International Journal of Advanced Smart Convergence Vol.7 No.2 7-21 (2018)
http://dx.doi.org/10.7236/IJASC.2018.7.2.7

IJASC 18-2-2

A study on Unifying Hanja Variant Groups of Korea and China for LGR (Label Generation Rule) of Internet Top-Level Hangeul Hanja Domain

Kyongsok Kim*

School of computer science and engineering, Pusan National University, Korea
gimgs@pnu.kr

Abstract

The author studied the process of unifying Hanja variant groups of Korea and China for LGR (Label Generation Rule) of Internet Top-Level Hangeul Hanja Domain and possible confusion between Hangeul syllable and Hanja character. Among 3518 Chinese variant groups, Korea and China need not review variant groups which include no or just one Korean Hanja character. Korea and China reviewed 304 Chinese variant groups (9% of the 3518 Chinese variant groups) which include two or more Korean Hanja characters. By doing so, Korea and China succeeded in efficiently unifying variant groups. Unification process of variant groups which is the main core of Korea-China coordination and almost final unification result is summarized in this paper. In addition, the author analyzed systematically whether some Hanja character could be confused with a Hangeul syllable and obtained a good result which was not expected at the beginning. Probably this kind of systematic analysis has not been performed in the past and seems the first attempt, which is one of the contributions of this paper. The author also reviewed how to express K-LGR in XML for submission to ICANN.

Keywords: *Unification of variant groups, Hangeul and Hanja, Label Generation Rule, Top-Level Domain, Korea-China*

1. Introduction

Currently ICANN (Internet Corporation for Assigned Names and Numbers, https://www.icann.org) is in the process of making LGR (Label Generation Rule) of Internet Top-Level Hanja Domain and Korean script domain, which started in 2013 [1, 2].

The author has been working on this job as a committee chair of KGP (Korean Script Generation Panel) from the beginning. ICANN asked Korea, China (including Taiwan, Hong Kong, Macau), and Japan that have interests in Hanja to make LGR of Internet Top-Level Hanja Domain agreed by three countries and to submit it to ICANN.

In the past, RFC (Request for Comments) related with CJK (China, Japan, Korea) domains [3] and RFC related with Hanja domain [4] were published.

Manuscript Received: Apr. 10, 2018 / Revised: Apr. 20, 2018 / Accepted: May. 1, 2018

Corresponding Author: gimgs@pnu.kr
Tel: +82-51-510-2292, Fax: +82-51-515-2208
School of computer science and engineering, Pusan National University, Korea

According to ICANN's request, since 2014, CJK has met and discussed at ICANN meetings which are held three times per year and at other CJK meetings. The author has attended these meetings continuously and discussed with CJ (China, Japan) experts. Probably in 2019, Hanja LGR agreed by CJK will be published.

There are two major issues with CJK Hanja domains. First, each country needs to make Table of Hanja characters to be used in its Hanja domain. Second, each country needs to make a list of variant groups of Hanja characters. The concept of variants and variant groups will be discussed below.

There is no problem with the first issue since no coordination among CJK is needed for the three lists of Hanja characters. However, with the second issue, CJK needs to coordinate since three lists of variant groups of CJK need to be unified (or merged) as one list of variant groups.

The author wrote a paper [5] dealing with how to unify variant groups when two lists of variant groups of KC (Korea, China) are given. This is a follow-up paper. In other words, the previous paper was a preliminary result at the beginning of coordination process and this paper shows almost final coordination results between Korea and China. By the way, Japan announced that Japan would not include its own variant groups of Hanja (Kanji) and will accept the coordination results between Korea and China. Therefore, the coordination was done between Korea and China for unifying variant groups.

In addition, the author analyzed systematically whether some Hanja character could be confused with a Hangeul syllable and obtained a good result which was not expected at the beginning. Probably this kind of systematic analysis has not been performed in the past and seems the first attempt. It is one of the contributions of this paper.

The author also reviewed how to express K-LGR (Korean LGR) in XML for submission to ICANN.

In Section 2, we will see the concept of variants and why variants are problematic. The main body of this paper starts in Section 3 where we will see how to select variant groups to be reviewed by Korea and China.

In Section 4, we will see the coordination process of unifying variant groups between Korea and China.

In Section 5, somewhat unfamiliar issue as to the possible confusion between Hangeul syllable and Hanja character will be discussed. In Section 6, representing K-LGR in XML will be discussed and, finally, in Section 7, conclusions will be given.

2. The Concept of Hanja Variant and why Is it Problematic in Internet Domains?

Since one of the two major issues to be dealt in this paper is unification of KC (Korea-China) Hanja variant groups, in this Section, we will see the concept of Hanja variant and why it becomes problematic in Internet domains. It is to be noted that this Section is a brief summary of the author's previous paper [5] with some additional explanations added.

2.1 What is a Hanja variant?

In general, for a given Hanja character, there can be another character whose glyph (shape) is somewhat different but the meaning is the same. Let's see an example of two Hanja variants in Korea.

峰 (U+5CF0) ("U+" indicates that the following (usually) four or five hexadecimal digits are code positions in ISO/IEC 10646 [6]. Therefore, Hanja 峰 is represented as 5CF0.)

峯 (U+5CEF)

The reading (sound) of the above two characters are "Bong" in Korea and their meaning is "mountain peak". These two characters are considered as variants in Korea. Their three components (山, 夂, 丰) are the same; however, their relative positions are different. In the first character, 山 is at left and, in the second character, 山 is at top.

2.2 A special class of variants in mainland China: Traditional and Simplified characters

Of course, variants explained in Section 2.1 also exist in China. However, there is a special class of variants in China in addition. They are Simplified (简化字, jiǎnhuàzì) and Traditional (繁体字, fántǐzì) characters.

China announced the Whole Table of 2235 Simplified characters (簡化字 總表) in 1964 [7]. Traditional characters are similar to the Hanja usually used in Korea (正字) and Simplified characters are similar to (but not the same as) the abbreviated form (略字). Chinese Simplified characters refer to the 2235 characters in the Whole Table of Simplified characters. The Table also shows (usually one, but sometimes two or three) Traditional character(s) corresponding to each Simplified character. About 1700 characters out of 2235 characters were created in 1960s (i.e., not used in the past). These characters are almost never used in Korea. The remaining about 500 characters were used in the past (i.e., not created recently) and were designated as Simplified characters in 1960s.

As an example, let's consider 東 (east) character. The following two characters are variants in China. We call the group of these two characters as a "variant group".

[東 (U+6771), 东 (U+4E1C)]

The first character, 東 (U+6771) is a familiar Traditional character and the second character 东 (U+4E1C) is an unfamiliar Simplified character. In China, these two are considered as variant characters.

2.3 Simplitional character (简繁字)

Characters that became Simplified in 1964 such as 机 are called "Simplitional" character. The term Simplitional came from the fact that a character is currently "SIMPLI"fied character but was tradi"TIONAL" character in the past.

For example, 东 (U+4E1C) was created and included in 1960s in the Whole Table of Simplified characters and is almost never used in Korea. However, in the following variant group, Simplified character 机 (U+673A) has been used both in Korea and in China for a long time before the Table of Simplified characters was announced in 1960s.

[機 (U+6A5F), 机 (U+673A)]

Currently, both characters mean "machine" in China. The only difference is that 機 (U+6A5F) is Traditional and 机 (U+673A) is Simplified. However, in Korea, 机 means desk (reading is "gwe") and 機 means machine (reading is "gi") and they are completely different characters. In China, before 1964, 机 meant desk as in Korea; however, after 1964, 机 became to mean machine.

Therefore, meaning and usage of one and the same character 机 is totally different in Korea and China.

The concept of Simplitional character is very important in unification of Hanja variant groups containing Simplitional character; however, there was no term referring to such characters. The author coined the term Simplitional which is becoming more widely used.

2.4 Why are variants problematic in Internet domains?

Let's consider two URL's using 東海 and 东海↵ as TLD (Top-Level Domain) and see why variants are problematic in domains.

www. 東海 (a domain with Traditional char)
www. 东海↵ (a domain with Simplified char)

When we enter each of these two addresses in the address window of Web browsers, we can think of two situations.

In situation 1, 東海 and 东海↵ are directed to the same IP address (it is highly probable that owners of 東海 and 东海↵ are the same). In situation 2, 東海 and 东海↵ are directed to different IP address (it is highly probable that owners of 東海 and 东海↵ are different).

Chinse users will consider 東海 and 东海↵ as the same domain and, therefore, situation 1 seems natural to Chines users but situation 2 is confusing to them. In situation 2, fake home page could be made.

Table 1. Two situations where www.東海 and www.东海 point to the same or different IP addresses

Situation domain	situation 1	situation 2
www.東海	site 111.11.111.11 accessed	site 111.11.111.11 accessed
www.东海		site 222.22.222.22 accessed

Due to these reasons, when a domain is applied for it is desirable to treat the domain and the other domains with characters replaced by variant characters as one "bundle". In the above example, if a Chinese applicant applies for 东海, both 东海 and 東海 are treated as one bundle.

There can be a few possibilities how to treat them as one bundle. First, one applicant owns both 东海↵ and 東海 and they are directed to the same IP address. Second, an applicant owns 东海↵ but not 東海 (in China). Third, an applicant owns 東海 but not 东海↵ (in Taiwan).

As another example, consider a domain 东海国 where each of 东 and 国 has one variant 東 and 國, respectively. Then the number of possible combinations of three Hanja characters for domain becomes 4 (= 2 x 1 x 2): 東海國, 东海国, 東海国, and 东海國. The third and fourth domains, 東海国 and 东海國, contain both Traditional and Simplified characters and, therefore, it is highly probable that they will not be actually used as domain. Although an applicant may not own all four domains, it is desirable that any of these four domains is not allocated to other people.

A variant group of 8 characters in C-LGR is shown below. Here, 8 is called the size of this variant group.

[(U+55A6 喦) (U+58E7 壧) (U+5CA9 岩) (U+5D52 嵒) (U+5DCC 巌) (U+5DD6 巖) (U+789E 碞) (U+7939 礹)]

Table 2. The difference between when two characters, 机 and 機, are unified (i.e., merged) in a variant group and when they are not

domain	Two chars [机, 機] are in a variant group			Two chars 机 and 機 are independent chars
	possibility 1	possibility 2	possibility 3	
wx机yz	owner1 owns both wx机yz and wx機yz	owner1 owns wx机yz only	nobody can own wx机yz	owner1 owns wx机yz only
wx機yz		nobody can own wx機yz	owner1 owns wx機yz only	owner2 owns wx機yz only

For a domain composed of several Hanja characters, there could be thousands of possible variant domains. As we saw above, depending on how characters are unified (i.e., merged) in variant groups, the number of permissible combinations (i.e., variant labels) for a given domain is determined. Therefore, the result of unification of variant groups between Korea and China has much effect on the number of variant labels (variant domains).

As shown in Table 2, if 机 and 機 are grouped in a variant group and if owner1 owns wx机yz, other people cannot own wx機yz. However, if 机 and 機 are not grouped in a variant group and if owner1 owns wx机yz, other people can still own wx機yz.

In the example of 机 and 機, Korea wants not to group these two characters in a variant group since these two are different characters in Korea. In other words, Korea wants to treat these two as independent. In contrast, China wants to group them in a variant group since they are variants. Due to this kind of conflicting requirements, in general, it was very hard to unify variant groups between Korea and China.

2.5 List of 19738 Hanja characters and 3518 variant groups of China (C-LGR) as of 2016.07.20.

The list of Hanja characters and variant groups of China changed several times [8-10]. The author's preliminary paper [5] analyzed C-LGR published in 2015.04.30. [8]. In this paper, C-LGR as of 2016.07.20 [9] was analyzed. As shown in Table 3, the number of Hanja characters and variants groups increased quite a lot.

Table 3. Publication date, numbers of characters and variant groups of C-LGR

Publication date of C-LGR	number of Hanja chars	number of variant groups
2015.04.30.	12563	3093
2016.07.20.	19738	3518

2.6 4819 Hanja characters and 50 variant groups in K-LGR v0.5 [12]

Like C-LGR, K-LGR changed several times [11-14]. In the author's previous paper [5], K-LGR v0.3 [11] was analyzed and, in this paper, K-LGR v0.5 [12] was analyzed. As shown in Table 4, the number of variant groups increased somewhat.

Table 4. Publication date, number of characters and variant groups of K-LGR

Publication date of K-LGR	number of Hanja chars	number of variant groups
2015.08.13. (v0.3)	4819	37
2016.09.28. (v0.5)	4819	50

3. Selecting Variant Groups out of 3518 C Variant Groups to be Reviewed by K and C

The main body of this paper starts here. Most versions of C-LGR have more than 3000 variant groups. The hardest issue in unifying KC variant groups was how to reduce the time needed to review variant groups for unification.

Table 5 shows the number of C variant groups for each of the size of 3518 C variant groups and the number of K characters therein.

Table 5. Table showing the number of C variant groups for each of the size of C variant groups and the number of K characters therein

size of C variant group	# of K chars in a C variant group					
	0 char	1 char	2 chars	3 chars	sub-total	sub-total
2	1242	1392	147		2781	3518
3	95	359	82	12	548	
4	13	80	32	5	130	
5	3	22	10	2	37	
6	1	4	6	2	13	
7		3	3	1	7	
8			2		2	
sub-total	1354	1860	282	22	3518	3518
			304			

Now, consider three C variant groups shown below. "X" indicates that the character, U+4F21 伡, as an example, is not K character (i.e., not included in K-LGR). "O" indicates that the character, U+4E07 万, as an example, is K character (i.e., included in K-LGR).

[(X U+4F21 伡), (X U+4FE5 俥)]
[(X U+4E05 丅), (O U+4E0B 下)]
[(O U+4E07 万), (O U+842C 萬)]

The first example is one of the 1242 variant groups in Table 5 where the size of C variant group is 2 and the number of K characters therein is 0.

Likewise, the third example is one of the 147 variant groups in Table 5 where the size of C variant group is 2 and the number of K characters therein is 2.

It was found that KC only need to review C variant groups containing 2 or more K characters. In other words, KC need not review C variant groups containing zero or one K character since such C variant groups need not be unified. It will be explained in more detail.

Let's consider a C variant group containing no K character. Since no character in the variant group is included in K-LGR, K and C does not need to review that C variant group. That character is used only by C. Therefore, we could safely ignore 1354 variant groups in Table 5 for the purpose of unifying variant groups between K and C.

Now consider a C variant group containing only one K character. Since the other character(s) in the variant group are not included in K-LGR, the K character is an independent character in K-LGR and, therefore, K and C do not need to review that C variant group. Therefore, we could safely ignore 1860 variant groups in Table 5 for the purpose of unifying variant groups between K and C.

By following this rule or principle, KC needed to review only the remaining 304 (about 9%) out of 3518 variant groups for unification. 304 is obtained by adding 282 and 22 in Table 5. This made it possible for K and C to review C variant groups very efficiently.

As shown in Table 5, KC needed to review and unify only C variant groups containing 2 or more K characters. The number of such C variant groups is 304. Therefore, KC could save much time in reviewing and unifying C variant group.

4. The Process of Unification of Chinese Variant Groups

Korea and China met, discussed, and prepared data for unification in years 2014 to 2016 for two and half years. Then almost final unification process was intensively carried out from Sep. 2016 to Feb. 2017.

4.1 Before unification process began (2016.09.28., before Taipei meeting)

Out of 304 C variant groups to be reviewed by Korea and China, 46 variant groups were accepted by both Korea and China (i.e., solved) and the remaining 258 were not solved.

Each of 258 variant groups belongs to one of four cases: [kC], [Kc], [kc], [KC]. The meanings of [k], [K], [c], and [C] are explained here. [k] indicates Korea is willing to yield (i.e., Korea is willing to accept Chinese variant group); [K] indicates Korea does not yield (i.e., Korea asks China to split Chinese variant group); [c] indicates China is willing to yield (i.e., China is willing to split Chinese variant group); [C] indicates China does not yield (i.e., China does not want to split Chinese variant group).

Now, let's consider four cases, ([k] or [K]) X ([c] or [C]) one by one: [kC], [Kc], [kc], and [KC].

1) [kC] Korea yields: Korea accepts Chinese variant group and this variant group is considered solved (i.e., unified). Usually, the number of [kC] variant groups is not zero.

2) [kc] Both Korea and China are willing to yield: Korea is willing to accept Chinese variant group and China is also willing to split Chinese variant group. However, it is not possible that both Korea and China yield. Only one of K and C can yield and, therefore, [kc] will eventually become either [Kc] or [kC]. The number of [kc] variant groups must become zero for final unification.

3) [KC] Neither Korea nor China is willing to yield: Korea asks China to split Chinese variant group and China does not want to split Chinese variant group. This is the most difficult case for Korea-China unification process.

[KC] will eventually become [Kc] or [kC] for final unification. Therefore, the number of [KC] variant groups must become zero for final unification.

4) [Kc] China yields: China accepts Korean request to split Chinese variant group and this variant group is considered solved (i.e., unified). Usually, the number of [Kc] variant groups is not zero.

There is one point to note in splitting Chinese variant group of size 3 or more.

First, Chinese variant group is fully split: after split, each K character in the variant group becomes an independent character. An example is shown below:

[(6DCB 淋) (75F2 痲) (9EBB 麻)]: a variant group of size 3
-->
[(6DCB 淋)], [(75F2 痲)], [(9EBB 麻)]: three independent characters

Second, Chinese variant group is partially split: C variant group of size 3 is split into a variant group of size 2 and one independent character. An example is shown below:

[(9762 面) (9EAA 麪) (9EB5 麵)]: a variant group of size 3
-->
A variant group of size 2 [(9762 面) (9EB5 麵)] + an independent character [(9EAA 麪)]

During Korea-China unification process, although most C variant groups of size 3 were partially split, a few were fully split. It needs be noted that there were no C variant groups with 4 or more K characters. If a Chinese variant group with 4 or more K characters is split, there can be several possibilities; however, such analysis is not explained in this paper.

4.2 After Taipei meeting (2016.09.30): out of 258 Chinese variant groups, 215 solved and 43 unsolved

Korea and China tried to find variant groups for which Korea or China is willing to yield (i.e., accept the other's request)

The criteria for "yield" are how much different the meanings of variant characters are, how frequently the variant characters are already used in domains, how frequently the variant characters are used in daily life, etc. Before Taipei meeting, 258 variant groups were unsolved. After Taipei meeting, 215 (83%) were solved tentatively and the remaining 43 (17%) remain unsolved. The detailed situation after Taipei meeting is shown in Table 6. Table 6 shows the number of variant groups belonging to [kc], [Kc], [kC], and [KC]. The same content is also shown in column 2) of Table 7.

Table 6. Coordination results for 258 C variant groups as of 2016.09.30, after Taipei meeting

China Korea	[c] C is willing to split C variant group	[C] C keeps C variant group	sub-total
[k] K is willing to accept C variant group	[kc] 56 var. group solved: Both K and C are willing to yield. Eventually will become either [Kc] or [kC].	[kC] 78 var. group solved: C variant group kept	K yields: 56 + 78 = 134

[K] K asks C to split C variant group	[Kc] 81 var. group solved: C variant group split	[KC] 43 var. group unsolved: Neither K nor K yields	K does not yield: 81 + 43 = 124
sub-total	C yields: 56 + 81 = 137	C does not yield: 78 + 43 = 121	total: 258 (215 solved, 43 unresolved)

4.3 Almost final unification

After Taipei meeting, several more meeting were held and, after the Beijing meeting on 2017.02.23, Korea and China almost finished unification of variant groups. It took about three and half years in total. This is summarized in column 3) of Table 7. The number of K variant groups is 168 (= 122 + 46).

Table 7. A table showing coordination results for variant groups (2016.09.28 ~ 2017.02.23)

Date, Meeting Name Who yields?	1) 2016.09.28. before Taipei meeting	2) 2016.09.30. after Taipei meeting	3)2017.02.23. after Beijing meeting
[kc] K&C willing to yield	0	56	0
[Kc] only C willing to yield	0	81	136
[kC] only K willing to yield	0	78	122
[KC] K, C: not willing to yield	258	43	0
solved: sub-total [kc + Kc + kC]	0	215	258
unsolved: [KC]	258	43	0
sub-total [kc + Kc + kC + KC]	258	258	258
[eq] K and C variant groups were equal from the beginning	46	46	46
total [kc + Kc + kC + KC + eq]	304	304	304

4.4 K-LGR v0.7 [13]

The number of Hanja characters in K-LGR v0.7 [13] is 4758 and the number of variant groups in K-LGR v0.7 is 152.

Previous versions of K-LGR included Hanja characters in DPRK standard KPS 9566 and some difficult characters in Hanja Proficiency Examination. Based on ICANN's IP (Integration Panel) request, these characters were removed from K-LGR, which explains why K-LGR v0.7 contains less Hanja characters and variant groups.

Table 8. Table showing the number of C variant groups for each of the size of C variant groups and the number of K characters therein

size of C var. group	# of K chars in a C variant group					
	0 char	1 char	2 chars	3 chars	subtotal	subtotal

2	1256	1464	81	--	2801	3475 (size of C var. group >= 2
3	96	367	38	5	506	
4	14	89	16		119	
5	3	25	7		35	
6	1	5	4		10	
7		3			3	
8			1		1	
subtotal	1370	1953	147	5	3475	3475
			152			

Table 8 shows the number of C variant groups in C-LGR (2017.03.31) [10] for each of the size of C variant groups and the number of K characters (K-LGR v0.7) therein. The number of variant groups in K-LGR v0.7 is 152 (= 147 + 5) and it is highly unlikely that 152 will change.

5. The Possibility of Confusion between Hangeul Syllables and Hanja Characters

The author analyzed systematically whether some Hanja character could be confused with a Hangeul syllable and obtained a good result which was not expected at the beginning. Probably this kind of systematic analysis has not been performed in the past and seems the first attempt, which is one of the contributions of this paper.

5.1 Confusion between Hanja and Hangeul syllable

After Korea-China unification process of variant groups is done, Korea and Japan need to take an additional step to define K-LGR and J-LGR, respectively. C-LGR treats only Hanja. However, Japan needs to include Kana in J-LGR in addition to Hanja (Kanji). Korea also needs to include Hangeul syllables in K-LGR in addition to Hanja so that domains such as ".한글", ".漢字", and ".한글과漢字" can be used.

Probably Korean people will not experience confusion between Hangeul syllables and Hanja. However, people unfamiliar with Hangeul and Hanja could be confused.

Most Hanja characters are quite different or too complex to be confused with Hangeul syllable. The systematic analysis shows that a few simple Hanja could be confused with Hangeul syllable.

Two method were used for the analysis. First, the author tried to find Hanja characters containing a Hanja component similar to Hangeul consonant letter and also a Hanja component similar to Hangeul vowel letter. This is a systematic approach. Second, the author tried to find a Hanja character as a whole that look similar to Hangeul syllable. Such Hanja character was not easily found using the first method.

5.2 Hanja components similar to Hangeul consonant or vowel letters

The author found that the following Hanja components are similar to Hangeul consonant or vowel letters. The Hangeul letter is shown within a pair of parentheses which look similar to the Hanja component.

1) 8 Hanja components similar to Hangeul consonant letters:

U+4E5A 乚 (U+1102 ᄂ); U+4EBA 人 (U+1109ᄉ); U+5165 入 (U+1109 ᄉ);

U+531A 匚 (U+1103 ᄃ); U+5338 匸 (U+1103 ᄃ); U+53E3 口 (U+1106 ᄆ);
U+56D7 囗 (U+1106 ᄆ); U+5DF1 己 (U+1105 ᄅ)

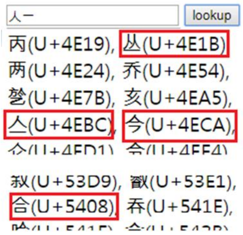
2) 9 Hanja components similar to Hangeul vowel letters:
U+4E00 一 (U+1173 ᅳ); U+4E04 丄 (U+1169 ᅩ); U+4E05 丅 (U+116E ᅮ);
U+4E0C 丌 (U+1172 ᅲ); U+4E28 丨 (U+1175 ᅵ); U+4E29* 丩 (U+116C ᅬ);
U+4E85 亅 (U+1175 ᅵ); U+4EA0 亠 (U+1169 ᅩ); U+535C 卜 (U+1161 ᅡ)

5.3 Hanja containing component similar to consonant letters and component similar to vowel letters

For example, U+4EBC "亼" contains U+4EBA 人 (similar to Hangeul consonant letter ㅅ) and U+4E00 一 (similar to Hangeul vowel letter ㅡ) and this Hanja could be confused with Hangeul syllable "스".

To find systematically, the author searched using each of 72 cases (8 components similar to consonant letters X 9 components similar to vowel letters). Hanja searching program was utilized. Out of 72 (= 8 X 9) cases, only three cases gave Hanja that could be confused with Hangul and their search results are shown below:

1) The author searched for Hanja characters containing both 人 (person) and 一 (one) components. About 6000 Hanja characters were found to contain both 人 (person) and 一 (one) components possibly in addition to other components. Most of them were too complicated or quite different from Hangeul syllables. The four Hanja characters enclosed in rectangle in Figure 1 are found to look similar to Hangeul syllable.



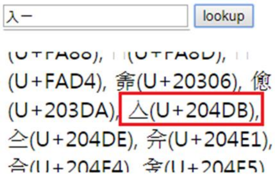
人一

丙(U+4E19), 丛(U+4E1B)
两(U+4E24), 乔(U+4E54),
乻(U+4E7B), 亥(U+4EA5),
亼(U+4EBC), 今(U+4ECA),

叙(U+53D9), 叡(U+53E1),
合(U+5408), 吞(U+541E),

Figure 1. Hanja characters which contain both 人 (person) and 一 (one) and also look similar to Hangeul syllables

2) About 250 Hanja characters were found to contain both 入 (enter) and 一 (one) components. Only one character looks similar to Hangeul syllable "스" (Figure 2).

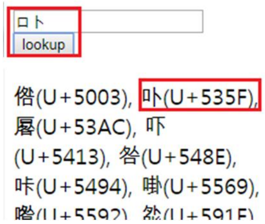


入一

(U+FAD4), 𠌆(U+20306), 𠏚
(U+203DA), 𠓛(U+204DB),
𠓞(U+204DE), 𠓡(U+204E1),

Figure 2. Hanja chars which contain both 入 (enter) and 一 (one) and also look similar to Hangeul syllables

3) About 50 Hanja characters were found to contain both 口 (mouth) and 卜 (fortune telling) components. Only one character looks similar to Hangeul syllable "□" (Figure3).



倃(U+5003), 卟(U+535F),
厬(U+53AC), 吓
(U+5413), 咎(U+548E),
咔(U+5494), 啩(U+5569),
喒(U+5592), 夞(U+591E)

Figure 3. Hanja chars which contain both 입 口 and 점 卜 and also look similar to Hangeul syllables

5.4 Hanja character as a whole which looks similar to Hangeul syllable

Some Hanja characters could not be decomposed into components. However, a certain Hanja character as a whole could look similar to Hangeul syllable. One such character was found.

U+723F 爿 (a piece of wood) - U+B258 뉘

5.5 Summary: Seven Hanja characters look similar to Hangeul syllables

Table 9 shows seven Hanja characters which is a summary of the results in Sections 5.3 and 5.4.

Table 9. Seven Hanja characters which look similar to Hangul syllables

Hanja char	Hangeul syllable
U+4E1B 丛	U+C4F0 쓰
U+4EBC 亼	U+C2A4 스
U+4ECA 今	U+C2A5 슥
U+5408 合	U+C2B4 슴
U+204DB 𠓛	U+C2A4 스
U+535F 卟	U+B9C8 마
U+723F 爿	U+B258 뉘

7 Hanja characters shown above can be classified into three cases:

1) The following two Hanja characters are included in K-LGR and in C-LGR.

[U+4ECA 今, U+C2A5 슥]
[U+5408 合, U+C2B4 슴]

2) The following three Hanja characters are included in C-LGR, but not in K-LGR.

[U+4E1B 丛, U+C4F0 쓰]

[U+535F 卟, U+B9C8 마]
[U+723F 爿, U+B258 뉘]

3) The following two Hanja characters are not included in K-LGR or in C-LGR.

[U+4EBC 亼, U+C2A4 스]
[U+204DB 𠓛, U+C2A4 스]

Any comments are welcome as to the validity of this approach and analysis result.

6. Representing K-LGR in XML according to RFC 7940

In general, all LGRs including K-LGR, C-LGR, and J-LGR, are read and processed by programs. Therefore, the list of characters and variant groups and other rules of LGRs are represented in XML according to RFC 7940 [15] for submission to ICANN.

The details will not be explained here, but some examples are shown below.

1) The fact that Hanja character 丁 U+4E01 is included in K-LGR is represented as follows:

```
<char cp="4E01" ref="0 101 102"/>
```

2) The fact that U+58F9 壹 is a variant of U+4E00 一 is represented as follows:

```
<char cp="4E00" ref="0 101 102">
  <var cp="58F9" ref="0 101 102" type="blocked"/> </char>
```

"blocked" indicates that, once a domain containing U+4E00 一 is allocated to an applicant, then other domains where U+4E00 一 is replaced with U+58F9 壹 cannot be allocated to other applicants.

3) The fact that U+4ECA 今 is a variant of Hangeul syllable U+C2A5 슥 is represented as follows:

```
<char cp="C2A5" ref="2">
  <var cp="4ECA" ref="0 101 102" type="blocked"/> </char>
```

7. Conclusions and Jobs to be done in the future

7.1 Activities of KGP

KGP (Korean Script Generation Panel) started its activity in Dec. 2013 and then, in Feb. 2016, the Korean community "formally" formed Korean Script Generation Panel for developing the Root Zone Label Generation Rules (RZ-LGR). KGP is still working. K-LGR v0.1 was announced in May 2015 and several versions of K-LGR such as v0.2, v0.3, v0.4, v0.5, and v0.6 were announced later. K-LGR v0.7 [13] was announced in Mar. 2017 and K-LGR v1.0 [14] was posted on ICANN web site for public comments in Jan. 2018.

7.2 Conclusions

Unification process of variant groups which is the main core of Korea-China coordination and almost final unification result are summarized in this paper. The CJK coordination started in 2013 and it is expected that K-LGR will be finalized and published probably in 2019. The author has been working as a chair of KGP from the beginning of this job. The author wrote a paper [5] showing a preliminary result at the beginning of Korea-China coordination and this is a follow-up paper. This paper shows almost final coordination results between Korea and China. By following the author's analysis and principle/rules, Korea and China could reduce drastically the number of variant groups to be reviewed by Korea and China. It was necessary for Korea and China to review only 304 C variant groups (9%) out of 3518 Chinese variant groups, which allowed efficient review and coordination process. In addition, the author analyzed systematically and found that seven Hanja characters could be confused with Hangeul syllables. Probably this kind of systematic analysis has not been performed in the past and is the first attempt, which is one of the contributions of this paper.

7.3 Jobs to be done in the future

There are two main issues to be solved in the future. K-LGR v1.0 [14] was posted on the ICANN web site for public comments in Jan. 2018. Currently K-LGR is being revised to accommodate the public comments. The main issue to be decided is whether Hangeul Hanja mixed domain such as ".한글과漢字" is needed. Some people argue that since Hanja is not used much in daily life, such domain must not be allowed. Other people argue that, even though the demand for such domains is not high, we should not forbid such domains for those who need them. This is the first issue.

The second issue is about variant groups composed of Hangeul syllable and Hanja character. ICANN's IP asked KGP to include such variant groups and KGP included them in K-LGR v1.0. Similarly, IP asked JGP to include variant groups composed of Hanja (Kanji) and Kana; however, JGP is not willing to include them in J-LGR. These two issues need to be solved in the future.

Acknowledgement

This work was supported by a 2-Year Research Grant of Pusan National University.

References

[1] Guidelines for Developing Script-Specific Label Generation Rules for Integration into the Root Zone LGR, Version 2015-04-24. https://www.icann.org/en/system/files/files/Guidelines-for-LGR-20150424.pdf.

[2] Procedure to Develop and Maintain the Label Generation Rules for the Root Zone in Respect of IDNA Labels, Version 2013-03-20b. https://www.icann.org/en/system/files/files/lgr-procedure-20mar13-en.pdf.

[3] K. Konishi, K. Juang, H. Qian and Y. Ko. RFC 3743, Joint Engineering Team (JET) Guidelines for Internationalized Domain Names (IDN) Registration and Administration for Chinese, Japanese, and Korean. Apr., 2004.

[4] X. Lee, W. Mao, E. Chen, N. Hsu and J. Klensin. RFC 4713, Registration and Administration Recommendations for Chinese Domain Names. Oct., 2006.

[5] K. KIM, “A study on a method of selecting variant groups to be reviewed for LGR (Label Generation Rule) of Internet Top-Level Hanja Domains,” KIPS Tr. Comp. and Comm. Sys. (Korea Information Processing Society), Vol. 5, No. 1, pp. 7~16, 2016.
DOI: https://doi.org/10.3745/KTCCS.2016.5.1.7

[6] ISO/IEC 10646, Information technology — Universal Coded Character Set (UCS), fifth edition, Dec. 2017.

[7] The Whole Table of Simplified Characters (简化字 总表), 1964, China Character Reform Committee, China Ministry of Culture, China Ministry of Education.

[8] CGP MSS 2015.04.30. Chinese repertoire of 12563 Hanzi characters and 3093 variant groups.

[9] C-LGR 2016.07.20. Chinese repertoire of 19738 Hanzi characters and 3518 variant groups.

[10] C-LGR 2017.03.31. Chinese repertoire of 19744 Hanzi characters and 3475 variant groups.

[11] K-LGR v0.3, Korean repertoire of Hangeul syllables and 4819 Hanja characters and 37 variant groups. Document number klgp171_4. 2015.08.13.

[12] K-LGR v0.5, Korean repertoire of Hangeul syllables and 4819 Hanja characters and 50 variant groups. Document number klgp200_51e. 2016.09.28.

[13] K-LGR v0.7, Korean repertoire of Hangeul syllables and 4758 Hanja characters and 152 variant groups. Document number klgp220_78g. 2017.03.03.

[14] Proposal for a Korean Script Root Zone LGR, LGR Version 1.0, Korean Script Generation Panel. Document number klgp220_101f. 2018.01.25.

[15] K. Davies and A. Freytag. RFC 7940, Representing Label Generation Rulesets Using XML. August 2016.