Investigation Into Using the Unicode Standard for Primitives of Unified Han Characters

Henry Larkin
Deakin University
Melbourne, Australia
henry.larkin@deakin.edu.au

Abstract

The Unicode standard identifies and provides representation of the vast majority of known characters used in today's writing systems. Many of these characters belong to the unified Han series, which encapsulates characters from writing systems used in languages such as Chinese, Japanese and Korean languages. These pictographic characters are often made up of smaller primitives, either other characters or more simplified pictography. This paper presents research findings of how the Unicode standard currently represents the primitives used in 4134 of the most common Han characters.

1 Introduction

The Unicode standard has made great strides in its ability to provide a single reference for indexing written characters in the world's languages. Several of these languages contain characters that are built up of other characters. This is especially true of the unified Han subset of the Unicode standard, which focuses on characters largely used within Japanese kanji, Chinese hanzi, and Korean hanja. These character sets are used in several languages in numerous regions in Asia. While the Unicode standard has been working towards creating a unified character set, from a research perspective there is an area of research open to explore what parts of characters might contain sub-characters (primitives), and how these primitives are represented. These primitives can be either whole characters in and of themselves, or primitive glyphs either in the form of simplified representations of actual characters, or common symbols which, by themselves, traditionally have only a vague or perhaps non-existent meaning. This is especially important to dictionary, research and language-learning projects, where the breakdown of primitives is greatly beneficial.

Some work has been done in this area before, particularly from the point of view of language learners. The work of Dr. Heisig [1][2] has made great strides in identifying common primitives within Chinese and Japanese characters. However, majority of these primitives are drawn as images and have no representation in the Unicode standard or are not referenced from the Unicode standard. Furthermore, previous research has not explored a comprehensive analysis of which primitives are used most commonly and in what positions of the character they are most commonly found. The purpose of this work is to explore the possibility of using the Unicode standard for all primitive characters.

2 Process

This research project looked at six Asian language character sets in order to investigate whether it is possible to use Unicode characters to describe the primitives that make up each character. Six language sets were considered in total.

28th Pacific Asia Conference on Language, Information and Computation pages 129–134

- JOYO is the official kanji character set as described by the government of Japan containing 2136 characters units when including the latest updates from 2010.
- JLPT (Japanese Language Proficiency Test) is a character set used specifically for learners of Japanese. It differs from the JOYO character set in that characters are given roughly in order of those most commonly used as opposed to those that are simplest to write as would be given in a Japanese language school. The JLPT set contains 2431 characters. JLPT has five levels.
- HSK (Hanyu Shuiping Kaoshi or Chinese Proficiency Test) is the official hanzi character set of mainland China covering 2804 characters. HSK has six levels.
- TOCFL (Test of Chinese as a Foreign Language) is the character set used for learners of traditional hanzi for years in Taiwan. It contains 2815 characters over five levels.
- Taiwan School System. 2809 characters are taken for the Taiwan educational system up to grade 7. In the case of traditional characters, there are a significant number of rarely used characters that are taught in advanced levels of the Taiwan high school system. These characters will not be considered as part of this research due to their rarity. It is also worth noting that the majority of advanced characters almost always consist of a subset of whole other characters as their primitives.
- Hong Kong School System. This contains 2929 traditional hanzi characters. Note that only up to grade six is included in this research for the same reasons that the more complex characters are rare and almost always consists of whole characters as primitives.

Korean hanja was not included as it is mostly only used in older and scholarly texts, as hangul is the most common form of writing in modern-day South Korea, and this research is considering common-use han characters.

Many of these character sets overlap greatly which is why the Unicode standard spent considerable time finding ways to unify character identification (although it is worth noting that there is some consideration to be given that different regions may consider some of their characters to not be able to be unified due to different styling of their characters and different meanings given to them). In total, 4134 characters were investigated at as part of this research. For each of these characters, each character was visually broken down into primitives based on the available characters present in the Unicode standard. This was done by hand. The majority of these primitives consisted primarily of characters that already existed as whole characters. It also consisted of glyphs used either as official simplifications or similar shapes.

Three examples are included below to demonstrate the types of primitives. In the first instance, *bright*, both primitives are complete characters in their own right. In the second instance, *fathom*, the primitive on the left is an official primitive, in the sense that it has a meaning (water), that is derived from the complete character 水. The right primitive is a whole character in its own right. In the third instance, *occupation*, the top primitive 𠂉 is not an official primitive. Any records of it being an official primitive have been lost over time, or are abstract in detail. Regardless of its lack of official meaning, the primitive still has a visual representation within the Unicode standard that occurs within the character. This research considers all cases when searching for visual representations, within the Unicode standard, for representing the primitives of each Han character within the six common character sets analyzed.

1. 明, bright, l 日, r 月
2. 測, fathom, l 氵, r 則
3. 营, occupation, t 𠂉, b 呂

The examples below demonstrate how this breakdown was achieved. Every character, for the purposes of this research, had an English term assigned to it for help with identification, although, this English term is not necessarily official, as different languages treat characters differently. It is worth noting, however, that in the vast majority of cases, the English term used to describe the

character was somewhat similar in meaning across most data sets.

For each entry, the primitives were then defined and described relative to their position. Character positions were broken up into four main directions: *top* (t), *bottom* (b), *left* (l), *right* (r), to describe where primitives belong visually within a parent character.

名, name, t 夕, b 口
明, bright, l 日, r 月
動, move, l 重, r 力
新, new, l 亲, r 斤
製, manufacture, t 制, b 衣
災, disaster, t 巛, b 火
仙, hermit, l 亻, r 山

Two special positions were also included. These are *outer* (o) and *inner* (i). *Outer* is used to describe where a primitive occurs outside the quadrant of others. *Inner* is used to describe where a primitive occurs inside an *outer* position. An example of *outer* and *inner* positioning is given for the character *wide* seen below. In this example, there are two primitives. One that belongs in the *outer* container and one that belongs technically *inside* the container.

広, wide, o 广, i 厶

Further to this, for complex characters, it is possible that there will be more than six positions of primitives. In many cases, there are multiple primitives within a position. To support this, indentation of splitting each grid position into sub-positions using subsequent letters was defined. For example, in the case of the character used for *brain* below, there is one character positioned on the left, and then on the right, there is another pseudo character consisting of three smaller primitives. This right hand side is then divided into top and bottom by simply indicating that there is a primitive on the right and in the top quadrant of the right side and two other primitives on the right hand side in the bottom component. Furthermore, in the right bottom components, this is split further into outer and inner sections.

腦, brain, l 月, rt 巛, rbo 口, rbi乂

Also note that primitives were split like this where a more complete primitive character did not exist within the Unicode standard. The primary aim was to determine if all characters could be represented by primitives in some form.

Where possible, all primitives used the most complex form possible. It is possible to represent a character, no matter how complex, using the most simple primitives, or some combination of simple and more complex and complete primitives. However, in this research, it was decided that the most detailed primitive would be used where possible. Take for example the character for wide above and the character for broaden below. *Broaden* makes use of two primitives. In this case, the right hand portion is the existing character *wide* and not the sub-components that *wide* consists of.

拡, broaden, l 扌, r 広

Furthermore, this research is focused on visual shapes entirely. So, where a character has a simplified form because of the way it is simplified visually inside another character, the simplified form is used. Table 1 below shows a sample of some of the most common characters and the simplifications.

food	食	飠	饣
water	水	氵	
going	行	彳	
gold	金	钅	
cow	牛	牜	
stream	川	巛	巜

Table 1: Example List of Official Character Simplifications

There were some instances where "official" primitives did not exist. In which case, liberties were made in selecting similar Unicode characters. A selection of which will be covered in Section 4 on Primitives with no Unicode Character. For the purposes of this research, all Unicode characters were considered as possibilities for primitives, though, in the majority of cases, the so called "official" primitives were used.

3 The Common Primitives

After all characters in the included character sets had their primitives identified and recorded, statistics were then calculated to determine information about how the primitives were being used. One of these was the common primitives in each character set. Table 2 below shows the breakdown of the common primitives and their frequency for each of the language sets investigated. Across all lists, the most common primitives are roughly the same in all cases. It is only when one gets further down the list that one starts to see new primitives that do not appears in other lists.

HK		HSK		JLPT		JOYO		TAIWAN		TOCFL	
口	249	口	219	口	147	口	152	口	238	口	242
氵	147	扌	151	木	131	氵	121	氵	137	氵	143
木	134	氵	142	氵	125	木	115	木	134	扌	140
扌	132	木	122	一	107	一	110	一	129	木	125
一	127	一	115	亻	102	亻	94	扌	127	一	120
亻	123	亻	105	土	90	土	90	亻	118	亻	116
土	104	土	87	日	86	扌	87	土	97	土	95
艹	84	日	85	扌	81	日	83	艹	91	日	83
日	83	月	78	艹	80	月	72	日	84	月	76
月	83	讠	72	言	67	言	71	月	75	言	75
言	79	艹	66	月	64	艹	69	言	74	艹	74
心	61	辶	59	辶	57	辶	54	辶	60	辶	61
辶	60	纟	57	心	54	心	53	心	58	心	59
十	58	十	52	十	49	宀	49	宀	57	宀	56
女	57	女	52	宀	49	十	46	十	56	十	54
宀	56	宀	51	阝	42	女	44	女	54	貝	53
糸	55	心	49	儿	41	田	42	糸	53	女	47
忄	53	丶	47	丷	40	阝	42	王	47	忄	47
貝	53	忄	47	女	40	丷	41	田	46	隹	46
隹	48	贝	46	糸	40	貝	40	冂	45	糸	45

Table 2: Top 20 Primitives per Character Set

Also interesting was the rapidly reducing frequency of primitive use. Figure 1 shows that the most common primitives appear far more commonly than any other character. The chart clearly shows a long tail style of frequency, where in the case of the HSK character set, only 45 primitives have an occurrence of more than 20 times with the top six primitives occurring 100 or more times. The frequency of primitive use drops off quite quickly, indicating that characters in each of these languages do have a common set of primitives. All language character sets had a very similar occurrence.

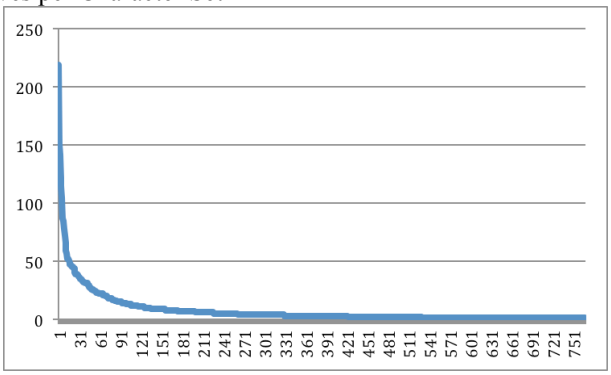

Figure 1: Frequency of Primitives in the HSK Set

It is also worth noting which position was more common in primitives. A sample of this data can be seen in Table 3 below. Each character is preceeded by a letter code to indicate its position within another character. The positions are: (l)eft, (r)ight, (t)op, (b)ottom, (i)nner, (o)uter. Across all

language sets, the most common position for any primitive is the *right* side, having vastly more occurrences than its nearest competitor, the *left* side. Following this, the *top* position is the most common and the *bottom* is least common across all languages, for the four main quadrants. The *outer* and *inner* positions were quite rare. What is extremely interesting about this data is that all languages had almost identical primitive positioning. This further supports the theory that there is a very common nature among Chinese style characters in Asian languages.

What is interesting to note about primitive positions is that while the *right* position was the most common for all primitives, the most popular primitives vastly favored the *left* and sometimes the *top*. This is due to the fact that the *right* position usually contained whole characters, which were not commonly used as primitives, but the *right* position was the most common positioning.

HK		HSK		JLPT		JOYO		TAIWAN		TOCFL	
l 氵	136	l 扌	150	l 氵	117	l 氵	112	l 氵	127	l 扌	139
l 扌	131	l 氵	134	l 亻	94	l 亻	88	l 扌	125	l 氵	131
l 亻	113	l 亻	100	l 扌	80	l 扌	86	l 亻	108	l 亻	106
l 口	76	l 口	89	l 木	75	l 言	64	l 口	75	l 口	77
l 木	71	l 讠	68	t 艹	63	l 木	59	l 木	73	l 言	64
l 言	66	l 木	67	l 言	60	l 辶	52	t 艹	68	l 木	63
t 艹	63	l 辶	58	l 辶	54	t 艹	48	l 言	61	l 辶	59
l 辶	58	t 艹	55	b 心	38	l 月	40	l 辶	58	t 艹	50
l 饣	53	l 纟	53	t 宀	37	b 心	38	l 月	42	l 饣	47
b 心	43	l 月	47	l 月	33	t 宀	35	t 宀	40	b 心	42
l 月	42	l 饣	46	l 糸	33	l 饣	32	b 心	39	l 月	41
l 糸	41	b 心	45	t 一	32	l 糸	32	l 糸	39	t 宀	41
l 女	37	t 宀	42	l 饣	31	t 一	32	l 女	38	l 金	37
t 宀	35	l 钅	34	l 纟	29	l 阝	30	l 饣	37	l 糸	35
t⺮	34	t 一	34	l 阝	29	l 口	29	l 金	35	o 广	31
l 金	33	l 阝	32	l 金	27	l 土	29	t⺮	31	l 女	30
l 土	31	l 土	31	l 女	26	l 女	28	l 阝	29	l 阝	30
r 刂	31	l 女	31	l 禾	26	l 金	28	t 一	29	t⺮	29
o 广	30	r 刂	28	l 土	25	r 刂	25	o 广	28	t 一	29
t 一	29	b 口	27	o 广	25	r 頁	24	l 土	27	r 刂	28

Table 3: Top 20 Primitives in Specific Positions

4 Primitives with no Unicode Character

Seven primitives were identified which had no Unicode representation that accurately took the shape. These are shown in Table 4 below, using the closest-matching character. All but two of these characters were taken from the Japanese hiragana and katakana alphabets. The primitives ⩛ and ‡ are Unicode symbols. They are not an accurate visual representation, but are the closest matching symbols found for those two commonly-used primitives.

Primitive	Examples
⩛	光 , 单 , 肖
フ	场 , 汤
メ	风 , 図, 肴 , 哎
マ	专 , 矛
ス	经 , 轻
ム	么 , 云 , 勾
‡	牛, 泽

Table 4: Missing Primitives

It is also worth mentioning that there is a severe lacking of font support for the primitives, which

can visually display the Unicode standard. This has been an issue among typeface users and designers for many years, and it is still an issue today. Even in creating this paper, several different fonts were used for displaying some of the more unique primitives.

5 Conclusion

In conclusion, this research has collated and documented the primitive breakdown of each character using Unicode primitives. The results of this research show that the Unicode standard does greatly support the identification and codifying of primitives as used in Han characters. There are only a few exceptions where character representation is not possible. Furthermore, what is interesting to note is that the most common primitives appear far more likely than any others. Also of note is that the most common positions for primitives were on the left, and also at the top. It would be interesting to see if further iterations of the Unicode standard will support the pseudo primitive characters for which there is currently no code point.

References

[1] James W. Heisig and Timothy W. Richardson. Oct 2008. Remembering Simplified Hanzi: Book 1. How Not to Forget the Meaning and Writing of Chinese Characters.

[2] James W. Heisig. Apr 2011. Remembering the Kanji: A Complete Course on How Not to Forget the Meaning and Writing of Japanese Characters.

[3] Etsuko Toyoda, Arief Muhammad Firdaus, and Chieko Kano. Identifying Useful Phonetic Components of Kanji for Learners of Japanese.

[4] James W. Heisig. 1987. Remembering the Kanji 2. Honolulu: University of Hawai'i Press.

[5] Hiroyuki Kaiho and Nomura Yukimasa. 1983. Kanji Joho Shori no Shinrigaku (The Psychology of Kanji Information Processing). Tokyo: Kyoiku Shuppan.

[6] Kano Chieko. 1993. Kanji no zoji seibun ni kan-suru ichi-kosatsu (Study on Basic Japanese Components of Kanji) (2). Bungei Gengo Kenkyu (Studies in Language and Literature) 24: 97–114.

[7] Masuda Hisashi and Saito Hirofumi. 2002. Interactive Processing of Phonological Information in Reading Japanese kanji Character Words and Their Phonetic Radicals. Brain and Language 81: 445– 453.

[8] Toshihiro Hayashi and Yoneo Yano. 1994. Kanji Laboratory: An Environmental ICAI System for Kanji Learning. IEICE Transactions on Information and Systems.

[9] Daniel Wagner and Istvan Barakonyi. 2003. Augmented Reality Kanji Learning. Proceedings of the 2nd IEEE/ACM International Symposium on Mixed and Augmented Reality.

[10] Mathieu Blondel, Kazuhiro Seki and Kuniaki Uehara 2010. Unsupervised Learning of Stroke Tagger for Online Kanji Handwriting Recognition. Pattern Recognition.

[11] OndĺEj Velek, Cheng-Lin Liu, Stefan Jaeger and Masaki Nakagawa. 2002. An Improved Approach to Generating Realistic Kanji Character Images from On-line Characters and its Benefit to Off-line Recognition Performance. Pattern Recognition.

[12] OndĺEj Velek, Cheng-Lin Liu, and Masaki Nakagawa. 2001. Generating Realistic Kanji Character Images from On-line Patterns. Document Analysis and Recognition.

[13] Jun Tsukumo. 1996. Handprinted Kanji OCR Development--What was solved in Handprinted Kanji Character Recognition? IEICE Transactions on Information and Systems

[14] Akiko Nagano and Masaharu Shimada. 2014. Morphological Theory and Orthography: Kanji as a Representation of Lexemes. Journal of Linguistics

[15] Ikumi Ota, Ryo Yamamoto, Takuya Nishimoto and Shigeki Sagayama. 2008. On-line Handwritten Kanji String Recognition Based on Grammar Description of Character Structures. Pattern Recognition

[16] Ondrej Velek and Masaki Nakagawa. 2002. The Impact of Large Training Sets on the Recognition Rate of Off-line Japanese Kanji Character Classifiers. Document Analysis Systems V.