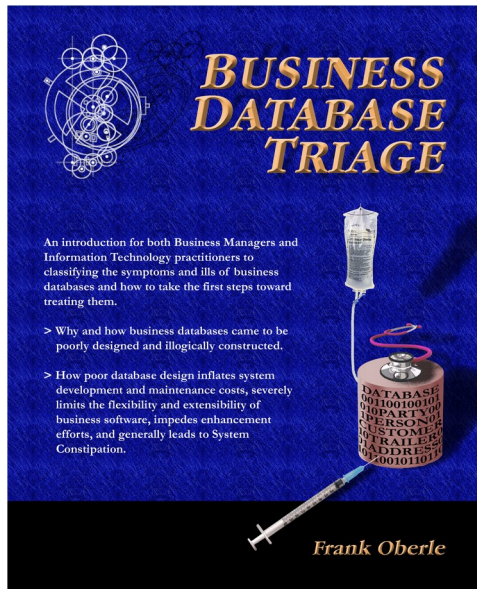
Note that the glyphs, Unicode values and on-disk UTF-8<sup>37</sup> representations for Hànzì and Kanji translations of “management” are identical.

	Database			Management	
Written Chinese (Simplified)	数	据	库	管	理
Written Pīnyīn	shù	jù	kù	guǎn	lǐ
Unicode Point	u6570	u636E	u5E93	u7BA1	u7406
UTF-8 Hex	E6-95-B0	E6-8D-AE	E5-BA-93	E7-AE-A1	E7-90-86

	Database					Management		
Written Japanese (Mixed script)	デ	ー	タ	ベ	ー	ス	管	理
Written Romaji	dē		ta	bē		su	kan	ri
Unicode Point	u30C7	u30FC	u30BF	u30D9	u30FC	u30B9	u7BA1	u7406
UTF-8 Hex	E3-83-87	E3-83-BC	E3-82-BF	E3-83-99	E3-83-BC	E3-82-B9	E7-AE-A1	E7-90-86

<sup>37</sup> The on-disk format of UTF-8 representations of Unicode values is explained with examples in the third handout in this series, *Exploring UTF-8*.

# Antikythera Publications



In addition to an ongoing series of Database Design Notes, Antikythera Publications recently released the book “*Business Database Triage*” (ISBN-10: 0615916937) that demonstrates how commonly encountered business database designs often cause significant, although largely unrecognized, difficulties with the development and maintenance of application software. Examples in the book illustrate how some typical database designs impede the ability of software developers to respond to new business opportunities – a key requirement of most businesses.

A number of examples of solutions to curing business system constipation are presented. Urban legends, such as the so-called object-relational impedance mismatch, are debunked – shown to be based mostly on illogical database (and sometimes object) designs.

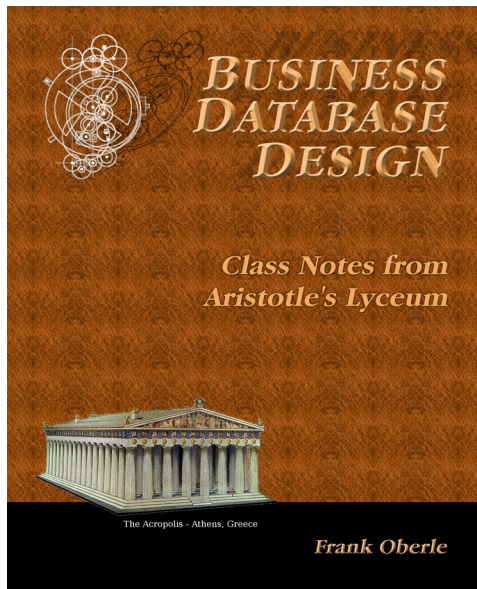
“*Business Database Triage*” is available through major book retailers in most countries, or from the following on-line vendors, each of which has a full description of the book on their site:

CreateSpace: <https://www.createspace.com/4513537>

Amazon:

[www.amazon.com/Business-Database-Triage-Frank-Oberle/dp/0615916937](http://www.amazon.com/Business-Database-Triage-Frank-Oberle/dp/0615916937)

More information and sample pages at: [www.AntikytheraPubs.com](http://www.AntikytheraPubs.com)



A follow-up book, “*Business Database Design – Class Notes from Aristotle’s Lyceum*” is due to be available in the latter part of 2014.

“*Business Database Design*” leads the reader through the logical design and analysis techniques of data organization in more detail than the earlier work – which concentrated more on understanding and identifying problems caused by illogical database design rather than their solutions.

These logical approaches to data organization, espoused by Aristotle and an “A-List” of his successors, have formed the basis for scientific discovery over more than 2,400 years, and directly led to the technology we deal with today, notably including both relational and object theory.

“*Business Database Triage*” explained the reasons why these principles were virtually impossible to apply during the early years of our transition to the use of computers in business, but since the technology is now sufficiently mature that such compromises can no longer be justified, the time has come to relearn logical data organization techniques and apply them to our businesses.