ISO/IEC JTC 1/SC22/WG20 N898

L2/02-008

Universal multiple-octet coded character set
International organization for standardization
Organisation internationale de normalisation

Title: String ordering weighting roadmap

Source: Kent Karlsson
Date: 2002-01-11
Status: Expert contribution **VERY DRAFT**
Document type: Working group document
Action: For consideration by the UTC and JTC 1/SC 22/WG 20

1 Introduction

2 Weighting scheme

2.1 Key terminator weight

0 (or nothing). The 0 is only needed between keys if several keys are concatenated, to form a multi-string collation key.

2.2 Subkey terminator weight

1 (or 0). These can in most cases be avoided easily, see below. However, in order to be able to reuse weight values at the fourth level, these are used after the third level weights when using the fourth level.

[variable width weights]

2.3 Third level weights

In order to avoid having to have a subkey terminator weight after the second level weights in a collation key, all the third level weights are lighter (come before) the second level weights.

[variable width weights]

2.4 Second level weights

In order to avoid having to have a subkey terminator weight after the first level weights in a collation key, all the second level weights are lighter (come before) the first level weights.

[variable width weights]

2.5 First level weights

2.5.1 Delimiter, punctuation, and symbol weights

It is sometimes useful to consider a string divided into substrings in various ways when ordering; e.g. sentence-by-sentence or word-by-word. Instead of actually dividing up the string into several strings, that are given keys that are then concatenated (with a key terminator weight inbetween each), an easier way is to just certain “part terminating” characters as having light primary weights. Indeed, the latter even allows for a hierarchy...

However, some scripts always use a clustered ordering. Hangul uses a clustered ordering where the clusters are sequences of consonants and sequences of vowels (*note that Hangul does* ***not*** *use orthographic syllable clustering in its ordering rules, despite popular claims to that effect*). The Brahmic derived scripts use a clustering in ordering that is based on the orthographic syllable clusters.

For the most part, we can weight the characters so that the appropriate ordering clustering falls out as a result. This is not fully the case for Hangul, however. Most of it can be achieved by assigning the collation weights in a good way. But one **must** still insert a lightweight terminator weight (that for ZWNBSP below) after each Choseong cluster. But none is needed after Jungseong or Jongseong clusters **if** the weights are assigned appropriately (see below).

Note that the clusters here are *not* directly related to the “grapheme clusters” elsewhere!

2.5.1.1 Top level termination characters

top delimitation characters (section/book end characters? FS/IS4?), [IGNORE by default]

2.5.1.2 Paragraph termination characters

paragraph delimitation characters (PS?, other paragraph end characters? GS/IS3?),

CR/LF/CR+LF/NEL? (esp. when doubled or has indent after), [IGNORE by default]

2.5.1.3 Sentence termination characters

sentence delimitation characters, (LS?, FF?, VT?,“.”, “!”, “?”, “;”, “,” ?, RS/IS2?

dashes ex. Hys?), CR/LF/CR+LF/NEL?, [IGNORE by default]

2.5.1.4 Word termination characters

word delimitation characters, (SP?, Zs?, HT?, NBSP?, ZWSP?, BPH?, US/IS1?, “)”-s, Hys?, other punctuation?), [IGNORE by default]; currency signs [IGNORE in CTT?]

2.5.1.5 Subword termination characters

subword delimitation characters, (SHY, NBH?, WJ?) [IGNORE by default]

2.5.1.6 Syllable or subsyllable termination character

cluster delimiter (ZWNBSP? Never IGNOREd! Needed exactly between leading Hangul consonants and Hangul vowels! Insert by prehandling!]

2.5.2 Symbols

symbols (except MINUS SIGN, PLUS SIGN, and INFINITY) [IGNORE by default]

2.5.3 Integral numeral weights

2.5.3.1 Negative infinity

contraction for <MINUS SIGN, INFINITY> (note: not a grapheme cluster)

2.5.3.2 Negative values

MINUS SIGN (don't IGNORE) [in addition 9's complement, or 10's complement must be done (by prehandling) to get the correct ordering (and prefixed exponent) for negative integer numerals and more for negative fractional numerals]

2.5.3.3 Positive values

digits(!) (and PLUS SIGN, if not IGNOREd)

[prehandling needed for proper numerical order (insert prefixed exponents); a tailoring that handles fractional numerals must insert a non-ignored delimiter character after each numeral (WJ?) if needed; script shifts...; super/subscript digits]

2.5.3.4 Positive infinity

INFINITY (don't IGNORE)

2.5.4 Script weights

2.5.4.1 Scripts (and Han lead weights), except "dependent" letters

various scripts in some order, including Hangul leading consonants (choseong, and compatibility consonants regardless of compatibility decomposition (prehandling **may** insert a CGJ before each actually trailing compat. consonant (regardless of decomposition)); *all with weighting for full jamo decomposition into single-letters*), as well as Indic, Khmer, and Tibetan non-combining letters, (Han lead **weights** somewhere in this group too?) Note that prehandling MUST canonically decompose precomposed Hangul Syllables, and must insert a ZWNBSP after each Choseong cluster, and should ideally do so also after a sequence of (truly) leading compatibility Hangul consonants.

2.5.4.2 Dependent vowels

Brahmic, Khmer, and Tibetan dependent vowels (combining)
(VIRIAM? (too often invisible/left-out, HALANT?, PHINTHU?, ...;
not necessarily last among the respective set of vowels)
Thai and Lao vowels too, also those that are not formally combining.
Note that some of the Thai and Lao vowels **require that the prehandling**
put them in logical sequence, instead of left-to-right sequence.

2.5.4.3 Letter gluers

VIRAMAs, COENG, COMBINING GRAPHEME JOINER (don’t IGNORE)
[Han trail weights? previous?]

2.5.4.4 Subjoined letters

Tibetan (and Khmer, if any) subjoined letters, Hangul trailing consonants (jongseong only, NOT any compat.; *all with weighting for full Jamo decomposition into single-letters*)

2.5.4.5 Hangul vowels

Hangul vowels (jungseong and compatibility vowel letters; all with weighting for *full Jamo decomposition into single-letters*), (Han trail **weights** here or with the previous; no restriction really!) Note again that Hangul does **not** use orthographic syllable clustering in its ordering rules, despite popular claims to that effect. Therefore the Hangul vowels are ordered here instead of among the other dependent vowels. Note the **prehandling is required** to insert a ZWNBSP (which is given a light, but not the lowest (see above), first-level weight. ZWNBSP must not be IGNORED due to this special use. ZWNBSP must not occur in actual (before prehandling) input.

2.6 Fourth level weights

Subkey terminator before the first fourth level weight in each key, so that fourth level weights can overlap with the other weight values.
No need for a subkey terminator after it.
PLAIN last!

3 Conclusions

4 References

ISO/IEC 10646-1:2000	Information Technology – Universal multiple-octet coded character set (UCS), Part 1, second edition.
Unicode 3.0	The Unicode standard, version 3.0.
UCD 3.1	Unicode character database, version 3.1.
ISO/IEC 14651:2001	International string ordering and comparison – Method for comparing character strings and description of the common template tailorable ordering.
UTS 10	Unicode technical standard 10, Unicode collation algorithm.
ISO/IEC JTC 1/SC22/WG20 N891R	Kent Karlsson, Hangul ordering rules. 2001-11-29.
ISO/IEC JTC 1/SC2/WG2 N2405R	Same as ISO/IEC JTC 1/SC22/WG20 N891R.
L2/01-469	Same as ISO/IEC JTC 1/SC22/WG20 N891R.
ISO/IEC JTC 1/SC22/WG20 N896	Kent Karlsson, Khmer ordering rules. 2001-12-19.
L2/01-476	Same as ISO/IEC JTC 1/SC22/WG20 N896.
	Segmental collation and the UCA. Mark Davis. “Very draft”, 2001-11-16.

--- end ---