Survey of Language Computing in Asia 2005

Sarmad Hussain
Nadir Durrani
Sana Gul

Center for Research in Urdu Language Processing
National University of Computer and Emerging Sciences

www.nu.edu.pk

www.idrc.ca

Published by

Center for Research in Urdu Language Processing
National University of Computer and Emerging Sciences
Lahore, Pakistan

Printed by Walayatsons, Pakistan

ISBN: 969-8961-00-3

This work was carried out with the aid of a grant from the International Development Research Centre (IDRC), Ottawa, Canada, administered through the Centre for Research in Urdu Language Processing (CRULP), National University of Computer and Emerging Sciences (NUCES), Pakistan.

Arabic

Arabic is a Semitic language spoken by about 206 million people across the world, especially in Middle East and North Africa, where it is also the national language of many countries [1]. There are many dialectical variations of Arabic across this region. It is widely used as a medium of communication in schools, government institutions and media in most of these Arabic speaking countries. Figure 1 below shows the linguistic lineage of standard Arabic [1].

Figure 1: Language Family Tree for Arabic [1]

Arabic script has evolved from the ancient Aramaic script, and has been in use since the 4th century AD. Earliest known Arabic inscriptions date back to 512 AD [2].

Character Set and Encoding

Unicode Arabic script block ranging from 0600-06FF is the standard character set encoding used for Arabic language. There is an additional set of code tables given for Arabic. Arabic Supplement (0750-0764) contains additional Arabic script characters used in African languages and Arabic Presentation Forms A and B (FB50-FDFF and FE70-FEFF respectively). Presentation forms have been introduced for backward compatibility and except for ligatures (FDFx) other codes are not recommended for current use.

ISO 8859-6 is also widely used, and is shown in Figure 2. This standard contains Arabic in addition to basic Latin characters and is an 8-bit standard.

These standards have been derived from earlier standards, e.g. ASMO 449, CODAR-U and ISO 9036. For a more comprehensive overview and reference list see [4]. Microsoft also used Arabic code page 1256 based on these earlier standards [5].

	00	01	02	03	04	05	06	07	08	09	0A	0B	0C	0D	0E	0F
00	NUL 0000	STX 0001	SOT 0002	ETX 0003	EOT 0004	ENQ 0005	ACK 0006	BEL 0007	BS 0008	HT 0009	LF 000A	VT 000B	FF 000C	CR 000D	SO 000E	SI 000F
10	DLE 0010	DC1 0011	DC2 0012	DC3 0013	DC4 0014	NAK 0015	SYN 0016	ETB 0017	CAN 0018	EM 0019	SUB 001A	ESC 001B	FS 001C	GS 001D	RS 001E	US 001F
20	SP 0020	! 0021	" 0022	# 0023	$ 0024	% 0025	& 0026	' 0027	(0028	) 0029	* 002A	+ 002B	, 002C	- 002D	. 002E	/ 002F
30	0 0030	1 0031	2 0032	3 0033	4 0034	5 0035	6 0036	7 0037	8 0038	9 0039	: 003A	; 003B	< 003C	= 003D	> 003E	? 003F
40	@ 0040	A 0041	B 0042	C 0043	D 0044	E 0045	F 0046	G 0047	H 0048	I 0049	J 004A	K 004B	L 004C	M 004D	N 004E	O 004F
50	P 0050	Q 0051	R 0052	S 0053	T 0054	U 0055	V 0056	W 0057	X 0058	Y 0059	Z 005A	[005B	\ 005C	] 005D	^ 005E	_ 005F
60	` 0060	a 0061	b 0062	c 0063	d 0064	e 0065	f 0066	g 0067	h 0068	i 0069	j 006A	k 006B	l 006C	m 006D	n 006E	o 006F
70	p 0070	q 0071	r 0072	s 0073	t 0074	u 0075	v 0076	w 0077	x 0078	y 0079	z 007A	{ 007B	| 007C	} 007D	~ 007E	DEL 007F
80																
90																
A0	NBSP 00A0				¤ 00A4								، 060C	00AD		
B0												؛ 061B				؟ 061F
C0		ء 0621	آ 0622	أ 0623	ؤ 0624	إ 0625	ئ 0626	ا 0627	ب 0628	ة 0629	ت 062A	ث 062B	ج 062C	ح 062D	خ 062E	د 062F
D0	ذ 0630	ر 0631	ز 0632	س 0633	ش 0634	ص 0635	ض 0636	ط 0637	ظ 0638	ع 0639	غ 063A					
E0	ـ 0640	ف 0641	ق 0642	ك 0643	ل 0644	م 0645	ن 0646	ه 0647	و 0648	ى 0649	ي 064A	ً 064B	ٌ 064C	ٍ 064D	َ 064E	ُ 064F
F0	ِ 0650	ّ 0651	ْ 0652													

Figure 2: ISO 8859-6 Code Page for Arabic [3]

Fonts and Rendering

Arabic fonts are widely available. In addition, these fonts are well supported on multiple platforms.

Microsoft Platform

Microsoft ships an exclusive version of Windows and Office products in Arabic language. Microsoft Arabic Windows includes a rich inventory of fonts for Arabic, many of which are not available in the English version of Microsoft Windows.

Linux Platform

Arabic script is fully supported on all Linux based applications. However, Open Type fonts do not exhibit satisfactory results. Arabic distributions provide support for rendering only basic four shaped fonts. Arabic text in Figure 3 is written in Open Office version 1.1.0.

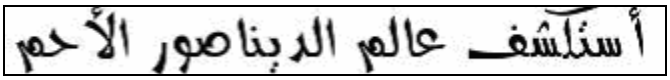


Figure 3: Ae_Dimnah Font in Open Office 1.1.0

Results of rendering Open Type fonts in Linux distributions vary with applications as different applications in Linux use rendering support from different rendering engines. Comparative font rendering results in GNOME, KDE, Open Office and Mozilla are presented below.

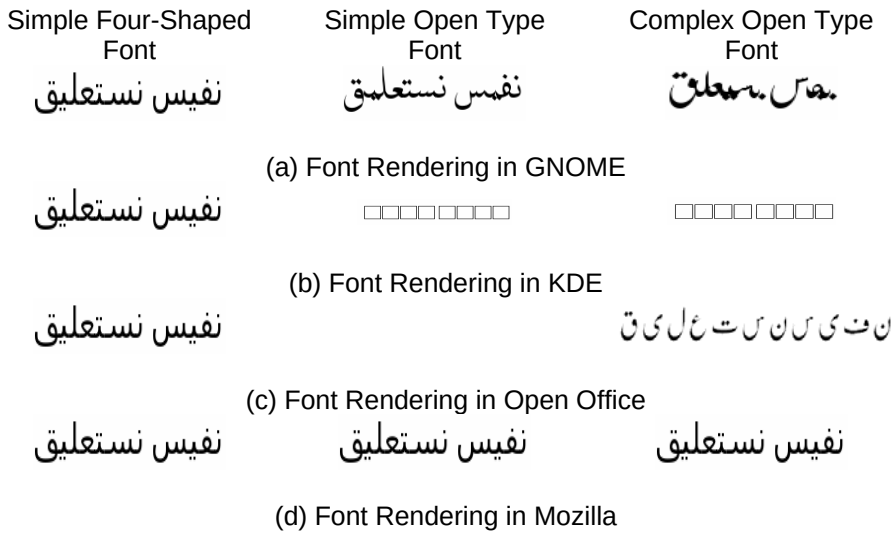


Figure 4: Font Rendering on Linux Platform

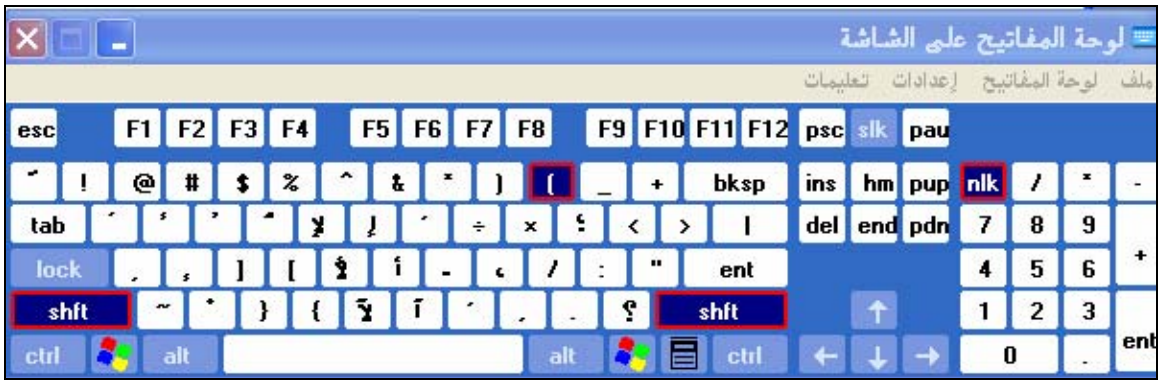
Figure 4 shows that GNOME displays the best results, but still does not work properly for Open Type fonts. KDE does not render Open Type fonts, Open Office only displays True Type fonts, and Mozilla defaults to a simple four shaped True Type font fonts because it fails to display Open Type fonts properly.

Keyboard

Microsoft Platform

Microsoft Windows XP provides an Arabic keyboard layout both in Arabic and English versions. Arabic keyboard layout is enabled by default in Arabic version. This layout is shown in Figure 5 below. Also see [6].

(a)



(b)

Figure 5: (a) Normal and (b) Shift Version of Keyboard on MS Windows

Linux Platform

Red Hat 9 provides an Arabic keyboard layout which can be used both with KDE and GNOME. Once enabled the Arabic keyboard can be used on all Unicode compliant applications.

Collation

Arabic collation is supported in Arabic Windows. LC_COLLATE for Arabic language has not been defined yet. Default sequence for sorting is used, which sorts data similar to original collation sequence for Arabic language.

Locale

Arabic (ar) locales are defined in IBM ICU library and CLDR 1.3 for different countries. They support Arabic date, time and number formats, currency symbol and collation. The locale is available for many countries where Arabic is spoken as a national language (see [7] for a complete list).

Microsoft Platform

Both versions of Microsoft Windows (Arabic and English version of Microsoft XP) provide locale settings for date, time and number formats. They also support currency symbol for some countries. Figure 6 displays the text box to set the country specific Arabic locale in Arabic Windows.

Figure 6: Available Locales of Different Countries in Arabic Windows

In addition, different calendars such as Mailadi and Hijri can also be enabled, as shown for Lebanese Arabic in Figure 7.

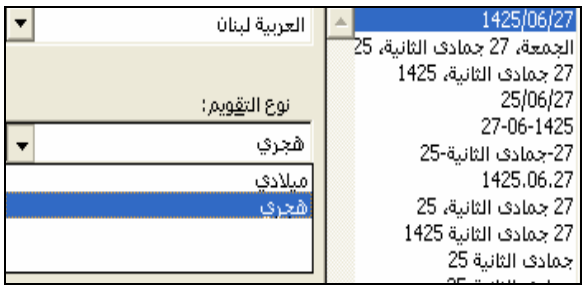
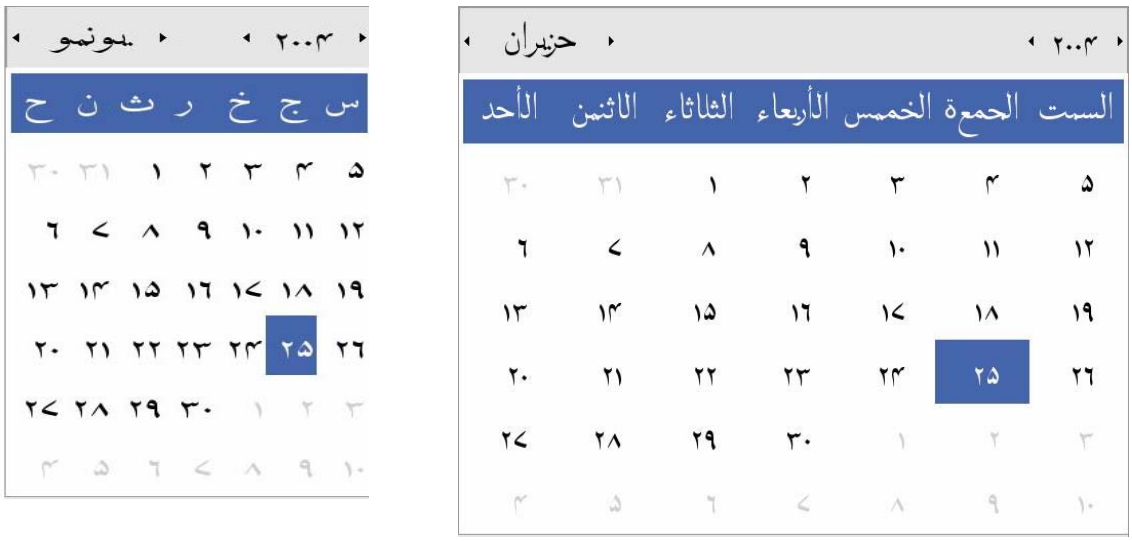


Figure 7: Calendar and Date Settings

Linux Platform

Eighteen locales have been defined in Glibc locale file for different Arabic speaking countries. Each locale includes localized settings for date, time, number formats, and collation specific to the country. Two Arabic locales, Egypt and Lebanon, are also available in Red Hat 9. Figure 8 shows date settings for Egypt and Lebanon from this distribution.



Arabic Egypt

Arabic Lebanon

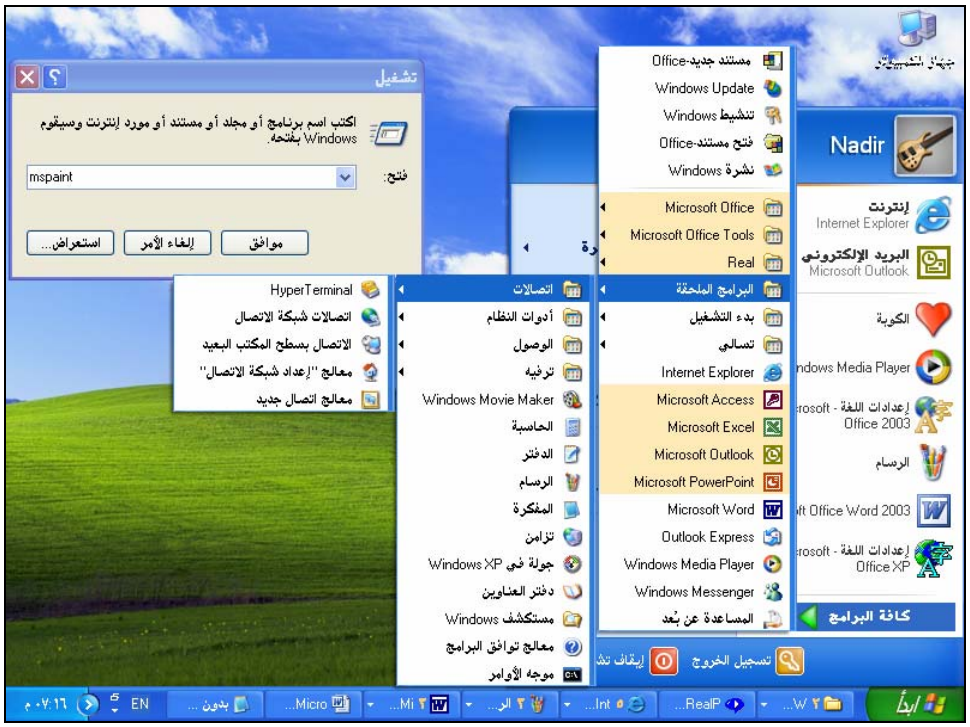
Figure 8: Calendar Settings, in GNOME for Egypt and Lebanon in Red Hat 9

Both Mozilla and Open Office (within Linux distribution) also provide locale support for Arabic.

Interface Terminology Translation

Microsoft Platform

As discussed before, Microsoft ships a complete Arabic version for Windows and Office products. This is different from other language versions, which are based on the English systems and then localized. Arabic version of Microsoft Windows is completely localized in Arabic language. All dialogue boxes, menu, messages and help files have been translated. This version has been available for a long time. Figure 9 below shows the localized version of Arabic Windows desktop and Internet Explorer.



(a)

(b)

Figure 9: (a) Localized Start-Up Menus in Arabic Windows, and (b) Arabic Internet Explorer

Linux Platform

Arrabix, a Debian based Linux distribution, has been localized for Arabic. Arrabix provides complete translation of the graphical user interface in Arabic. It has completely localized versions of all desktop applications including Open Office, Mozilla suite, spell checking utility (Duali), Arabic fonts, localized version of development tools like C/C++ environment, Python Development Kit and also the multimedia software and graphics tools. The following figures show desktop terminology translation in Arrabix CD.

Figure 10: GNOME Start-Up Menus in Arrabix

Figure 11: Terminal in Arrabix 0.8

Interface terminology translation for Arabic GNOME is 51% complete. Only basic desktop files, GNOME lib files, Nautilus browser files, etc. have been partially localized in Red Hat. Interface terminology translation for Arabic KDE is 99% complete. Mozilla 1.6 is partially localized in Arabic. However Mozilla version 1.4 provides maximum terminology translations for Arabic language. The latest version of Open Office has been completely localized in Arabic. All applications of Open Office, including Writer, Spread Sheet and Presentation etc. are available in Arabic.

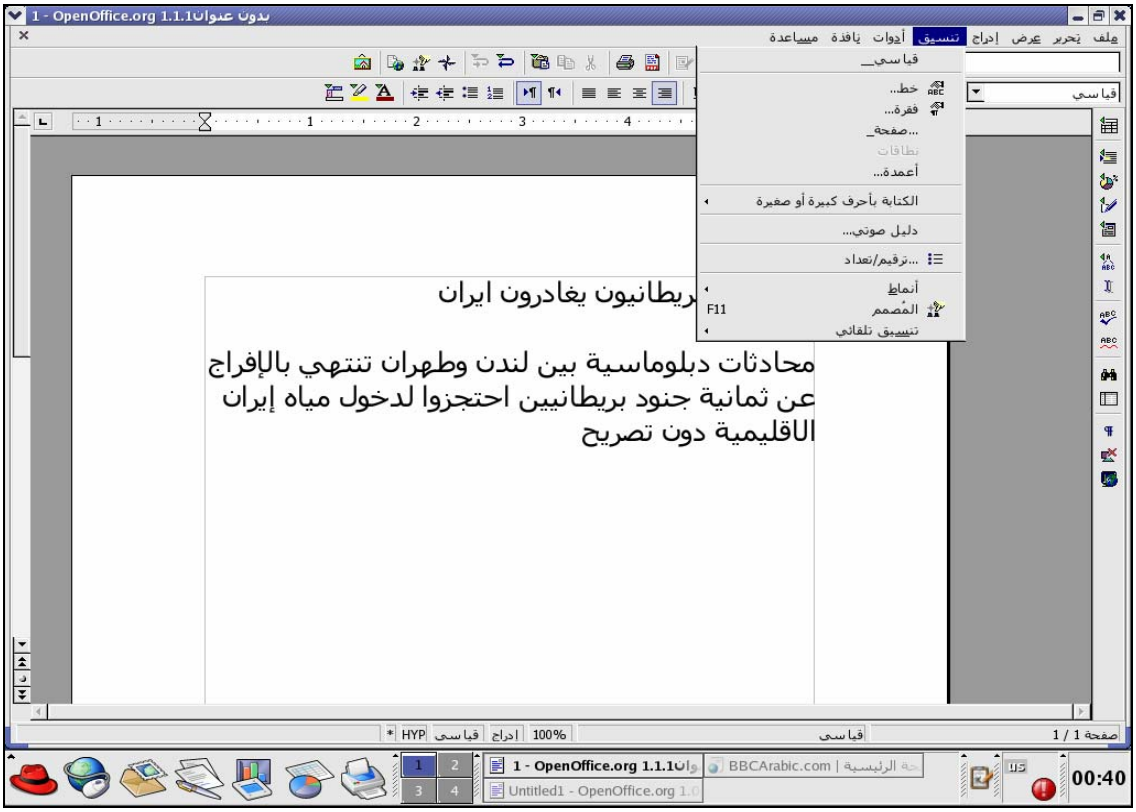


Figure 12: Open Office Writer

Status of Advanced Applications

A lot of work has been done and is currently underway across the world for development of Arabic language applications. Language applications available include encoding converters (e.g. see [8]), bidirectional writing utilities and spell checkers (e.g. Arabic Word Corrector, ARWOC [10], Duali (for Linux) [11], and products by Microsoft and Sakhr [12]).

Sakhr Automatic Reader [12] is a commonly used OCR. Other OCRs include Arabic OCR and ICR (intelligent character recognition) by CIYA [13]. Arabic lexicon, English-Arabic lexicon, Arabic text-to-speech system, Arabic speech recognition system, machine translation between English and Arabic and a variety of other applications are also available through Sakhr [12] and other vendors. For example, other vendor applications include text-to-speech systems by IBM, MBROLA, Lingvosoft and ArabTalk, Speech recognition systems by Natural Speech Communication, machine translation systems for English to Arabic by An-Nakel, ITRANS (Intelligent Translation System) WebTrans™ and Almisbar, and handwriting recognition systems by PackNet Arabizer Suite, IMAGiNET Arabic Writer and IFN/ENET. Some screen shots of these applications are given below.

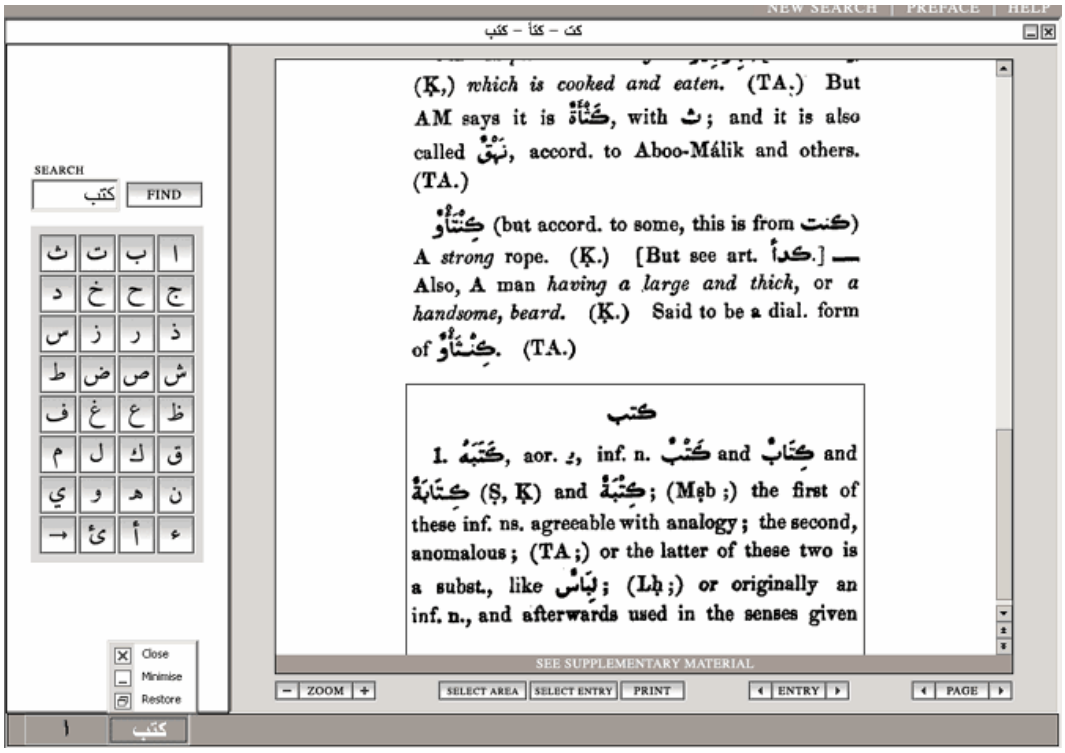


Figure 13: Lane's Arabic English Lexicon [14]

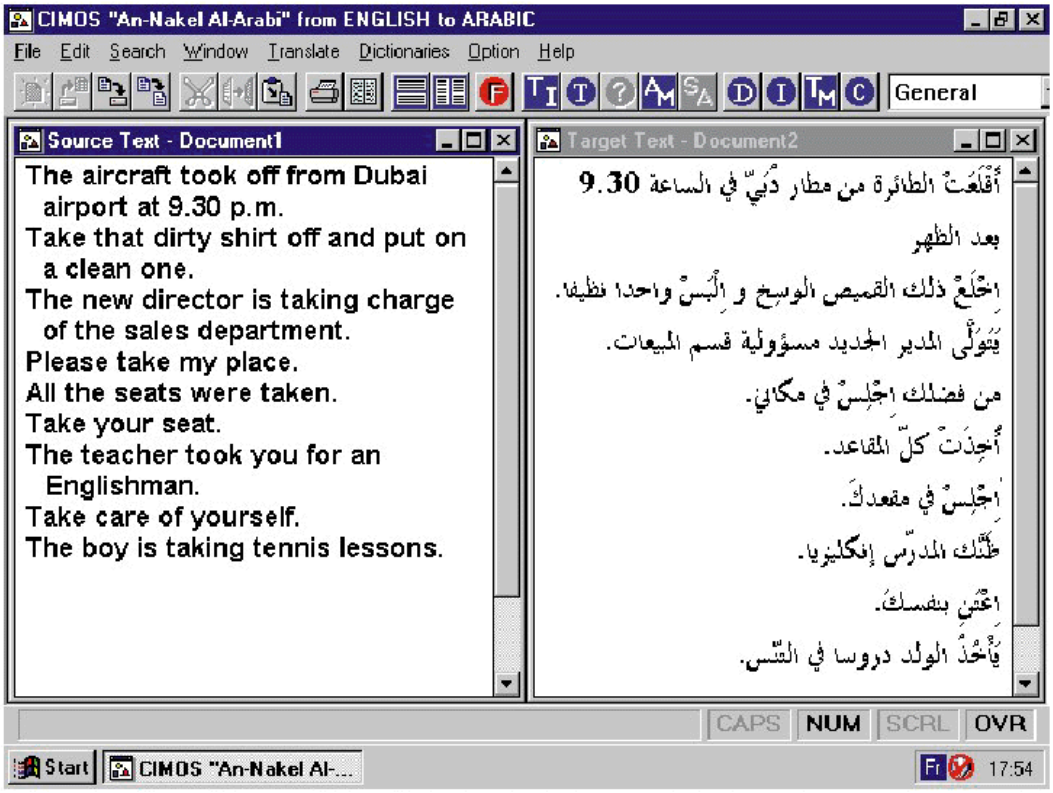


Figure 14: An-Nakel Machine Translation [15]

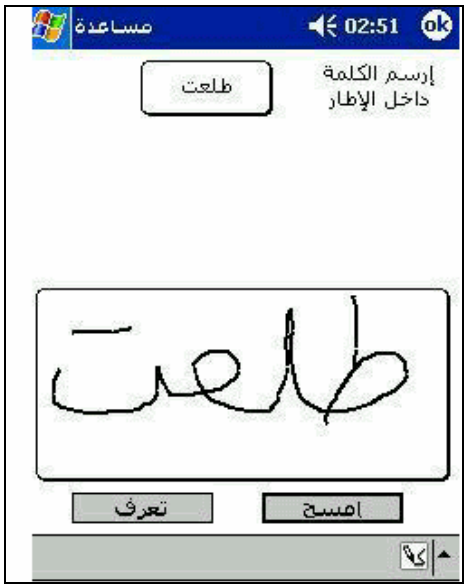


Figure 15: IMAGiNET Arabic Writer [16]

References

[1] http://www.ethnologue.com/show_language.asp?code=arb
[2] http://www.omniglot.com/writing/arabic.htm
[3] http://www.microsoft.com/globaldev/reference/iso/28596.mspx
[4] “Unicode and Arabic Script.” http://ead.staatsbibliothek-berlin.de/2003/unicode/arabic.pdf
[5] http://www.microsoft.com/globaldev/reference/sbcs/1256.mspx
[6] http://www.microsoft.com/globaldev/reference/keyboards.mspx
[7] http://www.unicode.org/cldr/version/1.3.html
[8] http://www.microsoft.com/middleeast/arabicdev/beta/converter/
[9] http://www.unicode.org
[10] http://www.coltec.net/new1/arwoc.html
[11] http://www.arabeyes.org/project.php?proj=duali
[12] http://www.sakh.com
[13] http://members.aol.com/gnhbos/ocroptions.htm
[14] http://www.fonsvitae.com/laneslexicon.html
[15] http://www.translation.net/nakel_translation.html
[16] http://www.pocketgear.com/software_detail.asp?id=10541