Source:	Yi Bai(白易), CheonHyeong Sim(沈天珩)
Title:	Proposal to consider adding CodeCharts support for kIRG_KPSource representative glyphs in Unicode
Date:	2022−10−16
Action:	To be considered by UTC

This document proposes the addition of kIRG_KPSource representative glyphs by a font from DPRK source to the CodeCharts, which is missing in current Unicode version, and will complete representative glyphs from all sources.

Background

Unlike other sources, very limited DPRK−source ideographs are included in current CodeCharts (8 in the Extension C block, 106 in the CJK Compatibility Ideographs block, 50 in the CJK Compatibility Ideographs Supplement block and 1 in the Extension H block). The difficulty comes from that DPRK has been inactive in Unicode community for more nearly two decades. Recently, we acquire a font from an Android application, namely Okpyon, from DPRK source, which includes all glyphs that has kIRG_KPSource.

After careful examination of the font, we found that the font includes exact the same number of ideographs as in the kIRG_KPSource, and the glyph shape in the font is identical with the ones in CodeCharts, i.e. it is the same font included in the current CodeCharts for DPRK's glyphs. Thus we propose the font to be included in future Unicode CodeCharts as representative glyphs of kIRG_KPSource.

Proposal

We propose including the font in the code chart, which will be an improvement to the code charts as it will complete the task of displaying representative glyphs from all sources.

The font is already used for CJK Ext−C, CJK Compatibility Ideographs, CJK Compatibility Ideographs Supplement, and CJK Ext−H block. Thus we propose the font to be included in other blocks, namely CJK Unified Ideographs (URO), CJK Ext−A and CJK Ext−B in future Unicode versions.

	KP0	KP1	*KPU*	CodeCharts Coverage (as of Unicode15.0)
URO	4652	10359		No
ExtA	1	3188		No
ExtB		5766		No
ExtC		8		Yes
ExtH		1		Yes
Comp.		106	*1*	Yes
Comp. Supp.		49	*1*	Yes
Unencoded		*83*		/

Reconstruction of font's mapping

As the order and blocks in the font is different from normal Unicode orders, the glyphs in the font are not a simple mapping to Unicode codepoints. To give a better understand of the font, we will introduce the reconstruction of font's mapping to Unicode here briefly.

1. Layout of Hanja-part in the font

The Hanja-part in the font can be divided into three parts:

The first part is CJK Radical Supplement, there is 1 Hanja as shown in the picture above; (Figure 1)

The second part is CJK Unified Ideographs and CJK Unified Ideographs Extension A, there are 18815 Hanjas, which has a full coverage of all the KP1-characters in BMP except for U+431B(鼬) and all the KP0-characters according to Unihan. There are 617 Hanjas(included in the 18815 Hanjas) without a KP-source, and it seems to be out of scope of our reconstruction;

The third part is U+AXXX and Private Use Area. There are 6014 Hanjas(range: U+A000-U+ABFF and U+E000-U+EB7D), including U+431B which is actually mapped to U+E211 in the font, the characters from CJK-ExtB, CJK-ExtC, CJK-Comp, CJK-Comp Supplement, and the characters unincluded in Unicode. These characters are all from KP1. And there are also some characters(not necessarily Hanja) in PUA after U+EB7E, but these are also useless for our reconstruction. The sketch maps of the Hanjas from the second and the third part will be shown below. (Figures 2-4)

(Figure 2)

(Figure 3)

(Figure 4)

As can be seen from the foregoing, there are 1+(18815−617)+6014=24213 Hanjas;

In KPS10721:2000, the first codepoint is 0x3400, and the last one is 0x9294. There are 0x9294−0x3400+1=24213 Hanjas.

According to Unihan, there are 4736 "holes"(codepoints are not continuous) in KP1. Meanwhile, there are 4653 Hanjas in KP0. If these Hanjas are filled into KP1, there will be 4736−4653=83 holes left;

For the Hanjas mapped to U+AXXX and PUA, 106 of them are encoded in CJK−Comp, 1 of them is encoded in CJK−ExtA, 5766 of them are encoded in CJK−ExtB, 8 of them are encoded in CJK−ExtC, 49 of them are encoded in CJK−Comp Supplement,

1 of them is encoded in CJK−ExtH. So there should be 6014−106−1−5766−8−49−1=83 holes left;

There are 94 characters in IRGN897, and 13 of them are later encoded to CJK−ExtC or horizontally extended to CJK−ExtA or CJK−ExtB. And there are also 2 characters withdrawn from the initial proposal of DPRK−Comp Characters(KP1−441D and KP1−510B, see WG2 N2573 for reference). IRGN897 came after that so these 2 characters was certainly not included in. So there should be 94−13+2=83 holes.

All the numbers listed above can match up, which shows the credibility of the font.

2. Procedure of the reconstruction

After knowing the credibility of the font, and we are able to conjecture that all the KP0−Hanjas are included in KP1 by the number of the holes, we can start doing the reconstruction.

First, we can easily discover that the Hanjas in KP1 are sorted by the radical and the residual strokes. When both the radical and the RS are same, the Hanjas are sorted by the Korean pronunciation(DPRK−Order). For example: (Radical Grass+6, 0x6D82−0x6DC5)

0x6D8X			茛간	茳강	茖격	𦰡곡	𦮽공	荍교	茭교	茩구	茥규	茋기	茣길	荂과	𦬦괄	茪광
0x6D9X	𦭂뇨		荅답	茼동	茢렬	荖로	𦰩루	荔리						荗술		茲자
0x6DAX	荢자	荃전	𦱶족	茽종	荵증	茡지	荐천	茜천	荈천		茺충	荎치	茬치	茝채	茞채	茦책
0x6DBX	莩표	茷폐	䓇혈		茠호	𦬼호	荄해			荁환		𦭵황	茚앙			茰유
0x6DCX	茟율	茙융		荋이												

The Hanjas with Radical Grass and RS 6 in KP0 are as follows: (Hanjas have already been arranged in order by the pronunciation in KP0)

荼다	茫망	茗명	茯복	茱수	荀순	茨자	草초	荊형	荇행	茴회	荒황	茹여	茸용	荑이	茵인	荏임

They can be filled into the holes just right.

0x6D8X			茛간	茳강	茖격	𦰡곡	𦮽공	荍교	茭교	茩구	茥규	茋기	茣길	荂과	𦬦괄	茪광
0x6D9X	𦭂뇨	荼다	荅답	茼동	茢렬	荖로	𦰩루	荔리	茫망	茗명	茯복	茱수	荀순	荗술	茨자	茲자
0x6DAX	荢자	荃전	𦱶족	茽종	荵증	茡지	荐천	茜천	荈천	草초	茺충	荎치	茬치	茝채	茞채	茦책
0x6DBX	莩표	茷폐	䓇혈	荊형	茠호	𦬼호	荄해	荇행	茴회	荁환	荒황	𦭵황	茚앙	茹여	茸용	茰유
0x6DCX	茟율	茙융	荑이	荋이	茵인	荏임										

Then we can do the filling for all the KP0−Hanjas.

Several known issues

1. About U+249D6(𤧖)

There is no glyph for this Hanja but U+746F(瑯) appears twice. It seems to be a mapping issue which led to a glyph mistake. See below:

0x594X								琙감	椇거	琕개	瑙노	瑖단	瑛돌	瑋대		瑁모
0x595X	瑂미	瑉민	瑊변	瑞서	瑄선	瑆성	瑟슬	瑑전	瑔전	瑅제	瑒창	瑃춘	瑕하	瑚호	琿훈	瑎해
0x596X	瑍환		瑝황	瑘야	瑌연	瑛영	瑀우	瑈유	瑜유	瑋위	瑗원	瑴각	瑰괴	瑭당	瑫도	
0x597X	瑬류	瑠류	瑮률	瑪마	瑭방	瑣쇄	瑣쇄	瑲장	璪조	瑱진	瑱진	瑨진	瑳차	瑹침	瑭퇴	瑩영
0x598X	瑦오	瑥온	瑤요	瑢용												

According to Unihan, KP1−594E should be U+249D6(環,환), but according to its pronunciation, it should be mapped to KP1−5961; According to Unihan, KP1−596F should be U+249E8(瑯), but there is not a same glyph in the font, we only have two "瑯" and a "瑯". While the glyph "瑯" refers to KP0−D7D7, we can conjecture that the glyph "瑯" is mapped to KP1−596F and U+249E8 for the cognition, but according to the RS, "瑯" seems to be more appropriate to be mapped to KP1−594E while "瑯" to KP1−596F.

2. Mapping issues

We discovered some mapping issues when doing the reconstruction. Some of them are completely wrong, and the rest of them are cognate to the current mappings but there are probably some better mappings. And we will divide these Hanjas into two parts to illustrate.

2.1. Wrong mappings

KP1−50FB(浫) is now mapped to U+23CD9(泘), but they are completely different.

0x50FX	沭술	泎색	泩생	泚자	沮저	泞저	沺전	泟정	注주	泜지	泲지	浫집	泲제	泎찰	泉천	沾첨

From the pronunciation, we could know that the glyph is not wrong but the mapping is wrong, because U+23CD9(泘) is read as 소. So it should be fixed to U+23CC0;

KP1−5B5D(痦) is now mapped to U+24D6A(瘖), but they are completely different.

0x5B5X	疲흡	疫역	疣우	痂가	疳감	痃거	痁고	痀구	痊근	疝과	痆닐	疸단	疼동	痦려	痋령	疺법

From the pronunciation, we could know that the glyph is not wrong but the mapping is wrong, because U+24D6A(瘖) is a character used for a person's name in Taiwan or a Chữ Nôm in Việt Nam, its pseudo−Korean−pronunciation should be 소 or 조. So it should be fixed to U+3F94;

KP1−7EF4(酦) is now mapped to U+9166(酦), but they are completely different. Although the pronunciation for "酦" and "酦" are both 발, firstly, from the two Hanjas illustrated before, it is more likely to be a mapping issue but not a glyph issue; secondly, "酦" is the simplified form of "醱" in the Mainland of China, it is unlikely to be used in DPRK, while "酦" is a traditional character without being simplified anywhere. So it should be fixed to U+48EE.

2.2. Suboptimal mappings

Some Hanjas are encoded separately with the different actual shape in Unicode, but they are cognate. For example, KP1−4A00(曚) is now mapped to U+66DA(曚), while its glyph is more close to U+232E1. But we cannot say it is wrong due to the cognition. They should probably be unified under the current unification rules. For these characters, we maintain a list of them in other sources[1].
(Remark: The table lists all the mapping issues regardless of whether it is a wrong mapping or a suboptimal mapping.

For KP1−5450 and KP1−8346, they have a similar case as "瑯" mentioned before, whose current mapping is not the best but the best mapping has already been occupied by another Hanja.)

(End of document)

Ref.

[1] CheonHyeong Sim. KP1−reconstitution.pdf[DB/OL]. http://cheonhyeong.com/PDF/KP1−reconstitution.pdf, 2022−06−19: PP.74−75.