Unicode in Education

Adil Allawi
Technical Director
adil@diwan.com

Why Education and Unicode:

- Cost of printing and distribution is high - paper is becoming a scarce resource
- Many classes only need part of a book - less waste
- More than any other kind of book - educational books need to be reused for different media - print/web/ebook etc
- Encoding must be right or text is lost for the future
- What else do you put on “one laptop per child” ?

So why not use PDF's?

- The good
  - a lot of the documents already exist in electronic form that prints to Postscript or can be scanned and saved as PDF
- The bad
  - limited to the feature set of the browser/library
  - royalties drive up the end user costs - and for education they have to be low
  - it sidesteps the real issue
  - Without properly handling the encoding and structure you end up with just an image of the document
  - text flow is lost, links between items are lost.

Encoding and rendering

- Take away everything and we have a story of encodings
- Books contain handwritten examples.
- Fonts created were built just to render the book in print
- Need to convert simple and practical encodings
- Trouble with Bidi Algorithms

Simple encodings

- Mostly easy to encode into Unicode
- No new ones being created
- For Arabic there are a few:
  - 2 Mac Arabic ones (System 6 and 7)
  - Windows Arabic
  - ISO-8859-6
  - DOS Arabic
  - 3B2 (an odd 16-bit encoding)

Practical encodings

- They work around system/application restrictions
- Font and Keyboard based
- Difficult to re-encode
- New ones are being created all the time
  - Unicode has only made this problem more complex

Practical Problems

- Based on a well-known DTP software

- From a well-known institution

FB53	FB54	FB55	FB56	FB57	FB58	FB59	FB5A	FB5B	FB5C	FB5D	FB5E	FB5F	FB60	FB61	FB62	FB63	FB64	FB65
وَلَمْ	نَكُ	نُطْعِمُ	مِسْكِينٍ	٤٤	كُنَّ	نَخُوضُ	مَعَ	ٱلْخَآئِضِينَ	٤٥	وَكُنَّا	نُكَذِّبُ	بِيَوْمِ	ٱلدِّينِ	٤٦	حَتَّىٰٓ	أَتَىٰنَا	ٱلْيَقِينُ	٤٧

- Arabic 8-bit ى = U+06CC or U+0649
  - In Mac Arabic it is the same for Alef Maksura or Egyptian-Arabic Yeh or Farsi Yeh.

Unicode Problems

- Many, many encoders use Arabic Presentation Forms A and B for encodings and placing characters
- is ١١/١٢ a date or a fraction?
- also ك ک are essentially the same letter
- so are ى and ي
- But the link is not made

Example - A Math Font

$$\frac{\text{قم}}{\text{ك}} = \text{جم}$$

$$\frac{\text{ع}^{\text{٢}}}{\text{نق}} = \text{جم}$$

Encoding Math Fonts

- Keep the practical encoding for data entry
  - simple - works with existing programs
- re-encode in the file format
  - some glyphs will encode as 2 characters

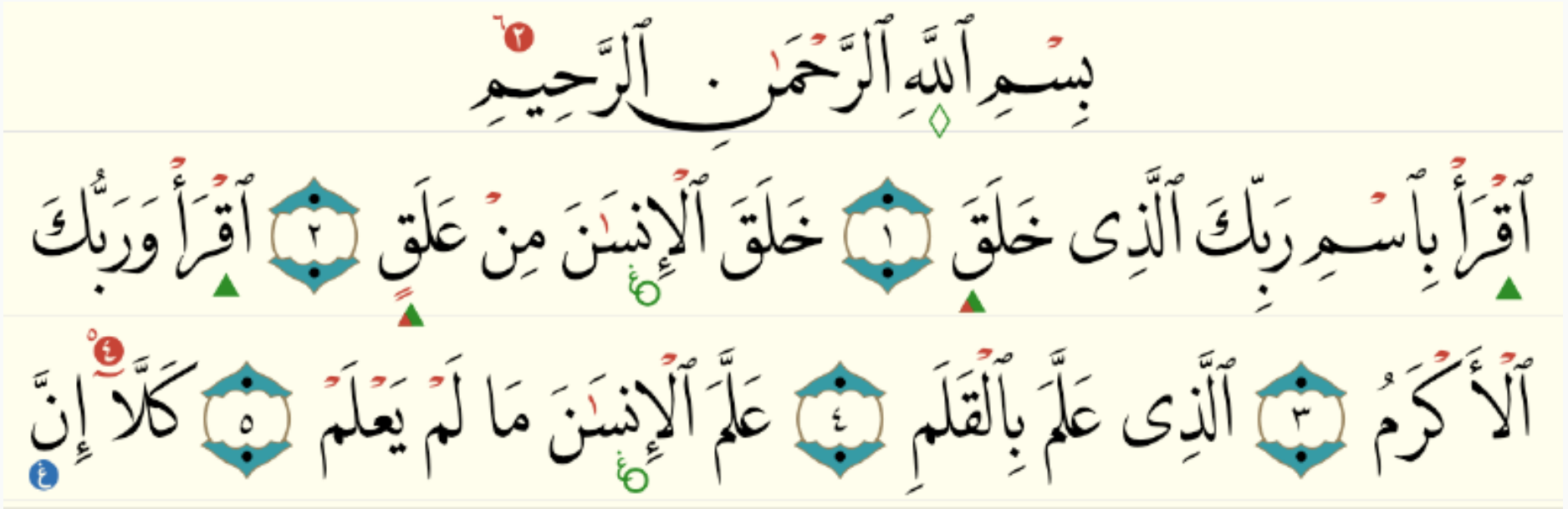
    e.g. initial Ha ـحـ = ح + zwj
  - some need separate markup

    e.g. subscript 2 _٢
- Try to use existing encoding

  Curly Arabic - just matches to normal Arabic
- But what about 

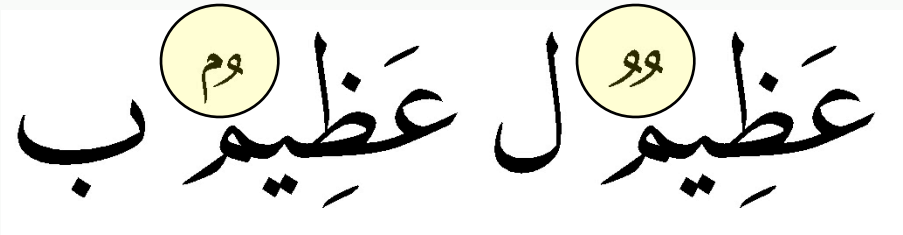
Encoding Quran in Unicode

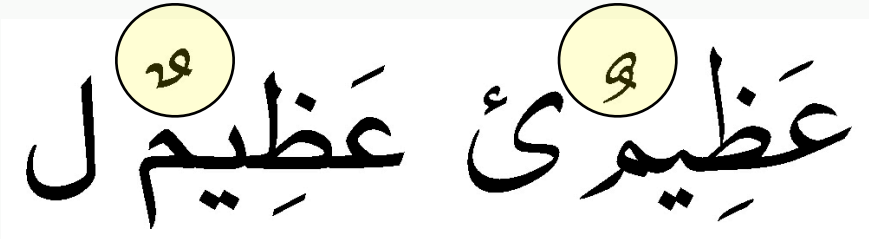
- Announced at the 17th IUC
- Extended to teaching Quran



Encoding Method

- Fitted encoding within the Unicode standard
- Example: Tanween ـًـٍـٌـ is one character but 3 display forms:

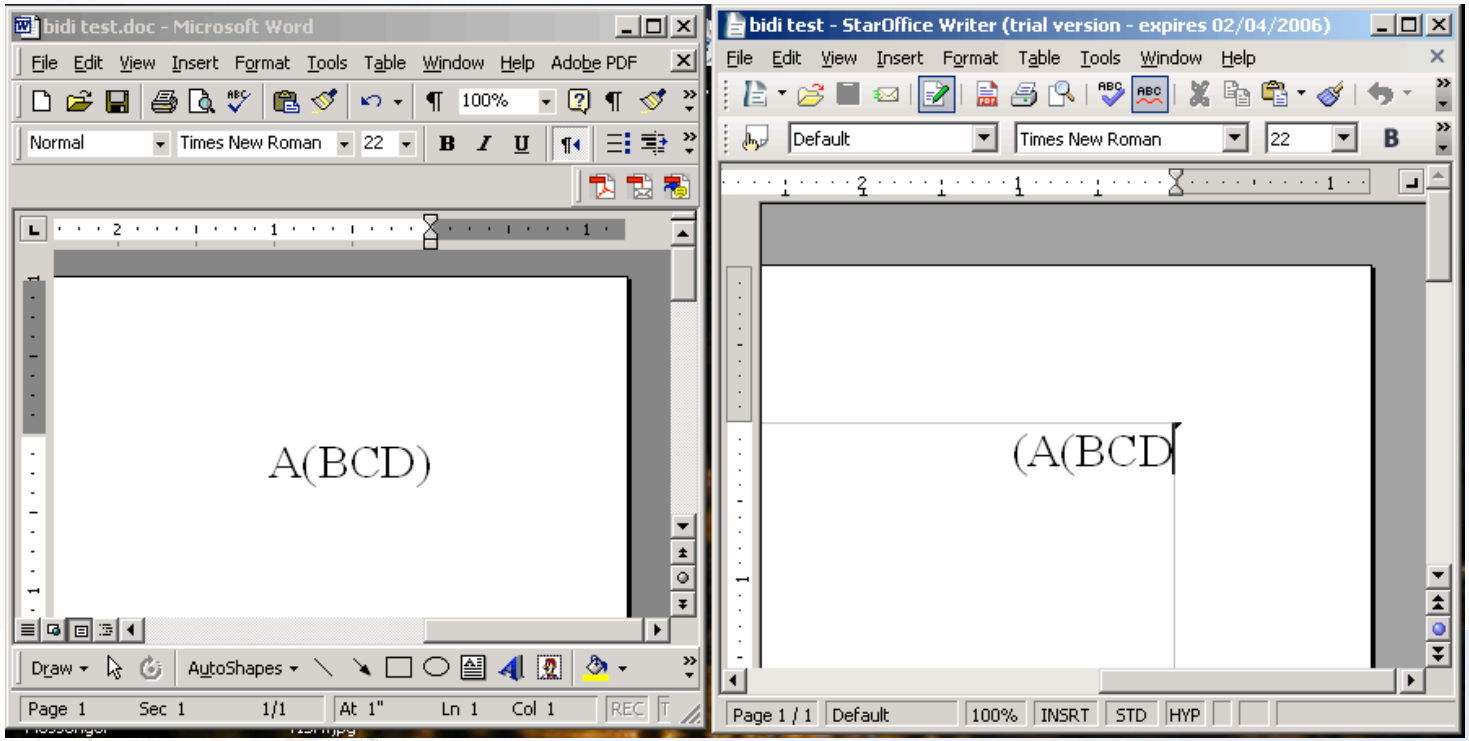




De-bidi the Bidi Algorithm

- Typical problem

  Word uses its own bidi algorithm



De-bidi the bidi algorithm

- There are 4 bi-di algorithms
  - Apple Mac OS 9
  - Windows 95
  - Unicode
  - Word for Windows
- And then there is text after printing
- Need to take apart bi-di'd text and reencode/reorder to work with Unicode Bidi.

Same issue different language

- Iñupiaq

  From Alaska Native Education Program:

  - 004C + 0323 Ḷ
  - 006C + 0200A +0323 ḷ
  - 0141 + 0323 Ł̣
  - 0141 + 0200A + 0323 ł̣
- Thin space is inserted because of rendering issues on web browsers

Solutions

- If the rendering engine does not fit your requirements - create your own.
- Use local knowledge to define the encoding and get it right first time. Don’t let the programmers create workarounds.
- If you really must bypass Opentype or Uniscribe or browser text renderers - create something to convert back to Unicode or all the effort will be lost in time.

Looking to the Future

- Unicode is important for education and for preserving knowledge. Yet very little effort is put into encoding for education outside the major languages.
- We are creating many more encoding issues for the future as people are 'fixing' problems with 'clever' workarounds. Unicode is not stopping this process.
- How will all this be indexed and searched?
- One day the memory will be lost for a particular workaround and the data and research will be damaged for the future.