Encoding the Holy Koran into Unicode

by Adil Allawi
Diwan Software Limited

www.diwan.com
E-mail: adil@diwan.com

Abstract

The central document of the Muslim religion, the Koran, also represents the highest standard for the representation of the Arabic language and typography. The text of the Koran is sacred to Muslims and as such it must be encoded without any errors and in its entirety. Also, a complete Koran should include all the symbols used to indicate its organisation and correct pronunciation. It is for these reasons that any approach in encoding the Koran fully on computers is a non-trivial task. It is not acceptable to compromise on any details from the design of the typeface to the encoding of the text. Diwan Software Limited have an ongoing project for the past three years to make a complete electronic Koran, starting with the creation of a suitable typeface to the completion of the text with all its annotations. The purpose of our paper will be to show how we have approached this subject and to demonstrate the results of our work. We also wish to use the opportunity to start a discussion on how the Unicode standard can best be applied to this purpose and what are its limitations.

Introduction

Throughout history, written Korans have always been very carefully defined to ensure that they remain faithful to the original and that the actual text is never changed. This strict definition makes the Koran a relatively easy document to encode on computers, as the actual letters used are well defined as are their meaning and usage.

Computer encoding of the Koran has run in parallel with the development of the Arabic language on computers. However, all the past and current encodings have been forced to make compromises due to the limitations of existing computer systems. Standard computer character-sets are built to deal with modern Arabic, which has several differences from the Arabic used in the Koran. These differences stem from the Koran having different spelling for certain words. These need to be indicated by the use of specific accents, which do not exist in classic computer encoding. Such encodings have only limited uses while most electronic representation of the Koran for education and publishing is done from photographs of hand-written text.

Unicode is the first standard encoding to include a rich set of characters specifically for Koranic text. This allows for a much more accurate encoding of Koranic text without resorting to compromises such as adding proprietary characters or making two separate characters share the same code point. This allows, for the first time, a fully computerised Koran and all the benefits that this can bring.

While it is possible to write custom software to render the Koran fully this would, in a way, defeat the main purpose of a computerised Koran. That is, to make it accessible to all people and for a wide variety of applications. For example, to make a Koran that can be transferred via the Internet as text and displayed on a client system using standard software. Technologies that will allow this are only now emerging as standard parts of modern operating systems. These are ‘intelligent’ font technologies (started by Apple Computer Inc. with QuickDraw GX) and system-wide Unicode support.

Diwan are not the first or only people to work on computer encoding of the Koran. I believe, however, our approach is unique in that we are attempting to make a computer encoding of the Koran that is as carefully defined as the text in written Korans alongside building typefaces that can render the Koran clearly and correctly. Our intent is for this encoding to form the standard from which all future electronic Korans can begin and to be used as a benchmark for testing typefaces and computer software that are to be used for Islamic applications. Also, we are making proposals for changes in the Unicode standard with an aim to make it possible for any Unicode aware software given the right fonts, to deal correctly with Koranic text without having to be written for this purpose.

Historical Background

One of the main characteristics of early Islam was the use of a revolutionary new technology for its time to bring together widespread and diverse peoples through a single message. At the birth of Islam Arabia was at a low level of technical and urban development. Written Arabic was at an elementary stage, those who could read and write were few and confined to the few small cities like Mecca and Yathrib (later Medina) in western Arabia. The message of Islam was to uproot the various existing beliefs and ways of tribal life replacing them with those of Islam. New ways of worship, social life and urban organisation had to be improvised and technologies borrowed or invented. This message of Islam was codified in a book, the Koran.

In those conditions the very process of transcribing the Koran and collecting the scattered verses into a single document was a historic act. It not only marked the birth of literacy in Arabia but also the technologies of standardisation of learning, book making and binding.

The Koran was first written on a variety of materials like camel shoulder plates, dried palm leaves, hides, slates and papyrus. It had not come as a whole complete book, but through a whole process of “Tanzil” which took about twenty one years. Save for a duration of three years the successive verses were dictated and copied in periods varying between ten to eighty one days.

The conditions existing then did not allow the compilation of a single master from which other copies were made and checked. This had to wait some years until the time of the third caliph, Othman, who, it was said, having seen the need for a unified text, brought together all the written chapters of the Koran and unified them into a single agreed master (Imam). From this other copies were made. The number of copies varies according to different reports, the most accepted is six. The original Imam was kept by the caliph at al-Medina, and the six copies (Masahif al Amsar) were dispatched to each of the six provincial centres (Amsar) to be masters there. From then on the number of Korans multiplied quickly. It was reported that during the famous battle at Siffeen, which took place seven years later, five hundred Korans were raised, in a move to stop the fighting.

The process of codifying the first master copy of the Koran in the seventh century AD meant that the word of the Koran has remained the same without changes to the text throughout the centuries.

The Style and Organisation of the Koran

The earliest script of the Koran was in the Kufi style. This was derived from the pre-Islamic Arab city Hiera. A city close to the later city of Kufa, in present day Iraq. Hiera used an Arabic style based on Syriac. The early Korans were written without diacritical marks and grammatical notations. These were added later.

The Kufi style remained for several centuries the basis for all the variations (about twenty of subtypes) of Koranic calligraphy. This situation changed during the tenth century AD, under the influence of the great calligrapher Ibn Muqla who devised an early form of the Naskh style. The latter style acquired its final shape in the twelfth century. It is the basis of modern Arabic and most if not all the modern Korans**.**

The present and most widely used is the Ottoman Naskh. It is this form that I will concentrate on for the purposes of this paper.

At a very early stage the Koran was divided into Sura’s, chapters, and Aya’s, single sentences. Later on, with the growth of Islam, it became apparent that the whole text needed to be organised into groupings so as to make it more convenient to read or to memorise. It is said that the first new division were made by the direction of al-Hadjaj, the governor of the province of Iraq in the eighth century AD. He ordered that the total number of characters be counted. When this was reported to be around 340,740 characters, Hadjaj directed to have the whole text divided into seven equal parts (Asba’). Several decades later, during the Abbasid times (after the 8th century AD), a new scheme was devised to have thirty parts (Juz’) instead of the seven. Each of the thirty parts was divided into two sub-parts (Hizb), and each of these into four quarters. The new scheme made it much easier and cheaper to reproduce these Juz’s separately as booklets and share them in mosques, as is still done now, especially during the month of Ramadhan. All the new divisions are now indicated on the margins of the pages and sometimes within the text.

The current style of a printed Koran presents four different kinds of information. First is the actual text of the Koran with the diacritical marks that indicate correct grammar and pronunciation. This text is divided into 114 Sura’s and the Sura’s are divided into 6236 Aya’s. Second are additional diacritical marks to indicate the correct recitation of the text. Thirdly are markers to indicate the divisions of the Ayah and the quarter of a Hizb. Lastly there is a marker which indicates an action for prayer called ‘Sajdah’.

A Unicode Koran

At its simplest level this is an encoding and a typeface. An ideal encoding should meet the following requirements:

- The text of the Koran should be plain Unicode text without any additional formatting. This is necessary so the Koran can be stored as a simple format that can be understood by any Unicode enabled system without the need for additional formatting markers. Additional formatting information should only need to be used to define design alternatives – for example the selection of the wide-form of a character or alternative ligatures.

- The typeface should support at least the minimum necessary features for Koranic calligraphy. Such a minimum needs to be well defined so that any typeface manufacturer can check their fonts against this to be able to claim that their font can be used for Koranic calligraphy.

A complete Unicode Koran would revolutionise the Islamic printing industry. Most printed Korans originate from hand-written text. By starting from a standard electronic text it would be possible to print Korans without having to go through the lengthy process of typing text and checking for errors. Also, an electronic Koran can be used in a wide variety of publications that need to print quotations from it. Currently, such publications rely on images of the Koran being pasted into the text or transliterating the text of the Koran into modern Arabic. With a standard text it would be possible to develop new computer typefaces that can verifiably support Koranic calligraphy and to develop applications that can work with any of these typefaces.

In the educational field, an electronic Koran can create the possibility for many new tools for teaching the Koran and for research into its meaning. For example, the text can be indexed and archived to a database with links to relevant references such as commentaries, translations and historical backgrounds.

Most importantly, a Unicode Koran makes it possible to work from a single and accurate text that correctly defines all the features of the printed Koran. Such a text can be check-summed and easily verified so that errors cannot be added by accident.

This is the intention of out work. However, I will argue in the following section that the current Unicode standard needs some additions and correction in order to meet these goals.

How We Did It

To be honest, the wrong way around.

Ideally we should have started with a full study of the text of the Koran, then created the encoding in Unicode. As part of this step one would work out what limitations exist in using Unicode and how to work around them. The next step would be to build an application to let us enter and correct the text. Finally one would make a typeface that will display the encoding correctly with its full calligraphy.

Diwan has started by developing a complex, calligraphic Arabic typeface then adjusted this to be able to render a complete Koran. We then went on to encode the Koran using an extended form of the 8-bit Arabic encodings ISO 8859-6 on Macintosh and the Microsoft Arabic code-page on Windows. We have finally gone on to defining the Koran as Unicode. This practical approach as opposed to a scientific approach has taken us longer to reach the goal of making a Unicode Koran. It has, however, resulted in us having a complete solution being available now. It also serves as a good example that can be used to illustrate the problems we have with using Unicode as an encoding.

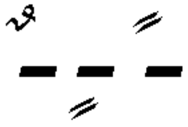
Before attempting this we did not have a calligrapher that understood both computer fonts and Koranic calligraphy. We had to train an existing calligrapher to build such a font. Also we had to train ourselves on how such calligraphy could be assembled by the computer. Finally this all needed to be built into an industry-strength publishing application so that we can make practical examples to our calligrapher and test the font.

To encode the Koran fully it is necessary to include the different kinds of information that I identified previously; the text including the essential diacritics; the diacritics for recitation; the Ayah markers; the quarter Hizb markers; and the Sajdah. However, it should be possible to select what information is needed. For example, when quoting passages from the Koran it is only necessary to include the text with essential diacritics and the Ayah markers. Also, for the purposes of indexing and computer processing one only needs the text while the other information should be ignored or marked separately from the text. However, if the text is to be used for recitation one needs all the recitation markers. In order to do this it is necessary to define what each character is so that it is clear what to omit for each purpose. Such a definition is not made in the Unicode standard and needs to be done separately.

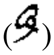
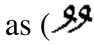
Finally, Koranic calligraphic rules differ from standard Arabic rules. A typeface that is designed to work with Koranic text needs the additional symbols for this purpose and the program that renders Koranic text needs to be aware of these rules. A clear definition of these rules should be included alongside the encoding.

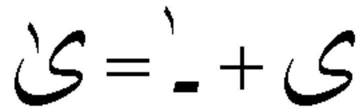
Issues when using Unicode for Encoding the Koran

Unicode provides a rich set of characters for encoding the Koran. However, in order to meet the requirement of having a text that does not require any additional formatting we have noticed some limitations. Following are the main problems we have experienced. I have included this section not for the purpose of debate as this is more appropriate for a standards committee to discuss but to illustrate the problems that have affected our approach to the encoding.

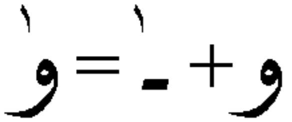


1. ‘Tanween’ – these are the characters Fathatan (U+064B), Kasratan (U+064D) and Dammatan (U+064C). These are diacritics that are only ever used at the end of a word. In the Koran these have alternate shapes depending on their pronunciation due to the sound of the following word. If the Tanween are drawn directly above each other (e.g.) then

the sound is complete. If they are drawn slightly apart (e.g.) then the sound has some variation according to the first letter of the next word. Although both cases are exactly the same character the use of the different form is well defined due to a different sound being. In the Koran the Dammatan is either drawn as: () or as () for the same reasons. This is a different presentation form to the Dammatan defined in Unicode ( ). It is debatable whether this distinction is enough reason to argue for additional code points. We are currently treating this as one letter and the alternate form as graphical variants defined by the formatting. I believe it would be preferable to define them without needing to use separate formatting as the different forms appear in specific places in the text of the Koran and are not optional.



2. Superscript Alef (U+0670) and Alef Maqsura (U+0649). This is drawn in a place where an Alef is pronounced but omitted from the word. It can be drawn like a diacritic or it is also drawn above the Alef Maqsura in its initial, medial and final positions. However the final Alef Maqsura is sometimes drawn without the superscript Alef. Using the Unicode definition I can define Alef Maqsura with a superscript Alef as two characters (U+0649, U+0670). However, here I am using two characters to define a single letter. Some existing Arabic fonts already define the final and isolated Alef Maqsura including the superscript Alef. To be able to include these encodings and for the purposes of easing searching and comparison, I propose that an additional character is defined called Alef Maqsura With Superscript Alef. This character would be defined as identical in meaning to the Alef Maqsura (U+0649). The pair (U+0649, U+0670) can then be considered an equivalence sequence of this new character. Alef Maqsura at the beginning and middle of a word does not exist in modern Arabic so this is replaced with an Alef character when transliterating the Koran in modern Arabic. Having a single letter that defines this will be helpful to aid this transliteration.



3. Superscript Alef (U+0670) and Waw (U+0648). A Waw followed by a superscript Alef is also a single character that is pronounced as an Alef but indicates the existence of a Waw in the root of the word. As above, the two shapes together also define a single letter and I would argue for the creation of a new character on the same terms. Having a unique character that is 'Waw with superscript Alef' will allow for that equivalence to be shown in the Unicode standard as this is interchangeable with the Alef without changing the meaning of the word. i.e. a word that contains Waw with superscript Alef is identical to a word that contains an Alef in the same position when performing a Unicode search of text that contains Koranic text as well as modern Arabic text. Another reason for a separate encoding is one of clarity. The superscript Alef is used in two ways in the Koran; to indicate a missing letter; and, in this case, as an accent above the Waw. It is possible to have a Waw, which is followed by the former form of the superscript Alef. Unless this case is well defined it can be unclear whether a letter is a Waw to be pronounced as an Alef or it is a Waw to be followed by an Alef.

4. In many existing typefaces the phrase “Bismillah ar-Rahman ar-Rahim” which heads

   بِسۡمِ ٱللَّهِ ٱلرَّحۡمَٰنِ ٱلرَّحِيمِ

   every Sura of the Koran is defined as a single character. It would be useful if this was also defined in Unicode.

Koranic Letter Names and their Definitions

The Unicode standard defines the names of most of the Koranic annotation signs (U+06D6 to U+06ED) by how they look as opposed to their correct name or function. The worst case is U+06E1 called “ARABIC SMALL HIGH DOTLESS HEAD OF KHAH”. This is really an alternate form of the Sukun (U+0652) and its relation to the Sukun should be noted. There are other examples but this goes beyond the scope of this paper.

When I described our approach to the encoding I defined four main types of information that is written in the Koran. While all the necessary characters exist in Unicode it would be helpful if the useage and meaning of the non-text characters was also mentioned in the standard. This would enable the writing of software to perform searches and computer processing of Koranic text without the need for specialised information about the Koran.

Defining a Koranic font

Arabic Presentation Forms (U+FB50 to U+FDFF AND U+FE70 to U+FEFF) were originally defined for compatibility with existing standards and the Unicode book recommends that they are used to allow conversion back to the standard forms of the letters. However, now these shapes are being used by many organisations to define the shapes for their basic Arabic shaping and display support. Apple’s ATSUI, Java 2 Text Layout and (I believe) Microsoft’s Unitype will use glyphs in a font associated with these code positions to render shaped Arabic text. In my work creating an Arabic typeface for the Siemens S35 mobile phone I also found this a useful way to define the Arabic glyphs and communicate glyph information to programmers otherwise ignorant of Arabic. Although this is an incorrect use of these characters they are now being used across the industry this way. This is a positive development but it should be formalised.

I would propose that a new encoding (possibly separate from Unicode) which defines the minimum glyph-set for proper rendering of the Koran. Such an encoding is important as it would help with the creation of typefaces that can properly define Koranic calligraphy and make it possible to use a Unicode Koran on many more system. For this, I would propose a reduced version of the Diwan Koranic typeface.

Conclusion

Building a Unicode Koran is not a task that can be taken lightly. Diwan have spent about 2 years developing a suitable typeface for Koranic calligraphy. Entering the text and formatting the typeface calligraphically (selecting alternate ligatures, positioning diacritics more optimally etc.) has taken another six months. Proper checking by an expert who specialises in checking Korans takes another four months which is the stage Diwan is at, at the time of writing this document. After this the new Koran must gain certification of the religious authorities such as Al-Azhar in Egypt.

Unicode definitely represents a very positive development for the encoding of the Koran on computers. The modifications I have suggested are minor and Unicode as it stands now does represent a workable solution for encoding the Koran. I hope this paper represents a start at making a full definition of the Koran for the computer age.