

Design and computer multilingualism: Case of diacritical marks

Mohamed Hssini* and Azzeddine Lazrek**

Department of Computer Science, Faculty of Sciences,
University Cadi Ayyad - Marrakech, Morocco

* m.hssini@ucam.ac.ma

** lazrek@ucam.ac.ma

Abstract— In a multilingual digital document, the problems of design are complicated by the presence of diacritical marks from various scripts and controlled by various typographic rules. This study is limited to Latin and Arabic case. In the first time, we compare the difficulty of processing information diacritical of both scripts and we study the limits of Latin resolution strategies applying for Arabic. In the end, we propose an approach for the resolution to the problem of positioning diacritical marks for multilingual fonts in TrueType format.

Keywords—Digital document; Diacritical marks; Arabic calligraphy; Fonts; Unicode; TrueType; OpenType; Graphite.

I. INTRODUCTION

In a multilingual digital document, the principles of design are risky by the likely conflict rules and mechanisms that control each of the writing. Diacritics are an

example. A diacritical mark is a sign accompanying a group or one letter, as the acute accent on the "e" product "é". Diacritics are often placed above the letter, but they can be placed below, in or through, before or after or around a glyph. Diacritical marks have common roles between the different languages of the world like:

- define playback;
- amend the phonetic value of a letter;
- avoid ambiguity between two homographs;
- etc.

However, the Arabic diacritical marks have an additional role, which is to fill the void: a task that is influenced by the effects of justification of Arabic text. This study focuses on to approximate a resolution to the problem of positioning of diacritics.

For that, we have taken three steps: in the first, we compared problems design of diacritical marks in the Arabic script with the design of diacritics for Latin script. In the second, we identified strategies to solve this problem and examine their ability in the Arabic case. In the third, we spend the last part to problem of positioning diacritical marks.

II. GENERAL INFORMATION

A. History about diacritics signs

The first diacritical mark appeared among the ancient Greeks and Romans. They were developed and distributed

in various European languages. The diacritical marks are often from letters that were written above another letter. For example, the tilde was originally a small "n". The addition of diacritics was an option from four to overcome the shortcomings of a language belonging to the Latin script [1]. The others were to add another letter, to combine two or more letters, or use the apostrophe. The origin of diacritical Latin script is evolutionary [2]. In periods of colonization, Latin diacritics have been used to expand the Latin alphabet for writing non-Roman languages: if there are more fundamentally different sounds (phonemes) in the language as there are letters base it invents new letters or they are taken to other alphabets. However, the most common solution is to add diacritical marks on the letters, often imitating the spellings of other languages [9].

Arabic is one of the Semitic languages, as Hebrew and Syriac. It's also cursive and written from right to left. The specialists are divided as to its origin. The majority believes it has developed down writing Nabatean. Others believe it comes from Al-Musnad also known as Al Hamiri (writing of the former Yemeni). A small group believes that writing is a pure divine production. The Holly Koran played a key role in the development of Arabic script. Before Islam, Arabic was little writing practiced, used primarily for commercial transactions or note contracts. Orally revealed to the Mohamed Prophet from 610, and its transcripts collected by 'Uthman on 653. The divine word brings a tremendous impetus to writing. The need to magnify the floor is so sacred and calligraphy, early Mushaf, is an essential component of the Islamic art. As the Koran was documented at the time of the Caliphs Rashid, about 700, the Arabic letters had no dots or punctuation. The dots are added as a succession during periods: the reading difficulties caused by confusion between the consonants of the same shape (the same sign can represent multiple letters) and the lack of scoring short vowels led to the invention of signs to facilitate reading. It was initially reported vowels by adding color points placed above or below letters. This usage has changed and led to the current practice of vowels noted by small signs or characters. This differentiation of consonants by diacritics existed in the oldest form of Mushaf fine or even points. Found in many Arabic calligraphy writing styles, each with their strict rules and their scope (illustration, architectural decoration, editing ...). Ali Ibn Moqlah (846-940), Minister of the three Caliphs Abassides Al-Moqtada (908-932), Al-Qahir

(932-934), Al-Radi (934-940), and his knowledge of science who introduced geometric the most important step in the development of Arabic calligraphy. Ibn Moqlah settled the task of drawing a cursive writing that is both beautiful and perfectly proportionate [6]. He established a comprehensive system of basic rules calligraphy based on the dot as the unit of measurement. It redesigned the geometric contour of letters and correct their shape and size through the point, the Alef and the circle. This is an Alef, which is measured with calligraphy and thought, and draw a circle whose diameter is Alef. Each letter was based on this circle [6].

In doing so, Ibn Moqlah has given the art of Arabic calligraphy precise scientific rules, whereby each letter, with a rigorous discipline, is attached to the three standard units that are the point, the Alef and the circle. This method of writing, called al-khatt al-Mansob, was perfected by his students the most famous is Ibn al-Bawbab (-1022). To understand the importance of Ibn Moqlah in the history of Arabic script, it is possible to cite Abdullah Ibn al-Zariji, which in the tenth century remarked: "Ibn Moqlah is a Prophet in the art of calligraphy. His gift is comparable to the inspiration of bees when they built the honeycombs."

B. Classification

There are three kinds of Arabic diacritical marks [7] (see figure 1 to 8 from WinSoft Pro font):

- **Language's diacritics:** composed on:
 - **Diacritics above**
It's a mark placed above a letter, as Fatha, Damma or Sukun.

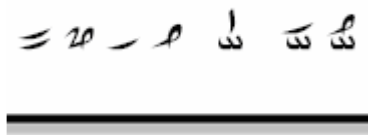


Figure 1. Arabic diacritics above

- **Diacritics below**
It's a mark placed under the base letter, as Kasra or Kasrattan.



Figure 2. Arabic diacritics below

- **Diacritics through**



Figure 3. Jarrat wasl through Alef

- **Aesthetics' diacritics**

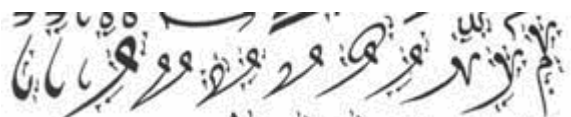


Figure 4. Kasra and Kasrattan

- **Explanatory diacritics**



Figure 5. Explanatory diacritics [10]

Latin diacritics can be classified according to their design, i.e. centered symmetric or not, or following their investment towards basic letters as follows:

- **Diacritics above**
The diacritical sup-script is placed above the letter to change.

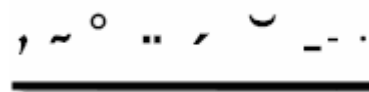


Figure 6. Diacritics above

- **Diacritics below**
There are made below the basic letter.



Figure 7. Diacritics below

- **Others**
Unlike diacritics over, most of those positioned through, before or after or around a glyph.

- ~ Tild
- ˘ Ogonek
- Ring
- ¨ Diaeresis
- ˆ Double acute
- ˆ Circumflex
- ˘ Caron
- ˘ Breve
- ˉ macron
- ˘ Acute
- ˘ Cedilla

Figure 8. Diacritical marks

III. DIACRITICAL MARKS IN UNICODE

Unicode is a character encoding that defines a consistent way of encoding multilingual texts and facilitates the exchange of textual data. It can encode all characters used by all the written languages of the world (more than one million characters are reserved for this purpose). All characters, regardless of the language in which they are used, are accessible without any escape sequence. The Unicode character encoding treats

alphabetic characters, ideographic characters and symbols in an equivalent manner, with the result that they can coexist in any order with equal ease. Unicode assigns to each of its character a unique numeric value and name. As such, it differs little from other standards or standards of character encoding. However, Unicode provides other information crucial to ensure that the encoded text will be readable: the case of coded characters, their properties and their directionality letter. Unicode also defines semantic information and includes correspondence tables of breakage or conversions between Unicode and directories of other important character sets.

A. Combinatorial characters and diacritics

Combining characters is a character to appear in association with another basic character. Unicode have two types of signs combinatorial: marks with space and non-spacing marks. The combinatorial non-spacing characters do not appear alone. However, the combination of the basic character to non-spacing character can occupy the space made more lateral that the base alone. Thus, an "i" hunts slightly more than a simple "i".

B. Composition and decomposition

In Unicode, character composition is the process of combining simpler characters into precomposed character such as the "n" character and the combining "~" character into the single "ñ" character. Decomposition is the opposite process, breaking precomposed characters back into their component pieces.

C. Bidirectionality

The bidirectional texts are written in two opposite directions. The bidirectional algorithm takes place in six steps:

- Determine the default direction of the paragraph;
- Process the Unicode characters that explicitly mark direction;
- Process numbers and the surrounding characters;
- Process neutral characters (spaces, quotation marks, etc.);
- Make use of the inherent directionality of characters;
- Reverse substrings as necessary.

IV. DESIGN AND MULTILINGUALISM

Many concepts underlie the field of design, as the balance, the rhythm, etc. The principles of design face in the case of mixture of different directions postings to change the rules of writing. It is in a somewhat similar situation when a multitude of styles in a monolingual Arabic text where the change of style indicates a title or section begins [7].

A. Space varieties

If characters are in a square imaginary languages for Latin, Hebrew, Chinese, etc... can align with the letter "x". In Arabic, heights [7] and forms of letters vary depending on the context:



Figure 9. Arabic letter Beh

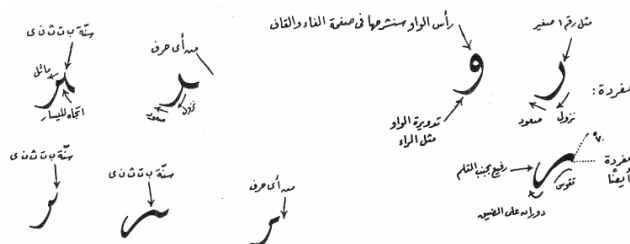


Figure 10. Arabic letter Reh

The spatial properties vary between Latin and Arabic scripts. The definition of "bold" depends, in Arabic, of style. The reduction in the density of letters is by layering or by reducing the body. Diacritics in the Thuluth style, unlike the Naskh, by a Qalam, pen, different from that used for the body of letters base. The harmonization of multilingual document is therefore influenced by the multitude of scripts or styles in the same language.

B. Justification of the Latin text

The justification of the Latin text makes itself while varying the space between the words and the characters, so that the line of text filled the inter-margin space. The value of the spacing varies between a minimal value and another maximal when the optimal value doesn't permit the justification of the text. The hyphenation permits to cut the word that arrives at the end of line in order to have a better visual within a text. A typographical rule imposes that we should not make more than three consecutive hyphenations. Avoid too many cuts in a text, it also means ensuring greater fluidity of reading.

Problems related to the justification of the text, especially a justification of the kind made by processing software word processing, without correction by a human operator are potentially many. Here, we will only raise the three most current: the problem of the hollow lines, the problem of the widows and the orphans, and the problem of cracks that cross the blocks of text [9].

1) The hollow lines

The hollow lines are the lines only including a syllable, an only word, or very few words, that finish a paragraph on a length lower to the third of the justification. He/it is counseled strongly to avoid them, in order to keep its aspect to the block of text. Today, one doesn't ask some so much, one can keep shorter lines than the third of the justification, but he/it is worth to avoid letting a syllable or a word isolated at the end of paragraph better.

2) The widows and the orphans

When working of layout, it is necessary to worry also of the unaesthetic aspects of the lines of paragraph end, isolated in top of page or column, and of the lines of paragraph beginning, isolated at the bottom of page or column. Some software, of desktop publishing or word processor, have a function that permits to determine the number of isolated lines tolerated in top or at the bottom of page. The most often, they allow a minimum of two lines.

Although some works give some different definitions, a widow is a line that is isolated at the bottom of a column or one page. This configuration is to avoid because it is unaesthetic, mainly on the long justifications. In principle, at the bottom of page, a new paragraph must include at least two lines. It is also valid, with greater reason, for a title, that must not be ever let alone at the bottom of page, for obvious reasons.

An orphan is an only word, or an isolated line, that is reported in top of a column or one page. This configuration is absolutely proscribed because not only it is unaesthetic, but again it disrupts the carving logical of the text, and therefore its reading. If one cannot make bring this orphan in the previous lines, it is necessary to shorten or to modify the text when this one permits it. For example, while adding some adjectival or some adverbs provided that this (innocent) "cheating" passes unobserved to the reader's eyes. One doesn't start a column or a new page with the only last line of a paragraph. A paragraph that ends in top of column or page must include, him also, at least two lines. If the last is hollow, three lines are preferable.

In the same way, a chapter that ends in top of column or page should include at least five lines of text.

3) *The cracks*

The cracks, known as rivers, are other phenomena unsightly, products at random from the disposal of a number of spaces between words of several overlapping lines. They form a white line sinua through a block of text or a kind of stream that flows across a page. One can often correct this by dividing whites differently, by changing the justification or the body of characters, or by amending the text. If the document contains graphics, they could be moved, or change there size, or also change the design of the entire text.

C. *Justification of the Arabic text*

In the Arabic writing, that is cursive, a word can be dilated by the kashida - specific to the Arabic writing - to cover much space [7] [8] and can be pressed by the use of the ligatures [7] [8]. It has other mechanisms of management of the Arabic line: graphic fillers (as the three points), reduction of the size of the characters, elongation of the letters, superposition of the letters, writing in the margin, etc. [7] [8]. These mechanisms influence on the measurements and the positioning of the Arabic diacritical marks [9].

V. DIACRITICS DESIGN

There are three challenges in the design of Latin diacritics:

- They must be harmonized with the basic glyphs;
- Do not cause problems with other basic glyphs;
- Respect the baseline.

In the Arabic case, there are aesthetic diacritics whose position depends on other diacritical marks. The interactive diacritics relationship with the mechanisms of justification requires resizing and repositioning diacritical word influenced by the effects of justification. Follow, we present the main issues of design diacritics as they have been cited in [2] and the specific issues to Arabic.

A. *Problem of asymmetry*

The balance is the stability resulting from the review of an image and a comparison with our ideas of the physical structure (such as mass, gravity, or the edges of a page). That is the arrangement of objects in a design specified according to their weight in the visual picture composition. The balance generally exists in two forms: symmetrical and asymmetrical. The symmetrical balance occurs when the weight of a graphic composition is evenly distributed around a central axis vertical or horizontal. The symmetrical balance is also known as formal balance. The asymmetrical balance occurs when the weight of the graphic composition is not spread evenly around a central axis. The asymmetrical balance is also known as informal balance. The size of a Latin diacritic and weight must be balanced with the glyph base with which it is used [2]. The horizontal alignment of diacritical glyph with the foundation should be such that there is balance the two views. For diacritic center symmetry with glyphs basic symmetrical, simply align the center of the bounding box of diacritic with the basic glyph [2]. If either one is asymmetrical other measures must be used.

1) *Case of symmetrical basic glyph*

The optical alignment is a tool to adjust the horizontal displacement of basic glyph or diacritic to focus on the diacritic glyph and maintain basic balance. One solution is to align the optical center of the letter with the mathematical center of space. The optical center is estimated by the center of the contour.

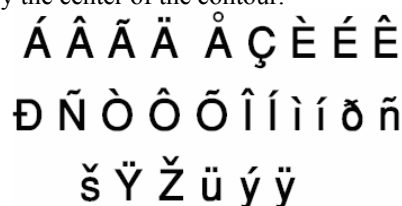


Figure 11. Symmetrical basic glyph

2) *Case of asymmetrical basic glyph*

In the case of asymmetrical basic glyph, the diacritic exchange up connection following the basic glyph. The optical alignment is not always used and other solutions are offered by new technologies such as OpenType and Graphite (see & VI).

B. *Problem of harmonization*

When the diacritics are sufficiently focused with the corresponding basic glyph, there are sometimes problems

with other basic glyphs. For example, the two "Diaeresis" and "Tild", in the following figure, enter in conflict with other glyphs base "d" and "b".

dīb dīb dīb

Figure 12. Conflict of diaeresis and tild with other glyph

One solution is to draw the diacritic specifically for each glyph basic reducing the space between the points or resizing. Another solution is the kerning.

C. Problem of vertical space

In fonts, the diacritical marks are aligned on a line parallel to the baseline. In other fonts, the distance between the diacritic and their base glyph is variant.

D. Multiple diacritics

Diacritics could cause multiple problems with the baseline or with other glyphs. Different techniques are used to solving this problem including: draw a glyph gathering all the diacritics multiple, etc.

E. Specific issues to Arabic

Arabic diacritics role is to fill the void, white space, in the word that there are specific diacritical marks, for aesthetics. There are three mechanisms for creating void in the Arabic word: kashida, extension glyphs and the interconnection between glyphs. In each case, the void is filled in two steps:

- The first, by resizing the Fatha in proportionality with the white;
- The second, by placing the aesthetics' and explanatory diacritics.

Diacritical marks lead, according to the language's function, to repeat the characteristics common to many of the glyphs.

The concept of symmetry in Arabic design is related to the line writing where the extensions are to balance the masses of other glyphs.

Arabic diacritics have a relationship with the mechanisms of justification. The diacritical marks are cosmetic compared to other signs respecting fill the void and not obscure the gray.



Figure 13. Arabic diacritics roles

VI. POSITIONING DIACRITICS AND NEW TECHNOLOGIES

We are studying the three font's formats: TrueType, OpenType and Graphite.

A. The GPOS table of OpenType

GPOS table manages the positioning of glyphs. We can put any diacritic on any glyph basic threw it [4]. Each

diacritic has a base. Diacritics are divided into several classes according to their behavior. Each basic glyph as attachment points that diacritic class.

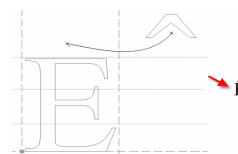
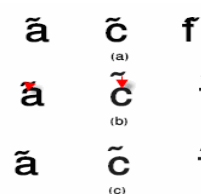


Figure 14. Diacritic position

B. Attachment and clusters in Graphite

The positioning of glyphs is done by two simple operations: moving and kerning, a simple tool: the points of attachment. If two glyphs "A" and "B" are attached, one-by-example "B" is attached to "A" and "A" is said base of "B". Another glyph "C" in turn can be attached to either "A" or "B", etc. [5].



Diacritics attachment points

The Figure 15 demonstrates the usefulness of attachment points. As shown in Figure 15 (a), a record of diacritics with a "not smart fonts" seems correct when they are attached to a tiny symmetrical centered as "a", but if not symmetric the diacritic is not centered correctly and comes into collision with the upper half of the glyph, or both. For Graphite font, stain is different: Figure 15 (b) shows the commitment indicated by small dots and arrows, and Figure 15 (c) shows the results with the correct record. The mechanism of base resolves the multiple diacritics problem, when the first diacritic is attached to a glyph base; it in turn is the basis of the following diacritic. The basic glyph and diacritic form a cluster. Graphite includes the ability to calculate metrics cluster or sub-cluster glyph individual for use in operations positioning [5].



Figure 15. Multiple diacritics attachment points



Figure 16. Examples of Arabic fonts

C. Diacritics positioning system

To place one or more diacritical marks relative to the base glyph, this system use a diacritic's bounding box and the base glyph's bounding box, in association with diacritic place data stored in the system[11]. The position data enables the diacritic positioning system to call associated functions that place multiple diacritics above and/or below a single base character without interfering with one another, e.g. to stack the diacritics. In addition, the information about the diacritic characters can be employed to prevent interference between a diacritic and the base character in special circumstances [11].

1) The architecture

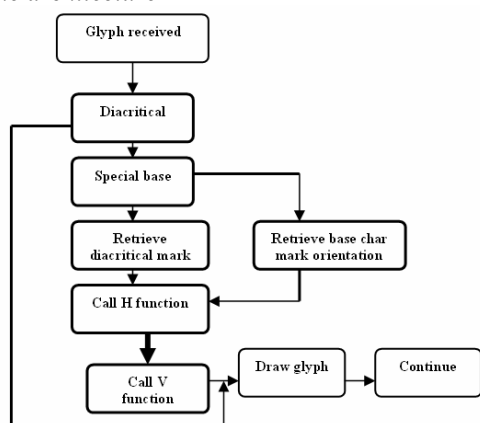


Figure 17. A diacritics positioning system

2) Description

When the system receives the information that the mark is to be placed over the base character, he looks up the orientation for this mark in the table that is stored in memory. This table [11] lists each diacritic by its name or their Unicode value. Based on this information in this step, the system calls a pair of functions H and V for properly positioning mark.

3) Commentary

Graphite and OpenType font formats have the advanced features to treat Arabic script. For this reason, we limit this study to the system for positioning diacritical mark in TrueType font format.

In the Arabic script, the position and dimension of diacritical mark Fatha and Fathattan are related to form of

base glyph and followed base glyph. So, to extend a system which operates under the same architecture as the diacritics positioning system three things to take into account:

- The functions H and V must have the ability to calculate the horizontal and vertical position of diacritic glyph relative to the base glyph and followed base glyph.
- The system must be able to substitute the diacritical mark if an extension takes place.

VII. CONCLUSION

Most of the fonts used to write Arabic do not have a deep tables and technologies of different formats, but we believe that the resolution of problems of diacritical in the multilingual digital document affects a layout engines. These problems have link with the problems of design of Arabic basic letters as the superposition of letters, the reduction of body and ligatures.

REFERENCES

- [1] J. C. Wells, "Orthographic diacritics and multilingual computing", *Language problems & language planning* ISSN, 2000, vol. 24, n° 3, pp. 249-272.
- [2] J. Victor Gaultney, "Problems of diacritic design for Latin script text faces", <http://www.sil.org/>, December 2008.
- [3] Yannis Haralambus, "Fontes et codage", O'Reilly, Paris, 2004.
- [4] R. Nicole, "Graphite Application Programmer's Guide", <http://www.sil.org/>.
- [5] <http://www.typographie.org/>, January 2009.
- [6] Mohamed Hssini, Azzeddine Lazrek and Mohamed Jamal Eddine Benatia, "Diacritical signs in Arabic e-document", CSPA'08, The 4th International Conference on Computer Science Practice in Arabic, Doha, Qatar, April 1-4, 2008 (in Arabic).
- [7] Vlad Atansiu, "Le phénomène calligraphique à l'époque du sultanat mamluk", PhD Thesis, Paris, 2003.
- [8] Mohamed Jamal Eddine Benatia, Mohamed Elyaakoubi, Azzeddine Lazrek, "Arabic text justification", TUGboat, Volume 27, Number 2, pp. 137-146, 2006.
- [9] <http://a1.esa-angers.educagri.fr/informa/>, February 2009.
- [10] H. Albaghdadi, "Korassat alkhat", Dar Alqalam, Beirut, 1980.
- [11] Chapman, Christopher J., "Diacritic positioning system for digital typography", <http://www.freepatentsonline.com/WO2008018977.html>, January 2009.