Open Letter to the Proposal for a Korean Script Root Zone LGR

Authors: Jaemin Chung, Seonghoon Kang, Shinjo Park

This is an open letter to the Proposal for a Korean Script Root Zone LGR, published as https://www.icann.org/en/system/files/files/proposal-korean-lgr-25jan18-en.pdf, made in public on 2018-01-25.

We are deeply concerned about the recent proposal for allowing both Hangul and Hanja in the Korean script (as opposed to Hangul only) in internationalized domain names (henceforth the Proposal). We believe the support will give negligible benefits to the Korean community while increasing the potential to confuse. It will also make similar-looking names open for exploit, and introduce new accessibility issues. We believe that the Proposal should be withdrawn.

Exaggerated Hanja Usage in Korea

The Proposal claims that Hanja usage is significant in Korea to warrant its support in IDNs. This is greatly misleading and we consider the evidence listed in the appendices of the Proposal to be intentionally chosen to exaggerate its usage. In this section, we will review them and show why they are not representative of what IDNs will encode.

Coexistence, Not Substitute

Today, Korean is written in Hangul. ***Hanja are sometimes provided in parentheses next to Korean words****, but only when the word in Hangul alone may be misunderstood due to its multiple meanings or when further clarification of a specific meaning is necessary.* ***In rare cases, words are written in Hanja and the particles and suffixes associated with the words are left in Hangul.***

(Emphases ours)

It is true that Hanja still enjoys the continuous usage as a supplementary script in the modern Korean language. But its usage has been greatly reduced to clarify homophones in Hangul and independent usages are much rarer, even mentioned in the Proposal itself. Unless IDNs allow annotation strings inside a label, the mere coexistence as a supplementary script is not a good fit for IDNs.

Registered Trademarks That Do Not Warrant Current Use

The Proposal shows multiple mixed-script trademarks seemingly registered in recent decades. This is misleading, as their origins could be traced back to several decades ago when the proprietor of trademarks actively used Hanja-based ones. While they are no longer actively used, the trademark renewal can be explained as a defensive measure against expiration or misuse for older trademarks that are irrelevant to current usage.

For example,

- Samsung Moolsan (more commonly known as Samsung C&T Corporation)
  This particular trademark traces back to at least 1975 (registration number 4000484540000). Of course, Samsung C&T no longer uses the mixed-script trademark today; it instead uses a trademark common to the Samsung group and a logotype for “삼성물산”.
- Samsung Jeonja (more commonly known as Samsung Electronics Co., Ltd.)
  This particular trademark traces back to at least 1985 (registration number 4001346350000). Again, this is not the current trademark used by Samsung Electronics.
- Hyundai Motors
  This particular trademark traces back to at least 1969 (registration number 4000191560000).

- Kumho
  This particular trademark traces back to at least 1986 (registration number 4100077620000).
- Taehwa Shopping
  This particular trademark traces back to at least 1983 (registration number 4100043980000). The company itself went bankrupt in 2001 and renamed to Judies Taehwa in 2003.

In most recognizable cases like Samsung and Hyundai, it is not hard to see that these companies no longer use Hanja trademarks in Korean publicly. One can argue that Hanja trademarks are used in Chinese and Japanese markets; we will later iterate that supporting Korean Hanja only is not sufficient for Chinese and Japanese speakers.

Decorative Nature of Hanja

It is not as prominent in signboards (due to the selective nature of examples), but many mixed-script trademarks are common in that the only portion written in Hanja is 주식회사 (株式會社, lit. *corporation*). This is not a coincidence. Being a script taught in school but rarely written, Hanja has acquired a supplementary role to decorate some words. We have chosen the word “decoration” for this purpose because its usage is limited to common words, like 남 (男, lit. *male*), 대 (大, lit. *big*; often short for *university* [대학교]), 가 (家, lit. *home*). If they were meant to be emphases, any word of importance would be written in Hanja.

We believe that such decorations are not to be written but to be only read, and an IDN label is not a place to reproduce decorations.

Other Bad Examples

Remaining examples comprise corporate registers and law books, both highly atypical as examples of a mixed-script IDN. The authors of the Proposal seem to have failed to find more convincing evidence of Hanja. We sympathize with the authors as it will be quite difficult.

In reality, Hanja usage is so low that neither Android nor iOS provides Hanja input methods in Korean by default. Please note that this is not just an oversight; smartphone penetration in Korea has reached 90% of the total population and remains extremely high across all ages, as observed in Gallup Korea’s 2017 survey[1] for example. If Hanja cannot be readily typed what’s the point in allowing Hanja IDN labels?

Hanja Will Be Used in Parallel with Hangul in Primary School Textbooks?

Appendix H.6.2 cites an article from The Hankyoreh[2] mentioning that Hanja will be included in primary school textbooks, starting from 2019. However, later article from the same newspaper[3] dated 9th January 2018 states that the inclusion of Hanja in primary school textbooks will not be implemented. Given that the Proposal was published on 25th January 2018, two weeks after the announcement, the validity of the entire section of the appendix needs to be questioned.

Potential Ambiguities

One may still recognize even the slightest use of Hanja and argue for its support in IDNs. We believe that, however, in addition to low usage, mixed-script IDNs will be actively harmful to the overall Korean community, due to confusion not only within Hanja, but also between Hangul and Hanja.

---

1 http://www.gallup.co.kr/gallupdb/reportContent.asp?seqNo=813. 2017-02-15, Retrieved 2018-02-12.

2 http://www.hani.co.kr/arti/society/schooling/776746.html. 2016-12-30, Retrieved 2018-02-12.

3 http://www.hani.co.kr/arti/society/schooling/827055.html. 2018-01-10, Retrieved 2018-02-12.

Hangul-Hanja Confusion

The Proposal lists five variant groups of similar-looking Hangul and Hanja, only two of them in the repertoire. This is a massive underestimation and we believe that there can be tens or even hundreds of them; it is so well known that it got a name Yaminjeongeum (야민정음) on Korean Internet, a pun using the name of the supposed origin and the original name of Hangul, Hunminjeongeum. Namuwiki, a community-driven encyclopedic wiki, has a great list[4] of them.

Below is the non-exhaustive list of Hanja characters included in MSR-2 which have the similar visual appearance with Hangul characters (either jamo or syllable), which are not included in the Proposal's section 6.2:

- 勹 (U+52F9) ↔ ㄱ (U+1100)
- 廿 (U+5EFF) ↔ ㅂ (U+1107)
- 刁 (U+5201) ↔ ㅋ (U+110F)
- 大 (U+5927) ↔ ㅊ (U+110E)
- 刀 (U+5200) ↔ ㄲ (U+1101)
- 丗 (U+4E17) ↔ ㅃ (U+1108)
- 从 (U+4ECE) ↔ ㅆ (U+110A)
- 金 (U+91D1) ↔ 숲 (U+C232)
- 長 (U+9577) ↔ 틍 (U+D2BD)
- 辛 (U+8F9B) ↔ 푸 (U+D478)
- 卒 (U+5352) ↔ 쭈 (U+CB48)
- 끄 (U+536B) ↔ 고 (U+ACE0)
- 告 (U+544A) ↔ 솜 (U+C19C)
- 丕 (U+4E15) ↔ 조 (U+C870)
- 奏 (U+5940) ↔ 좇 (U+C887)
- 笑 (U+7B11) ↔ 쑷 (U+C47B)
- 豆 (U+8C46) ↔ 묘 (U+BB18)
- 号 (U+53F7) ↔ 묵 (U+BB35)
- 뭇 (U+5434) ↔ 뭊 (U+BB4A)
- 크 (U+5F50) ↔ 크 (U+D06C)

Even with a simple search on the Internet, we can easily find look-alike characters which are not listed in the Proposal. Given the fact that this list is non-exhaustive containing less than 1% of characters included in MSR-2, we are afraid that there are much more pairs unknown to us.

Same-Script Confusion

Chinese characters are generally noted for their visual complexity, which suggests higher potential of same-script confusion.

The complexity of Hanja can also contribute to phishing attacks. Let's take an example of 검찰청 (檢察廳, Prosecutors' Office). The Hangul spelling 검찰청 is used daily, while some news headlines use the initial character's Hanja equivalent 檢 as a symbol. However, all three characters in its Hanja spelling have quite large numbers of strokes, making them hard to distinguish especially in low-resolution and/or small display. If Prosecutors' Office wants to register its Hanja-only IDN, it would be 檢察廳.한국. Assume that an attacker manages to register 檢祭廳.한국, which shares the first and third characters while only the second one is different (察 (U+5BDF) vs. 祭 (U+796D)). Unlike Chinese and Japanese speakers who use Hanzi or Kanji every day, a substantially high number of Korean speakers will believe 檢祭廳.한국 is also the correct IDN since the full Hanja spelling is not used on a daily basis.

Another example is 청구서 (請求書, invoice). The Hanja spelling 請求書 is not used daily, and unlike the example of 검찰청, none of the Hanja characters are used as a symbol. Suppose that legitimate companies could use their name with

4 https://namu.wiki/w/%EC%95%BC%EB%AF%BC%EC%A0%95%EC%9D%8C?rev=1559#s-2.7 야민정음 (r1559 판) – 나무위키 (Yaminjeongeum (revision 1559) – Namuwiki). 2018-02-12, Retrieved 2018-02-12.

請求書 suffixed as a Hanja IDN for their electronic invoice system. If an attacker wants to phish as an electronic invoice system, they can buy the Hanja IDN 請求晝. This IDN shares the first and second characters with the legitimate one, while only the third one is different (書 (U+66F8) vs. 晝 (U+665D)). Unlike the Prosecutors' Office example, even a partial Hanja spelling is not used daily, therefore making users more vulnerable to same-script confusion.

In fact, Korea already paid a lot of social resources to counter domain-name-based phishing even without IDNs. Prosecutors' Office[5] and invoices[6], listed as examples here, can easily become real victims of domain-name-based phishing. An extra countermeasure will have to be added to defend against Hanja-based phishing.

Although chances of same-script confusion also exist in Chinese and Japanese IDNs, due to the low daily usage of Hanja in Korean, it will create much more confusion when Hanja is introduced into Korean IDNs.

Multiple Representations of Mixed-Script IDNs

When a name consists of multiple Korean words, there can be multiple combinations of Hangul and Hanja for that name. Let's take an example of 대전도시공사 (Daejeon City Corporation, doing city infrastructure management like German Stadtwerk plus real estate development). This name is comprised of three individual words: 대전 (大田, city of Daejeon), 도시 (都市, city), and 공사 (公社, state-owned enterprise). In Hangul-Hanja mixed-script IDNs, these three words can be written in either Hangul or Hanja. All of these domain names are possible: 大田都市公社.한국, 大田도시공사.한국, 대전都市公社.한국, and so on. Unless all possible domain names are taken by the legitimate owner, someone else can register any possible domain names that are not taken and pretend that they are the real corporation. To avoid such a problem, owners of multi-word domain names need to register all possible combinations of Hangul-Hanja mixed-script IDNs, or only register the one that is likely to be most commonly used and hope for the best. The difficulty of choosing one from all possible combinations can be a detriment to safety and discourage the usage of Korean IDNs altogether.

Other Problems

We have found other problems in the Proposal, which further weaken the validity of the entire proposal.

Incompleteness of Repertoire

The K portion of IICore comprises 4744 (not 4743) characters, which consists of KS X 1001 (without 14 characters) and additional 138 characters. According to Dr. Ken Lunde[7], an expert on CJKV information processing, the details of how these 138 particular characters were selected are unknown. He also says that there is no document or report explaining how the K portion of IICore was prepared, and the people who compiled the K portion of IICore either passed away or are no longer participating in the Korean National Body. It is questionable why the Proposal should include IICore characters.

Accessibility Issues

We believe that IDNs in general need to be platform-neutral. To this end, Korean IDNs are not meant to be used only on Microsoft Windows computers with the KS X 5003 Korean keyboard. Appendix H.5 in the Proposal shows how to enter Hanja with the Korean IME on Microsoft Windows. According to StatCounter[8], there are about 15% of users who are not using Microsoft Windows as their PC operating system. Even though widely used input method applications in macOS and Linux support Hanja input, the Proposal fails to show PC operating systems other than Microsoft Windows.

---

5 http://www.ilyoseoul.co.kr/news/articleView.html?idxno=211304 – The article mentions that the phisher gave a fake website of Prosecutors' Office. 2017-11-14, Retrieved 2018-02-07.

6 http://blog.alyac.co.kr/1353 – Figure 2 in the article shows an example of a phishing email with Apple's invoice, which redirects to a fake website containing "Apple" in their domain name. 2017-09-28, Retrieved 2018-02-07.

7 https://blogs.adobe.com/CCJKType/2018/02/exploring-iicore-part-1.html. 2018-02-05, Retrieved 2018-02-07.

8 http://gs.statcounter.com/os-market-share/desktop. 2018-01, Retrieved 2018-02-07.

We believe that the Proposal should have included Hanja input methods for all these three major PC operating systems for the sake of completeness.

Although the Korean Standard KS X 5003 includes dedicated keys for Hanja conversion and Hangul/Alphabet toggle, not all keyboards sold in Korea are equipped with those keys, especially for laptops where they are shared with the right Ctrl and Alt keys respectively. Similarly, Apple keyboards sold in Korea do not have them at all[9].

Moreover, web browsers are not only installed on PCs today. But the Proposal also fails to address Hanja input on non-PC environments. There are other media devices with web browsers such as smartphones, set-top boxes, smart TVs and game consoles. They are more resource-constrained than a PC, and Hanja input requires a big dictionary like Chinese and Japanese input methods. Designers usually leave only Hangul input functionality in Korean locale to avoid the overhead introduced by Hanja data. While Hanja input requires candidate selection – which could be time-consuming without efficient candidate navigation interface – Hangul input methods can directly translate QWERTY or ten-key keyboard keystrokes to final precomposed characters, making Hangul input more efficient than Hanja input in an environment without the traditional PC keyboard. For those reasons, it is impractical to implement Hanja input in addition to Hangul input on every such device. Given increased penetration of Internet-enabled devices without Hanja input, chances of utilizing Hanja IDNs on such devices will be very low. Presenting Hanja input only on Microsoft Windows PC is not a proof that Hanja input is universally available for all Korean speakers.

Domain names are also used in spoken language. Alphabet-only and Hangul-only domain names have no or minor difficulties in spoken language. In contrast to both types of domains, dictating Hanja in a spoken conversation is hard. People may use different approaches to describe Hanja, such as

- meaning (훈) + reading (음),
- word-based (e.g., the character A in the word AB), and
- shape-based (e.g., component C on the left and component D on the right; or component E which looks like something).

But all of them are inefficient and inconvenient in a spoken conversation. Besides, even if we ignore the fact that these are inefficient and inconvenient, they cannot always work (note that the examples below are all from KS X 1001).

- A single combination of meaning and reading can correspond to more than one Hanja. For example, “클 석” can mean either 奭 (U+596D) or 碩 (U+78A9); “옥돌 민” can mean either 玟 (U+739F) or 珉 (U+73C9).
- There may not be well-known words. Sometimes there might be no word at all. For example, 贇 (U+8D07) is not used in any Korean word (only appears in personal names).
- There are characters whose shapes cannot easily be described, such as 寯 (U+5BE1), 隶 (U+79C9), and 肅 (U+8085).

Also, when a listener has no background knowledge on Hanja, or the domain name consists of an uncommon or ambiguous Hanja character, it will be much more difficult to reconstruct a Hanja IDN from a spoken conversation.

To make things even worse, Hanja IDNs will greatly inconvenience visually impaired users, as they need to be read aloud by a screen reader or presented in braille. Korean screen readers can read aloud Hanja inside plain text or in a Hanja candidate window of an input method application by only its reading or its meaning + reading, depending on the user preference[10]. If a Hanja IDN is read aloud just by its reading by a screen reader, it is not possible to distinguish whether the given IDN is using Hangul or Hanja without changing user preference. Even when both meaning and reading are read aloud by a screen reader, the ambiguity issue mentioned in the previous section still applies.

While there are two unofficial incompatible Hanja orthographies for Korean braille[11] (there is no “official” supervision of denoting Korean Hanja in braille according to the 2017 Korean braille orthography[12]), the report states that the demand for Hanja braille among visually impaired Koreans is low. As a reflection of this low demand, a Korean

9 https://support.apple.com/en-us/HT201794. 2017-06-06, Retrieved 2018-02-07.

10 https://www.freelists.org/post/nvda-korean/NVDA-1 NVDA용 한자 데이터 1차로 완성했습니다. (Finished an initial work on Hanja data for NVDA) – the email mentions that the NVDA data file contains multiple formats of read-aloud Hanja in Korean text. 2012-09-08, Retrieved 2018-02-07.

braille terminal named 한소네 (Hansone) developed by Selvas Healthcare only supports Hangul and Latin alphabet input without Hanja input capability[13]. Unlike Chinese and Japanese, Hangul-only IDNs require no processing of Hanja – which is inevitable in mixed-script or Hanja-only IDNs – on screen readers and braille displays. This Hanja processing problem can be avoided by not introducing Hanja into IDNs in the first place.

Information Exchange with Chinese and Japanese Speakers?

One possible argument for allowing Hanja in Korean IDNs is information exchange with Chinese and Japanese speakers. In fact, Hangul-Hanja mixed-script IDNs and IDNs with Korean Hanja will likely cause confusion to Chinese and Japanese speakers rather than helping them due to lexical and regional differences.

Let's take the city of Seoul as an example. Owing to its etymology, Seoul cannot be written in Korean Hanja. Due to this fact, Seoul has exonyms 汉城 (Simplified Chinese Hanzi) and ソウル (Japanese Katakana). Given the Chinese exonym caused confusion within Korean speakers, the city proposed 首尔 as a transliteration of the Korean pronunciation into Simplified Chinese in 2005. When encoding a domain name of Seoul City Hall, one hypothetical mixed-script IDN could be 서울市廳.한국. This actually causes confusion to Chinese and Japanese speakers: 1) they can neither understand nor enter Hangul '서울', '한국'; 2) Chinese and Japanese use different words for "city hall" – 市政府 and 市役所 respectively.

If you still don't grasp lexical differences, let's take a look at this Chinese word: 足球. If 足球 or its Korean reading (족구) is presented to Korean speakers, they would have no idea that this means "soccer," as 족구 (足球) in Korean is a domestic sport distinct from soccer ("soccer" in Korean is 축구 (蹴球)). You cannot blindly assume that anything written in Chinese characters will be understood with the same meaning regardless of languages. It is not always true.

If Korean speakers want to target Chinese and Japanese speakers, they need to register IDNs in Chinese (首尔市政府 and 足球) and Japanese (ソウル市役所 and サッカー), not the ones in Korean written in (Hangul and) Hanja (서울市廳 and 蹴球).

Chinese and Japanese speakers input non-Korean variants of Chinese characters. Even though Chinese characters are called CJK "Unified" Ideographs in Unicode and ISO/IEC 10646, there are lots of characters separately encoded due to the Source Separation Rule. The same abstract shape is sometimes unified and sometimes separately encoded. Let's take a look at these examples: 友情.한국 and 青年.한국 (these are commonly used words). Chinese and Japanese speakers will have no problems with inputting 友情, but will face a problem when inputting 青年. Why? This is because 忄青 and 忄青 are unified (情 U+60C5) but 青 and 青 are separately encoded (青 U+9751 vs. 青 U+9752) due to the Source Separation Rule (青 is used in Korean, and 青 is used in Chinese and Japanese; this also applies to 情). There are lots of examples like these (e.g., 僧 unified, 增/增 separated; 愉 unified, 兪/俞 separated; 慨 unified, 槪/概 separated; 朗 unified, 郎/郎 separated; etc.). Ordinary users are not aware of the Source Separation Rule and don't know which ones are unified and which ones are separated.

We believe that Korean Hanja domain alone is not helping information exchange with Chinese and Japanese speakers.

---

11 http://www.korean.go.kr/common/download.do?file_path=bookData&c_file_name=e5ab7da6-8b09-4f2f-9d6d-595d8e9ea1e6_0.pdf&o_file_name=%ED%8A%B9%EC%88%98%EC%96%B8%EC%96%B4-01-10.pdf&downGubun=bookDataView&book_seq=312 한자점자 규정 제정에 관한 기초연구 (A preliminary research on Korean Hanja braille orthography). 2012-12, Retrieved 2018-02-07.

12 http://www.korean.go.kr/common/download.do?file_path=etcData&c_file_name=5b0b6d52-0375-4fc8-98f0-912b26c0ee47_1.pdf&o_file_name=2017%20%ED%95%9C%EA%B5%AD%20%EC%A0%90%EC%9E%90%20%EA%B7%9C%EC%A0%95.pdf 2017 Korean braille orthography. Retrieved 2018-02-07.

13 http://www.himsintl.co.kr/download/hasone5_manual_170628.docx 한소네 5 사용자 설명서 (Hansone 5 user manual). Retrieved 2018-02-07.

Conclusion

We believe that the Proposal has to be withdrawn, given the technical and practical flaws. Although Hanja is used as a supplementary script in the Korean language, its usage does not simply fit into what IDNs need to encode. The Proposal misrepresents the usage of Hanja and why it should be coded into Korean IDNs. It also underestimates possible confusion between Hangul and Hanja and within Hanja, especially for Korean speakers. Mere addition of Hanja in Korean IDNs will introduce new accessibility issues, especially for visually impaired users. We believe that the benefit of using Hanja in Korean IDNs is much lower than the harm caused by it.