L2/10-045

Proposal for changes to ArabicShaping.txt to allow machine generation of Arabic fonts and glyphs

by Adil Allawi, Diwan Software Limited
adil@diwan.com

Introduction

One of the big problems for Arabic text rendering on computers is the large number of characters in the Arabic script blocks. There are about 200 separate characters with each character having up to four display forms. With some systems requiring Arabic presentation forms to be supported as well, this means a font needs to support around 1500 glyphs to cover the full range of Arabic scripts. As a result all Arabic characters are not fully supported in many fonts and systems. The basic Arabic character shapes can be defined with only 90 consonant forms as well as 20 different dots and marks. With automation, it is possible to construct all these glyphs with relatively little human interaction.

The Unicode Character Database (UCD) file Arabic Shaping.txt provides nearly all the information needed for this kind of automated Arabic font creation but has some inconsistencies which hinder its use in font software.

This proposal is to make minor modifications to this file to allow its machine processing for creating Arabic fonts without breaking its compatibility with existing systems.

A. Generating Arabic glyphs from the Schematic Name

The Method

Like the Roman script, the Arabic script is built from of a small group of basic shapes that are used with various marks to make each character. ArabicShaping.txt includes a 'Schematic Name' for each character. The schematic name defines what marks are drawn on a basic shape for each character.

So for the Unicode character U+067A, the Unicode name is “ARABIC LETTER TTEHEH”. The Schematic Name is: “TEH WITH 2 DOTS VERTICAL ABOVE”.

The great majority of these schematic names are made with a simple consistent grammar. It is possible for simple lexical analyzer to parse these names into a set of tokens that can be used to construct a glyph.

e.g. for the line in “ArabicShaping.txt”:

```
068F; DAL WITH 3 DOTS ABOVE DOWNWARD; R; DAL
```

the name, “DAL WITH 3 DOTS ABOVE DOWNWARD”, is processed into the following tokens:

<ARABIC LETTER DAL>,
<ABOVE>, and
<THREE DOTS DOWNWARDS>

These tokens can be processed into a definition to be used as input for a program to generate the new glyphs (e.g. the Apple font tool ftxenhancer). The above example is composed of: U+062F (Dal) combined with a glyph named “threeDotsDown” and the combining glyph is placed above the main glyph. In XML it looks like this:

```
<newGlyph name="u068f.dalWith3DotsAboveDownward" Unicode="U+068f">
    <pieceGlyph glyphRefID="dal">
        <position X="0" Y="0"></position>
    </pieceGlyph>
    <pieceGlyph glyphRefID="threeDotsDown" linkToPrior="yes">
        <position X="3" Y="5" useZones="yes"></position>
    </pieceGlyph>
</newGlyph>
```

The Problem

The great majority of the schematic names can be processed but there are several names that are inconsistent with the common grammar of the schematic names that hinders machine processing.

For Example:

I list below a some character with their schematic name from Arabic Shaping.txt and the visual appearance.

Character	Schematic Name	Appearance
U+062A	TEH	ت
U+0679	TEH WITH SMALL TAH	ٹ
U+067C	TEH WITH RING	ټ
U+0768	NOON WITH SMALL TAH	ڻ
U+0760	FEH WITH 2 DOTS BELOW	ٿ

Character	Schematic Name	Appearance
U+06A3	FEH WITH DOT BELOW	ڣ

(table 1)

Each of these schematic names follow the same grammar <basic shape> <with> <mark> <position>. However, in each case the basic shape implies dots (TEH has two dots above, NOON and FEH have one dot above) but there is no consistent way to know if the mark replaces the existing dots or not. TEH WITH SMALL TAH does not have the two dots of the TEH but NOON WITH SMALL TAH keeps the dot of the NOON; FEH WITH 2 DOTS BELOW does not have a dot above but FEH WITH DOT BELOW does have a dot above, and so on. This means that any machine processing must include many special cases.

The proposal

If the schematic name is to be used for machine processing then it needs to be built from a grammar based on a the following principles:

1. consistent descriptive names
2. a set of well-defined assumptions
3. simple grammar

1. The schematic name for U+06CC is DOTLESS YEH. This is a problem as U+06CC has dots in the initial and medial contextual forms. I suggest the correct schematic name is FARSI YEH.

There are some cases where different words meaning the same thing are used. This proposal would be to use only one word. E.g. some names use “HORIZONTALLY” and some use “HORIZONTAL”. I would rename to only use “HORIZONTAL”

2. Assumptions :

(Listed with the Unicode value followed by the Schematic name)

Have one dot below:
- U+0628 BEH,
- U+062c JEEM

Have two dots above:
- U+0629 TEH MARBUTA,
- U+062a TEH,
- U+0642 QAF

Have three dots above:
- U+062b THEH
- U+0634 SHEEN

Have one dot above:

U+062e KHAH
U+0630 THAL
U+0632 ZAIN
U+0636 DAD
U+0638 ZAH
U+063a GHAIN
U+0641 FEH
U+0646 NOON

Has two dots below:

U+064a YEH

Has two dots below in initial and medial forms only:

U+06cc FARSI YEH

If a character that is built from one of the above characters has dots added in the same position as the assumed dots the new dots replace the assumed ones.

If a character has dots added in a different location but no dots in the basic shape, then the basic shape should be called “DOTLESS”. So U+0760 is currently called FEH WITH 2 DOTS BELOW. As there is no dot above (see table 1), it should be renamed to DOTLESS FEH WITH 2 DOTS BELOW. While the name may sound contradictory, as the FEH is normally a character with a dot above the point will be clear that the FEH is dotless.

A mark is assumed to be above if the position is not defined (except for RING which has to be treated as a special case). So, NOON WITH SMALL TAH, the SMALL TAH is assumed to be above the NOON.

Three dots will point up above a base glyph and point down below a base glyph.

3. Limit the grammar to:

```
<schematic name> ::= <glyph expression> | "mark" ON/UNDER <glyph expression>
| “mark” <glyph expression>
<glyph expression> ::= "base glyph" | "base glyph" WITH <mark expression>
<mark expression> ::= "mark" | "mark" ABOVE/BELOW  |  <mark expression> AND
<mark expression>
```

On the basis of these rules only 31 schematic names out of the 205 Arabic names need to be changed. These are as follows:

Unicode	Original Name	New Name	Reason
075F	AIN WITH 2 DOTS VERTICALLY ABOVE	AIN WITH 2 DOTS VERTICAL ABOVE	Vertical / Vertically consistency
0675	HIGH HAMZA ALEF	HIGH HAMZA ON ALEF	Grammar - add ON

Unicode	Original Name	New Name	Reason
0679	TEH WITH SMALL TAH	DOTLESS TEH WITH SMALL TAH	There are no dots above so use DOTLESS
067D	TEH WITH 3 DOTS ABOVE DOWNWARD	TEH WITH 3 DOTS POINTING DOWNWARDS ABOVE	“Pointing Downwards” is easier to parse
067E	TEH WITH 3 DOTS BELOW	BEH WITH 3 DOTS BELOW	Teh has dots above
0750	BEH WITH 3 DOTS HORIZONTALLY BELOW	BEH WITH 3 DOTS HORIZONTAL BELOW	Horzontal / Horizontally consistency
0755	BEH WITH INVERTED SMALL V BELOW	DOTLESS BEH WITH INVERTED SMALL V BELOW	There is no dot so use DOTLESS
0756	BEH WITH SMALL V	DOTLESS BEH WITH SMALL V	There is no dot so use DOTLESS
0759	DAL WITH 2 DOTS VERTICALLY BELOW AND SMALL TAH	DAL WITH 2 DOTS VERTICAL BELOW AND SMALL TAH	Vertical / Vertically consistency
06A2	FEH WITH DOT MOVED BELOW	DOTLESS FEH WITH DOT BELOW	There is no FEH dot so use DOTLESS
0760	FEH WITH 2 DOTS BELOW	DOTLESS FEH WITH 2 DOTS BELOW	There is no FEH dot so use DOTLESS
06A5	FEH WITH 3 DOTS BELOW	DOTLESS FEH BASE WITH 3 DOTS BELOW	There is no FEH dot so use DOTLESS
0761	FEH WITH 3 DOTS POINTING UPWARDS BELOW	DOTLESS FEH WITH 3 DOTS POINTING UPWARDS BELOW	There is no FEH dot so use DOTLESS
06AB	KAF WITH RING	KEHEH WITH RING	Consistency - KAF implies KAF shaping
0683	HAH WITH MIDDLE 2 DOTS	HAH WITH 2 DOTS BELOW	Middle can be assumed to be below for HAH
0684	HAH WITH MIDDLE 2 DOTS VERTICAL	HAH WITH 2 DOTS VERTICAL BELOW	Middle can be assumed to be below for HAH
0686	HAH WITH MIDDLE 3 DOTS DOWNWARD	HAH WITH 3 DOTS BELOW	Middle can be assumed to be below for HAH + 3 Dots below can be assumed to point downwards
0687	HAH WITH MIDDLE 4 DOTS	HAH WITH 4 DOTS BELOW	Middle can be assumed to be below for HAH
06BF	HAH WITH MIDDLE 3 DOTS DOWNWARD AND DOT ABOVE	HAH WITH 3 DOTS BELOW AND DOT ABOVE	Middle can be assumed to be below for HAH + 3 Dots below can be assumed to point downwards
076F	HAH WITH SMALL TAH AND 2 DOTS	HAH WITH SMALL TAH BELOW AND 2 DOTS ABOVE	Clarity - dots and Small Tah do not have a position
06C3	TEH MARBUTA GOAL	HEH GOAL WITH 2 DOTS ABOVE	Proper schematic name
06FF	HEH WITH INVERTED V	KNOTTED HEH WITH INVERTED V	Proper schematic name

Unicode	Original Name	New Name	Reason
076B	REH WITH 2 DOTS VERTICALLY ABOVE	REH WITH 2 DOTS VERTICAL ABOVE	Vertical / Vertically consistency
075B	REH WITH STROKE	REH WITH BAR	Stroke/Bar consistency
076D	SEEN WITH 2 DOTS VERTICALLY ABOVE	SEEN WITH 2 DOTS VERTICAL ABOVE	Vertical / Vertically consistency
06C0	HAMZA ON HEH	HAMZA ON AE	Proper schematic name
063D	FARSI YEH WITH INVERTED V	DOTLESS YEH WITH INVERTED V	Does not draw with two dots below in initial and medial forms
063E	FARSI YEH WITH 2 DOTS ABOVE	DOTLESS YEH WITH 2 DOTS ABOVE	Does not draw with two dots below in initial and medial forms
063F	FARSI YEH WITH 3 DOTS ABOVE	DOTLESS YEH WITH 3 DOTS ABOVE	Does not draw with two dots below in initial and medial forms
06CC	DOTLESS YEH	FARSI YEH	Dotless implies no dots in all forms - but farsi yeh draws with dots in initial and medial forms
06CE	YEH WITH SMALL V	DOTLESS YEH WITH SMALL V	Does not draw with two dots below in initial and medial forms

B: Changes to the Joining Group

These are more sensitive to modification. However there is one glaring inconsistency that really needs to be addressed:

U+077A and U+077B have the joining group BURUSHASKI YEH BARREE.

This should be: YEH BARREE.

also the header to ArabicShaping.txt notes Joining_Group HAMZA ON HEH GOAL is anachronistically named for historical reasons. If these reasons no longer apply then it should be replaced with the joining group HEH GOAL.

C. Generating font shaping data

The Problem

For AAT, OpenType and other text layout methods, it is important to know which Arabic characters are combiners so that shaping is not broken at these characters. To get these one must parse two files, UnicodeData.txt and DerivedJoiningType.txt. First one would parse DerivedJoiningType.txt to get the ranges of characters with “Join-

ing_Type=Transparent" then to parse UnicodeData.txt to get the individual names of these characters.

A much simpler route would be to list the Arabic and Syriac combiners in ArabicShaping.txt. Also by listing the combining characters it would be clear which characters must be transparent for shaping in that language.

The proposal

I suggest adding the following lines to define all the relevant combining characters:

```
0610; SIGN SALLALLAHOU ALAYHE WASSALLAM; T; No_Joining_Group
0611; SIGN ALAYHE ASSALLAM; T; No_Joining_Group
0612; SIGN RAHMATULLAH ALAYHE; T; No_Joining_Group
0613; SIGN RADI ALLAHOU ANHU; T; No_Joining_Group
0614; SIGN TAKHALLUS; T; No_Joining_Group
0615; SMALL HIGH TAH; T; No_Joining_Group
0616; SMALL HIGH LIGATURE ALEF WITH LAM WITH YEH; T; No_Join-
ing_Group
0617; SMALL HIGH ZAIN; T; No_Joining_Group
0618; SMALL FATHA; T; No_Joining_Group
0619; SMALL DAMMA; T; No_Joining_Group
061A; SMALL KASRA; T; No_Joining_Group
064B; FATHATAN; T; No_Joining_Group
064C; DAMMATAN; T; No_Joining_Group
064D; KASRATAN; T; No_Joining_Group
064E; FATHA; T; No_Joining_Group
064F; DAMMA; T; No_Joining_Group
0650; KASRA; T; No_Joining_Group
0651; SHADDA; T; No_Joining_Group
0652; SUKUN; T; No_Joining_Group
0653; MADDA ABOVE; T; No_Joining_Group
0654; HAMZA ABOVE; T; No_Joining_Group
0655; HAMZA BELOW; T; No_Joining_Group
0656; SUBSCRIPT ALEF; T; No_Joining_Group
0657; INVERTED DAMMA; T; No_Joining_Group
0658; MARK NOON GHUNNA; T; No_Joining_Group
0659; ZWARAKAY; T; No_Joining_Group
065A; VOWEL SIGN SMALL V ABOVE; T; No_Joining_Group
065B; VOWEL SIGN INVERTED SMALL V ABOVE; T; No_Joining_Group
065C; VOWEL SIGN DOT BELOW; T; No_Joining_Group
065D; REVERSED DAMMA; T; No_Joining_Group
065E; FATHA WITH TWO DOTS; T; No_Joining_Group
0670; ALEF ABOVE; T; No_Joining_Group
06D6; SMALL HIGH LIGATURE SAD WITH LAM WITH ALEF MAKSURA; T;
No_Joining_Group
```

```
06D7; SMALL HIGH LIGATURE QAF WITH LAM WITH ALEF MAKSURA; T;
No_Joining_Group
06D8; SMALL HIGH MEEM INITIAL FORM; T; No_Joining_Group
06D9; SMALL HIGH LAM ALEF; T; No_Joining_Group
06DA; SMALL HIGH JEEM; T; No_Joining_Group
06DB; SMALL HIGH THREE DOTS; T; No_Joining_Group
06DC; SMALL HIGH SEEN; T; No_Joining_Group
06DE; START OF RUB EL HIZB; T; No_Joining_Group
06DF; SMALL HIGH ROUNDED ZERO; T; No_Joining_Group
06E0; SMALL HIGH UPRIGHT RECTANGULAR ZERO; T; No_Joining_Group
06E1; SMALL HIGH DOTLESS HEAD OF KHAH; T; No_Joining_Group
06E2; SMALL HIGH MEEM ISOLATED FORM; T; No_Joining_Group
06E3; SMALL LOW SEEN; T; No_Joining_Group
06E4; SMALL HIGH MADDA; T; No_Joining_Group
06E7; SMALL HIGH YEH; T; No_Joining_Group
06E8; SMALL HIGH NOON; T; No_Joining_Group
06EA; EMPTY CENTRE LOW STOP; T; No_Joining_Group
06EB; EMPTY CENTRE HIGH STOP; T; No_Joining_Group
06EC; ROUNDED HIGH STOP WITH FILLED CENTRE; T; No_Joining_Group
06ED; SMALL LOW MEEM; T; No_Joining_Group

# Syriac characters

070F; ABBREVIATION MARK; T; No_Joining_Group
0711; LETTER SUPERSCRIPT ALAPH; T; No_Joining_Group
0730; PTHAHA ABOVE; T; No_Joining_Group
0731; PTHAHA BELOW; T; No_Joining_Group
0732; PTHAHA DOTTED; T; No_Joining_Group
0733; ZQAPHA ABOVE; T; No_Joining_Group
0734; ZQAPHA BELOW; T; No_Joining_Group
0735; ZQAPHA DOTTED; T; No_Joining_Group
0736; RBASA ABOVE; T; No_Joining_Group
0737; RBASA BELOW; T; No_Joining_Group
0738; DOTTED ZLAMA HORIZONTAL; T; No_Joining_Group
0739; DOTTED ZLAMA ANGULAR; T; No_Joining_Group
073A; HBASA ABOVE; T; No_Joining_Group
073B; HBASA BELOW; T; No_Joining_Group
073C; HBASA-ESASA DOTTED; T; No_Joining_Group
073D; ESASA ABOVE; T; No_Joining_Group
073E; ESASA BELOW; T; No_Joining_Group
073F; RWAHA; T; No_Joining_Group
0740; FEMININE DOT; T; No_Joining_Group
0741; QUSHSHAYA; T; No_Joining_Group
0742; RUKKAKHA; T; No_Joining_Group
0743; TWO VERTICAL DOTS ABOVE; T; No_Joining_Group
0744; TWO VERTICAL DOTS BELOW; T; No_Joining_Group
```

```
0745; THREE DOTS ABOVE; T; No_Joining_Group
0746; THREE DOTS BELOW; T; No_Joining_Group
0747; OBLIQUE LINE ABOVE; T; No_Joining_Group
0748; OBLIQUE LINE BELOW; T; No_Joining_Group
0749; MUSIC; T; No_Joining_Group
074A; BARREKH; T; No_Joining_Group

# N'Ko Characters

07EB; NKO SHORT HIGH TONE; T; No_Joining_Group
07EC; NKO SHORT LOW TONE; T; No_Joining_Group
07ED; NKO SHORT RISING TONE; T; No_Joining_Group
07EE; NKO LONG DESCENDING TONE; T; No_Joining_Group
07EF; NKO LONG HIGH TONE; T; No_Joining_Group
07F0; NKO LONG LOW TONE; T; No_Joining_Group
07F1; NKO LONG RISING TONE; T; No_Joining_Group
07F2; NKO NASALIZATION MARK; T; No_Joining_Group
07F3; NKO DOUBLE DOT ABOVE; T; No_Joining_Group
```

References

ArabicShaping.txt: http://unicode.org/Public/UNIDATA/ArabicShaping.txt

DerivedJoiningTypes.txt: http://unicode.org/Public/UNIDATA/extracted/DerivedJoiningType.txt

UnicodeData.txt: http://unicode.org/Public/UNIDATA/UnicodeData.txt

AAT font specification: http://developer.apple.com/textfonts/TTRefMan/RM06/Chap6.html

OpenType font specification: http://www.microsoft.com/typography/otspec/

Apple Font Tools manual:
http://developer.apple.com/textfonts/FontToolsDocs/Apple%20Font%20Tool%20Suite.pdf

With thanks to The Apple Type Group and Peter Lofting.

Table of Changes:

Modification	Date	Modified By
Fixed 073C; HBASA–ESASA DOTTED	4 Jan 2010	Adil Allawi
Replaced BASE with DOTLESS	10 Jan 2010	Adil Allawi
Minimized list of changes	25 Jan 2010	Adil Allawi
Replaced Socratic dialogue with reasoned argument	25 Jan 2010	Adil Allawi