A. Matveev

SAKHR BILINGUAL OCR (AL-QARI' AL-ALI). A USER'S INITIAL IMPRESSIONS

In this paper I would like to record some initial impressions from working with the Sakhr bilingual OCR system known as Al-Qari' al-Ali, to comment certain specific features of the program and to suggest a number of ways in which it may be improved.

It is perhaps appropriate to start with a few remarks on the origin of the product. It was first mentioned in 1990, when Dr. Efim Rezvan of the St. Petersburg Branch of the Oriental Institute of the Russian Academy of Sciences proposed the development of such a program in his report "Computer Methods in Qur'anic Studies" presented at the 2nd Conference and Exhibition on Bilingual Computing in Arabic and English in Cambridge. Originally the program was conceived as a powerful tool to facilitate the preparation of critical editions of Arabic sources by means of transferring large amounts of printed Arabic texts to computer files for subsequent processing. The immediate objective was the preparation by Valeriy V. Polosin of a critical edition of the famous "Fihrist" by Ibn al-Nadim. Dr. Rezvan considered this to be an excellent opportunity to develop and apply new techniques and software, and managed to interest a group of talented young programmers who had worked in the former Soviet high-tech military industry in the project. For a year Alexander Staryh, Mikhail Beregov, Alexander Popov and Fedor Bikov, in collaboration with Efim Rezvan, devoted nearly all their free time to the development of the DOS prototype of the program, which was given the name MULTREC (Multi-Lingual Text Recognizer). The program was demonstrated in 1993 at the 3rd International Conference and Exhibition on Multi-lingual Computing held at Durham, where it aroused considerable interest, since it was virtually the only working program of its type. At this time the software company al-Alamiah became interested in the program, and subsequent to a visit to St. Petersburg by al-Alamiah's General Manager Dr. Ashraf Zaki, the preparation of a new Arabized version of the program was planned. The new version combined the achievements of the Russian programmers with important contributions made by specialists at al-Alamiah.

The first commercial version of Al-Qari' al-Ali was marketed in 1994. This product, although quite useful, has not yet become wide-spread, on the one hand because of its recent appearance and on the other because of its relatively high price and the powerful hardware it requires (a Pentium processor and a scanner with 600 dpi resolution are recommended). Hoping to introduce the product to my colleagues in Arabic studies who may not have had the opportunity to use it yet, I would like to report briefly on some characteristics of the program and how it may be applied.

Characteristics and area of use

Al-Qari' al-Ali works under the operating system "An-Nawafidh al-'Arabiya" 4.01 (or later), which, in turn, is installed over a Windows 3.1 operating system. It allows the transfer of scanned images of printed Arabic materials into text format, yielding 8-bit encoded text files which can be processed with al-Alamiah's word processor "al-Ustadh" or, for example, with the Arabic version of Microsoft Word for Windows 6.0. The program can be used for recognizing any Arabic printed matter. But if the text contains numerous ligatures, which is characteristic of older printed texts [1], errors at "Recognition" are practically inevitable, so the user has to correct them later during "Spell Checking". The best results are obtained from well printed modern texts with a minimum of ligatures. It is possible to transfer rather quickly a modern book or magazine into computer text with few errors (no more than 1%). As for poorly printed older books with a great many ligatures not included on the training keyboard and a variety forms for a given character, the process of recognition is regularly accompanied by errors. With such materials the production of a computer text file is extremely time consuming because of the need for careful correction of the recognized text (at first with the help of the built-in spell checker, and then by checking the corrected text in Word 6.0 or some other word processor). Even so, the production of an Arabic text is much faster than by typing, though it requires a