

# Arabic Orthography vs. Arabic OCR

## Rich Heritage Challenging A Much Needed Technology

Mohamed Attia

**H**aving consistently been spoken since more than 2000 years and on, Arabic is doubtlessly the oldest among the major contemporary languages. Over that amazingly long history, this language has been able to respond to the civil needs of consecutive ages, and also to react with the geographical and ethnic expansions of its speakers from only the limited inhabitants of the Arabic peninsula to currently 300 million native speakers (called collectively Arabs) plus numerous tens of millions of non native ones among almost 1 billion non Arab Muslims.

Although the basics of phonology, morphology, grammar, ..., and the other Arabic discriminant components remained essentially the same, some aspects of Arabic have naturally and continuously been evolving with the aforementioned time passing and expansion of the speakers base. One such aspect is the Arabic orthography which is also used nowadays to transcript other widely spoken languages such as Persian (official language of Iran) and Urdu (official language of Pakistan), and also used to be the transcription format of others like Turkish (until the thirties of the twentieth century).

### A. Historical Background

**I**n the very early stages; writing was not so common among the people of the world in general and among Arabs in special who used to mainly communicate via speaking. Being - by that time - mainly Bedouin troops isolated in the severe deserts of the Arabic peninsula with a superior talent of composing and memorizing poetry and with a little need for official documentation, the minority of Arabs who had the writing and reading ability were satisfied by a relatively simple orthographic scheme.

This scheme - detailed in the next section - is based on 28 alphabetical characters {Alif, Baa, Taa, ..., Ha, Waw, Yaa} each represented mainly by a basic shape (and variants) called *grapheme*. The (complicating) simplification was that sets of different characters are represented by the same grapheme! While such a scheme with only 15 (or 16) graphemes is obviously quite ambiguous, the educated minority of olden talented Arabs had no problem communicating with it among each other. If one of them were to write the sentence "*Translation is a basic means of the mutual exchange of civilizations among the peoples over the ages*", the result would look like figure 1.

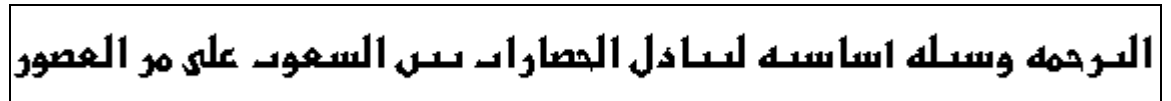


Fig. 1 ([BareArabicOrthography.bmp](#)); A sample sentence written in bare Arabic orthography used in the early ages of the language.

**W**ith the emergence of Islam by the early seventh century AD in the Arab peninsula, *Qur'aan*; the holy book of Muslims and the basic source of their

jurisprudence has been revealed in Arabic. The early Muslims who were mainly Arabs carefully documented the holy Qur'aan in the aforementioned bare style which was good enough for them. Few decades later, when Islam became the religion of many non Arab peoples, mistakes at reading the holy Qur'aan were experienced. Serious as it was, the threat of misconstruing the holy Qur'aan hardly alerted the Arab linguists of the time to the ambiguity of the bare orthographic style.

It was then logical for the (less ambiguous) *dotted* orthographic scheme to replace the bare one resorting to the rule that "Each character is represented by a basic grapheme which has a unique shape". In order to comply with its antecedent, the dotted scheme *cleverly* added discriminating dots over and under ambiguous graphemes. Using the latter scheme, the transcript of our sample sentence looks like figure 2.

**الترجمة وسيلة أساسية لتبادل الحضارات بين الشعوب على مر العصور**

Fig. 2 ([DottedArabicOrthography.bmp](#)); The same sample sentence written with dotting to remove character identification ambiguity.

**T**o this point the Arabic orthography could perfectly describe the spelling of text, but the phonetic transcription was still to be inferred by a knowledgeable reader who had - among other tasks - to supply the short vowels and differentiate whether a character belonging to {Alif, Waw, Yaa} is a consonant or a long vowel. Again, the new comers to Islam from an ever expanding area outside the Arabic peninsula were troubled by all these jobs while reading the holy Qur'aan written in the dotted orthographic scheme.

In response to this problem, Arabic linguists devised later an elaborate Arabic orthographic scheme containing many *diacritical* marks (or simply *diacritics*) and punctuators as well as a wide set of reading rules that all completely and unambiguously determine the exact phonetic transcription of the holy Qur'aan in special, and any written Arabic text in general. This new scheme was called the *Ottoman* orthography and became to date the exclusively approved style for transcribing the holy Qur'aan from which we show a sample page on figure 3.



Fig. 3 (OttomanOrthography.bmp); One page of the holy Qur'aan - describing the creation phases of the human being - written in the Ottoman orthography.

**W**ith the maturation of the respective Muslim states (actually empires) from Abbasids to Ottomans, the extensive need for all kinds of official, intellectual, technical, ..., etc. documentation turned the Arabic orthography into a rigorous science which produced very early the concept of *font* in a wide range of variety from the practical (see figure 4) to the artistic (see figure 5). Arabic orthography was then doubtlessly so ready for the age of printing, and later for the age of digital computers.

الترجمة وسيلة أساسية لتبادل الحضارات بين الشعوب على مر العصور

Fig. 4 (PracticalFont.bmp); An example of a practical Arabic font (now called Traditional Arabic) from the Naskh family.

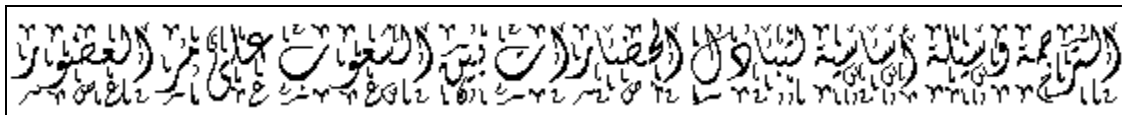


Fig. 5 (DiwaniFont.bmp); An example font of the Diwani fonts family with artistic effects.

## B. Challenges Of Modern Arabic Orthography To OCR Technology

Equipped with the background presented above, we are in a good place to spot the most challenging features of Arabic orthography to the OCR technology.

- 1- **The connectivity challenge:** Whether handwritten or typewritten, Arabic text can only be scripted in connected (or *cursive*) mode; i.e. graphemes are connected to one another within the same word with this connection interrupted at few certain characters or at the end of the word. This necessitates any Arabic OCR to do not only the traditional grapheme recognition task, but also another tough grapheme segmentation one (see figure 6). To make things even harder, both of these tasks are mutually dependent and must hence be done simultaneously.

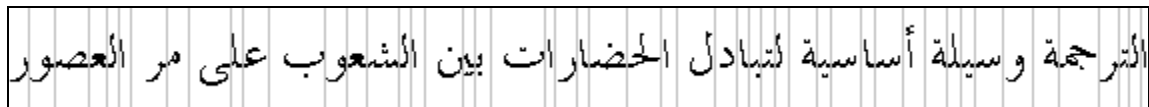


Fig. 6 ([GraphemeSegmentation.bmp](#)); Grapheme segmentation process illustrated by manually inserting vertical gray lines at appropriate grapheme connection points.

- 2- **The dotting challenge:** As stated before; dotting is extensively used to differentiate characters sharing similar graphemes. According to figure 7 where some example sets of dotting-differentiated graphemes, it is apparent that the digital differences between the members of the same set are small. Whether the dots are eliminated before the recognition process, or recognition features are extracted from the dotted script, dotting is a significant source of confusion – hence recognition errors – in Arabic typewritten OCR systems especially when run on noisy documents; e.g. those produced by photocopiers. On the contrary, dotting may be helpful for Arabic handwritten OCR systems as dots are usually sensed as separate short strokes.

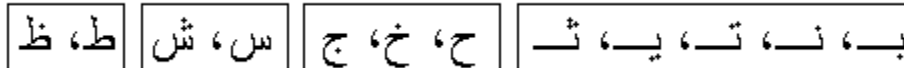


Fig. 7 ([DottingOnSimilarCharacters.bmp](#)); example sets of dotting-differentiated graphemes.

- 3- **The multiple grapheme cases challenge:** Due to the mandatory connectivity in Arabic orthography; the same grapheme representing the same character can have multiple variants according to its relative position within the Arabic word segment {Starting, Middle, Ending, Separate} as exemplified by the 4 variants of the Arabic character "Ein" highlighted in red in figure 8.

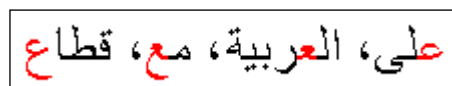


Fig. 8 ([MultipleCasesOfGrapheme.bmp](#)); The 4 cases; Starting, Middle, Ending, and Separate cases of the grapheme representing character "Ein" highlighted in red.

- 4- **The ligatures challenge:** To make things even more complex, certain compounds of characters at certain positions of the Arabic word segments are represented by single atomic graphemes called *ligatures*. Ligatures are found in almost all the Arabic fonts, but their number depends on the involvement of the specific font in use. Figure 9 illustrates some ligatures in the famous font "Traditional Arabic" highlighted in red.

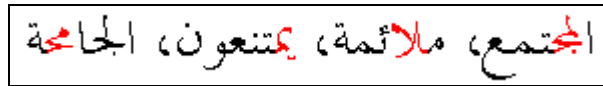


Fig. 9 (Ligatures.bmp); Some ligatures in the Traditional Arabic font highlighted in red.

- 5- **Broad graphemes set:** Multiple grapheme cases as well as the occurrence of ligatures directly lead to broad grapheme sets; e.g. a common highly involved font like Traditional Arabic contains around 190 graphemes, and another common less involved one (with less ligatures) like Simplified Arabic contains around 95 graphemes. Compare this to English where 40 or 50 graphemes are enough! Again, a broader grapheme set means higher ambiguity, and hence more confusion.
- 6- **The diacritics challenge:** Unless the reader is knowledgeable enough, each character in Arabic strictly needs one or more diacritical marks to be drawn over or under the corresponding grapheme in order to ensure the intended phonetic transcription and hence the correct pronunciation. Apart from the teaching purposes, Arabic diacritics are used in practice only when they help in resolving linguistic ambiguity of the text. The problem of diacritics with typewritten Arabic OCR is that their direction of flow is vertical while the main writing direction of the body Arabic text is horizontal from right to left. (See figure 10) Like dots; diacritics – when existent – are a source of confusion of typewritten OCR systems especially when run on noisy documents, but due to their relatively larger size they are usually preprocessed.

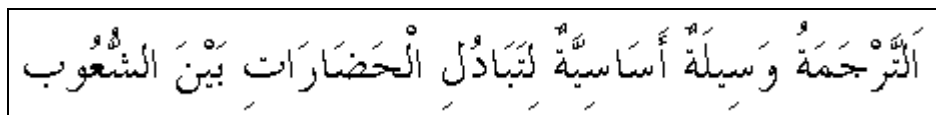


Fig. 10 (DiacritizedText.bmp); Diacritics added to Arabic text.

### C. Current state of the art

Among numerous applications of typewritten OCR systems comes Document Management Systems (DMS) as the largest industrial consumer. In such systems scanned images of electronically unavailable documents are archived by the DMS, meanwhile OCR are run on each scanned image. While the images are used for viewing the document, the text resulting from OCRing the images is used for all kinds of Information Retrieval (IR) and Knowledge Management (KM) purposes which are insensitive to the inevitable error rate of the OCR process as long as this rate is kept small enough (< 4% of the word rate is a rational criterion).

As the Arabic market – esp. in the Gulf countries – is currently a hot one, there is a quite high need for reliable typewritten Arabic OCR engines to be integrated in such DMS systems. Perhaps the – by far – most ready and best equipped system is Automatic Reader<sup>®</sup> 7.0 provided by Sakhr<sup>®</sup>. Affording document retrieval, omni Arabic OCR, learning mode, 95% word-level accuracy rate, SDK for integration, and being able to deal with bilingual Arabic-Latin documents, this system is the best-to-choose-now for heavy duty serious applications. For more details on this system; the reader can visit:

[http://www.sakhr.com/Sakhr\\_e/Products/OCR\\_Off.htm?Index=2&Main=Products&Sub=OCR](http://www.sakhr.com/Sakhr_e/Products/OCR_Off.htm?Index=2&Main=Products&Sub=OCR)

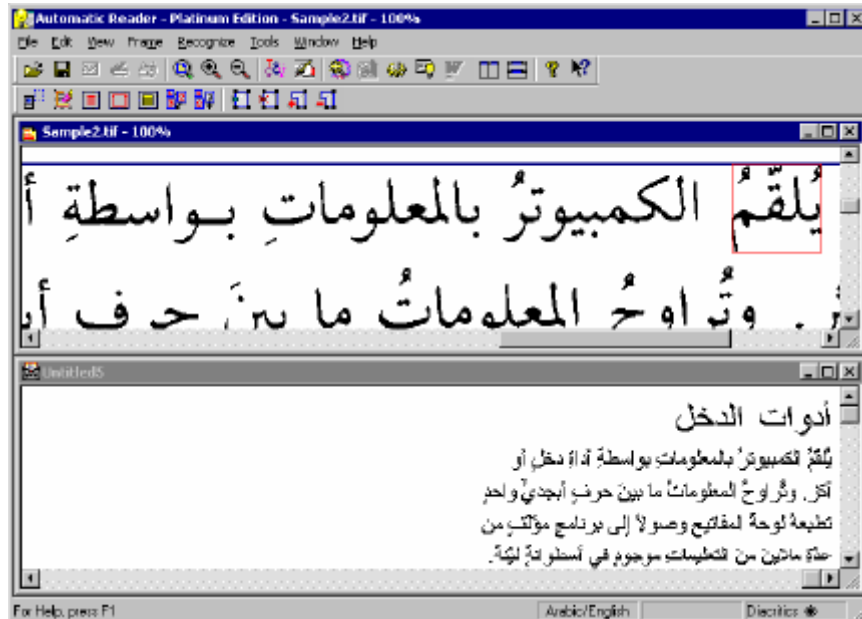
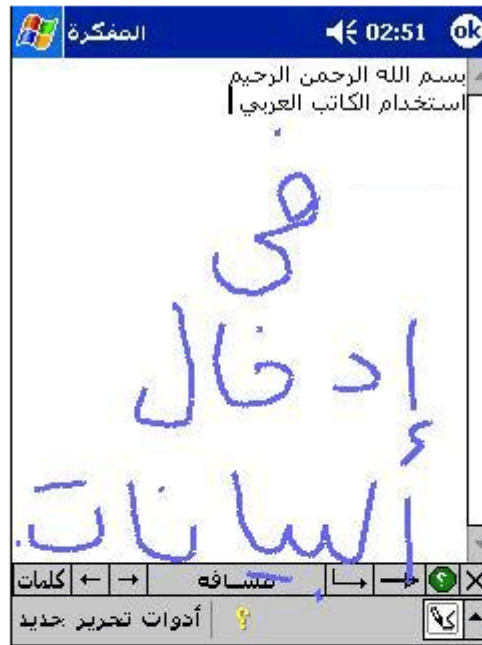


Fig. 11 (AutomaticReader\_Sakhr.bmp); A screen capture of Automatic Reader<sup>®</sup> 7.0 (Platinum Edition) from Sakhr<sup>®</sup>

Automatic Reader<sup>®</sup> 7.0 from Sakhr<sup>®</sup> is essentially based on a huge set of ad hoc orthographic rules and tips put in a work frame of AI searching techniques to decide on each of the Arabic OCR phases; preprocessing of tiny blocks like dots and diacritics, segmentation, and grapheme recognition. The last phase of synthesizing the recognized text is cleverly guided by Sakhr's Arabic NLP tools for filtering out nonsensical results.

On the other hand, online handwritten OCR has turned into a real business with the booming of the keyboardless hand held devices. Beyond the academic pilots, practically functional Arabic handwritten OCR systems are rare, and the product of Arabic Writer<sup>®</sup> from ImagiNet<sup>®</sup> can be selected as a representative one. The underlying methodology of this system is to train and deploy artificial Neural Networks to decide on the most likely character sequences corresponding to the dynamically sensed features sequences of curvature, with a preprocessing of short strokes corresponding to dots and diacritics. For more details on this system; the reader can visit: <http://www.imagnet-software.com/index.aspx>



**Important Note:**

The objective of this product is to provide Arabic Pocket PC users a more comprehensive and easier method than pinpointing the keyboard and writing isolated characters.

**It's important in order to avoid over expectation to mention that this current version is not a free handwriting recognition system.**

There are some few guidelines regarding Arabic handwriting styles which when followed will guarantee high accuracy and will make the system more usable than the on-screen keyboard.

Please read the guidelines before using the system: [click here for Arabic handwriting guidelines.](#)

Future Versions will accommodate a wider range of writing styles and there will be FREE upgrade.

Fig. 12 ([ArabicWriter\\_ImagiNet.bmp](#)); A snapshot of Arabic Writerr<sup>©</sup> from ImagiNet<sup>®</sup>. Note the caution box!

**F**or such a needy language like Arabic; there is a wide room for enhancement either to lower the error rate of typewritten systems and/or allow for completely free hand writing style of online ones. The research group of Prof. Mohsen A. A. Rashwan and his post graduate students in the faculty of engineering of Cairo University-Egypt may be regarded as a representative one. They are trying a fully mathematical approach based on an analogy to the ASR (Automatic Speech Recognition) where (phoneme-grapheme) segmentation and recognition are done simultaneously using HMM techniques applied on feature-vector sequences extracted via a sliding window in the writing direction. Besides being a cleaner architecture, this promising approach for Arabic typewritten – and also online – OCR has the virtue of realizing an enhancing accuracy and noise immunity with the increase of training data as is the case with ASR.

**Mohamed Attia** is the Arabic NLP team leader in The Engineering Company for The Development of Computer Systems; RDI, [www.RDI-eg.com](http://www.RDI-eg.com), and is also a PhD student in the Faculty of Engineering, Cairo University, Egypt. He can be contacted at [m\\_Atteya@RDI-eg.com](mailto:m_Atteya@RDI-eg.com) or [m\\_Atteya2004@Yahoo.com](mailto:m_Atteya2004@Yahoo.com)

