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-Nawafidh 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