L2/04-290

Updating the Arabic Shaping normative data

Kent Karlsson
2004-07-16

For consideration by the UTC.

1. The ArabicShaping.txt data for Unicode 4.0 contains a number of oddities that should be corrected.

   The character names (second data field) in the list are not the normative Unicode names. This is confusing, and sometimes misleading. Do one of:

   a) The “short names” should be replaced by the normative ones. The additional information that the “short names” may carry should be placed elsewhere (if there is any suitable elsewhere).

   b) The list could be complemented by the normative name in a comment after each data line. In this case the “short names” should be turned into being *fully descriptive* (telling number of dots or other adornments, where and how they are placed, and movement of dots between different shapes.

   Note: this edit is not done in the data file quoted below, to make it easier to compare to the existing ArabicShaping.txt.

2. Some of the characters have oddly assigned joining type or joining group. In some cases to such a degree that canonically equivalent substrings have different joining behaviour, in other cases so that adding dots changes the joining behaviour, which would be unexpected, especially where the use of the dotted versus undotted character is a spelling option. The characters that should be changed w.r.t. this are: ARABIC LETTER TEH MARBUTA, ARABIC LETTER HEH WITH HAMZA, ARABIC LETTER HEH GOAL WITH HAMZA, ARABIC LETTER TEH MARBUTA GOAL, and ARABIC LETTER AE. *Note that the latter could have the same joining group as HEH, while only being R-joining.*

3. The “Cf” **graphic**(!) characters currently have the joining metagroup <no shaping> which is not very descriptive of their true behaviour, since they are shaped. Change to the new joining metagroup <span> since these span under (around in one case) a following sequence of digits (or any sequence of non-spaces?). The SYRIAC ABBREVIATION SIGN is similar (<span>), but has the joining type T.

4. The Arabic vowel marks currently have the default joining properties (“T; <no shaping>”). But apparently it is ok to put the vowel (harakat, tashkeel?) on a subsequent lne (a tatweel glyph, autoinserted during line justification), or just isolated (in case there is no join between the preceding and following base characters, and the word is “stretched” for line justification). Other combining marks do not behave that way. Change the joining metagroup to <tashkeel> or <harakat> (whichever is the better designation), to indicate this behaviour.

5. At least one character should be considered medial-joining: ARABIC LETTER HIGH HAMZA. This optionally applies also to the harakat, when they are not placed on the consonant, but instead on a tatweel or SPACE (INVISIBLE LETTER...). **This is a new joining type, M.**

6. Enclosing and reordrant combining characters should break the joining not only where they occur, but also before the letter to which one is applied. **This is a new joining type, B.**

Below is an edited ArabicShaping.txt file where the changes mentioned in points 2-6 above are done, including some updates to introductory comments.

```
##########################################################################
ORIGIN:  # ArabicShaping-4.0.1.txt
#
# This file is a normative contributory data file in the
# Unicode Character Database.
#
# This file defines the shaping classes for Arabic and Syriac
# positional shaping, repeating in machine readable form the
# information printed in Tables 8-3, 8-7, 8-8, 8-11, 8-12, and
# 8-13 of The Unicode Standard, Version 4.0, but updated for
# Unicode 4.x
#
# See sections 8.2 and 8.3 of The Unicode Standard, Version 4.0
# for more information.
#
# Each line contains four fields, separated by a semicolon.
#
# Field 0: the code point, in 4-digit hexadecimal form, of an Arabic or
#    Syriac character (there are no such characters on higher planes).
#
# Field 1: the Unicode character name.
############## NOTE: THIS EDIT IS STILL TO BE DONE FOR EASE OF COMPARISON
############## WITH THE CURRENT ArabicShaping.txt FOR THE OTHER CHANGES.
#
# Field 2: defines the joining type
#   R right-joining,
#   L left-joining, there are no characters of type L defined in Unicode,
#   D dual-joining,
#   M medial-joining, like T, but there are isolated and medial glyphic forms,
#   C join-causing,
#   U non-joining,
#   B enclosing/reordrant combining character, makes the combining sequence U,
#   T transparent, join processing ignores such a character.
#      See the Arabic block description for more information on these types.
#
# Field 3: defines the joining group. The following joining groups are defined:
#      Arabic: AIN, ALEF, BEH, DAL, FEH, HAH, HEH, HEH GOAL, KNOTTED HEH,
#              HIGH HAMZA, KAF, SWASH KAF, GAF, LAM, MEEM, NOON, QAF,
#              REH, SAD, SEEN, WAW, YEH, YEH BARREE, YEH WITH TAIL.
#
#      Syriac: ALAPH, BETH, DALATH RISH, E, FE, GAMAL, HE, HETH,
#              KAPH, KHAPH, LAMADH, MIM, NUN, PE, REVERSED PE, QAP,
#              SADHE, SEMKATH, FINAL SEMKATH, SHIN, SYRIAC WAW,
#              TAW, TETH, YUDH, YUDH HE, ZAIN, ZHAIN.
#
#      Meta:   <no shaping>, <tashkeel>, <span>. <tashkeel> is used for vowels,
#              that optionally may be shaped using isolated and medial forms
#              (the latter on a tatweel), but normally are placed on the
#              consonant. Note that double vowels (e.g. <kasra, shadda>) are
#              placed together. <span> is used for characters that span
#              under (or around) a following sequence of letters or digits.
#              <span> characters are visible characters in all circumstances,
#              despite that they have the Cf General_Category property.
#
#      Retired (mistaken) joining groups: TEH MARBUTA and HAMZA ON HEH GOAL.
#
# Note: Code points that are not explicitly listed in this file are either of
# type T, B, or U:
#
# - Those that are not explicitly listed below that are of General Category
#   Mn or Cf have joining type T, and are of joining metagroup <no shaping>.
#
# - Those that are not explicitly listed below that are of General Category
#   Me or Mc have joining type B, and are of joining megagroup <no shaping>.
#
# - All others not explicitly listed below have joining type U, and are of
#   joining metagroup <no shaping>, including the Arabic presentation forms.
#
# For an explicit listing of characters of joining type T, see
# the derived property file DerivedJoiningType.txt.
#
# #############################################################

# Code point; Character name; Joining type; Joining group

# Arabic characters

0600; ARABIC NUMBER SIGN; U; <span>
0601; ARABIC SIGN SANAH; U; <span>
0602; ARABIC FOOTNOTE MARKER; U; <span>
0603; ARABIC SIGN SAFHA; U; <span>
      #0621; HAMZA; U; <no shaping>  ### has default values; need not be listed.
0622; MADDA ON ALEF; R; ALEF
0623; HAMZA ON ALEF; R; ALEF
0624; HAMZA ON WAW; R; WAW
0625; HAMZA UNDER ALEF; R; ALEF
0626; HAMZA ON YEH; D; YEH
0627; ALEF; R; ALEF
```

```
0628; BEH; D; BEH
0629; TEH MARBUTA; D; HEH      ### was:  R; TEH MARBUTA
062A; TEH; D; BEH
062B; THEH; D; BEH
062C; JEEM; D; HAH
062D; HAH; D; HAH
062E; KHAH; D; HAH
062F; DAL; R; DAL
0630; THAL; R; DAL
0631; REH; R; REH
0632; ZAIN; R; REH
0633; SEEN; D; SEEN
0634; SHEEN; D; SEEN
0635; SAD; D; SAD
0636; DAD; D; SAD
0637; TAH; D; TAH
0638; ZAH; D; TAH
0639; AIN; D; AIN
063A; GHAIN; D; AIN
0640; TATWEEL; C; <no shaping>
0641; FEH; D; FEH
0642; QAF; D; QAF
0643; KAF; D; KAF
0644; LAM; D; LAM
0645; MEEM; D; MEEM
0646; NOON; D; NOON
0647; HEH; D; HEH
0648; WAW; R; WAW
0649; ALEF MAKSURA; D; YEH
064A; YEH; D; YEH
064B; ARABIC FATHATAN; T or M; <tashkeel>
064C; ARABIC DAMMATAN; T or M; <tashkeel>
064D; ARABIC KASRATAN; T or M; <tashkeel>
064E; ARABIC FATHA; T or M; <tashkeel>
064F; ARABIC DAMMA; T or M; <tashkeel>
0650; ARABIC KASRA; T or M; <tashkeel>
0651; ARABIC SHADDA; T or M; <tashkeel>
0652; ARABIC SUKUN; T or M; <tashkeel>
066E; DOTLESS BEH; D; BEH
066F; DOTLESS QAF; D; QAF
0670; ARABIC LETTER SUPERSCRIPT ALEF; T or M; <tashkeel>
0671; HAMZAT WASL ON ALEF; R; ALEF
0672; WAVY HAMZA ON ALEF; R; ALEF
0673; WAVY HAMZA UNDER ALEF; R; ALEF
0674; HIGH HAMZA; M; HIGH HAMZA  ### was:   U; <no shaping>
0675; HIGH HAMZA ALEF; R; ALEF
0676; HIGH HAMZA WAW; R; WAW
0677; HIGH HAMZA WAW WITH DAMMA; R; WAW
0678; HIGH HAMZA YEH; D; YEH
0679; TEH WITH SMALL TAH; D; BEH
067A; TEH WITH 2 DOTS VERTICAL ABOVE; D; BEH
067B; BEH WITH 2 DOTS VERTICAL BELOW; D; BEH
067C; TEH WITH RING; D; BEH
067D; TEH WITH 3 DOTS ABOVE DOWNWARD; D; BEH
067E; TEH WITH 3 DOTS BELOW; D; BEH
067F; TEH WITH 4 DOTS ABOVE; D; BEH
0680; BEH WITH 4 DOTS BELOW; D; BEH
```

```
0681; HAMZA ON HAH; D; HAH
0682; HAH WITH 2 DOTS VERTICAL ABOVE; D; HAH
0683; HAH WITH MIDDLE 2 DOTS; D; HAH
0684; HAH WITH MIDDLE 2 DOTS VERTICAL; D; HAH
0685; HAH WITH 3 DOTS ABOVE; D; HAH
0686; HAH WITH MIDDLE 3 DOTS DOWNWARD; D; HAH
0687; HAH WITH MIDDLE 4 DOTS; D; HAH
0688; DAL WITH SMALL TAH; R; DAL
0689; DAL WITH RING; R; DAL
068A; DAL WITH DOT BELOW; R; DAL
068B; DAL WITH DOT BELOW AND SMALL TAH; R; DAL
068C; DAL WITH 2 DOTS ABOVE; R; DAL
068D; DAL WITH 2 DOTS BELOW; R; DAL
068E; DAL WITH 3 DOTS ABOVE; R; DAL
068F; DAL WITH 3 DOTS ABOVE DOWNWARD; R; DAL
0690; DAL WITH 4 DOTS ABOVE; R; DAL
0691; REH WITH SMALL TAH; R; REH
0692; REH WITH SMALL V; R; REH
0693; REH WITH RING; R; REH
0694; REH WITH DOT BELOW; R; REH
0695; REH WITH SMALL V BELOW; R; REH
0696; REH WITH DOT BELOW AND DOT ABOVE; R; REH
0697; REH WITH 2 DOTS ABOVE; R; REH
0698; REH WITH 3 DOTS ABOVE; R; REH
0699; REH WITH 4 DOTS ABOVE; R; REH
069A; SEEN WITH DOT BELOW AND DOT ABOVE; D; SEEN
069B; SEEN WITH 3 DOTS BELOW; D; SEEN
069C; SEEN WITH 3 DOTS BELOW AND 3 DOTS ABOVE; D; SEEN
069D; SAD WITH 2 DOTS BELOW; D; SAD
069E; SAD WITH 3 DOTS ABOVE; D; SAD
069F; TAH WITH 3 DOTS ABOVE; D; TAH
06A0; AIN WITH 3 DOTS ABOVE; D; AIN
06A1; DOTLESS FEH; D; FEH
06A2; FEH WITH DOT MOVED BELOW; D; FEH
06A3; FEH WITH DOT BELOW; D; FEH
06A4; FEH WITH 3 DOTS ABOVE; D; FEH
06A5; FEH WITH 3 DOTS BELOW; D; FEH
06A6; FEH WITH 4 DOTS ABOVE; D; FEH
06A7; QAF WITH DOT ABOVE; D; QAF
06A8; QAF WITH 3 DOTS ABOVE; D; QAF
06A9; OPEN KAF; D; GAF
06AA; SWASH KAF; D; SWASH KAF
06AB; KAF WITH RING; D; GAF
06AC; KAF WITH DOT ABOVE; D; KAF
06AD; KAF WITH 3 DOTS ABOVE; D; KAF
06AE; KAF WITH 3 DOTS BELOW; D; KAF
06AF; GAF; D; GAF
06B0; GAF WITH RING; D; GAF
06B1; GAF WITH 2 DOTS ABOVE; D; GAF
06B2; GAF WITH 2 DOTS BELOW; D; GAF
06B3; GAF WITH 2 DOTS VERTICAL BELOW; D; GAF
06B4; GAF WITH 3 DOTS ABOVE; D; GAF
06B5; LAM WITH SMALL V; D; LAM
06B6; LAM WITH DOT ABOVE; D; LAM
06B7; LAM WITH 3 DOTS ABOVE; D; LAM
06B8; LAM WITH 3 DOTS BELOW; D; LAM
06B9; NOON WITH DOT BELOW; D; NOON
```

```
06BA; DOTLESS NOON; D; NOON
06BB; DOTLESS NOON WITH SMALL TAH; D; NOON
06BC; NOON WITH RING; D; NOON
06BD; NOON WITH 3 DOTS ABOVE; D; NOON
06BE; KNOTTED HEH; D; KNOTTED HEH
06BF; HAH WITH MIDDLE 3 DOTS DOWNWARD AND DOT ABOVE; D; HAH
06C0; HAMZA ON HEH; D; HEH   ### was:   R; TEH MARBUTA
06C1; HEH GOAL; D; HEH GOAL
06C2; HAMZA ON HEH GOAL; D; HEH GOAL ### was:  R; HAMZA ON HEH GOAL
06C3; TEH MARBUTA GOAL; D; HEH GOAL ### was:  R; HAMZA ON HEH GOAL
06C4; WAW WITH RING; R; WAW
06C5; WAW WITH BAR; R; WAW
06C6; WAW WITH SMALL V; R; WAW
06C7; WAW WITH DAMMA; R; WAW
06C8; WAW WITH ALEF ABOVE; R; WAW
06C9; WAW WITH INVERTED SMALL V; R; WAW
06CA; WAW WITH 2 DOTS ABOVE; R; WAW
06CB; WAW WITH 3 DOTS ABOVE; R; WAW
06CC; DOTLESS YEH; D; YEH  ## has two dots below in init and medi forms
06CD; YEH WITH TAIL; R; YEH WITH TAIL
06CE; YEH WITH SMALL V; D; YEH
06CF; WAW WITH DOT ABOVE; R; WAW
06D0; YEH WITH 2 DOTS VERTICAL BELOW; D; YEH
06D1; YEH WITH 3 DOTS BELOW; D; YEH
06D2; YEH BARREE; R; YEH BARREE
06D3; HAMZA ON YEH BARREE; R; YEH BARREE
06D5; AE; R; HEH    ### was: 06D5; AE; R; TEH MARBUTA
06DD; ARABIC END OF AYAH; U; <span>
      #06E5; ARABIC SMALL WAW; M; <tashkeel>??? no
      #06E6; ARABIC SMALL YEH; M; <tashkeel>??? no
06EE; DAL WITH INVERTED V; R; DAL
06EF; REH WITH INVERTED V; R; REH
06FA; SEEN WITH DOT BELOW AND 3 DOTS ABOVE; D; SEEN
06FB; DAD WITH DOT BELOW; D; SAD
06FC; GHAIN WITH DOT BELOW; D; AIN
06FF; HEH WITH INVERTED V; D; KNOTTED HEH    ### KNOTTED?


# Syriac characters

070F; SYRIAC ABBREVIATION MARK; T; <span>
0710; ALAPH; R; ALAPH
0712; BETH; D; BETH
0713; GAMAL; D; GAMAL
0714; GAMAL GARSHUNI; D; GAMAL
0715; DALATH; R; DALATH RISH
0716; DOTLESS DALATH RISH; R; DALATH RISH
0717; HE; R; HE
0718; WAW; R; SYRIAC WAW
0719; ZAIN; R; ZAIN
071A; HETH; D; HETH
071B; TETH; D; TETH
071C; TETH GARSHUNI; D; TETH
071D; YUDH; D; YUDH
071E; YUDH HE; R; YUDH HE
071F; KAPH; D; KAPH
0720; LAMADH; D; LAMADH
```

```
0721; MIM; D; MIM
0722; NUN; D; NUN
0723; SEMKATH; D; SEMKATH
0724; FINAL SEMKATH; D; FINAL SEMKATH
0725; E; D; E
0726; PE; D; PE
0727; REVERSED PE; D; REVERSED PE
0728; SADHE; R; SADHE
0729; QAPH; D; QAPH
072A; RISH; R; DALATH RISH
072B; SHIN; D; SHIN
072C; TAW; R; TAW
072D; PERSIAN BHETH; D; BETH
072E; PERSIAN GHAMAL; D; GAMAL
072F; PERSIAN DHALATH; R; DALATH RISH
074D; SOGDIAN ZHAIN; R; ZHAIN
074E; SOGDIAN KHAPH; D; KHAPH
074F; SOGDIAN FE; D; FE

# Other

200D; ZERO WIDTH JOINER; C; <no shaping>
200C; ZERO WIDTH NON-JOINER; U; <no shaping>
```