

Micro tutorial

What's the difference between Simplified & Traditional Chinese, and are they separate in Unicode?

These notes attempt to provide answers to the following questions:

- Is it correct that simplified and traditional Chinese are not completely separate sets of code entries in Unicode?
- If so, are they simply like two different fonts for the same Unicode point?
- Would I have to have a simplified and a traditional font installed?
- One traditional character may correspond to several simplified ones, right?

what is simplified chinese?

Variant forms of a given Chinese character have developed over time. For example Japanese has many simplified forms, such as 國 (country) which derives from the Chinese 國. The number of Chinese characters kept growing too. In the 1950's, Mainland China decided to reform the Chinese writing system. They simplified the shapes of many of the more common characters in use. For example, they chose the same form of 'country' as used in Japanese to replace the previous form. However, not all the simplifications adopted were simply taken from existing variants. The following shows a few examples:

Traditional	Japanese	Simplified	
廣	広	广	all different
東	東	东	simplified Chinese different
缺	欠	缺	Japanese different
國	国	国	traditional Chinese different
界	界	界	all same

The simplification process also simplified certain *components* that occur in many characters. For example the component derived from 言 in 語 becomes 讠 in the simplified form of the same character, 语.

Simplification also attempted to define a relatively smaller set of characters for common usage than had traditionally been the case. In many cases, this meant that a single character from the simplified set was used in place of several characters from the larger traditional set. For example, 干 was used in the simplified character set in place of the following four characters from the traditional set, 干 幹 乾 and 榦.

Traditional Chinese is still used to write characters in Taiwan and Hong Kong, and much of the Chinese diaspora. Simplified Chinese is used in Mainland China and Singapore. It is important to stress that people speaking many different, often mutually unintelligible, Chinese dialects would use one or other of these scripts to write Chinese – ie. the characters do not necessarily represent the sounds. There are also a few local characters, such as for Cantonese in Hong Kong, that are not in widespread use.

han unification in unicode

Next we turn to how these characters are encoded in Unicode, and we have to start with a

short word about 'unification' in Unicode.

Unicode provides a superset of most character sets in use around the world, but tries not to duplicate characters unnecessarily. For example, there are several ISO character sets in the 8859 series that all duplicate the ASCII characters. Unicode doesn't have as many codes for the letter 'a' as there are character sets - that would be ridiculous. The same principle applies for Han (Chinese) characters. The initial set of sources for Han encoding in Unicode laid end to end comprised 121,000 characters, but there were many repeats, and the final Unicode tally for all these after elimination of duplicates was 20,902.

(It is said that Chinese people typically use around 3-4,000 characters for most communication, but a reasonable word processor would need to support at least 10,000. Unicode now supports over 70,000 Han characters.)

If Han characters had different meanings or etymologies, they were not unified in Unicode. Han characters, however, are highly pictorial in nature. So the (dis-)unification process had to take into account the visual forms to some extent. Where there was a significant visual difference between Han characters that represented the same thing they were allotted to separate Unicode codepoints. (This was a pretty sophisticated process, in fact, carried out over a long period by many East-Asian experts.)

Factors such as the following prevented unification:

Factor	Examples	
Different number of components	崖	厓
Same components but different position	峰	峯
Different components	祕	秘

(Note that the last example also disunifies traditional-simplified pairs where the radical is simplified, such as 語 and 语.)

What is left for unification are characters representing the same thing but exhibiting no visual differences, or relatively minor differences such as different sequence for writing strokes, differences in stroke overshoot and protrusion, differences in contact and bend of strokes, differences in accent and termination of strokes, etc.

The codes that remained after unification were all lumped together and sorted by 'radical'. A radical is one of 214 named character components. Nearly all Han characters include one of these radicals. (This is a very simplified view, but I don't think it's necessary to bore you with all the gory details. Also as more characters were added to the initial 20,000, new characters were stored in different areas of the code space. But lets keep this simple for now.)



simplified vs. traditional character sets

So now, coming back to the question about simplified vs. traditional character sets...

The Chinese national GB standard defines a basic set of (around 6,000) characters for use with Simplified Chinese writing that does not include many of the characters in the Taiwanese industry standard for Traditional Chinese called Big 5 (around 13,000 characters in the basic set). Unicode is however a superset of both with all duplication removed down to the level of detail described above.

So the characters for 'country' in Simplified and Traditional Chinese, 国 and 國 respectively, are stored as separate codes and you cannot simply switch between the two by using a different font. On the other hand, the character for 'the world' in both Simplified and Traditional writing looks the same, 界, and both writing systems do share the same code point. Then there are characters such as 雪 ('snow') which share the code point because they are not significantly different in appearance, but may typically exhibit systematic differences in stroke overshoot and rotation of minor strokes between simplified and traditional writing systems. To

see these correctly you need to apply the right font, eg. a Song font for simplified and a Ming font for traditional.

If you have the fonts I use, you will see the difference here in the second horizontal stroke from the bottom (alternatively, look at the [pdf version](#)): simplified  (uses simsun font), traditional  (uses mingliu).

converting between the two

There are applications that attempt to convert between simplified and traditional Chinese. (It's generally easier from traditional to simplified, since you are more likely to be mapping from many to one character than the other way around.)

Over the years, translation between simplified and traditional Chinese has been complicated by the divergences in language usage between the two communities. Grammatical differences are not generally considered to be major, but there are terminological differences such as that for the word 'computer', which is different in the PRC and Taiwan.

Last modified 26 April 2003 by Richard Ishida.