L2 08-396

Date: 2008-10-28

Comments on L2 /08-343
GHOST CHARACTERS: Atomization- Combination Theory, by Dr. Attash Durrani

Source: IBM (Tarek Abou Aly, Waleed Oransa, Mohamed Mohie, Marwa Aboulfadl, IBM GcoC Arabic Competency Centre, Egypt))

We don't recommend to decompose the Arabic characters with dots as in the proposal for the following reasons:

1. Compatibility and equivalence issues will occur with this new set, including conversion between Unicode and different code pages. The proposal adds an unnecessary level of complexity to the Unicode normalization forms described in annex #15 (http://www.unicode.org/reports/tr15/).

2. Existing Arabic shapes in Unicode could not represent possible combination "20460 as in the document". Shaped Arabic characters are represented Arabic presentation forms A and B (Unicode range uFBxx and uFExx).

3. Keyboard entry and presentation for these characters/dots will be very difficult to user, or would need a significant change to the infrastructure to support the proposed encoding.

4. Sorting and Searching algorithms will be very complex and will include much more processing. An algorithm will need to be defined to find both characters 1- existing Arabic data is stored in 06xx range and 2- The equivalent presentation stored using the suggested proposal.

5. Representing Arabic text using the suggested proposal leads to a significant increase in the Buffer length and storage used. For example the single Unicode character u062B Arabic letter THEH would be decomposed, according to the proposed example in page 9, to **Three** different characters leading to tripling the size of the used memory.

6. The proposal introduces a flexibility that can result in erroneous characters combination from a linguistic perspective. This can easily arise from combining marks to base characters resulting in characters that do not exist in the real world. A strong validation and mistake prevention mechanism needs to be defined to avoid introducing unknown letters resulting from the combination of (letter + some dots)) in the Arabic-script based text.

7. The concept of combining marks already exists in Unicode, like in the following examples:
   Tashkeel or diacritic marks from u064B to u0652.

8. Some of the suggested combining marks are already represented in Unicode, like for example:
   - u065A
   - u065B
   - u065C

9. In page 13 and 14 The proposal suggests restricting the combination for some composed characters with other combining marks. Will there be an algorithm/rules or will it be loosely defined?