ア NB! *Correspondence Round table Arabic/Farsi OCR*

CONTENTS

Color plates: *Muraqqa'*. Album of the Indian and Persian Miniatures of the 16—18th Centuries
and the Models of the Persian Calligraphy of the Same Period (see p. 63—67).

Front cover:
Fol. 17a. Portrait of a Man by Riḍā-yi 'Abbāsī, 11.8 × 8.2 cm.

Back cover:
Plate 1. Fol. 16a. Portrait of Tīmūr Khān Turkmān by Ṣādiqī beg Afshār, 19.3 × 11.6 cm.
Plate 2. Fol. 36a. The Darvishes Picnic in the Mountains. Probably Isfahan school, 25.5 × 14.5 cm.
Plate 3. Fol. 6a. The Shaykh and the Harlot by Muḥammad Yūsuf Muṣavvir, 18.2 × 11.3 cm.
Plate 4. Fol. 1a. Portrait of Mīrzā Jalālā by 'Alī Qulī beg Jabbādār, 16.0 × 9.1 cm.

Manuscripta Orientalia

International Journal for Oriental Manuscript Research

Vol. **1** No. **3** December 1995

J. Bell & P. Zemanek

TEST OF TWO ARABIC OCR PROGRAMS

From the fourteenth to the eighteenth of December we met in Bergen to experiment with the two of the OCR programs for Arabic that were available in the software market as of November 1994. One of these was TextPert 3.7 Arabic, produced by CTA, Inc., which runs on the Macintosh Arabic system (system 7.1 was used in the test) [1]. The other was al-Qari' al-Ali (Arabic "Automatic Reader") 1.0, upgraded to 1.1, a version of the program known as MULTREC. It is produced by al-Alamiah Software Co. and runs on al-Nawafi<u>dh</u> al-Arabiya, the Arabization program for Windows from the same company [2]. Taking part with us were administrative assistant and librarian Awni Taki Musa and undergraduate student Navid Saminasab.

The limited time and means at our disposal did not allow us to try out a third program, ICRA 4.0, which is an application for Windows (with Arabic Support) produced by Arab Scientific Software & Engineering Technologies (cf. the communications by Jan Hoogland, Discussion Forum on Personal Computers Arabization, Dec. 21, 1994; Itisalat, Jan. 5, 1995). Subsequently Jan Hoogland was himself able to compare al-Qari' al-Ali with ICRA, and found al-Qari' al-Ali to be superior (1) in character recognition, (2) in training for ligatures, (3) in the fact that the spelling checker is linked (initially) to the original image, and (4) in that the batch mode does not require confirmation after every page (cf. Itisalat, May 4, 1995).

Another program which has been discussed recently, one using neural-net based software from Mitek Systems in San Diego, was as of late November not yet available, and the company could provide no comparison results.

Both of the programs we tested were able to recognize certain computer printed texts of good quality with a reasonable degree of accuracy considering the difficulties of the Arabic script. Both were many times slower than comparably priced programs for Latin OCR, also when reading Latin.

TextPert

TextPert is a program which is extremely easy to use, but which offers in the normal version no means of influencing character recognition other than adjustment of resolution, brightness, and contrast on the scanner. Thus it was not possible to choose, or to train for, the fonts we were scanning. On very good and simple texts the results were approaching acceptable standards, but on more complicated fonts the program recognized virtually nothing. Moreover, on the computers we used (a PowerBook 180 with 14 Mb of memory and an LC III with 8 Mb of memory), the program was not always able to follow the paths between the automatically established zones on the document to be read. When it could not do this, the Macintosh would crash. There is a much faster and three or four times more expensive version of Arabic TextPert which uses a RISC board. We have been told by the company that it does not perform essentially differently from the cheaper version except for speed, but that they may allow access to the engine for certain purposes the user may require. For Macintosh users who only want to scan certain kinds of computer produced documents, TextPert may offer something approaching an acceptable solution, but it is to be hoped that future versions will take into account the need to train for different fonts.

Al-Qari' al-Ali

This program is based on a very powerful algorithm which seems to combine vector and bit-map analysis. In its first upgraded version it offers a number of means, although still not quite enough, of controlling recognition performance. Thus it is possible to select desired level of accuracy and to train for the majority of fonts in Arabic and in most other scripts. The results of an OCR operation can be controlled with a spelling checker that, while far from what one might hope for, is surprisingly good, particularly for controlling words that have run together. To facilitate comparison between the original scanned image and the text document, the spelling checker highlights problem areas simultaneously in both.

The texts on which we tried al-Qari' al-Ali were for the most part photocopies from works printed in the late nineteenth century in relatively complex fonts (for example Shaykh'zadah's *Hashiyah* on al-Baydawi printed in Constantinople in 1306/1888—1889). There were quite a few

breaks between letters, and spaces as often as not occurred in the middle of words, rather than between them. The results were none the less impressive, although anyone interested in scanning texts of this type must be prepared to invest a great deal of time both before and after scanning.

The text documents we produced using al-Qari' al-Ali were later converted for the Macintosh using a conversion table we made in Paradigma 2.0, a program designed by Espen Aarseth at the University of Bergen. The PC Arabic system handles ligatures and initial and final forms differently from the Macintosh, and word boundaries in the text document produced by al-Qari' al-Ali were often clear on the PC even when there was no actual space between the words. These boundaries disappeared when the text was converted for the Macintosh. Since at this stage in the program's development the adding and subtracting of spaces has to be done manually, it is probably better to carry out this part of the correction process on a PC, even if one intends to continue working on a Macintosh later on. We understand that a Macintosh version of the program is under development, but we have no information about how this particular problem will be handled. Perhaps the best results can be achieved, once preliminary spellchecking has been carried out, by converting the text from the Alamiah Nawafidh Windows code page to that used by Arabic Windows 3.1 or 3.11 by means of the utility al-Muhawwil that comes with these version of Arabic Windows, and then continuing correction of the text in the Arabic version of Microsoft Word for Windows 6.0.

The very considerable amount of time it takes to train for new fonts, especially hand set fonts with many ligatures, is one of the main problems with al-Qari' al-Ali. Even when teaching Latin fonts the process went slower and the operations were more cumbersome than, for example, in the bit-map program ProLector, which, however, is considerably more expensive. Quicker routines for training fonts would be a great improvement. A feature that al-Qari' al-Ali has which is not in ProLector, is the possibility of editing bit-map models within the program and inserting them into a set of previously trained models. Al-Qari' al-Ali has an English menu option, so it can in fact easily be used by persons unfamiliar with Arabic. The Arabic menus in al-Alamiah's Nawafidh Windows constitute only a slight problem for experienced Windows users who do not know the language.

Because the program, although slow, seems so powerful and so promising, we would like to note some problems which we hope the developers will take into account in future upgrades.

✓ **Manual.** Although the manual may look nice, it contains only very superficial information and needs to be entirely redone. An English version would also be helpful.

✓ **Image rotation.** We did not find, within the program, a tool for gradual rotation of the images to be scanned or read. Such a feature would make it easier to maintain a constant alignment of scanned images so that the program always sees the characters it is to learn or read from the same angle.

✓ **Recognition blocks.** Al-Qari' al-Ali places groups of connected letters into a green frame and what it thinks are individual letters within the group between horizontally adjustable red lines inside the green frame. Neither the width nor the height of the green frame can be manually

adjusted, which means that characteristic elements of a block are on occasion excluded or extraneous information included. Within the green frame, the program lets one know what it is taking as characteristic of a letter by outlining it in blue. It would be helpful, if it is possible, to have a means, in addition to the red lines, of activating or deactivating the blue outline where the program has made a mistake. The program will have certain difficulties with complex fonts until these problems are remedied. For the moment the best guideline seems to be not to override the program's choice when training any more than necessary, since it is not unlikely that it will make the same choice again anyway. When the program has seen a medial letter or ligature as one in isolation because of breaks in the word, for example, the "in isolation" choice at times has to be accepted. Otherwise the program may fail to read the letter or ligature, or read it as something else.

✓ **Fonts.** The program comes with few pre-trained fonts, and those it does provide are computer fonts with few ligatures. Given the amount of time needed to train fonts, the library of pre-trained fonts, particularly non-computer fonts, needs to be greatly expanded. Further, the program lacks an efficient means of visually comparing fonts to be read with the pre-trained fonts, since the font display window in the "create/emend" font library dialogue box gives an inadequate image of small fonts. Lastly, there is in the present version no means of scaling up or down previously trained fonts, which means that every font in every size has to be trained separately. However we have been told by the company that in the next version it will be possible to reproduce fonts in other sizes (plus or minus 2 points in either direction).

✓ **Confusing messages.** One problem we experienced with al-Qari' al-Ali was that when all the places allowed for the variants of a character in a given position had been used up, the warning that appeared was not always the same. A character may have eleven variants in each position (initial, medial, final, or in isolation). When we tried to teach a twelfth variant, the message occasionally stated that we had exceeded some other limit. The problem may have been insufficient memory in the computer we were using, or it may be in the program. In any event, when using the current version of al-Qari' al-Ali one should be aware of the possibility of inappropriate messages appearing.

✓ **Ligature dialogue boxes.** The dialogue boxes for certain combinations of letters, such as "fii", offer only the normal position option, in this case "in isolation" or "final", when in fact in some fonts other positions occur. An "other" button is needed here to allow for the less common options.

✓ **Ligature list.** The window listing optional ligatures gives them in order of creation rather than alphabetically, which in most cases makes it more difficult to find the ligatures one is after. The current method, however, makes it easier to correct mistakes one has just made. If possible, the ligature window should include an optional alphabetical sorting button.

✓ **Limited number of ligature possibilities.** The possibilities for creating (coding for) new ligatures are far too limited for fonts of any complication. It is possible to train new fonts just for ligatures, but this is a complex and

laborious process that inevitably involves duplication, unless a separate record is kept for all ligatures. We need up to 1000 more possibilities just for fairly complicated older fonts. If the possibilities were almost infinite, the program could be used to recognize an almost infinite number of signs, images, and symbols, and al-Alamiah would, as a result, have an almost infinite market for their product.

✓ **Space markers.** Because of the problem with spaces between and within words alluded to above, the al-Muharrir word processor that comes with al-Qari' al-Ali should have an option for marking spaces between groups of letters.

✓ **Stability.** The stability of the program, especially when communicating with the scanner, seems to need improvement. The problem may have been in Arabic Windows or in our hardware. When writing the first version of this review we were using a modest Olivetti 486SX/25 Mhz with 8 Mb RAM and a Hewlett-Packard ScanJet IIcx. Subsequently we have used a Compaq XL 590 with a Pentium 90 processor and 16 Mb RAM. This machine may be a little too fast for the program. Recognition speed has increased by a factor of about nine, but clicking with the mouse does not always stop the recognition feature in the training mode as it should.

We tested incidentally some of al-Alamiah's other software, in particular the word processor Al Ostaz, the Koran database for Arabic Windows, and the hadith databases for Arabic DOS. All of these were impressive products which should receive a warm welcome in any milieu, academic or religious, with a special interest in the Arabic and Islamic heritage.

This review was first made available on the Internet on the lists Reader (14.01.95) and Itisalat (17.01.95), and the original version is preserved in electronic form and in hard copy in the Archive of Electronic Publications of the Section for Middle Eastern Languages and Cultures, University of Bergen (http://www.hf-fak.uib.no/institutter/midtspraak/aep.htm).

Notes

1. Al-Qari' al-Ali 1.1

Producer:

al Alamiah Software Company
al Alamiah Building
Freezone, Nasser city
Cairo, Egypt.
☎ 20-2-2749929
fax: 20-2-2740044
e-mail: alamiah@intouch.com

Requirements:

❑ *IBM compatible; minimum 386 processor;*
486 or Pentium is to be recommended
❑ *4 Mb RAM, 8 Mb recommended, 16 Mb still better*
❑ *mouse*
❑ *Windows 3.1*
❑ *al-Nawifidh al-Arabiya Arabic interface for Windows, version 4.01 or later (same producer)*

2. TextPert 3.7 Arabic

Producer:

CTA, S.A.
c/Joan d'Austria, 68
08005 Barcelona
Spain.
☎ 34-3-4850410
fax: 34-3-4855327
e-mail: textpert.int@applelink.apple.com

Requirements:

❑ Apple Macintosh; all more recent models
❑ 1 Mb RAM minimum, 2 Mb or more recommended
❑ at least 2.5 Mb on HD
❑ Arabic system 6.1 or higher, including 7.1
❑ hardware protection unit