

Arabic scientific e-document typography
Mohamed Elyaakoubi and Azzeddine Lazrek
{m.elyaakoubi, lazrek}@ucam.ac.ma
University Cadi Ayyad, Faculty of Sciences
B.O. Box 2390, Marrakech, Morocco
Phone: +212 44 43 46 49 Fax: +212 44 43 74 09
<http://www.ucam.ac.ma/fssm/rydarab>

Arabic scientific e-document typography
Mohamed Elyaakoubi and Azzeddine Lazrek
{m.elyaakoubi, lazrek}@ucam.ac.ma
University Cadi Ayyad, Faculty of Sciences
B.O. Box 2390, Marrakech, Morocco
Phone: +212 44 43 46 49 Fax: +212 44 43 74 09
<http://www.ucam.ac.ma/fssm/rydarab>

Abstract:

The main goal of this contribution consists on exploring the Arabic handwriting rules for the purpose of their formalization. The development of Arabic scientific and technical documents processing tools can not be undertaken without such formalization.

Keywords: Multilingual, Arabic calligraphy, e-document, Mathematical expression visual rendering, Digital typography.

1 Introduction

Since many years, tools authoring for e-documents composition was a central activity of several companies and organizations. In the beginning, these tools were tailored for the English script needs. The adaptation of these tools to the Western writings was not a difficult task, since these writings are different from the English writing, in their majority, only by some accented letters, some specific letters or other minor alternatives. The basic typographical rules were almost the same. The localization of these tools to the Arabic alphabet based writings is not an easy business, because of the great typographical and structural characteristic differences.

Nowadays, documents composition isn't just a typographer's task; as it was in the past. The final shape of document until the smallest detail is in charge of the author. The scientific text author, for example mathematical text, in a written language with the Arabic alphabet is confronted with the inexistence of solutions for document composition. A great number of Arabic mathematical documents are still written by hand. Millions of learners are concerned in their daily learning by the availability of systems for typesetting and structuring mathematics. Arabic is a native language for roughly three hundred million people. Moreover, the Arabic script is used, in various slightly extended versions, to write many major languages such as Urdu (Pakistan), Persian (Iran), or other languages such as Berber (North Africa), Sindhi (India), Uyghur, Kyrgyz (Central Asia), Pashto (Afghanistan), Kurdish, Jawi, Baluchi, and several African languages. A great part of humanity is concerned by the problem of e-document composition tools development.

We will present some tools for Arabic scientific documents composition, especially in the context of mathematical text, and we will look into the typographical components with their various structures.

2 Arabic writing particularities

2.1 Direction of writing

The Arabic writing spread out from right-to-left. Numbers also are written and read according to the natural direction of the Arabic script, from the smallest number to the biggest (units, tens, hundreds, thousands, etc.). Arabic is of course a unidirectional script. The traditional literature seems to be the last bastion of this property. Nowadays, numbers are written and read without obeying to any given order. Indeed, the number 123 is currently read following the right direction (1) then left (3) then medium (2). That can lead to believe that Arabic writing is bidirectional because of the current mixing of text and numbers directions.

In some contexts, Arabic mathematical document adopts Latin and Greek alphabetic symbols. Typesetting documents with these symbols yields to the problem of mixing right-to-left strings with left-to-right expressions. In some other contexts, mathematical document use special symbols and Arabic alphabetic symbols. The writing then spreads out from right-to-left. Samples of such documents can be found in handbooks in Middle East and even in some popular newspapers used to publish mathematical exercises when examinations are organized. Does it mean that composition difficulties are less? Certainly, but other big problems remain to solve such as writing cursivity and Arabic calligraphy rules witch are challenges to rise.

2.2 Cursivity

Since the Roman era, Latin script based languages adopted two fundamental styles: the cursive one and that with independent and isolated characters. The Latin based typography is based on the use of independent characters. In Arabic only the cursive style is allowed. This cursivity implies four different morphologies for the same letter according to its position in the word: initial, middle, final and isolated¹ except for the six letters, Alef, Dal, Thal, Reh, Zain² and Waw³ witch have only the two last forms. This way of writing is strongly dependent on the context and leads to a great variety of graphics. Each letter depends contextually, but also aesthetically, of the surrounding letters. The Arabic writing is very rich in aesthetic ligatures; these ligatures are abundant in the Arabic calligraphy [Haralambous, 1997].

If the Arabic writing is a cursive writing, the Persian one is even more. The Persian is an Arabic alphabet based writing. It is single in the world by the diagonal writing of each word. The lines of the text are held from right-to-left. Each word begins in top right and finishes in left below the baseline. Each word thus has its clean small local baseline. The letters take thus more than four alternatives of position, even if, grammatically, there are only four forms.

In terms of tools design for text processing, the cursivity brings new constraints. For example, the algorithms of justification of text must be completely re-examined. Instead of inserting white spaces or blanks between the characters, Kashida, a small stretching must be used. In Arabic, the hyphenation of the words is not allowed⁴, as it is the case in the Latin languages. This is perhaps due to the morphosyntactic structure of the Arabic language and the general absence of the vocalization. If the hyphenation of words in Arabic was accepted, the legibility of the text would be considerably affected.

2.3 Diacritic dot and letters weight

The Naskh style employs a reed pen (qalam) with the working point cut on an angle. The feather's head is a flat rectangle of width w and thickness t with $t = w/6$. The rectangle is maintained with an inclination angle of approximately 70° with the baseline. In Arabic calligraphy, the size of the feather's head in use is a determinant factor. Arabic calligraphy, thus, is the art of beautiful and elegant handwriting as exhibited by the correct formation of characters, the ordering of the various parts, and the harmony of proportions. The letter's weight or the horizontal and vertical metrics of the surface delimited by the contour are regularized using a particular unit of measurement: the diacritic dot that is presented as a rhombus marked by the feather in use in a precisely slope direction. Thus the Arabic letter Alef of the Naskh style has a height of six diacritic dots and a length of one dot and a declivity of a half dot.

2.4 Kashida

The Kashida is the stretching of the character. The Kashida is used either to respect the constraints of calligraphy or, in a second time, for the text justification (cf. Figure 1 and Figure 2⁵)

Figure 1: Letter stretching in a word

Figure 2: Symbol stretching in a mathematical expression

The Kashida is performed by a glyph which is not a character. Sometimes, the stretching is a part of the letter rather than the links between two letters. The glyph's shape of Kashida depends on the letters and on their contextual forms. Often, the typesetting software draws Kashida as horizontal linear segments not as curves connecting two letters. Indeed, it is much easier to draw segments than to calculate Bezier curves, in real time, with the required properties. Characters extensibility according to calligraphic rules requires the development of what can be called dynamic fonts.

Designers of computer-aided typography [DecoType] have developed a digital version of the Arabic Naskh script and a system that takes account of some calligraphic effects such as the Kashida [Milo, 2003]. An extension of the system ditroff/ffortid [Berry, 1999] makes it possible to carry out horizontal extensions of characters. However, the composition of e-documents also requires vertical extensions. The CurExt system [Lazrek, 2003] allows the composition of Kashida in a horizontal stretching with the T_EX system, using the two font generators METAFONT and PostScript. It also makes

it possible to compose the extensible mathematical symbols in the vertical direction. It allows in particular, to compose automatically curvilinear variable-sized as curvilinear braces or integrals signs. There is a relation of dependence between the degree of the extensibility (vertical extension) and the degree of concavity (horizontal extension). Such dependence obeys to the following constraints:

- The horizontal extension does not exceed a maximum of 12 diacritic dots;
- The vertical extension depends on the horizontal one, and does not exceed a maximum of a half diacritic dot.

In a recent contribution to appear of A. Bayar et al. we find a mathematical modelling of the parameters which determine the Kashida, particularly the dependence between the degree of the extensibility and the degree of concavity.

2.5 Justification

Some text justification software carry out some handling of hyphenation and therefore some inter-word spaces stretching. In the case of T_EX, the hyphenation is a formal algorithm used to determine all the possible points where a given word is supported to be broken. The T_EX algorithm is based on the solution proposed by Liang F. M. [1983]. The key idea underlying Liang's method is to look for patterns in the word. The algorithm works quickly and finds nearly all the legitimate places to insert hyphens. On the other hand, it is extremely difficult to obtain a uniform typographical grid. The main reason is that it is not possible to ensure an equal value of inter-word spaces in various lines. This leads to a heterogeneous optical density. Hàn Thê Thành [1999] proposed a solution for the improvement of the typographical quality of T_EX. The solution is to slightly modify the width of characters in order to limit the elasticity of inter words spaces, instead of only changing inter-words spaces. This modification is implemented by a horizontal scaling in PDF. This method can appreciably improve the aspect of the produced typography.

In Arabic, the problem is even more complex, because it is necessary initially to determine if the separation between two letters must be marked by means of a Kashida and/or a space. The Kashida can be stretched. In Arabic text justification, the Kashida is a typographical effect that allows the lengthening of letters in some carefully selected points on the line with determined parameters in order to produce the paragraphs alignment.

We propose an extension of CurExt, the typesetting system of curvilinear variable-sized symbols according to calligraphic rules requirements and mathematical documents constraints in an Arabic presentation. There are many constraints for determining the extensibility points that makes the system eager to lengthen a letter.

2.6 Diacritic signs

To specify the pronunciation of the purely consonant text, short vowels⁶ and other orthographic signs⁷ are added above or below the consonants. The Arabic text can be completely, partially or not vocalized. The vocalization leads to diacritic signs positioning problem, with respect to basic letters. The vocalization signs take different heights, not only with respect to basic glyphs but also according to the other contextual elements. The Arabic letter can be compared to a magnet for the diacritic mark [Haralambous, 2004]. From a technical point of view, some font formats offer functionalities for positioning the diacritic at varied heights. The aesthetic ligatures make the exact positioning of the vocalization mark practically impossible. The diacritics do not have a fixed size; they also stretch with the letters in presence of the Kashida.

Arabic script is used, to write some other languages, some languages borrowed the totality of the letters; it is the case of Persian, Urdu, Pashto, etc. Others, borrowed only some letters, it is the case of the Kurdish which uses only 21 letters. To make some unknown Arabic phonemes⁸, it was necessary to invent new letters by the adjunction of dots to Arabic letters that have the nearest pronunciation⁹. The use of the same character in two languages using the same writing system can lead to some confusion. For example, that happens if a word has two meanings in Arabic and Persian. The encoding system does not give any indication on the origin of this word (Arabic or Persian) which will be subject of treatment, spell checking, translation to another language, etc.

2.7 Calligraphy

In the Arabic world, calligraphy has traditionally been held in high regard. The high esteem accorded to copying the Koran, and the aesthetic energy that was devoted to it, raised Arabic calligraphy to the status of an art. The major styles of Arabic calligraphy are: Farisy, Koufy, Maghriby, Naskh, Thuluth, Ruqaa and Dywany (cf. Figure 3). Of course, the presence of a considerable number of calligraphic styles with very strict rules gives place to a diversity of glyphs, ligatures, rules, etc.

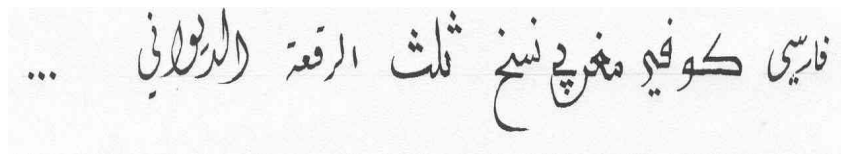


Figure 3: Calligraphic Arabic styles examples

Among the most relevant phenomenon of calligraphy, we can quote the treatment of the allograph [Tanasiu, 2003]: two glyphs representing the same position variant of the same letter¹⁰. The choice of an allograph is more than an automatic contextual treatment. It deals with aesthetic reasons. Some Arabic writing allograph took their places as independent characters, because they are used in two different meanings¹¹. That's the reason why, it is necessary to be careful in the use of an allograph substitution for aesthetic or justification, and to be sure that the modifications carried out do not have any semantic repercussion [Haralambous, 2004].

3 Arabic scientific documents

3.1 Mathematical expression

The Arabic scientific text can use two symbolic notation systems: Arabic mathematical expressions¹² using specific symbols and written from right-to-left in compliance with the direction of the Arabic writing, or Latin expressions¹³ from left-to-right, according to the French model as it is the case in Morocco, or according to the English model as it is the case in high level schools in Egypt.

Generally, in the “Maghreb Model” (cf. Figure 4¹⁴), the function is denoted by the symbol λ initial of the word λ دالة (function); the summation uses the Sigma reversed symbol \sum and symbols as well as the abbreviations of usual functions are written without diacritic dots.

$$\left. \begin{array}{l} \text{إذا كان } x > 0 \\ \text{إذا كان } x \geq 0 \\ \text{غير ذلك (مع } \pi \simeq 3,141) \end{array} \right\} = (x) \text{ د}$$

Figure 4: Mathematical expression in the Maghreb Model

Generally, in the “Machrek model” (cf. Figure 5), the function is denoted by the symbol λ initial of the word λ تابعة (function); Arab-Indic numerals are used; the summation uses the alphabetical symbol \sum initial of \sum مجموع (sum) and the abbreviations of usual functions are written with diacritic dots.

$$\left. \begin{array}{l} \text{إذا كان } x > 0 \\ \text{إذا كان } x \geq 0 \\ \text{غير ذلك (مع } \pi \simeq 3,141) \end{array} \right\} = (x) \text{ ت}$$

Figure 5: Mathematical expression in the Machrek model

Generally, in the “Moroccan model” (cf. Figure 6), mathematical expressions are based on the French model. The connecting words, in addition to the text, are in Arabic. The document is bidirectional.

$$f(x) = \begin{cases} \sum_{i=1}^x x^i & \text{إذا كان } x < 0 \\ \int_1^x x^i dx & \text{إذا كان } x \in E \\ \text{tg } \pi & \text{غير ذلك (مع } \pi \simeq 3,141) \end{cases}$$

Figure 6: Mathematical expression in the Moroccan model

Such differences also exist between French and English. Generally, in the “French model” (cf. Figure 7); the set is noted using the symbol E initial of “Ensemble” (set); the separation of the decimal part of a number is done using the comma and the abbreviation of the tangent function is noted tg.

$$f(x) = \begin{cases} \sum_{i=1}^s x^i & \text{si } x < 0 \\ \int_1^s x^i dx & \text{si } x \in E \\ \text{tg } \pi & \text{sinon (avec } \pi \simeq 3,141) \end{cases}$$

Figure 7: Mathematical expression in the French model

Generally, in the “English model” (cf. Figure 8); the set is denoted by the symbol S initial of the word “Set”; the separation of the decimal part of a number marked with a dot, the abbreviation of the tangent function is noted tan and the alphabetical variable i is without dot.

$$f(x) = \begin{cases} \sum_{i=1}^s x^i & \text{if } x < 0 \\ \int_1^s x^i dx & \text{if } x \in S \\ \tan \pi & \text{otherwise (with } \pi \simeq 3.141) \end{cases}$$

Figure 8: Mathematical expression in the English model

3.2 Mathematical expressions composition

Now, what about Arabic mathematical documents composition tools? Among these tools, one can quote: the ArabT_EX [Lagally, 1992] system which makes it possible to write Arabic text. It doesn't allow the typesetting of Arabic mathematical expressions. One can also refer to the multilingual system Ω [Haralambous et al, 1997] that allows also the composition of Arabic text but without Arabic mathematical expressions. The RyDArab [Lazrek, 2001] system makes it possible to compose Arabic mathematical expressions with specific symbols and in a homogeneous direction of writing.

Arabic scientific document composition requires a homogeneous symbolic environment, in particular, a proper mathematical font. The creation of such font is at the same time a technical and artistic complex task. In the following, we will describe a font family designed in the compliance with the rules of Arabic calligraphy as well as dynamic fonts for the composition of curvilinear extensible symbols and the composition of Kashida through calculating corresponding Bezier curves in real time.

3.3 Mathematical fonts

The RyDArab system is a T_EX extension. It was designed for the composition of Arabic symbolic expressions. In order to avoid making new fonts, the RyDArab system, in its first versions, adopted the Arabic font (nash or xnsh) of Naskh style developed by K. Lagally for the system ArabT_EX as well as the Computer Modern font family made using METAFONT for T_EX. RyDArab adopted, in particular, the symbols of Computer Modern Symbol and the Extensions Computer Modern, an extension containing different glyphs; necessary for symbols construction of various sizes after mirroring.

The use of symbols coming from various font families in a mathematical document leads to many heterogeneities concerning: size, level of boldness, etc. The simple horizontal reflection of the symbols, designed for the left-to-right writing, does not give better results. Indeed, by taking account of the movement direction of the feather's head in Arabic calligraphy, one wonders whether the glyph \sum , and not Σ , is really the ideal representative of the symbol Arabic mathematics sum, corresponding to \sum , according to the thickness variation. Another aspect of this problem is the mirror image of \notin which is \notin and not \notin . In general, in Arabic mathematical notation, the bar of negation is slanted to the right, as in Latin. In addition, the METAFONT format, generating bitmap fonts, remains always prisoner inside the T_EX environment. Whereas, the font's production technology tends to the vectorial one in PostScript, TrueType, OpenType or SVG. For this reason, one quickly feels the need to build a font made in a universal format which meets the need for homogeneity of the writing and fullness of the font.

3.4 The mathematical font RamzArab

RamzArab is the first Arabic mathematical font in the OpenType format. It contains a complete set of mathematical symbols in Arabic presentation [Banouni et al, 2004]. Some of symbols of this font are currently submitted to be included in the Unicode standard [Lazrek, 2005]. The package RamzArab is already used by the systems RyDArab and CurExt to compose the mathematical expressions, with extensible symbols. The principal motivations to build such a font were:

- The inexistence of an Arabic mathematical font in the OpenType format;
- The Heterogeneity and incompleteness that appear when different families coming from various environments are imported;
- The low level of the aesthetic aspect and lack of compliance with the rules of Arabic calligraphy.

In Arabic, there are two ways to arrange the alphabet: al-alefba'ya and al-abajadiya. The first order is essential for the effective search for information in an Arabic text, particularly in the dictionaries. In mathematical mode, the order in use is the second. The alphabetical symbols of the RamzArab font are arranged in this order. The alphabetical symbols are used in the six forms in the handbooks of the Middle-East. These various forms are: isolated, initial, tailed, stretched, looped and double-struck¹⁵. As it is the case of Latin mathematics, where the alphabetical symbols are without accents, dots and other diacritics, the Arabic alphabetical symbols are in general without dots or marks of vocalization. The RamzArab font adopts Naskh style. In order to eliminate the confusion which can appear between some alphabetical symbols and some numerals, a solution consists on the use of the Ruqaa style to draw these symbols as it is used. Particular forms were added to provide a maximum of symbols and to allow denoting some functions such as the limit, the product, the factorial and other functions. Moreover, the RamzArab font offers the figures used in Machrek as those used in the Maghreb, the punctuation marks, the accents, the ordinary operators, some symbols that are presented as the mirror image of the Latin mathematical symbols, etc.

3.5 Dynamic fonts with the CurExt system

In the section 2.4, we mentioned the problem of the composition of Kashida in the process of composition of the Arabic text. In fact, the composition of the Kashida is only one aspect of a more general problem in the composition of the dynamic font. Indeed, the extensible mathematical symbols, contrary to the symbols with fixed size, change size according to the context, in other words, according to the expression covered by the symbol. This extensibility can be linear as in the case of the brackets, or curvilinear as in the case of parenthesis or braces. CurExt is an application on the T_EX system which makes it possible to manage vertical extensibility while composing automatically extensible curvilinear delimiters. It also makes it possible to manage horizontal extensibility by composing Kashida. The current version of CurExt is compatible with the two generators of fonts METAFONT and PostScript.

4 Conclusion

Arabic typography still seeks its way; the majority of Arabic scientific e-documents are currently imperfect. This state of the things is due mainly to the difficulty of the production tools adaptation. If the adaptation of typesetting tools to the context of the Latin writings has started by restructuring the Latin writing itself, Arabic typography rules are not easily tameable to the requirements of the electronic publishing.

¹ Ex. م, م, م and م respectively for the letter Meem.

² Letters names used are those standardized by Unicode in its English version.

³ ا, ا, ا and ا respectively.

⁴ Previously, the hyphenation was allowed in Arabic writing.

⁵ These stretched letters and symbols are composed with CurExt system.

⁶ Fatha, Damma and Kasra.

⁷ Sukun, Tanwin (Fathatan, Dammatan and Kasratan), Shadda, Wasla and Madda.

⁸ Such as “p”, “v” and “g”.

⁹ ب, ف, and ح respectively.

¹⁰ The glyphs ك and ك are two allographs of the Arabic letter Kaf. Also, the glyphs م and م are two allographs of the Arabic letter Meem.

¹¹ Thus, normal Kaf ك is used for “gh” in Sindhi and zinadi Kaf ك is used for “k”.

¹² The adjective Arabic is used to abbreviate Arabic alphabet based notation.

¹³ The adjective Latin is used to abbreviate Latin alphabet based notation.

¹⁴ These mathematical expressions are composed using the RyDArab system.

¹⁵ Ex. ج, جـ, جـ, and ج respectively for the symbol Jeem.

Bibliography

Banouni, M., Elyaakoubi, M. & Lazrek, A., (2004). Dynamic Arabic mathematical fonts, LNCS 3130, pp. 149–157, International Conference on T_EX, XML and Digital Typography, TUG2004, Xanthi, Greece, <http://www.springerlink.com/index/URHRT2EYKYHH1RPA>.

Berry, D. M., (1999). Stretching Letter and Slanted-baseline Formatting for Arabic, Hebrew and Persian with ditroff/ffortid and Dynamic PostScript Fonts, Software Practice & Experience, no. 29:15, pp. 1417-1457.

Designers of Computer-aided Typography DecoType, <http://www.decoType.com>.

Haralambous, Y., (1997). Tour du monde des ligatures, Cahiers Gutenberg, n° 22, pp. 69–80.

Haralambous, Y. & Plaice, J., (1997). Multilingual Typesetting with Ω, a Case Study: Arabic, Proceedings of the International Symposium on Multilingual Information Processing (Tsukuba), pp. 137–154.

Haralambous, Y., (2004). Fontes & codages, Glyphes et caractères à l'ère numérique, Éditions O'REILLY, Paris, 2004.

Lagally, K., (1992). ArabT_EX - Typesetting Arabic with Vowels and Ligatures, EuroT_EX92, Prague.

Lazrek, A., (2001). A package for typesetting Arabic mathematical formulas, Die T_EXnische Komödie, DANTE e.V. 13. (2/2001), pp. 54-66, <http://www.ucam.ac.ma/fssm/rydarab/doc/communic/dtk201.pdf>.

Lazrek, A., (2003). CurExt, Typesetting variable-sized curved symbols, EuroT_EX2003, 14th European T_EX Conference, Brest, France, pp. 47–7, <http://www.ucam.ac.ma/fssm/rydarab/doc/communic/curext.pdf>.

Lazrek, A., (2005). Arabic mathematical symbols for Unicode, <http://www.ucam.ac.ma/fssm/rydarab/english/unicode.htm>.

Liang, F. M., (1983). Word Hy-phen-a-tion by Comput-er, Ph. D. Thesis, Department of Computer Sciences, Stanford University.

Milo, T., (2003). ALI-BABA and the 40 Unicode Character – Towards the Ideal Arabic Working Environment, New input output concepts under Unicode, EuroT_EX2003, 14th European T_EX Conference, Brest, France, pp. 97–102.

Tanasiu, V., (2003). Allographic Biometrics and Behavior Synthesis, EuroT_EX 2003, 14th European T_EX Conference, Brest, France, pp. 103–108.

Thành, H. T., (1999). Améliorer la typographie de T_EX, Cahiers GUTenberg, Lyon, France.