Title: Ordering rules for Hangul; CTT suggestion

Source: Kent Karlsson
Date: 2004-03-09
Status: Expert contribution
Document Type: Working group document, regarding 14651 amd 2
Action: For consideration by the UTC and JTC 1/SC 2

1 Introduction

"In the winter of our year 1443-4, our King [Seycong] originated and designed the twenty eight letters of the Correct Sounds. The letters are simple and fine and very easy to learn; their shifts and changes in function are endless; and there are no [Korean] sounds that cannot be written."

[Ceng Inci, in Hwunmin Cengum Haylyey, 1446; as translated in ‘The Korean Language’ by Ho-Min Sohn, Cambridge University Press, 1999; the additions in brackets are my clarifications]

The Hangul script is very elegantly designed. There are just a small number of letters (28, plus a small number of variant letters introduced later, but the latter have fallen out of use) and even a design philosophy for the shapes of the letters.

However, its incarnation in 10646/Unicode is far from elegant. This paper is about restoring the elegance of Hangul, as much as it can be restored, for the process of string ordering (collation). This results in an ordering of Hangul texts that is general, independent of various Hangul letter clusters that have been encoded. Basically, it’s back to basics for Hangul.

In summary, all Hangul Jamo characters are ordered as if decomposed into their basic Hangul Jamo letter sequences, taking into account the grouping into lead consonants (one or more), vowels (one or more), and trail consonants (zero or more) letters of each orthographic syllable. This way all representations available for a Hangul syllable are treated as equivalent. Hangul strings are, in the ordering given below, ordered in the currently established dictionary order.
1.1 Letter Hangul jamo characters

A letter Hangul jamo character represents a basic Hangul letter, or a variant of such a letter. There are 17 basic consonant letters, and 11 basic vowel letters. Some of the consonant letters have variants, that were invented after the invention of Hangul, for denoting sounds in Chinese. The variants, as well as some of the original basic letters have fallen out of use. In addition, there is a vowel filler character, for encoding consonants (or consonant clusters) as pseudo-syllables, and a consonant filler character for encoding vowels (or vowel clusters) as pseudo-syllables.

The letters for an orthographic syllable are grouped into a “syllable block”, typographically the size of a Hàn ideograph. In practice, there are at most (in total, however represented, see below) three consonant letters in a consonants cluster, and at most (in total, however represented) three vowel letters in a vowel cluster. Note that some vowel combinations look very much the same (like e.g. A-I and I-EO, EU-YO and YU-EU), but for each such same-looking pair of vowels apparently only one is allowed in Korean.

The encoding as characters for Hangul jamo employ a little coding trick to determine syllable boundaries: (most of) the consonants are encoded twice, leading and trailing. Other ways that could have been used would include (a) using a terminator/separator character (a similar approach is sometimes used for the Hangul compatibility letters), or (b) using combining characters for the Hangul letters following the first one in a syllable (this was the original Unicode design for Hangul, it is somewhat similar to the approach chosen).

A possible problem here is that the variant consonant letters are only allocated as choseong (leading), with no jongseong (trailing) counterpart. There may also be missing, as encoded characters, some historic variants of the letters, as well as a few historic punctuation marks for Hangul.

1.2 Basic composition of Hangul syllables

A Hangul syllable has the following syntax (disregarding precomposed Hangul syllable characters, but see below):

\[
\text{Hangul-syllable ::= } \ L^+ \ V^+ \ T^* \ M^* \]

where \( L \) is a leading consonant jamo letter, \( V \) is a vowel jamo letter, \( T \) is a trailing consonant jamo letter, and \( M \) is any combining mark, in particular a Hangul tone mark \([U+302E, U+302F]\). The tone mark (if any, or more generally, the sequence of combining characters) applies to the entire preceding syllable, not just the last part of it, since the Hangul syllable components, including the precomposed Hangul syllable characters (see below), are conjoining characters, not base characters. The tone mark glyphically appears at the left of a syllable, so for a \( L \ V \ T \ M \) syllable, where \( M \) is a Hangul tone mark, the glyph for \( M \) is to be rendered to the left of the (possibly dynamically composed) glyph for \( L \ V \ T \), not to the left of the (sub)glyph for \( T \).

An addition to the encoded repertoire are the filler jamos: choseong filler (Lf) and jungseong filler (Vf). They do not stand for any letter, but are used as a “placeholder” for a missing letter where one is formally required to form an orthographic syllable (note that there has to be at least one lead consonant and at least one vowel in a syllable according to the syntax above). The filler characters are a bit special, in that they are not letters at all, and if approach (b) as mentioned in section 1.1 above, a space, or no-break space, would have been used for Lf and the empty string would
have been used for \textit{Vf}. The filler characters should not occur in Hangul jamo letter clusters of two or more characters.

What has been presented so far is fully sufficient for representing any text in Hangul, historical (except for as yet missing historic variants), modern, and future (unless new letters are invented and gain use). What follows next here is several sections on additions that are strictly speaking unnecessary for the representation of Hangul texts, but have been added for various other reasons, such as compatibility with older standards. They generally introduce a number of difficulties for all processes handling Hangul. Still, they are commonly used, especially the precomposed Hangul syllable characters.

1.3 \textbf{Letter cluster Hangul jamo characters}

Letter cluster Hangul jamo characters represent either clusters of two or three consonants, or clusters of two or three vowels. Cluster jamo characters for most (not all) occurring consonant and vowel clusters are allocated. They work as \textit{L}, \textit{V}, and \textit{T} respectively in the syllable syntax above. One can preferably represent the sequence of consonants or sequence of vowels using single-letter Hangul jamo characters.

\textit{In the ordering rules below, we will assume that a vowel cluster jamo will only occur at the leftmost position in the vowels part of a Hangul syllable, and also assume that only “modern” letter clusters occur last (if at all) in the leading consonants part. Using these assumptions, decomposition of letter cluster jamos can be avoided; a decomposition that otherwise would be required unless very many (over 6000, instead of the around 350 as used below) contractions are defined.}

At one point the cluster jamos had compatibility decompositions into single-letter jamos. But now there is, unfortunately, no longer any formal decomposition of the cluster jamos into single-letter jamos. Not having such decompositions leads to multiple possible representations of the exact same piece of Hangul text, multiple representations that are not normalised via the Unicode normalisation forms. Ideally, as is done below, whenever possible the cluster Hangul jamos should be treated as if they had canonical decompositions into the corresponding sequence of single-letter Hangul Jamos.

Note that typographic features, such as cluster ligatures, variant (sub)glyph selection, and syllable layout should be handled by font mechanisms.

Ordering of Hangul syllables should be based on a weighting scheme that orders each cluster character as the sequences of its constituent letters. We will below additionally assume that vowel cluster jamos occur first in a sequence of vowel jamos within a Hangul syllable and that only single letter jamos or modern cluster jamos occur at the end of a sequence of lead consonant jamos. That implies that a cluster vowel jamo does not occur after an LV Hangul Syllable character, though single letter vowel jamos still can do so.

1.4 \textbf{Hangul compatibility letters}

The Hangul compatibility letters and half-width letters encode the consonants and some of the consonants clusters only once each (no separation into lead and trail). The Hangul compatibility letters are normally rendered as spacing characters without any conjoinment. In addition the compatibility Hangul letters have also FILLER characters, 3164 \textsc{Hangul Filler} and \textsc{FFA0 Halfwidth Hangul Filler}, that work differently from the jamo fillers. Note that the choseongness or jongseongness of
the compatibility mappings for the compatibility Hangul consonants are incorrect and irrelevance. (See below for a better treatment.) Let C be a (possibly half-width) Hangul consonant(s) letter, W a (possibly half-width) Hangul vowel(s) letter, and H is a (possibly half-width) FILLER character.

Each Hangul compatibility letter should be seen as a compatibility encoding of a pseudo-syllable, with a filler character. For consonant Hangul compatibility characters (C) they should be seen as lead consonant jamos followed by a vowel filler jamo (C → L + Vf) and for vowel Hangul compatibility characters (W) they should be seen as a consonant filler character followed by vowel jamos (W → Lf V+). That leaves HANGUL (HALFWIDTH) FILLER as rather useless characters (should be seen as Lf Vf). Note that the normal forms NFKD and NFKC do not do this conversion but returns a completely incorrect result for these characters.

Another, not recommended, way of handling the Hangul compatibility letters is to use them as they were intended in KS X 1001. When converting from Hangul compatibility letter sequences to proper conjoining Hangul letters, the Hangul compatibility syllables then have the syntax like the following (KS X 1001 does not really allow for consonant and vowel sequences, nor for mixing fullwidth and halfwidth):

\[
\text{Hangul-compatibility-syllable ::= } H (C+|H) (W+|H) (C+|H) M^* \\
\]

Note again that the normal forms NFKD and NFKC do not do this conversion but returns a completely incorrect result for these characters.

However, compatibility Hangul characters are not expected to display as conjoined Hangul syllables, but display as free-standing. Similarly, circled and parenthesised Hangul compatibility characters are freestanding. All of them are treated as free-standing, ignoring KS X 1001, in the ordering described below.

As a historic side-note, there have been experiments with writing Hangul “linearly”. The Hangul compatibility letters can be used for representing such texts.

### 1.5 Hangul syllable characters

A lot of Hangul syllables have a character of their own character in the range AC00–D7A3. They each have an arithmetic canonical decomposition into other conjoining Hangul characters.

The arithmetically specified decompositions for precomposed Hangul syllable characters are best described as follows:

Each Hangul precomposed syllable character of \text{Hangul Sylable Type LV} has a canonical decomposition mapping into \text{L} and \text{V} Hangul jamos:

<table>
<thead>
<tr>
<th>LV</th>
<th>L in 1100–1112</th>
<th>V in 1161–1175</th>
</tr>
</thead>
<tbody>
<tr>
<td>s</td>
<td>LBase + (((s – SBase) div NCount) * NCount)</td>
<td>VBase + (((s – SBase) mod NCount) div TCountP1)</td>
</tr>
</tbody>
</table>

Each Hangul precomposed syllable character of \text{Hangul Sylable Type LVT} has a canonical decomposition mapping into a \text{LV} Hangul syllable character and a \text{T} Hangul jamo:

<table>
<thead>
<tr>
<th>LVT</th>
<th>LV</th>
<th>T in 11A8–11C2</th>
</tr>
</thead>
<tbody>
<tr>
<td>s</td>
<td>SBase + (((s – SBase) div NCount) * NCount)</td>
<td>TBaseM1 + (((s – SBase) mod TCountP1)</td>
</tr>
</tbody>
</table>

Note: This description is slightly different from that in The Unicode Standard 3.0 and 4.0, but the net result is the same. There was separate proposal for updating the
formal decompositions to the ones above (and where the constants are defined). That proposal has been accepted by the Unicode Technical Committee, and the decompositions above will be incorporated into the Unicode standard.

The Hangul syllable characters alone can represent most modern Hangul words (and all in the official orthography). They cannot represent historic Hangul words (Middle Korean), nor modern/future Hangul words using syllables not preallocated. However, all Hangul words can elegantly be represented by sequences of single-letter Hangul Jamo characters plus optional tone mark (historic). (This is with an exception for still missing historic Hangul letter variants and historic Hangul punctuation.)

1.6 Full rule for composition of Hangul syllables

A Hangul syllable, allowing for precomposed syllable characters) has the following syntax (see page 53 of The Unicode Standard version 3.0, with adjustment for tone marks):

```
Hangul-ext-syllable ::=    L+  V+  T*  M*    |    L*  LV  V*  T*  M*    |    L*  LVT  T*  M*
```

where LV is a precomposed consonants-vowels syllable character (Hangul_syllable_type LV), and LVT is a precomposed consonants-vowels-consonants syllable character (Hangul_syllable_type LVT). Note that we will here assume that only the first V in the V+ sequence can be a vowel letter cluster jamo, and that all other vowel jamo occurrences are single letter vowel jamos.

1.7 Circled and parenthesised Hangul letters and syllables

All of the parenthesised or circled Hangul characters should be treated as compatibility characters with a compatibility mappings to a Hangul syllable (and parentheses where applicable), not to individual Hangul jamo letters. This holds even for the single letter characters of this kind; a filler jamo should be considered to be part of the (collation) decomposition in these cases.

2 Suggestions for the Hangul part of ISO/IEC 14651

Current (ISO/IEC 14651:2001) ordering for Hangul handles ‘modern’ Hangul well, provided that the text is represented in such a way that a syllable is is represented with just a single precomposed Hangul syllable, or always composed of a ‘modern’ leading, ‘modern’ vowel, and optionally ‘modern’ trailing Hangul Jamo, where the Jamos may be cluster Jamos. It does not handle ‘historic’ Hangul Jamo characters well, nor does it handle well Hangul syllables where the consonant or vowel clusters are composed from multiple single Hangul Jamo letters according to the general syntax above.

This proposal is intended to remedy this, by handling Hangul in a way that is similar to how other alphabetic scripts are handled. The result is that "historic" Hangul characters (and compositions for those) are ordered among the "modern" Hangul letters as expected, and that compositions from single-letter Jamos are ordered as expected. In order to keep ISO/IEC 14651 and UTS 10 in synchrony, corresponding changes to UTS 10 are also suggested.

A difficulty is that the ordering of Hangul is cluster based. One way to deal with this is to insert "low" weights after each sequence of lead consonant jamos, after each sequence of vowel jamos, and after each sequence of trail consonant jamos. This method is rather expensive though, both in terms of processing for computing the keys
for a string, as well as in the (nominal) length of the keys (adding two occurrences of the light weight for each Hangul syllable). Though the latter can be compressed, that compression adds to the processing time for computing the keys. The proposal here is much more direct, and produces one weight (per level) per letter in a Hangul string.

2.1 Ordering Hangul strings

The syllable and cluster based ordering that is used for Hangul has been done by assigning the weights as implied by the desired order as follows (X is any independent character that cannot be part of a Hangul syllable):

- \(L_1V < L_1LV\) implies that (initial V) < L.
- \(L_1V_1L < L_1V_1T\) implies that L < T.
- \(LVX < LVT\) implies that X < T.
- \(L_1V_1T < L_1V_1V\) implies that T < (non-initial V).
- Basic Hangul letters, within each group as above, are by default ordered in modern order.

In summary: (initial V) < L, and (X or L) < T < (non-initial V).

Note, however, that if non-initial Ls are weighted as initial Ls (which is convenient), then by implication, \(LL_1...\) incorrectly comes before \(LT...\) But the latter (LT) is an improperly constructed Hangul syllable (there is no V, not even a filler). Contraction could be added to handle these cases. The correctly constructed \(LVfT\), with a filler vowel, gets ordered before \(LL_1...\), as expected. Ordering of incorrectly constructed Hangul syllables are not prioritised, so initial and non-initial Ls will be ordered the same, and there are no contractions below to fix this.

The true problem here is that (initial V) < (non-initial V). In properly constructed Hangul syllables, the syllable initial V is always directly after an L. So we can use contractions between L and V, which are weighted so that the initial V gets a weight lighter than any L. Non-initial V occurrences are weighted after all Ts, which in turn are weighted after all independent characters.

Now we require contractions between each L and each V. That is 91 Ls times 67 Vs. Which would be 6097 contractions. That is a bit much... Notice first that the fillers should normally be ignored, except a lead consonant filler followed by a non-filler vowel. That means 91 fewer contractions (for this), i.e. 6006 contractions left. Still very many... Just considering the basic letters and variants, excluding letter cluster jamos, we get 20 times 11 (= 220) contractions. This would require that all cluster letter jamos are decomposed into their constituent letter jamos. These decompositions are unfortunately not part of the Unicode character database. If one does not want to do these decompositions, but still have a reasonable number of contractions, the following is a possible compromise: add contractions for modern lead consonant clusters (which result from arithmetically decomposing precomposed Hangul jamo syllables) and basic vowel jamos. Vowel letter cluster jamos are weighted as the first constituent letter get a [initial V] weight, and the rest get a [non-initial V] weight. This assumes that vowel letter cluster jamos never occur as non-first vowel in a Hangul syllable, and that non-modern lead consonant clusters always end with a non-cluster Hangul jamo letter (though vowel letter cluster jamos are already handled anyway). It would still be best to never use historic letter cluster jamos, but instead use sequences on non-composite jamos, not only for historic syllables, but also modern ones.
## 2.2 Suggested new Hangul excerpt for the 14651 CTT

%%% BEGIN SUGGESTED HANGUL EXCERPTS FOR 14651 CTT

%%% Arithmetic decomposition of Hangul Syllable characters into Hangul Jamos must be
%%% done as a prehandling. Alternatively, the weighings of the Hangul syllable characters
%%% can be precomputed, in which case no prehandling is needed for all of modern and known
%%% historic Hangul text strings.

%%% Full decomposition of cluster Hangul Jamos into single-letter Jamos must also be
%%% done as a prehandling if unusual combinations of Hangul letters are used. Alternatively,
%%% a selection of additional contractions can be added.

%%% Not all contractions needed to cover all possible cases are included.
%%% This is done in order to keep the number of collating elements (contractions) down
%%% to around 350 rather than more than 6000. However, vowel cluster Jamo characters are
%%% weighted with the assumption that they are always the first vowel Jamo in a syllable.
%%% This way they do not need to have any contractions for first vowel letter in a syllable.

% Declaration of collating symbols for Hangul Jamo single-letters (similar to how this
% is done for other alphabetic scripts) (order here is arbitrary):

%%% Collating symbols for the Hangul vowels, first in syllable. We will use the ones for the 11
%%% basic Hangul vowels, plus the vowel filler.
collating-symbol <V1160>..<V119E> % HANGUL JUNGSEONG FILLER..HANGUL JUNGSEONG ARAEA

%%% Collating symbols for the lead Hangul consonants. We will use the ones for the 17 basic Hangul
%%% consonants, plus the lead consonant filler.
collating-symbol <S1100>..<S115F> % HANGUL CHOSEONG KIYEOK..HANGUL CHOSEONG FILLER

%%% Collating symbols for the trail Hangul consonants. We will use the ones for the 17 basic
%%% Hangul consonants.
collating-symbol <S11A8>..<S11F9> % HANGUL JONGSEONG KIYEOK..HANGUL JONGSEONG YEORINHIEUH

%%% Collating symbols for the Hangul vowels, not first in syllable. We will use the ones for the
%%% 11 basic Hangul vowels.
collating-symbol <S1161>..<S119E> % HANGUL JUNGSEONG A.. HANGUL JUNGSEONG ARAEA

%%% Declaration of contractions:
collating-element <U1100_1160> from "<U1100><U1160>" % HANGUL CHOSEONG KIYEOK, HANGUL JUNGSEONG
%%% FILLER
collating-element <U1100_1161> from "<U1100><U1161>" % HANGUL CHOSEONG KIYEOK, HANGUL JUNGSEONG A
%%% collating-element <U1100_1163> from "<U1100><U1163>" % HANGUL CHOSEONG KIYEOK, HANGUL JUNGSEONG
%%% YA
%%% collating-element <U1100_1165> from "<U1100><U1165>" % HANGUL CHOSEONG KIYEOK, HANGUL JUNGSEONG
%%% EO
%%% collating-element <U1100_1167> from "<U1100><U1167>" % HANGUL CHOSEONG KIYEOK, HANGUL JUNGSEONG
%%% YEO
%%% collating-element <U1100_1169> from "<U1100><U1169>" % HANGUL CHOSEONG KIYEOK, HANGUL JUNGSEONG
%%% O
%%% collating-element <U1100_116D> from "<U1100><U116D>" % HANGUL CHOSEONG KIYEOK, HANGUL JUNGSEONG
%%% YO
%%% collating-element <U1100_116F> from "<U1100><U116F>" % HANGUL CHOSEONG KIYEOK, HANGUL JUNGSEONG
%%% YE
%%% collating-element <U1100_1172> from "<U1100><U1172>" % HANGUL CHOSEONG KIYEOK, HANGUL JUNGSEONG
%%% U
%%% collating-element <U1100_1173> from "<U1100><U1173>" % HANGUL CHOSEONG KIYEOK, HANGUL JUNGSEONG
%%% EU
%%% collating-element <U1100_1175> from "<U1100><U1175>" % HANGUL CHOSEONG KIYEOK, HANGUL JUNGSEONG
%%% I
collating-element <U1100_119E> from "<U1100><U119E>" % HANGUL CHOSEONG KIYEOK, HANGUL JUNGSEONG ARAEA

collating-element <U1101_1160> from "<U1101><U1160>" % HANGUL CHOSEONG SSANGKIYEOK, HANGUL JUNGSEONG FILLER
collating-element <U1101_1161> from "<U1101><U1161>" % HANGUL CHOSEONG SSANGKIYEOK, HANGUL JUNGSEONG A
collating-element <U1101_1163> from "<U1101><U1163>" % HANGUL CHOSEONG SSANGKIYEOK, HANGUL JUNGSEONG YA
collating-element <U1101_1165> from "<U1101><U1165>" % HANGUL CHOSEONG SSANGKIYEOK, HANGUL JUNGSEONG EO
collating-element <U1101_1167> from "<U1101><U1167>" % HANGUL CHOSEONG SSANGKIYEOK, HANGUL JUNGSEONG YEO
collating-element <U1101_1169> from "<U1101><U1169>" % HANGUL CHOSEONG SSANGKIYEOK, HANGUL JUNGSEONG O
collating-element <U1101_116D> from "<U1101><U116D>" % HANGUL CHOSEONG SSANGKIYEOK, HANGUL JUNGSEONG YO
collating-element <U1101_116E> from "<U1101><U116E>" % HANGUL CHOSEONG SSANGKIYEOK, HANGUL JUNGSEONG U
collating-element <U1101_1172> from "<U1101><U1172>" % HANGUL CHOSEONG SSANGKIYEOK, HANGUL JUNGSEONG YU
collating-element <U1101_1173> from "<U1101><U1173>" % HANGUL CHOSEONG SSANGKIYEOK, HANGUL JUNGSEONG EU
collating-element <U1101_1175> from "<U1101><U1175>" % HANGUL CHOSEONG SSANGKIYEOK, HANGUL JUNGSEONG I
collating-element <U1101_119E> from "<U1101><U119E>" % HANGUL CHOSEONG SSANGKIYEOK, HANGUL JUNGSEONG ARAEA

collating-element <U1102_1160> from "<U1102><U1160>" % HANGUL CHOSEONG NIEUN, HANGUL JUNGSEONG FILLER
collating-element <U1102_1161> from "<U1102><U1161>" % HANGUL CHOSEONG NIEUN, HANGUL JUNGSEONG A
collating-element <U1102_1163> from "<U1102><U1163>" % HANGUL CHOSEONG NIEUN, HANGUL JUNGSEONG YA
collating-element <U1102_1165> from "<U1102><U1165>" % HANGUL CHOSEONG NIEUN, HANGUL JUNGSEONG EO
collating-element <U1102_1167> from "<U1102><U1167>" % HANGUL CHOSEONG NIEUN, HANGUL JUNGSEONG YEO
collating-element <U1102_1169> from "<U1102><U1169>" % HANGUL CHOSEONG NIEUN, HANGUL JUNGSEONG O
collating-element <U1102_116D> from "<U1102><U116D>" % HANGUL CHOSEONG NIEUN, HANGUL JUNGSEONG YO
collating-element <U1102_116E> from "<U1102><U116E>" % HANGUL CHOSEONG NIEUN, HANGUL JUNGSEONG U
collating-element <U1102_1172> from "<U1102><U1172>" % HANGUL CHOSEONG NIEUN, HANGUL JUNGSEONG YU
collating-element <U1102_1173> from "<U1102><U1173>" % HANGUL CHOSEONG NIEUN, HANGUL JUNGSEONG EU
collating-element <U1102_1175> from "<U1102><U1175>" % HANGUL CHOSEONG NIEUN, HANGUL JUNGSEONG I
collating-element <U1102_119E> from "<U1102><U119E>" % HANGUL CHOSEONG NIEUN, HANGUL JUNGSEONG ARAEA

collating-element <U1103_1160> from "<U1103><U1160>" % HANGUL CHOSEONG TIKEUT, HANGUL JUNGSEONG FILLER
collating-element <U1103_1161> from "<U1103><U1161>" % HANGUL CHOSEONG TIKEUT, HANGUL JUNGSEONG A
collating-element <U1103_1163> from "<U1103><U1163>" % HANGUL CHOSEONG TIKEUT, HANGUL JUNGSEONG YA
collating-element <U1103_1165> from "<U1103><U1165>" % HANGUL CHOSEONG TIKEUT, HANGUL JUNGSEONG EO
collating-element <U1103_1167> from "<U1103><U1167>" % HANGUL CHOSEONG TIKEUT, HANGUL JUNGSEONG YEO
collating-element <U1103_1169> from "<U1103><U1169>" % HANGUL CHOSEONG TIKEUT, HANGUL JUNGSEONG O
collating-element <U1103_116D> from "<U1103><U116D>" % HANGUL CHOSEONG TIKEUT, HANGUL JUNGSEONG YO
collating-element <U1103_116E> from "<U1103><U116E>" % HANGUL CHOSEONG TIKEUT, HANGUL JUNGSEONG U
collating-element <U1103_1172> from "<U1103><U1172>" % HANGUL CHOSEONG TIKEUT, HANGUL JUNGSEONG YU
collating-element <U1103_1173> from "<U1103><U1173>" % HANGUL CHOSEONG TIKEUT, HANGUL JUNGSEONG EU
collating-element <U1103_1175> from "<U1103><U1175>" % HANGUL CHOSEONG TIKEUT, HANGUL JUNGSEONG I
collating-element <U1103_119E> from "<U1103><U119E>" % HANGUL CHOSEONG TIKEUT, HANGUL JUNGSEONG ARAEA

collating-element <U1104_1160> from "<U1104><U1160>" % HANGUL CHOSEONG SSANGTIKEUT, HANGUL JUNGSEONG FILLER
collating-element <U1104_1161> from "<U1104><U1161>" % HANGUL CHOSEONG SSANGTIKEUT, HANGUL JUNGSEONG A
collating-element <U1104_1163> from "<U1104><U1163>" % HANGUL CHOSEONG SSANGTIKEUT, HANGUL JUNGSEONG YA
% % Scripts before Hangul here %%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
The 17 basic Hangul consonants, lead copy, modern order, plus lead consonant filler:

<1100> ㄱ, HANGUL CHOSEONG KIYEOK
<1102> ㄴ, HANGUL CHOSEONG NIEUN
<1103> ㄷ, HANGUL CHOSEONG TIKEUT
<1105> ㄹ, HANGUL CHOSEONG RIEUL
<1106> ㅁ, HANGUL CHOSEONG MIEUM
<1107> ㅂ, HANGUL CHOSEONG PIEUP
<1109> ㅅ, HANGUL CHOSEONG SIOS
<1110> ㅇ, HANGUL CHOSEONG IEUNG
<1114> ㅈ, HANGUL CHOSEONG SIOS
<110B> ㅊ, HANGUL CHOSEONG YEORINHIEUH
<110E> ㅋ, HANGUL CHOSEONG IEUNG
<110F> ㅌ, HANGUL CHOSEONG YEORINHIEUH
<1110> ㅍ, HANGUL CHOSEONG YEORINHIEUH
<1112> ㅎ, HANGUL CHOSEONG YEORINHIEUH
<1118> ㅏ, HANGUL CHOSEONG KIYEOK
<111A> ㅑ, HANGUL CHOSEONG NIEUN
<111C> ㅓ, HANGUL CHOSEONG TIKEUT
<111E> ㅕ, HANGUL CHOSEONG RIEUL
<1120> ㅗ, HANGUL CHOSEONG MIEUM
<1122> ㅛ, HANGUL CHOSEONG PIEUP
<1124> ㅜ, HANGUL CHOSEONG SIOS
<1126> ㅠ, HANGUL CHOSEONG IEUNG
<1128> ㅡ, HANGUL CHOSEONG YEORINHIEUH
<112A> ㅣ, HANGUL CHOSEONG YEORINHIEUH

Various scripts after Hangul here

The 17 basic Hangul consonants, trail copy, modern order:

<11A8> ㄱ, HANGUL CHOSEONG KIYEOK
<11AB> ㄴ, HANGUL CHOSEONG NIEUN
<11AE> ㄷ, HANGUL CHOSEONG TIKEUT
<11AF> ㄹ, HANGUL CHOSEONG RIEUL
<11B7> ㅁ, HANGUL CHOSEONG MIEUM
<11B8> ㅂ, HANGUL CHOSEONG PIEUP
<11BA> ㅅ, HANGUL CHOSEONG SIOS
<11EB> ㅈ, HANGUL CHOSEONG YEORINHIEUH
<11EC> ㅊ, HANGUL CHOSEONG IEUNG
<11F0> ㅋ, HANGUL CHOSEONG YEORINHIEUH
<11FD> ㅌ, HANGUL CHOSEONG YEORINHIEUH
The 11 basic Hangul vowels, not first vowel in syllable, modern order:

- ㅏ, HANGUL JUNGSEONG A
- ㅑ, HANGUL JUNGSEONG YA
- ㅓ, HANGUL JUNGSEONG EO
- ㅕ, HANGUL JUNGSEONG YEO
- ㅗ, HANGUL JUNGSEONG O
- ㅛ, HANGUL JUNGSEONG YO
- ㅜ, HANGUL JUNGSEONG U
- ㅠ, HANGUL JUNGSEONG YU
- ㅡ, HANGUL JUNGSEONG EU
- ㅣ, HANGUL JUNGSEONG I
- ㆍ, HANGUL JUNGSEONG ARAEA

Dependent vowels for Indic scripts here. (Including Khmer, Tibetan, and maybe the vowels for Thai, Lao, and Polynesian scripts too)

Viramas (and silent dependent vowels) here.

% Weighting table for Hangul, except Hangul syllables (which are collated by their NFD form).
% The order here is arbitrary (except for the fourth level weight, which is unimportant).

% Hangul tone marks (handled as accents):
% % Hangul single dot tone mark (applies to entire syllable)
% % Hangul double dot tone mark (applies to entire syllable)
The 17 Hangul consonants, plus variants, clusters, and compatibility forms. The Hangul vowel compatibility characters are included here too, since they are seen as having a lead consonant filler.
<U1109> "<S1109><V1160>";"<BASE><BASE>;"<COMPAT><COMPAT>;"<U1109> % HANGUL LETTER PUEP-CIEUC
<U1128> "<S1109><V1160>";"<BASE><BASE>;"<MIN><MIN>;"<U1128> % HANGUL CHOSEONG PUEP-CIEUC
<U1129> "<S1109><V1110>";"<BASE><BASE>;"<MIN><MIN>;"<U1129> % HANGUL CHOSEONG PUEP-THIEUTH
<U1177> "<S1109><V1110>";"<BASE><BASE>;"<COMPAT><COMPAT>;"<U1177> % HANGUL LETTER PUEP-THIEUTH
<U112A> "<S1109><V1111>";"<BASE><BASE>;"<MIN><MIN>;"<U112A> % HANGUL CHOSEONG PIEUP-PHEUPH

<U1109> "<S1109><V1160>";"<BASE><MIN>;"<U1109> % HANGUL CHOSEONG SIOS (SIEUS)
<U113C> "<S1109><V1160>";"<BASE><VRNT1>;"<MIN><MIN>;"<U113C> % HANGUL CHOSEONG CHITUEUMSIOS (IMORISORISIEUS)
<U113E> "<S1109><V1160>";"<BASE><VRNT2>;"<MIN><MIN>;"<U113E> % HANGUL CHOSEONG CHITUEUMSIOS (IMCSRISORISIEUS)

<U1109_1160> "<S1109><V1160>";"<BASE><BASE>;"<COMPAT><COMPAT>;"<U1109_1160> % HANGUL CHOSEONG SIOS, HANGUL JUNGSEONG FILLER
<U113C_1160> "<S1109><V1160>";"<BASE><VRNT1><BASE>;"<MIN><MIN>;"<U113C_1160> % HANGUL CHOSEONG SIOS, HANGUL JUNGSEONG FILLER

<U1109_1161> "<S1109><V1161>";"<BASE><BASE>;"<MIN><MIN>;"<U1109_1161> % HANGUL CHOSEONG SIOS, HANGUL JUNGSEONG A
<U113C_1161> "<S1109><V1161>";"<BASE><VRNT1>;"<MIN><MIN>;"<U113C_1161> % HANGUL CHOSEONG SIOS, HANGUL JUNGSEONG A

<U1109_1163> "<S1109><V1163>";"<BASE><BASE>;"<MIN><MIN>;"<U1109_1163> % HANGUL CHOSEONG SIOS, HANGUL JUNGSEONG YA
<U113C_1163> "<S1109><V1163>";"<BASE><VRNT1>;"<MIN><MIN>;"<U113C_1163> % HANGUL CHOSEONG SIOS, HANGUL JUNGSEONG YA

<U1109_1165> "<S1109><V1165>";"<BASE><BASE>;"<MIN><MIN>;"<U1109_1165> % HANGUL CHOSEONG SIOS, HANGUL JUNGSEONG EO
<U113C_1165> "<S1109><V1165>";"<BASE><VRNT1>;"<MIN><MIN>;"<U113C_1165> % HANGUL CHOSEONG SIOS, HANGUL JUNGSEONG EO

<U1109_1167> "<S1109><V1167>";"<BASE><BASE>;"<MIN><MIN>;"<U1109_1167> % HANGUL CHOSEONG SIOS, HANGUL JUNGSEONG YEO
<U113C_1167> "<S1109><V1167>";"<BASE><VRNT1>;"<MIN><MIN>;"<U113C_1167> % HANGUL CHOSEONG SIOS, HANGUL JUNGSEONG YEO
% HANGUL CHOSEONG PHIEUPH, HANGUL JUNGSEONG YA
% HANGUL CHOSEONG PHIEUPH, HANGUL JUNGSEONG EO
% HANGUL CHOSEONG PHIEUPH, HANGUL JUNGSEONG YEO
% HANGUL CHOSEONG PHIEUPH, HANGUL JUNGSEONG A
% HANGUL CHOSEONG PHIEUPH, HANGUL JUNGSEONG YO
% HANGUL CHOSEONG PHIEUPH, HANGUL JUNGSEONG U
% HANGUL CHOSEONG PHIEUPH, HANGUL JUNGSEONG EU
% HANGUL CHOSEONG PHIEUPH, HANGUL JUNGSEONG I
% HANGUL CHOSEONG PHIEUPH, HANGUL JUNGSEONG ARAEA
% HANGUL CHOSEONG PHIEUPH, HANGUL JUNGSEONG KAPYEOUNPHIEUPH (PHIEUPH-IEUNG)
% HANGUL LETTER KAPYEOUNPHIEUPH
% HANGUL CHOSEONG HIEUH, HANGUL JUNGSEONG FILLER
% PARENTHESEIZED HANGUL LETTER HIEUH
% CIRCLED HANGUL HIEUH
% HALFWIDTH HANGUL LETTER HIEUH
% HANGUL JUNGSEONG YA
% HANGUL JUNGSEONG EO
% HANGUL JUNGSEONG YEO
% HANGUL JUNGSEONG A
% HANGUL JUNGSEONG YO
% HANGUL JUNGSEONG U
% HANGUL JUNGSEONG EU
% HANGUL JUNGSEONG I
% HANGUL JUNGSEONG ARAEA
% HANGUL LETTER SSANGHIEU
% PARENTHESEIZED HANGUL LETTER HIEUH A
% CIRCLED HANGUL HIEUH A
% HANGUL JUNGSEONG ARAEA
% HANGUL CHOSEONG YEORINPHIEUPH
% HANGUL LETTER SSANGHIEU
% HANGUL CHOSEONG YEO
% HANGUL JUNGSEONG FILLER
% HANGUL LETTER YEORINPHIEUPH
% HANGUL CHOSEONG YEO
% HANGUL JUNGSEONG ARAEA
%% <UAC1C> "<S1100><V1161><S1175>";"<BASE><BASE><BASE>";"<MIN><MIN><MIN>"; <UAC1C> % HANGUL SYLLABLE GAE

%%% ...

%%% <UD7A3> "<S1112><V1175><S11C2>";"<BASE><BASE><BASE>;"<MIN><MIN><MIN>";<UD7A3> % HANGUL SYLLABLE HIH

##### SCRIPTS AFTER HANGUL HERE ######

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

%%%% DEPENDENTS FOLLOW %%%%%

%%

Hangul trail consonants.

<U11A8> "<S11A8><BASE><MIN><U11A8> % HANGUL JONGSEONG KIYEOK (KIEUK)
  <U11A9> "<S11A9><S11A9><BASE><MIN><U11A9> % HANGUL JONGSEONG SSANGKIYEOK
  <U11C3> "<S11A8><S11A8><BASE><MIN><U11C3> % HANGUL JONGSEONG KIYEOK-RIEUL
  <U11AA> "<S11A8><S11BA><BASE><MIN><U11AA> % HANGUL JONGSEONG KIYEOK-SIOS
  <U11C4> "<S11A8><S11BA><S11A8><BASE><MIN><MIN><MIN><U11C4> % HANGUL JONGSEONG KIYEOK-SIOS-KIYEOK
  <U11AB> "<S11AB><BASE><MIN><U11AB> % HANGUL JONGSEONG NIEUN
  <U11C5> "<S11AB><S11A8><BASE><MIN><MIN><MIN><U11C5> % HANGUL JONGSEONG NIEUN-KIYEOK
  <U11C6> "<S11AB><S11AE><BASE><MIN><MIN><MIN><U11C6> % HANGUL JONGSEONG NIEUN-TIKEUT
  <U11C7> "<S11AB><S11BA><BASE><MIN><MIN><MIN><U11C7> % HANGUL JONGSEONG NIEUN-SIOS
  <U11C8> "<S11AB><S11EB><BASE><MIN><MIN><MIN><U11C8> % HANGUL JONGSEONG NIEUN-PANSIOS
  <U11AC> "<S11AB><S11BD><BASE><MIN><MIN><MIN><U11AC> % HANGUL JONGSEONG NIEUN-CIEUC
  <U11AE> "<S11AB><S11C2><BASE><MIN><MIN><MIN><U11AE> % HANGUL JONGSEONG NIEUN-HIEUH
  <U11CA> "<S11AE><S11A8><BASE><MIN><MIN><MIN><U11CA> % HANGUL JONGSEONG TIKEUT-KIYEOK
  <U11CB> "<S11AE><S11AF><BASE><MIN><MIN><MIN><U11CB> % HANGUL JONGSEONG TIKEUT-RIEUL
  <U11AF> "<S11AF><BASE><MIN><U11AF> % HANGUL JONGSEONG RIEUL
  <U11B0> "<S11AF><S11A8><BASE><MIN><MIN><MIN><U11B0> % HANGUL JONGSEONG RIEUL-KIYEOK
  <U11B2> "<S1105><S1107><BASE><MIN><MIN><MIN><U11B2> % HANGUL JONGSEONG RIEUL-PIEUP
  <U11CC> "<S11A8><S11BA><BASE><MIN><MIN><MIN><MIN><U11CC> % HANGUL JONGSEONG RIEUL-KIYEOK-SIOS
  <U11CD> "<S11A8><S11AB><BASE><MIN><MIN><MIN><MIN><MIN><U11CD> % HANGUL JONGSEONG RIEUL-NIEUN
  <U11CE> "<S11AF><S11AE><BASE><MIN><MIN><MIN><MIN><MIN><MIN><U11CE> % HANGUL JONGSEONG RIEUL-TIKEUT
  <U11CF> "<S11AF><S11AC><BASE><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><U11CF> % HANGUL JONGSEONG RIEUL-KIYEOK-NIEUN
  <U11D0> "<S11AF><S11AB><BASE><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN><MIN>
<U11D7> "<S11AF><S11EB>";"<BASE><BASE>";"<MIN><MIN>"; % HANGUL JONGSEONG RIEUL-PANSIOS
<U11D8> "<S11AF><S11BF>";"<BASE><BASE>";"<MIN><MIN>"; % HANGUL JONGSEONG RIEUL-KHIEUKH
<U11B4> "<S11AF><S11C0>";"<BASE><BASE>";"<MIN><MIN>"; % HANGUL JONGSEONG RIEUL-THIEUTH
<U11B5> "<S11AF><S11C1>";"<BASE><BASE>";"<MIN><MIN>"; % HANGUL JONGSEONG RIEUL-PHIEUPH
<U11B6> "<S11AF><S11C2>";"<BASE><BASE>";"<MIN><MIN>"; % HANGUL JONGSEONG RIEUL-HIEUH
<U11D9> "<S11AF><S11F9>";"<BASE><BASE>";"<MIN><MIN>"; % HANGUL JONGSEONG RIEUL-YEORINHIEUH

<U11B7> "<S11AF><S11EB>";"<BASE><BASE>";"<MIN><MIN>"; % HANGUL JONGSEONG MIEUM
<U11DA> "<S11B7><S11A8>";"<BASE><BASE>";"<MIN><MIN>"; % HANGUL JONGSEONG MIEUM-KIYEOK
<U11DB> "<S11B7><S11AF>";"<BASE><BASE>";"<MIN><MIN>"; % HANGUL JONGSEONG MIEUM-RIEUL
<U11DC> "<S11B7><S11B8>";"<BASE><BASE>";"<MIN><MIN>"; % HANGUL JONGSEONG MIEUM-PIEUP
<U11DD> "<S11B7><S11BA>";"<BASE><BASE>";"<MIN><MIN>"; % HANGUL JONGSEONG MIEUM-SIOS
<U11DE> "<S11B7><S11BA><S11BA>";"<BASE><BASE><BASE>";"<MIN><MIN><MIN>"; % HANGUL JONGSEONG MIEUM-SSANGSIOS
<U11DF> "<S11B7><S11EB>";"<BASE><BASE>";"<MIN><MIN>"; % HANGUL JONGSEONG MIEUM-PANSIOS

<U11E0> "<S11B7><S11BE>";"<BASE><BASE>";"<MIN><MIN>"; % HANGUL JONGSEONG PIEUPH
<U11E1> "<S11B7><S11C1>";"<BASE><BASE>";"<MIN><MIN>"; % HANGUL JONGSEONG PIEUP-HIEUH
<U11E3> "<S11B8><S11AF>";"<BASE><BASE>";"<MIN><MIN>"; % HANGUL JONGSEONG PIEUP-RIEUL
<U11E6> "<S11B8><S11BA>";"<BASE><BASE>";"<MIN><MIN>"; % HANGUL JONGSEONG PIEUP-SIOS
<U11E6> "<S11B8><S11BC>";"<BASE><BASE>";"<MIN><MIN>"; % HANGUL JONGSEONG KAPYEOUNPIEUP (PIEUPH-IEUNG)

<U11E4> "<S11B8><S11C1>";"<BASE><BASE>";"<MIN><MIN>"; % HANGUL JONGSEONG PIEUP-PHIEUPH
<U11E5> "<S11B8><S11C2>";"<BASE><BASE>";"<MIN><MIN>"; % HANGUL JONGSEONG PIEUP-HIEUH

<U11E7> "<S11B9><S11AC>";"<BASE><BASE>";"<MIN><MIN>"; % HANGUL JONGSEONG SIEUS
<U11E8> "<S11B9><S11AD>";"<BASE><BASE>";"<MIN><MIN>"; % HANGUL JONGSEONG SIOS-SIOS
<U11E9> "<S11B9><S11AF>";"<BASE><BASE>";"<MIN><MIN>"; % HANGUL JONGSEONG SIOS-RIEUL
<U11EA> "<S11B9><S11B8>";"<BASE><BASE>";"<MIN><MIN>"; % HANGUL JONGSEONG SIOS-PIEUP
<U11EB> "<S11B9><S11BA>";"<BASE><BASE>";"<MIN><MIN>"; % HANGUL JONGSEONG SIOS-SIOS

<U11E0> "<S11B7><S11BE>";"<BASE><BASE>";"<MIN><MIN>"; % HANGUL JONGSEONG PANISIOS (PANISORI)

<U11B8> "<S11AF><S11EB>";"<BASE><BASE>";"<MIN><MIN>"; % HANGUL JONGSEONG IEUNG
<U11E7> "<S11B7><S11AC>";"<BASE><BASE>";"<MIN><MIN>"; % HANGUL JONGSEONG IEUNG-KIYEOK
<U11E8> "<S11B7><S11AD>";"<BASE><BASE>";"<MIN><MIN>"; % HANGUL JONGSEONG IEUNG-SIOS
<U11E9> "<S11B7><S11AF>";"<BASE><BASE>";"<MIN><MIN>"; % HANGUL JONGSEONG IEUNG-RIEUL
<U11EA> "<S11B7><S11B8>";"<BASE><BASE>";"<MIN><MIN>"; % HANGUL JONGSEONG IEUNG-PIEUP
<U11EB> "<S11B7><S11BA>";"<BASE><BASE>";"<MIN><MIN>"; % HANGUL JONGSEONG IEUNG-SIOS

<U11F0> "<S11F0><S11BA>";"<BASE><BASE>";"<MIN><MIN>"; % HANGUL JONGSEONG YESIEUNG (EOKEUMISORI)
<U11F1> "<S11F0><S11BB>";"<BASE><BASE>";"<MIN><MIN>"; % HANGUL JONGSEONG YESIEUNG-RIEUL
<U11F2> "<S11F0><S11BF>";"<BASE><BASE>";"<MIN><MIN>"; % HANGUL JONGSEONG YESIEUNG-PANSIOS

<U11F0> "<S11F0><S11BA>";"<BASE><BASE>";"<MIN><MIN>"; % HANGUL JONGSEONG YEORINHIEUH

<U11BD> "<S11BD>";"<BASE><BASE>";"<MIN><MIN>"; % HANGUL JONGSEONG CIEUC (JIEUJ)

<U11F9> "<S11F9>";"<BASE><BASE>";"<MIN><MIN>"; % HANGUL JONGSEONG YEORINHIEUH
3 Tailorings

By tailoring the order of the basic Hangul letters we can get other orders for Hangul strings. With the above weightings in the CTT, it suffices to tailor the single-letter jamos.

In the suggested excerpt (above) for Hangul in the CTT (and in the UCA default ordering), the letter variants are handled specially. Whether Chitueum- and Ceongchieum- variants really should be ordered with the same first level weight as the original letter they are based on or not is not entirely clear. Maybe the special handling of these is best left to a tailoring, having a simpler approach in the CTT. E.g., letting the Chitueum- and Ceongchieum- variants collate as separate letters.

To get fully correct handling of the Kapyeoun- constructs, there need to be contractions and special extra weights. In this suggestion, this is left for a tailoring (see below).

3.1 Tailoring for original order

The original order in which the Hangul letters were presented differs from the order they most often have today. The following tailoring gives the order implied by the original document describing the, then new, Hangul script. No real order was
presented, and in particular, no indication of collation clustering was presented, so those details are modern interpretations.

It is interesting to see that this initial ordering is very “featuralistic”; adding and removing strokes, or adding one or two dots on either side of a single stroke. There are other “featuralistic” interpretations of Hangul, but the original one is particularly interesting, since it gives some insight into the design ideas for the script.

Note again that the “fillers” are modern coding inventions, not part of the original Hangul. Likewise for the division into lead and trail codes for the consonants. Other approaches would have been viable for the digital representation of Hangul.

An interesting option is be to collate by default according to the original order in which the Hangul letters were presented (including the modern interpretation of clustered ordering), and handle other orderings of Hangul strings as tailorings.

% The 11 basic Hangul vowels, first vowel in syllable, original order, plus vowel filler:
<V1160>  % HANGUL JUNGSEONG FILLER
<V119E>  % , HANGUL JUNGSEONG ARAEA
<V1173>  % ㅡ, HANGUL JUNGSEONG EU
<V1175>  % ㅣ, HANGUL JUNGSEONG I
<V1169>  % ㅗ, HANGUL JUNGSEONG O
<V1161>  % ㅏ, HANGUL JUNGSEONG A
<V116E>  % ㅜ, HANGUL JUNGSEONG U
<V1165>  % ㅣ, HANGUL JUNGSEONG EO
<V116D>  % ㅑ, HANGUL JUNGSEONG YO
<V1163>  % ㅐ, HANGUL JUNGSEONG YA
<V1172>  % ㅠ, HANGUL JUNGSEONG YU
<V1167>  % ㅒ, HANGUL JUNGSEONG YEO

% The 17 basic Hangul consonants, lead copy, original order, plus lead consonant filler:
<S1100>  % ㄱ, HANGUL CHOSEONG KIYEOK (KIEUK) (k, g)
<S110F>  % ㅋ, HANGUL CHOSEONG KHIEUKH (kh)
<S114C>  % ㆁ, HANGUL CHOSEONG YESIEUNG (EOKEUMISORI) (ng, both when lead and when trail)
% the glyph for yesieung should have a clearly visible vertical stroke on top
<S1103>  % ㄷ, HANGUL CHOSEONG TIKEUT (TIEUT) (t, d), and variant of tikeut
<S1110>  % ㅌ, HANGUL CHOSEONG THIEUTH (th), and variant of thieuth
<S1102>  % ㄴ, HANGUL CHOSEONG NIEUN (n), and variant of nieun
<S1107>  % ㅅ, HANGUL CHOSEONG PIEUP (p)
<S1111>  % ㅈ, HANGUL CHOSEONG PHIEUPH (ph, f?)
<S1106>  % ㅁ, HANGUL CHOSEONG MIEUM (m), and variant of mieum
<S110C>  % ㅂ, HANGUL CHOSEONG CIEUC (JIEUJ) (c, j), and variants of cieuc
<S110E>  % ㅍ, HANGUL CHOSEONG CHIEUCH (ch, jh), and variants of chieuch
<S1109>  % ㅎ, HANGUL CHOSEONG SIOS (SIEUS) (s, sh), and variants of sios
<S1159>  % ㅋ, HANGUL CHOSEONG YEORINHIEUH (MOKKUNEONGTHEOCHIMSORI)
% The 17 basic Hangul consonants, trail copy, original order:

% HANGUL CHOSEONG HIEUH
% HANGUL CHOSEONG IEUNG
% *later* used for silence (when lead) and ng (when trail) [may have been
% silent for both lead and trail; and could be omitted for trail...]

% HANGUL CHOSEONG RIEUL
% HANGUL CHOSEONG PANSIOS
% HANGUL CHOSEONG FILLER

% The 17 basic Hangul consonants, trail copy, original order:
% HANGUL CHOSEONG HIEUH
% HANGUL CHOSEONG IEUNG
% *later* used for silence (when lead) and ng (when trail) [may have been
% silent for both lead and trail; and could be omitted for trail...]

% HANGUL CHOSEONG RIEUL
% HANGUL CHOSEONG PANSIOS
% HANGUL CHOSEONG FILLER

% The 11 basic Hangul vowels, non-first occurrences in syllable, original order:
% HANGUL JOSUNG ARAEA (ARAIA)
3.2 Kapyeoun-tailoring

order-after <S11F9> % ☸ , HANGUL JONGSEONG YEORINHIEUH

collating-element <U1105_110B> from "<U1105><U110B>" % HANGUL CHOSEONG RIEUL-IEUNG (KAPYEOUNRIEUL)
collating-element <U1106_110B> from "<U1106><U110B>" % HANGUL CHOSEONG MIEUM-IEUNG (KAPYEOUNMIEUM)
collating-element <U1107_110B> from "<U1107><U110B>" % HANGUL CHOSEONG PIEUP-IEUNG (KAPYEOUNPIEUP)
collating-element <U1107_1107_110B> from "<U1107><U1107><U110B>" % HANGUL CHOSEONG SSANGPIEUP-IEUNG (KAPYEOUNSSANGPIEUP)
collating-element <U1111_110B> from "<U1111><U110B>" % HANGUL CHOSEONG PHIEUPH-IEUNG (KAPYEOUNPHIEUPH)
collating-element <U111B_110B> from "<U111B><U110B>" % HANGUL JONGSEONG RIEUL-IEUNG (KAPYEOUNRIEUL)
collating-element <U111D_110B> from "<U111D><U110B>" % HANGUL JONGSEONG MIEUM-IEUNG (KAPYEOUNMIEUM)
collating-element <U111D_111D_110B> from "<U111D><U111D><U110B>" % HANGUL JONGSEONG SSANGPIEUP-IEUNG (KAPYEOUNSSANGPIEUP)
collating-symbol <K115F> % used for JONGSEONG KAPYEOUN- % should not be in the CTT <K115F> % used for JONGSEONG KAPYEOUN

reorder-after <SFFFF>

order_start forward;forward;forward;forward

< U1105_110B > "<S1105><K115F><S110B>";"<BASE><BASE><BASE>";"<MIN><MIN><MIN>";<U1105_110B>
% HANGUL CHOSEONG RIEUL-IEUNG (KAPYEOUNRIEUL)
< U111B > "<S1105><K115F><S110B>";"<BASE><BASE><BASE>";"<MIN><MIN><MIN>";<U111B>
% HANGUL CHOSEONG KAPYEOUNRIEUL (RIEUL-IEUNG)
< U1106_110B > "<S1106><S115F><S110B>";"<BASE><BASE><BASE>";"<MIN><MIN><MIN>";<U1106_110B>
% HANGUL CHOSEONG MIEUM-IEUNG (KAPYEOUNMIEUM)
< U111D > "<S1106><S115F><S110B>";"<BASE><BASE><BASE>";"<MIN><MIN><MIN>";<U111D>
% HANGUL CHOSEONG KAPYEOUNMIEUM (MIEUM-IEUNG)
< U3171 > "<S1105><S115F><S110B><V1160>";"<BASE><BASE><BASE><BASE>";"<COMPAT><COMPAT><COMPAT><COMPAT>";<U3171>
% HANGUL LETTER KAPYEOUNMIEUM
< U1107_110B > "<S1107><S115F><S110B>";"<BASE><BASE><BASE>";"<MIN><MIN><MIN>";<U1107_110B>
% HANGUL CHOSEONG PIEUP-IEUNG (KAPYEOUNPIEUP)
< U112B > "<S1107><S115F><S110B>";"<BASE><BASE><BASE>";"<MIN><MIN><MIN>";<U112B>
% HANGUL CHOSEONG KAPYEOUNPIEUP (PIEUP-IEUNG)
< U3178 > "<S1107><S115F><S110B><V1160>";"<BASE><BASE><BASE><BASE>";
"<COMPAT><COMPAT><COMPAT><COMPAT>";<U3178>
% HANGUL LETTER KAPYEOUNPIEUP
< U1107_1107_110B > "<S1107><S115F><S1107><S110B>";"<BASE><BASE><BASE><BASE>";
"<MIN><MIN><MIN><MIN><MIN>";<U1107_1107_110B>
% HANGUL CHOSEONG SSANGPIEUP-IEUNG (KAPYEOUNSSANGPIEUP)
< U112C > "<S1107><S115F><S1107><S110B>";"<BASE><BASE><BASE><BASE>";
"<MIN><MIN><MIN><MIN><MIN>";<U112C>
% HANGUL CHOSEONG KAPYEOUNSSANGPIEUP
< U3179 > "<S1107><S115F><S1107><S110B><V1160>";"<BASE><BASE><BASE><BASE><BASE>";
"<COMPAT><COMPAT><COMPAT><COMPAT><COMPAT><COMPAT>";<U3179>
% HANGUL LETTER KAPYEOUNSSANGPIEUP
3.3 Suggested tailoring for R. of Korea ordering

The official ordering appears to be (Han'gŭl matchumpŏp, 1988):

For modern leading consonants and leading consonants clusters:

ㄱ ㄲ ㄴ ㄷ ㄸ ㄹ ㅁ ㅂ ㅃ ㅅ ㅆ o ㅈ ㅉ ㅊ ㅋ ㅌ ㅍ ㅎ

For modern trailing consonants and trailing consonant clusters:

ㄱ ㄲ ㄳ ㄴ ㄵ ㄶ ㄷ ㄹ ㄺ ㄻㄼㄽㄾㄿㅀㅁㅂㅄㅅㅆ o ㅈ ㅊ ㅋㅌㅍㅎ

For modern vowels and vowel clusters:

ㅏ ㅐ ㅑ ㅒ ㅓ ㅔ ㅕ ㅖ ㅗ ㅘ ㅙ ㅚ ㅛ ㅜ ㅝ ㅞ ㅟ ㅠ ㅡ ㅢ ㅣ

This order is a subset of the ordering that results from the above suggested CTT excerpt without any tailoring. However, the kapyeoun-tailoring above may be applicable.

3.4 Suggested tailoring for Han'gŭl match'umpŏp t'ong'iran ordering

The Han'gŭl match'umpŏp t'ong'iran (1933) ordering appears to systematically order clusters of two after singletons, and clusters of three thereafter. The intended ordering appears to be:

ㅏ ㅐ ㅑ ㅒ ㅓ ㅔ ㅕ ㅖ ㅗ ㅘ ㅙ ㅚ ㅛ ㅜ ㅝ ㅞ ㅟ ㅠ ㅡ ㅢ ㅣ
For modern leading consonants and leading consonants clusters:

ㄱ ㄴ ㄷ ㄹ ㅁ ㅂ ㅅ o ㅈ ㅊ ㅋ ㅌ ㅍ ㅎ

ㄲ ㄸ ㅃ ㅆ ㅉ

For modern trailing consonants and consonants clusters:

g ㄴ ㄷ ㄹ ㅁ ㅂ ㅅ o ㅈ ㅊ ㅋ ㅌ ㅍ ㅎ

ㄲ ㄳ ㄵ ㄶ ㄺ ㄻ ㄼ ㄽ ㄾ ㄿ ㅀ ㅄ ㅆ

For modern vowels and vowels clusters:

ㅏ ㅑ ㅓ ㅕ ㅗ ㅛ ㅜ ㅠ ㅡ ㅣ

ㅐ ㅒ ㅔ ㅖ ㅘ ㅚ ㅝ ㅟ ㅢ ㅙ ㅞ

This order is a subset of the ordering that results from the suggested CTT excerpt together with the following tailoring. Some, as yet unused, weights are needed for this tailoring. In order to get this ordering the following prehandling is needed:

Insertion of cluster size pseudo-characters (used just for Hangul) L2, L3, V2, V3, T2, and T3 in the following positions in a Hangul syllable: L+ V+ T* M* → (Ln)?L+ (Vn)?V+ (Tn)?T* M*. They are inserted in this way: L2 is inserted in front of clusters of two lead consonants, L3 is inserted in front of clusters of three (or more) lead consonants, V2 is inserted in front of clusters of three (or more) vowels, V3 is inserted in front of clusters of three (or more) vowels, T2 is inserted in front of clusters of two trail consonants, and T3 is inserted in front of clusters of three (or more) trail consonants. Clusters that are single letters do not get any pseudo-characters inserted in front of them.

The weighting for singleton letter characters are not changed, nor is there a need for more contractions (V2, V3 only occur as first “vowel”, and are directly weighted accordingly). However, all Hangul letter vowel cluster jamos are reweighted with just “heavy” weights. The pseudo-characters (some non-character codes can be chosen) inserted by the prehandling are weighted thus: L2 and L3 are ordered just after lead consonants; V2 and V3 are ordered just after first-vowels; T2 and T3 are ordered just after trail consonants.

However, an approximation that does not require any modified prehandling would be interesting. For this, we need to limit the correctly ordered syllables to the ones that are expressed as L V or L V T. The weighting builds in the pseudo-character weights of the general approach into the weightings for the letter cluster jamos, like for instance:

\[
\text{<U1176> "<V2><S1161><S1169>";}<\text{BASE}>\text{BASE}>;"<\text{MIN}>\text{MIN}>;<\text{U1176}> \% \text{HANGUL JUNGSEONG A-O}
\]

\[
\text{<U1177> "<V2><S1161><S116E>";}<\text{BASE}>\text{BASE}>;"<\text{MIN}>\text{MIN}>;<\text{U1177}> \% \text{HANGUL JUNGSEONG A-U}
\]

\[
\text{<U1162> "<V2><S1161><S1175>";}<\text{BASE}>\text{BASE}>;"<\text{MIN}>\text{MIN}>;<\text{U1162}> \% \text{HANGUL JUNGSEONG AE (A-I)}
\]

\[
\text{<U116A> "<V2><S1169><S1161>";}<\text{BASE}>\text{BASE}>;"<\text{MIN}>\text{MIN}>;<\text{U116A}> \% \text{HANGUL JUNGSEONG WA (O-A)}
\]

\[
\text{<U116B> "<V3><S1169><S1161><S1175>";}<\text{BASE}>\text{BASE}>;"<\text{MIN}>\text{MIN}>;<\text{U116B}> \% \text{HANGUL JUNGSEONG WAE (O-A-I)}
\]

...
etc. for all letter cluster Hangul jamos; or simply ordering them all in the desired order within the respective group (syllable-first vowels, lead consonants, trail consonants).

### 3.5 DPR of Korea ordering

The DPRK ordering may be a modification of the Han'gŭl match'umpŏp tong'iran ordering. The ordering appears to be:

- **For leading consonants**: (note that IEUNG is last, maybe because it is silent when leading, where neither of the lists in SC2/WG2/N2246 put it, but how it is ordered in JTC1/N5999; apart from that the ‘tense’ consonants are last, other sources [Sohn] put IEUNG right after SIOS):
  
  |ㄱ|ㄴ|ㄷ|ㄹ|ㅁ|ㅂ|ㅅ|ㅈ|ㅊ|ㅋ|ㅌ|ㅍ|ㅎ|
  |ㄲ|ㄸ|ㅃ|ㅆ|ㅉ|

- **For trailing consonants**: (note the mixed approach; the ‘tense’ (orthographically: double-written) consonants are last, but those that should be read as multiple consonants are not specially treated):
  
  |ㄱ|ㄳ|ㄴ|ㄵ|ㄶ|ㄷ|ㄹ|ㄺ|ㄻ|ㄼ|ㄽ|ㄾ|ㄿ|ㅀ|ㅁ|ㅂ|ㅄ|ㅅ|ㅈ|ㅊ|ㅋ|ㅌ|ㅍ|ㅎ|
  |ㄲ|ㅆ|

- **For vowels**: (either from JTC1/N5999, and SC2/WG2/N2246 page 3, this agrees with [Sohn]):
  
  |ㅏ|ㅑ|ㅓ|ㅕ|ㅗ|ㅛ|ㅜ|ㅠ|ㅡ|ㅣ|
  |ㅐ|ㅒ|ㅔ|ㅖ|ㅒ|ㅖ|ㅒ|ㅖ|ㅒ|ㅖ|ㅒ|ㅖ|ㅒ|ㅖ|

Once the details are settled, such an ordering can be achieved in similar manners to how the Han'gŭl match’umpŏp tong’iran (1933) ordering can be achieved (see above). Preferably, a general approach with prehandling inserting pseudo-characters that are weighted for the desired ordering. Or alternatively, a more limited approach, where only L V and L V T Hangul syllables are handled correctly, building in the extra weights into the weighting of some of the letter cluster Hangul jamo characters, or simply ordering them in the desired order.

### 4 Requirements fullfillment

In SC22/WG20 N1037 (L2/03-081), Hangul Collation Requirements (quoted in italics below), Mark Davis gives some requirements that Hangul collation should fulfill.

**R1. Tailorability.** *It must be possible to tailor the UCA so as to meet the following requirements for Korean.*

Fullfilled by the proposed solution. However, tailoring to get DPRK ordering requires the use of prehandling to insert “length marks” before each consonant cluster (both lead and trail) and each vowel cluster, or alternatively use contractions for all occurring consonant clusters (both lead and trail) and all occurring vowel clusters. The suggested default Korean ordering is for modern ordering of Hangul in the RoK.
fashion. Tailoring for historic orders is relatively easy, given that in most cases only the basic letters’s order need be tailored (and possibly also the KAYEOUN contractions). Letter clusters, syllables and compatibility letters will then by implication also be tailored since their weightings are expressed in terms only of the weights for the basic letters.

**R2. Canonical Equivalence.** Whenever text is canonically equivalent, it must sort the same. In particular, one string made up of HANGUL SYLLABLE characters will sort the same as a string made up of the equivalent JAMO characters. Example: the following sequences must sort the same:

- `<U+AC01 (ㄱ) HANGUL SYLLABLE GAG>`
- `<U+1100 (ㄱ) HANGUL CHOSEONG KIYEOK, U+1161 (ㅏ) HANGUL JUNGSEONG A, U+11A8 (ᆨ) HANGUL JONGSEONG KIYEOK>`

Fullfilled by the proposed solution. This is done by formally requiring canonical expansion, or rather, work as if there is canonical expansion, of Hangul precomposed syllable characters.

**R3. Performance/Size.** Whatever approach is taken, it must be possible to sort strings containing the composite 11,172 Hangul Syllables (which form the vast majority of Korean text) without undo performance/storage costs. In particular, when restricted to such text, sort keys must be reasonably short and incremental comparison reasonably fast, in comparison to current good implementations of Korean sorting. A reasonable limit is no more than about 3 times worse. See Performance.

I have no performance measurements. However, if it is known that the text consists exclusively of precomposed Hangul syllable characters, the weightings for these can be precomputed (each syllable character would get 2-7 level 1 weights) and used without actually doing the expansion at the character level. (The reason this cannot be done for mixed precomposed and Jamo texts is that a trail KAPYEOUN consonant cluster can be formed if a precomposed syllable characters is followed by a trail IEUNG Jamo character. and KAPYEOUN consonant clusters are specially ordered.)

**R4. Existing Korean Syllable-Blocks.** It is only important that existing standard Korean syllable-blocks and equivalent text sorts correctly, as long as for other characters the ordering is still determinate. That is:

The proposed solution orders all Hangul texts correctly according to modern RoK ordering. (And other letter based orderings can easily be tailored, except that DPRK tailoring is less trivial.)

**R3a.** There is no requirement on "degenerate" syllables (such as L₁T₁).

No attention has been paid to degenerate syllables.
R3b. It is only a requirement to sort existing syllables, not all theoretically possible syllables. See Desired Features. It must, however, be possible to tailor additional syllables.

The proposed solution orders all Hangul texts (all theoretically possible syllables) correctly according to modern RoK ordering if full Hangul decomposition (into individual letters, not letter clusters; NFD plus additional decompositions of Hangul letter cluster characters) is done as a prehandling. The requirement is fulfilled for modern Hangul if NFD is performed as a prehandling.

R5. Syllables as Primary Units. Standard Korean syllable-blocks sort as primary units, meaning that if two syllables are different, no further characters in the string make any difference (in ordering). Examples:

- if $S_1 < S_2$, then $S_1x < S_2y$, where $x$ and $y$ are any other characters that would not extend the syllables.

This is fulfilled by the proposed solution, by assigning the weights in a “clever” way (see above, section 2.1).

R6. Clusters as Primary Units. Within a standard Korean syllable-block, elements of a cluster sort as primary units, meaning that if two such elements are different, no further characters in the syllable make any difference (in ordering). Examples:

- if $L_1 < L_2$, then $L_1x < L_2y$, where $x$ and $y$ are any non-L.
- if $V_1 < V_2$, then $V_1x < V_2y$, where $x$ and $y$ are any non-V.
- if $T_1 < T_2$, then $T_1x < T_2y$, where $x$ and $y$ are any non-T.

All three are trivially fulfilled. (Is a different set of requirements intended?)

R7. Shorter Clusters Sort First. A shorter, but otherwise identical, cluster sorts before a longer one, no matter what is later in the string. Examples:

- $L_1x < L_1L_2$, where $x$ is any non-L
- $V_1x < V_1V_2$, where $x$ is any non-V
- $T_1x < T_1T_2$, where $x$ is any non-T

First one is fulfilled if the non-L is a V (see 3a above). The second one is fulfilled if the non-V is a T or a non-dependent letter/ideograph (similar to 3a above). The third is fulfilled if the non-T is a non-dependent (not a T, V, virama, or dependent Indic vowel).

There are lots of other requirements that have to be fulfilled; in order to get expected dictionary order. Some of these requirements are some of the requirements mentioned by Professor Gim (while he also has some requirements that seem superfluous, not needed to get proper dictionary order, and indeed in contradiction to getting dictionary order without having extraneous weights (for end cluster markers or similar)). The proposed solution fulfills the dictionary order requirements (but not any extraneous, non-dictionary, requirements).
5 Additional properties

- Exactly one weight per level per letter. Same as for all other alphabetic scripts.
- Historic and modern Hangul handled together, intracollating historic Hangul according to the letters in the string, just like for modern Hangul.
- Independence of decompositions. Even cluster letter jamo decomposition (which is no longer part of the Unicode database).
- The only technical restriction is that vowel letter cluster jamos must occur only as the first character in a sequence of vowel jamos. This is no limitation with regard to which Hangul strings may be written (using the encoded jamo letters), since one can always use just single-letter jamos. (There may still be some single-letter jamos missing in the encoding.)
- No prehandling is needed for strings using Hangul jamos.
- If the weighting for Hangul Syllable characters are precomputed, there is not even any need for arithmetic decomposition of Hangul syllables as a prehandling. Thus avoiding the need for prehandling altogether.

6 Acknowledgements

Many thanks to Jungshik Shin for providing an excerpt from SOHN Ho-Min's paper Orthographic divergence in South and North Korea: Toward a unified spelling system, for helping the author understand the basic ordering requirements for Hangul, as well as for the references to texts in Korean (which are not directly accessible to me). Also thanks to Mark Davis for comments that resulted in substantial changes and improvements. Any remaining errors or shortcomings are of course mine.

7 References


Unicode 2.0  The Unicode standard, version 2.0. (In particular the compatibility decompositions in the UCD and (algorithmic) canonical decomposition of Hangul syllables.)

Unicode 3.0  The Unicode standard, version 3.0. (In particular the (algorithmic) canonical decomposition of Hangul syllables.)

UCD 2.0.0  Unicode character database, version 2.0.

UCD 4.0.0  Unicode character database, version 4.0.0.


UTS 10  Unicode technical standard 10, Unicode collation algorithm.
<table>
<thead>
<tr>
<th>ISO/IEC JTC1 N5999</th>
<th>DPRK, Proposal for a New Work Item, Amendment to the part concerning Korean characters in ISO/IEC 10646-1:1993. (NWI rejected.)</th>
</tr>
</thead>
<tbody>
<tr>
<td>ISO/IEC JTC1/SC22/WG20 N858</td>
<td>(Same as SC2/WG2 N2249R.)</td>
</tr>
<tr>
<td>ISO/IEC JTC1/SC22/WG20 N876</td>
<td>KIM, Kyongsok, Comments regarding N867, the DPRK proposed delta file for IS 14651.</td>
</tr>
<tr>
<td>L2/02-008R</td>
<td>(Same as ISO/IEC JTC 1/SC22/WG20 N898R.)</td>
</tr>
<tr>
<td>L2/02-109</td>
<td>Kent Karlsson. Additional Hangul character decompositions for the proper support of Hangul.</td>
</tr>
<tr>
<td>L2/02-mmmm</td>
<td>KIM, Kyongsok, New canonical decomposition and composition processes for Hangeul.</td>
</tr>
<tr>
<td>Han'gŭl match'umpŏp t'ongi'iran</td>
<td>(A proposition for unified han'gŭl spelling conventions), Chosŏnŏ Hakhoe (Korean Language Association), Seoul, 1933.</td>
</tr>
</tbody>
</table>
Orthographic divergence in South and North Korea: Toward a unified spelling system

The Korean Language

A History of Korean Alphabet and Movable Types
Ministry of Culture and Information, Republic of Korea, 1970. (Contains Hun Min Jong Um as photographic plates.)

—— end ——