Typesetting the Qur'an and its specific challenges to the TEX family*

Hossam A. H. Fahmy
Electronics and Communications Department,
Faculty of Engineering, Cairo University, Egypt
hfahmy@arith.stanford.edu

1 Peculiarity of Arabic typography

The arabic alphabet has been adopted for use by many languages in Africa and Asia including: Arabic, Dari, Farsi, Jawi, Kashmiri, Pashto, Punjabi, Sindhi, Urdu, and Uyghur. The arabic script is also used for a number of other languages either to present how the language used to be written historically (as for Turkish) or how some write it in an unofficial manner (as for Hausa in western Africa).

Similar to the latin alphabet, with its adoption to several languages, the arabic alphabet acquired new symbols to represent the sounds that do not exist in Arabic. In contrast to the latin alphabet that has dots only on the 'i' and 'j', the arabic alphabet uses dots extensively both above and below the letter shapes to distinguish the different characters. This explicit distinction between the different letters in written text using dots was in itself an addition to the original script which had no dots. In the original arabic script, the distinction between بيت (house) and بنت (girl) when represented as ٮٮٮ was understood from the context. In general, the developed symbols for the other languages follow the same idea as Arabic and use more dots (up to four) and special marks on the original shapes of the arabic letters.

Arabic being a semitic language, only the consonants are usually written in a word. The equivalent of short vowel sounds are written as additional marks on top of the letters. Obviously, the different languages have different vowels and need different symbols to mark them. In addition to that, since the geographical area covered by the arabic script historically is quite vast, different regions of the world developed different symbols. The result that we have today is a plethora of additional marks developed historically.

A beginner learns that Arabic is written from *right to left* and must practice writing each letter and its connection rules to other letters. Because of the cursive nature, any letter may connect to the previous and following letters. Hence, a beginner learns the general *four basic forms* of a letter: at the start of a word, at the middle, at the end, and isolated. The simplest example of this rule is the equivalent of 'b' in Arabic: ب ـب ـبـ بـ. A reader with a sensitive eye would notice that the four shapes of the same letter differ in their width, height above the line, and depth below it. The same structural shape is used for the equivalent of 't' (ت ـت ـتـ تـ), and 'th' (ث ـث ـثـ ثـ). The equivalent of 'n' and 'y' share the same shapes as 'b' in the initial and medial forms (ـنـ نـ and ـيـ يـ) but not in the final or the isolated forms (ن ـن and ى ـى). In traditional Arabic writing styles (but with the exception of thuluth, riqā', and tawqī' styles [13]), the letters و ر د ا and their siblings with dots or marks do not connect to the following letter but only to the preceding one.

The cursive nature of arabic script adds another characteristic; many letters combine together to produce new shapes as in الحج becoming الحج. In the latin script this phenomena occurs infrequently and when it happens, a *ligature* is used to improve the appearance of the problematic letter combinations such as 'ff' and 'ffl'. In arabic typography, on the other hand, the presence of combined letters is abundant *but optional* in many cases. Haralambous [2] gives a long list (yet not exhaustive) of possible 'ligatures' in arabic. While speaking about the history of arabic typography, Milo [9] explains that "each letter can have a different appearance in *any* combination, something that can only be crudely imitated with ligatures". According to Milo [9], most modern books present the connected letter groups "as 'ligatures' and 'artistic expressions' without so much as a hint at traditional morphographic rules". Mackay [8] discusses the range of context evaluation in Arabic and concludes that clusters of four, five, six, and sometimes more letters may combine into a unique shape. Mackay then proposes the use of virtual fonts as an adequate solution.

Another feature in the Arabic script is its reliance on subtle changes to the letter shape to aid the reader in identifying the beginning and end of each letter within a combination. The letter ـسـ has three "teeth" (vertical pen strokes) similar to the

[*] A graduation project under the author's supervision by: Ibrahim R. Mohamed, Ahmed Z. Ameen, Ahmed M. Amin, Akram A. Mojahed, Kareem O. Sharawi, and Hisham Shihab

teeth in ﺑ and ﺗ. When ﺳ is connected to ﺑ, the tail of the ﺳ may be elongated to alert the reader to the correct grouping of the teeth. Furthermore, the two words سبع and تسع present different heights for the teeth of the ﺑ and ﺗ to help the reader as well. This difference in width and height is a type of encoding to prevent a misreading of the word and to aid the trained eye in quickly catching the letter combination. That encoding helps in other cases as well. If due to any reason the dots fade away, a reader faced with ٮسع can guess its correct origin. This encoding to emphasize the different letters by raising some teeth is essential in words such as تبينت and تسببت. However, the raising of the teeth is only possible by investigating the group of letters. So the height of the tooth for ﻨ in تبينت and تبنت is not a feature of the individual letter but of the whole combination. The same goes for the height of the dot on top of that same letter as in سنجح or سن.

In traditional (manual) writing, the "skeleton" of the letter combination is written first then the dots and the other marks are provided. So, a writer probably progresses from the skeleton to the dotted to the vocalized form as سں → سن → سُنَّ.

In the arabic script, a good calligrapher justifies the lines not just by stretching the spaces between the words but mainly by using optional ligatures or wider forms of some letters as in changing ك to ڪ and writing ڪتب instead of كتب. The use of optional ligatures and wider forms is the preferred method in high quality works. Another method is to add an elongation to the tail of some letters by using the taṭwīl or kashīdah symbol 'ـ' such as ســـبب instead of سبب. This second method has been widely abused in newspapers and low quality materials using mechanical typewriters.

The arabic script has a large number of writing styles that were developed traditionally to accommodate the different languages and different purposes. Latin scripts use bold, italic, or larger fonts for section headings and for emphasis. Traditional arabic writings vary the typeface instead. The printings of the Qur'an as well as of most books almost always use the naskh typeface for the main body. The headings, the introductory materials, and the back materials frequently use other typefaces such as the thuluth, ta'līq, and ruq'ah.

In summary, the points just mentioned are:

1. Arabic is written from right to left.
2. Characters in general have four different forms.
3. These forms are of different width, height, and depth.
4. The shape (the height of the teeth for example) of a specific form depends on its context.
5. There are additional marks that are put on top or below the character.
6. The horizontal and vertical location of the dots and marks on the characters is not at the same position always but depends on the character *and* its context.
7. There are too many letters that combine ("ligatures").
8. Ligatures and variable width forms of the letters are used to justify the lines.
9. Several typefaces are needed for special materials in a work of good quality.

With all of these issues, to find a suitable position for the dots and marks on the letter combinations is sometimes a real challenge even for a human let alone a machine.

2 Automated typesetting of Arabic

Arabic script does not enjoy the same luxury that latin script has when it comes to automated typesetting on computers. Milo correctly asserts [9] that the use of individual letters as the building block is not suitable for Arabic. Both Mackay [8] and Milo [10] argue that a layered approach is a better solution. In such a layered approach, some basic elements are provided in the font to represent the skeleton of some letter combinations, an individual letter's skeleton, or even a part of a letter. These elements are combined first to give the correct skeleton of the word with all the needed shaping for the teeth or other style requirements. On top of that skeleton, a second layer for the dots is added. Then, the vowel marks and any other marks come in subsequent layers. Milo [11] developed a system with a layered approach for his company, DecoType. It is a proprietary system used by a number of commercial software tools. Due to its proprietary nature, the full details and the extent of the capabilities of this system are not widely known.

Our goal is to provide a freely available system capable of typesetting the Qur'an, other traditional texts, and any publications in the languages using the arabic script. The Qur'an is one of the most demanding arabic texts from a typographical point of view. However, there is a long historical record of excellent quality materials (manuscripts and recent printings) to guide the work on a system to typeset it. Such a system, once complete, can easily typeset any work using the arabic script including those with mixed languages.

Knuth and MacKay [5] were the first to present a working solution for including right to left text (for Arabic and Hebrew) in the TeX family. Their proposed TeX-XeT system is an extension of TeX that produces a different DVI file. The enhanced mode of ε-TEX allows bidirectional text processing and produces regular DVI files but ε-TEX does not provide any arabic fonts or any specific functionalities that ease the typesetting of arabic books. Within the TEX extensions, both Ω [4] and ArabTEX [6, 7] have been used for Arabic and have met some of the basic requirements to varying degrees.

With the historical trend to extend TEX, Ω evolved as an implementation allowing multilingual text processing. As an offshoot of the work on Ω, Al-Amal system [2] is an early attempt to typeset the Qur'an specifically. Unfortunately, it is not freely available and its output (as shown in the example published in the paper describing it) falls really short of the desires of a native reader.

Due to its various attractive features, Ω was the first choice to achieve our goal. However, the result of our early experiments with the available arabic font provided with Ω were not satisfactory. None of the people whom we asked for their opinion liked the font. That font is suitable for a simple publication but definitely not for a high quality work using the arabic script. Ω in its current state does not easily lend itself to the layered approach described earlier. The modification of Ω is not an easy task since it is a very large system and such a modification means the creation of a new system that is not compatible with the existing base of TEX. The newer developments to Ω [3] —once they are stable, widely available, and documented— may help in implementing the layered approach necessary for typesetting high quality texts in arabic.

Lagally in ArabTEX [6] preferred to stay within the stable TEX standard and perform all the necessary processing with TEX macros. That decision allowed ArabTEX to be portable to any TEX implementation. However, ArabTEX had to compromise on the issue of line breaking. For right to left text, ArabTEX is forced to handle the line breaking by itself in a slow and complicated algorithm bypassing one of the best parts of TEX available for the latin script.

Although not a simple program, ArabTEX is confined to a number of style files each performing a specific task. ArabTEX implements a layered approach where each character is represented by a skeleton and additional modifiers (dots and vowels). According to the collected opinions, the quality of its font is much better than that of Ω. The font still needed improvements but it was an acceptable start. ArabTEX uses the LATEX license and hence we changed the name of our work to AlQalam (the pen in Arabic).

3 Implementation

Our goal of typesetting the Qur'an and traditional texts means a few more challenging requirements in addition to those of the general arabic typesetting. To assist the reader in recitation, several indicators for vowels, joints, text structure, and pausing locations have been added historically to the text of the Qur'an. We present here a few symbols.

Signs of pause (علامات الوقف) :

- The ۗ sign : the reader may continue but it is better to pause.
- The ۖ sign : possible to pause but it is better to continue.
- The waqf jā'iz sign ۚ : equal possibility to pause or continue.

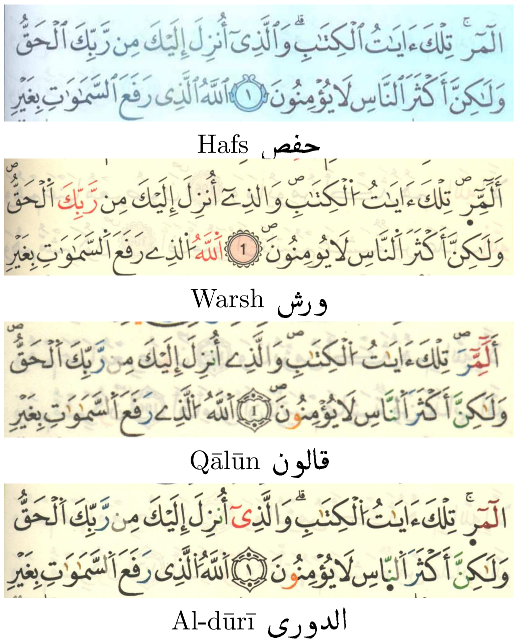
Additional diacritics :

- Ra's khā' ـۡ corresponds to sukūn.
- The madda ـٓ appears in the Qur'an on many letters such as in كٓهيعٓصٓ.
- The small 'ـۦ' in ﴿أَن يُشۡرَكَ بِهِۦ﴾ and 'ۥ' in ﴿وَٱللَّهُ عِندَهُۥ حُسۡنُ﴾.

Furthermore, there are different "narrations" of the Qur'an that differ in the pronunciation in some locations and hence lead to a plethora of additional marks needed. The vast majority of the printed copies of the Qur'an are in the narration known as Hafs. Only three other narrations (with their special marks for the special pronunciations) are printed in the whole muslim world. The remaining narrations (sixteen remaining for a total of twenty) are still in manuscript form.

Fig. 1 shows an example of the four narrations that exist in print. To make the comparison easier, we present the same two lines from the four narrations written by the same calligrapher and printed by the same press: King Fahd's complex for printing the Qur'an in Madinah. A simple look at the first word (top right in each example) reveals some of the different symbols needed. Those additional symbols fit well in a layered approach but would be quite difficult to accommodate otherwise. The ۚ symbol appearing on the first word of the top most narration belongs to the pausing signs. In AlQalam, we introduced a layer for the pausing signs beyond the layer of the dots and the layer of short vowels.

We use the existing symbols in the original font of ArabTEX where they are sufficient. We have improved the shape of the madda 'ٓ' and dagger alif 'ٰ'. The symbols that we have added so far are:



الٓمٓر تِلْكَ ءَايَٰتُ ٱلْكِتَٰبِ وَٱلَّذِىٓ أُنزِلَ إِلَيْكَ مِن رَّبِّكَ ٱلْحَقُّ
وَلَٰكِنَّ أَكْثَرَ ٱلنَّاسِ لَا يُؤْمِنُونَ ۝١ ٱللَّهُ ٱلَّذِى رَفَعَ ٱلسَّمَٰوَٰتِ بِغَيْرِ

Hafs حفص

أَلَٓمَٓرَ تِلْكَ ءَايَٰتُ ٱلْكِتَٰبِ وَٱلذِےٓ أُنزِلَ إِلَيْكَ مِن رَّبِّكَ ٱلْحَقُّ
وَلَٰكِنَّ أَكْثَرَ ٱلنَّاسِ لَا يُومِنُونَ ۝١ ٱللَّهُ ٱلذِے رَفَعَ ٱلسَّمَٰوَٰتِ بِغَيْرِ

Warsh ورش

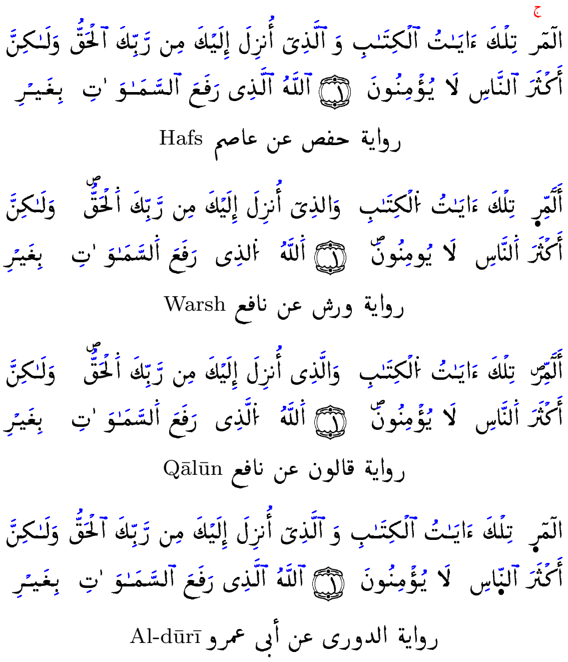
أَلَٓمَٓرَ تِلْكَ ءَايَٰتُ ٱلْكِتَٰبِ وَٱلَّذِےٓ أُنزِلَ إِلَيْكَ مِن رَّبِّكَ ٱلْحَقُّ
وَلَٰكِنَّ أَكْثَرَ ٱلنَّاسِ لَا يُؤْمِنُونَ ۝١ ٱللَّهُ ٱلَّذِے رَفَعَ ٱلسَّمَٰوَٰتِ بِغَيْرِ

Qālūn قالون

الٓمٓر تِلْكَ ءَايَٰتُ ٱلْكِتَٰبِ وَٱلَّذِىٓ أُنزِلَ إِلَيْكَ مِن رَّبِّكَ ٱلْحَقُّ
وَلَٰكِنَّ أَكْثَرَ ٱلنَّاسِ لَا يُؤْمِنُونَ ۝١ ٱللَّهُ ٱلَّذِى رَفَعَ ٱلسَّمَٰوَٰتِ بِغَيْرِ

Al-dūrī الدورى

Figure 1: The first two lines of surat al-ra‘d. An example from the four printed narrations.



الٓمٓر تِلْكَ ءَايَـٰتُ ٱلْكِتَـٰبِ وَ ٱلَّذِىٓ أُنزِلَ إِلَيْكَ مِن رَّبِّكَ ٱلْحَقُّ وَلَـٰكِنَّ
أَكْثَرَ ٱلنَّاسِ لَا يُؤْمِنُونَ ۝ ٱللَّهُ ٱلَّذِى رَفَعَ ٱلسَّمَـٰوَ ٰتِ بِغَيْرِ

Hafs رواية حفص عن عاصم

أَلَٓمَٓرَ تِلْكَ ءَايَٰتُ أَلْكِتَٰبِ وَالذِے أُنزِلَ إِلَيْكَ مِن رَّبِّكَ أَلْحَقُّ وَلَٰكِنَّ
أَكْثَرَ أَلنَّاسِ لَا يُومِنُونَّ ۝ أَللَّهُ أَلذِے رَفَعَ أَلسَّمَٰوَٰتِ بِغَيْرِ

Warsh رواية ورش عن نافع

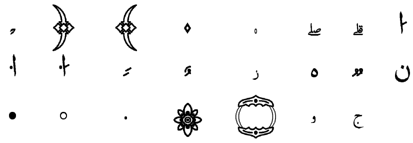
أَلَٓمَٓرَ تِلْكَ ءَايَٰتُ أَلْكِتَٰبِ وَالَّذِے أُنزِلَ إِلَيْكَ مِن رَّبِّكَ أَلْحَقُّ وَلَٰكِنَّ
أَكْثَرَ أَلنَّاسِ لَا يُؤْمِنُونَّ ۝ أَللَّهُ أَلَّذِے رَفَعَ أَلسَّمَٰوَٰتِ بِغَيْرِ

Qālūn رواية قالون عن نافع

الٓمٓر تِلْكَ ءَايَـٰتُ ٱلْكِتَـٰبِ وَ ٱلَّذِىٓ أُنزِلَ إِلَيْكَ مِن رَّبِّكَ ٱلْحَقُّ وَلَـٰكِنَّ
أَكْثَرَ ٱلنَّاسِ لَا يُؤْمِنُونَ ۝ ٱللَّهُ ٱلَّذِى رَفَعَ ٱلسَّمَـٰوَ ٰتِ بِغَيْرِ

Al-dūrī رواية الدورى عن أبى عمرو

Figure 2: Start of surat al-ra‘d by AlQalam.



The use of transliteration for a purely arabic document that is a few hundreds of pages long is obviously not practical nor desired. Hence, the default assumption for AlQalam is an input file with the characters coded in Unicode-UTF8. Bi-directional editors such as emacs and gedit (we used both) are good options. Editors have their own limitations though. If a symbol has a unicode point associated with it but there is no key combination mapped to that symbol or no glyph in the editor's font to represent it, another facility must be used. In a case such as ۖ the user might supply the unicode as `^^db^^96`. AlQalam interprets this code as belonging to the "signs of pause" category then raises it to its correct position such as in ٱلْحَقُّ ۖ. In addition to that, for some frequently occurring symbols we provide a macro such as `\l` for ٱ. The improvement of the input method is one of the major steps in future developments.

Another feature of typesetting the Qur'an is the use of colors. In some printings, certain letters, marks, or sometimes complete words take a different color usually to remind the reader of a pronunciation rule. Educational texts for young children often use similar color encoding schemes to stress new reading concepts and to train their little eyes in picking up the distinctive features of the script. A complete system for dealing with the arabic text should be able to color a piece of a letter combination or some specific marks.

To color the text, we use our modified version of the `acolor` package which Karol Mokry had originally written for ArabTEX. With a few modifications to the `aboxes` and `awrite` components of ArabTEX, AlQalam allows the user to color the diacritics and the pausing signs of the Qur'an through commands such as **`\coldia{blue}`**.

4 Results and future work

Fig. 2 presents the current results of AlQalam for the same narrations shown earlier in Fig. 1.

The initial phase of our work produced a usable system with known problems in the case of some symbols which we are currently solving. In addition to the four narrations of Fig. 2, we completed the work for another ten narrations and wrote a few lines as an example for each one. These results are the first step in a long study of the typesetting requirements for the ten main "readings" of the Qur'an and their expansion into the twenty narrations.

Some of the differences between the narrations constitute simple rules that may be programmed.

As an example, every plural pronoun ending with مْ and not followed by ا becomes ,مُ in the reading of 'abī ja'far. In the future, we hope to write macros for all the programmable rules.

Font development is a must. The current results of Fig. 2 are still quite far from the level of Fig. 1. With the vast repertoire of ligatures needed to represent the letter combinations in each typeface (naskh, thuluth, ta'līq, . . .), this task is quite complicated. The macros to detect the combinations and produce the appropriate ligatures for the skeletons (which might be changed for line justification) followed by the accumulation of the additional layers of dots and marks is a hard design problem as well.

For line justification, AlQalam inherits the ability to stretch letters using a kashīdah from ArabTeX. However we do not have yet a facility to use the better method of alternative ligatures and wider forms for some letters. In his system, Milo [10] attempts to provide for the use of wider forms. Berry [1] presents another solution (but without high quality fonts and ligatures) to stretch the letters for Arabic, Hebrew, and Persian using troff. Within the TeX family, Tánh [12] presents what he calls "selective use of multiple glyph" for the latin script although wider forms are not necessarily needed there. Our initial assessment indicates that line breaking for Arabic is an area that needs much more research.

Besides these research problems, AlQalam must extend its capabilities inherited from ArabTeX such as footnotes, construction of a table of contents, marginal notes, more LaTeX commands and environments,...

In conclusion, we provided a summary of the arabic typesetting requirements as well as the first steps of a system to fulfill them. This study is useful for those working on any of the languages written in the arabic script and specifically those supporting the inclusion of Qur'anic quotations. A layered approach is a must for high quality typography. The first version of AlQalam is ready for use but with very limited documentation. Improvements are in process for subsequent versions. AlQalam is free. We would like to share it with anyone interested in testing and providing us with feedback. To the best of our knowledge, there is no other software system that serves the requirements of the different Qur'anic narrations.

References

[1] Daniel Berry. Stretching letter and slanted-baseline formatting for Arabic, Hebrew, and Persian with ditroff/ffortid and dynamic POSTSCRIPT fonts. *Software—Practice and Experience*, 29(15):1417–1457, 1999.

[2] Yannis Haralambous. Typesetting the holy Qur'an with TeX. In *Multi-lingual computing: Arabic and Roman Script: 3rd International conference — Durham, UK*, December 1992.

[3] Yannis Haralambous and Gábor Bella. Omega becomes a sign processor. In *EuroTeX 2005: Proceedings of the* 15^{th} *Annual Meeting of the European TeX Users, Pont-à-Mousson, France*, pages 8–19, March 2005.

[4] Yannis Haralambous and John Plaice. Multilingual typesetting with Ω, a case study: Arabic. In *Proceedings of the International Symposium on Multilingual Information Processing, Tsukuba*, pages 63–80, March 1997.

[5] Donald E. Knuth and Pierre A. MacKay. Mixing right-to-left texts with left-to-right texts. *TUGBoat*, 8(1):14–25, 1987.

[6] Klaus Lagally. ArabTeX — typesetting Arabic with vowels and ligatures. In Jiří Zlatuška, editor, *EuroTeX 92: Proceedings of the 7th European TeX Conference*, pages 153–172, Brno, Czechoslovakia, September 1992. Masarykova Universita.

[7] Klaus Lagally. ArabTeX: a system for typesetting arabic. In *Multi-lingual computing: Arabic and Roman Script: 3rd International conference — Durham, UK*, page 9.4.1, December 1992.

[8] Pierre A. MacKay. The internationalization of TeX with special reference to Arabic. *Proceedings of the IEEE International Conference on Systems, Man and Cybernetics*, pages 481–484, November 1990. IEEE catalog number 90CH2930-6.

[9] Thomas Milo. Arabic script and typography: A brief historical overview. In John D. Berry, editor, *Language Culture Type: International Type Design in the Age of Unicode*, pages 112–127. Graphis, November 2002.

[10] Thomas Milo. Authentic arabic: A case study. right-to-left font structure, font design, and typography. *Manuscripta Orientalia*, 8(1):49–61, March 2002.

[11] Thomas Milo. Ali-baba and the 4.0 unicode characters. *TUGBoat*, 24(3):502–511, 2003.

[12] Hàn Thê Thánh. Micro-typographic extensions to the TeX typesetting system. *TUGBoat*, 21(4):317–434, 2000.

[13] Mohamed Zakariya. أنماط الحرف العربى. *Al-Computer, Communications and Electronics Magazine* الكمبيوتر والاتصالات والإلكترونيات, 22(8):48–53, October 2005.