klgp175\_1

**Responses to IP feedback to the Korean script root zone LGR**

**Source: KLGP (Korean Language Generation Panel)**

**Date: 2015.09.10.**

**Responses are added below the feedback from IP.**

==========================================================

Considerations concerning the repertoire and variant set of

the Korean script root zone LGR

**Source: Integration Panel**

**Date: June 21st2015**

1. Summary

This document analyses the *KR's repertoire and variants list ( v0.1, first draft, 2015.05.11.), file K2-KLGR-1.xlsx* dated May 11th 2015 authored by MinjeeKim.

The repertoire contains 6017 Hanja characters but no Hangul syllables. It includes the basic Korean Hanja set (K0), and a subset of other K sources repertoires for which there seems to be missing a rationale, which may be understandable since it is a preliminary LGR. Additionally, out of these 6017 code points 14 appear to have no K sources. There are also 66 variant pairs which might probably be reduced if the KLGR repertoire was restricted to the K0 set. It is also suggested that some evidence of current usage or at least planned usage of Hanja in second level domain is desirable.

1. Definitions
2. References used in this document

The following references are used in this document.

* [KLGR] *KR's repertoire and variants list ( v0.1, first draft, 2015.05.11.), file K2-KLGR-1.xlsx* dated May 11th2015 authored by MinjeeKim
* [Procedure] Procedure to Develop and Maintain the Label Generation Rules for the Root Zone in Respect of IDNA Labels, <http://www.icann.org/en/resources/idn/variant-tlds/draft-lgr-procedure-20mar13-en.pdf>
* Hanja K sources (as documented by the Unicode Standard 8.0 and ISO/IEC 10646:2014:
* K0 KS C 5601-1987 (now known as KS X1001:2004)
* K1 KS C 5657-1991 (now known as KS X1002:2001)
* K2 PKS C 5700-1 1994 (Reedited and standardized as KS X1027-1:2011)
* K3 PKS C 5700-2 1994 (Reedited and standardized as KS X1027-2:2011)
* K4 PKS 5700-3:1998 (Reedited and standardized as KS X1027-3:2011)
* K5 Korean IRG Hanja Character Set 5th Edition: 2001 (Reedited and standardized as KS X1027-4:2011)

1. Variant mapping references

KLGR contains 132 lines with variants definition constituting 66 pairs. Some examples:

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **No** | **Glyph** | **Code** | **VP1** | **Code** |  | **VP2** |  |  |
| 5468 | 鑑 | 09451 | 鑑 | 09451 | A | 鑒 | 09452 | A |
| 5469 | 鑒 | 09452 | 鑑 | 09451 | A | 鑒 | 09452 | A |
| 0370 | 减 | 051CF | 減 | 06E1B | A | 减 | 051CF | A |
| 2694 | 減 | 06E1B | 減 | 06E1B | A | 减 | 051CF | A |

No definitions for VP1, VP2, or ‘A’ were provided in the preliminary KLGR draft.

==> VP1: Code position of variant1, VP2: Code position of variant2, A = allocatable

1. Repertoire considerations

First it is worth mentioning the distribution of K sources (Hanja characters) encoded in Unicode 6.3 in MSR-2:

**Table 1: Distribution of K sources characters in MSR-2**

|  |  |  |  |
| --- | --- | --- | --- |
| **Source name** | **Full Count** | **Included in MSR-2** | **Excluded from MSR-2** |
| K0 | 4620 | 4620 |  |
| K1 | 2856 | 2731 | 125 |
| K2 | 7911 | 7462 | 449 |
| K3 | 1834 | 20 | 1814 |
| K4 | 172 | 4 | 168 |
| K5 | 404 | 0 | 404 |
| K source overall | **17797** | **14837** | **2960** |

This clearly shows that only K0 is fully represented in MSR-2. The next table shows the distribution of KLGR into K sources:

**Table 2: Distribution of K sources in KLGR**

|  |  |  |  |
| --- | --- | --- | --- |
| **Source name** | **Full Count** | **Included in KLGR** | **Excluded from KLGR** |
| K0 | 4620 | 4620 |  |
| K1 | 2856 | 840 | 2016 |
| K2 | 7911 | 538 | 7373 |
| K3 | 1834 | 5 | 1829 |
| K4 | 172 | 0 | 172 |
| K5 | 404 | 0 | 404 |
| K source overall | **17797** | **6003** | **11794** |

The discrepancy between the number found above (6003) and the number of the CJK characters in KLGR (6017) is created by the fact that of this 6017 code points, 14 have no K sources. The following table shows these code points, using as representative glyph the G source:

**Table 3: KLGR code points with no K source**

|  |  |  |
| --- | --- | --- |
| U+5AB4 | 媴 | G, H, T |
| U+741F | 琟 | G, H, T, KP |
| U+7494 | 璔 | G, H, T, KP |
| U+74A4 | 璤 | G, H, T |
| U+74B6 | 璶 | G, H, T, KP |
| U+74B8 | 璸 | G, H, T, KP |
| U+7682 | 皂 | G, H, T |
| U+7807 | 砇 | G, T, KP |
| U+7A66 | 穦 | G, T, KP |
| U+7B30 | 笰 | G, H, T, J, KP |
| U+7FF6 | 翶 | G, T |
| U+8CEB | 賫 | G, H, T, KP |
| U+9667 | 陧 | G |
| U+9ED9 | 黙 | G, T, J |

Findings:

KLGR is fully included in MSR-2. This was not the case in its first iteration but is now established, and we are pleased to see this.

KLGR covers completely the Korean basic set: K0 which is also fully included in MSR. This was expected because K0 is much alike the Japanese J0 set. One would expect it to be in any Hanja minimal set such as J0 is in any minimal Kanji set.

The case for the extra characters (840 + 538 + 5 = 1383) still needs to be made. Some of these may be needed for out of repertoire (in a K source meaning) variants only.

==>

- Explanation of marks/signs used in the Venn Diagram.

. K = KS X 1001, P = KPS 9566, I = IICORE/K, T = Hanja Test (한검 시험)

1) K0 4620 -> The number of Hanja chars in K0 (KS X 1001) is 4620.

2) P0 4653 -> The number of Hanja chars in P0 (KPS 9566) is 4653.

3) IK 4743 -> The number of Hanja chars in K column of IICORE is 4743.

4) HT 4641 -> The number of Hanja chars in QTKHP is 4641.

\* QTKHP: Qualifying Test of Korean Hanja Proficiency, 한국 한자 능력 검정 시험, http://www.hanja.re.kr/

5) Note 1. Hanja characters in IICORE/KP is exactly the same as those in P0 (KPS 9566) and therefore IICORE/KP is not shown explicitly in the Venn Diagram.

Note 2. Size of (~K)∩(~P)∩(~I)∩(~T) is 0

Note 3. Size of (K∪P∪I∪T) is 4819 (= The same as the number of Hanja chars in K-LGR-1, v0.3)

- Explanation of regions in the Venn Diagram of four sets:

Example 1: "KPIT 4594" -> The number of characters in K∩P∩I∩T is 4594.

Example 2: "K 2", -> The number of characters in K∩(~P)∩(~I)∩(~T) is 2.

Example 3: "KP 3" -> The number of characters in K∩P∩(~I)∩(~T) is 3.



It is also surprising to find 14 non-K source code points in the KLGR set. One would expect all of them to have a K source value if they are going to express Hanja text. Note however that 8 of these code points have a KP source (North Korea or DPRK).

==> 4819 Hanja characters in K-LGR v0.3 are extracted from four (or five) sources:

|  |  |
| --- | --- |
| Source of Hanja Character Set | # chars |
| 1) KS X 1001 (268 Comptb. Hanja characters excluded) | 4,620 |
| 2) KPS 9566 | 4,653 |
| 3) IICORE - K column marked | 4,743 |
| 4) IICORE - KP column marked  \* This is exactly the same as 4,653 Hanja characters in 2) KPS 9566 | 4,653 |
| 5) QTKHP: Qualifying Test of Korean Hanja Proficiency  (한국 한자 능력 검정 시험)  \* 9 Compatibility Hanja characters not in MSR-1 are excluded. | 4,641 |
| K-LGR-1, Hanja list v0.3 (2015.08.13.) | 4,819 |

1) The following seven code positions are kept in K-LGR v0.3:

05AB4 0741F 074A4 074B8\* 07807 07A66 09ED9

- The above 7 code positions are marked in K column of IICORE.

- The other 7 code positions were in K-LGR v0.2, but NOT in K-LGR v0.3.

|  |  |  |  |
| --- | --- | --- | --- |
| U+5AB4 | 媴 | G, H, T |  |
| U+741F | 琟 | G, H, T, KP |  |
| U+7494 | 璔 | G, H, T, KP | in K-LGR v0.2, but not in K-LGR v0.3 |
| U+74A4 | 璤 | G, H, T |  |
| U+74B6 | 璶 | G, H, T, KP | in K-LGR v0.2, but not in K-LGR v0.3 |
| U+74B8 | 璸 | G, H, T, KP |  |
| U+7682 | 皂 | G, H, T | in K-LGR v0.2, but not in K-LGR v0.3 |
| U+7807 | 砇 | G, T, KP |  |
| U+7A66 | 穦 | G, T, KP |  |
| U+7B30 | 笰 | G, H, T, J, KP | in K-LGR v0.2, but not in K-LGR v0.3 |
| U+7FF6 | 翶 | G, T | in K-LGR v0.2, but not in K-LGR v0.3 |
| U+8CEB | 賫 | G, H, T, KP | in K-LGR v0.2, but not in K-LGR v0.3 |
| U+9667 | 陧 | G | in K-LGR v0.2, but not in K-LGR v0.3 |
| U+9ED9 | 黙 | G, T, J |  |

If we consider basic Hanja at this point, it may be appropriate to limit the repertoire to K0 augmented by any characters in MSR to provide full variant transitivity.

1. Variants considerations

The KLGR variant set has 132 entries with 66 pairs. The Integration Panel (IP) did not use a tool to analyze this, partially because it is not clear what the LGR variant entries VP1 and VP2 mean in term of variants types. A quick check with entries located in the .asia zh set and the Korean set shows that the KLGR entries are mostly a subset of the .zh variants with however enough differences in the pair sets that it would require some important adjustment in both sides to make a consolidated set. The IP would rather wait to have an updated version of the .zh LGR in term of repertoire and variant types before going too deep in comparing with another set.

Out of the 14 KLGR with no K sources, 2 (U+7682 and U+8CEB) have variants pairs (with U+7681 and U+9F4E respectively). This is consistent with the .zh variant set. For these, one could possibly see an out-of-repertoire variant relationship (in K context), that still leaves the issue unresolved about the other 12 code points in the KLGR with no K source.

At least half of the variant pairs mix K0 with another K source (disproportionately with their ratio in the set). That means that a K0 only set will have much less variants. The IP have not examined all 66 pairs but enough to have a good assessment of the situation.

1. Conclusion

It would be sensible to have a KLGR containing K0 (and only K0) if the IP can get reasonable evidence that a Hanja set with such content is being considered as a second level domain repertoire in Korea.

==> We've started a discussion on whether to include Hanja characters in 2nd level domain under .kr/.한국 and hope to deliver Korean internet community decision soon.

Moreover, a final KLGR would be expected to contain the Hangul syllables.

==> K-LGR v0.3 (2015.08.13.) includes 11,172 Hangeul (Hangul) syllables.

\* \* \*