PAN
Localization

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

IDRC CRDI
Canada

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.

Khmer

Khmer is the official language of the Kingdom of Cambodia. It is spoken by about 13 million speakers, mostly residing in Cambodia, Vietnam, Laos, Thailand, China, France and the USA [1]. Khmer belongs to Mon-Khmer group of Austro-Asiatic languages (see Figure 1), and shares many features and vocabulary with Thai as a result of centuries of two-way borrowing. Khmer also has significant influence of Sanskrit, Pali, French, and Chinese languages [2].

```
Austro-Asiatic
        Mon-Khmer
                Eastern Mon-Khmer
                           Khmer
                                KHMER, CENTRAL
```

Figure 1: Language Family Tree of Khmer [1]

Khmer alphabet is derived from Brahmi script and resembles Thai and Lao writing systems. The earliest known inscriptions in Khmer, found at Angkor Borei (in Takev Province south of Phnom Penh), dates back to 611 AD [2].

Character Set and Encoding

Unicode chart 1780-17FF [3] is the standard encoding for Khmer character set. This encoding is being increasingly used at the national level. However other ad hoc 8-bit encodings are also being used nationally. Among the few more popular encodings used are Limon, Khek and ABC. Encoding converters between Unicode and these fonts are available through PAN Localization Cambodian component [4] and Khmer OS Project [5].

Fonts and Rendering

As mentioned above, most fonts currently being used are based on ad hoc 8-bit encoding schemes. However, now increasing number of Unicode based fonts are available, which can work on both Microsoft and Linux systems, if rendering support is available. For a list see [5, 7].

Microsoft Platform

Microsoft Office 2003 with Service Pack 1 have now included support for rendering Khmer text, but Khmer fonts are not shipped. Microsoft has published some guidelines to develop Khmer fonts based on Unicode [6].

Linux Platform

Khmer OS has been working on developing rendering support for Khmer on Linux. Qt shipped with KDE 3.3, now supports complete rendering for Khmer in KDE. Patch for Pango has also been developed for support in GNOME [5].

Keyboard

No standard keyboard layout exists for Khmer. AZERTY keyboard layout is commonly used. This is shown in Figure 2 below.

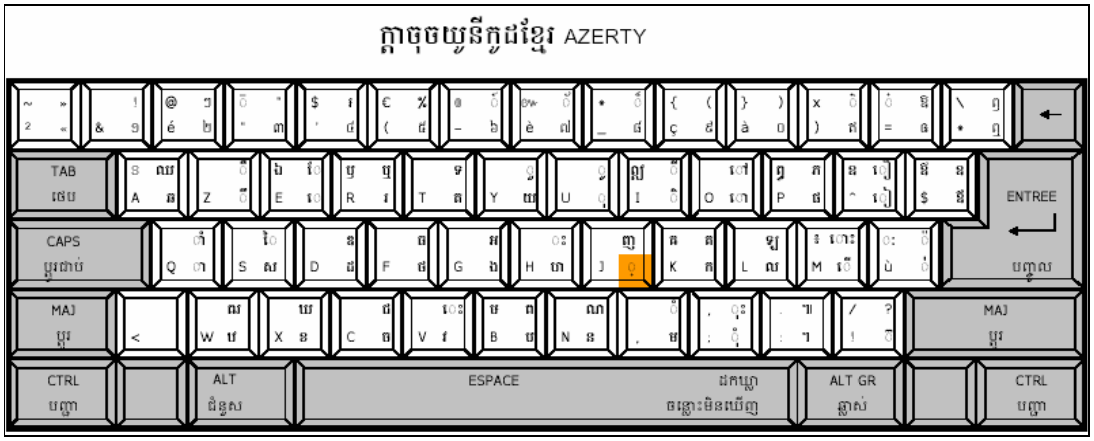
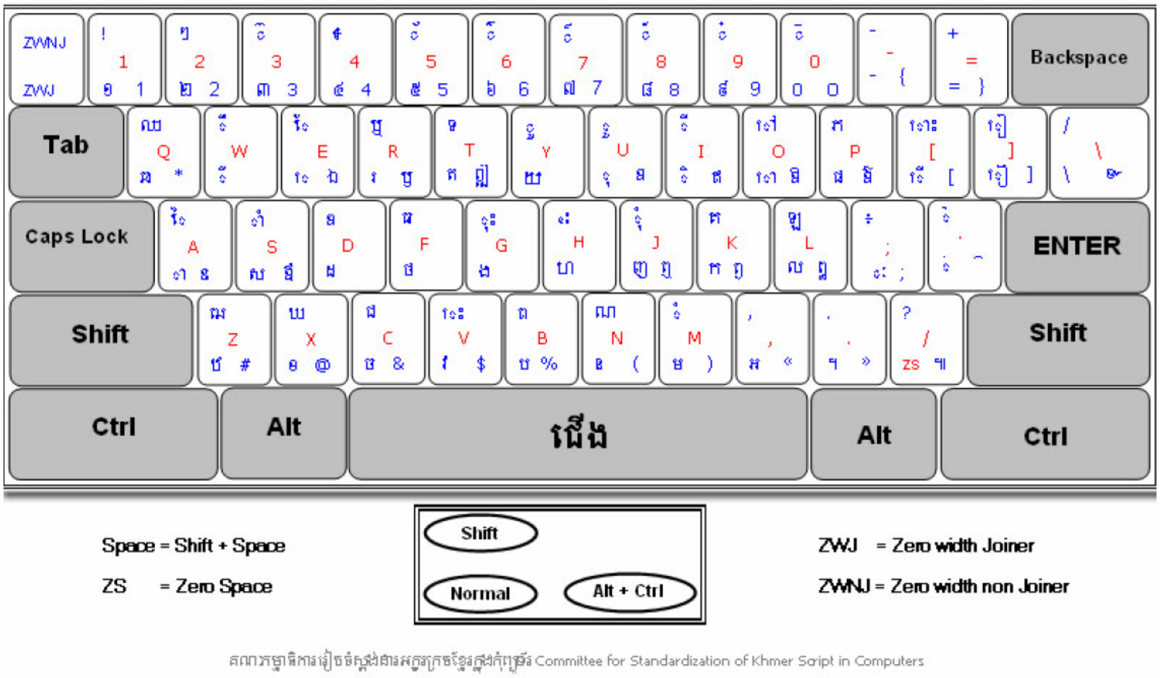
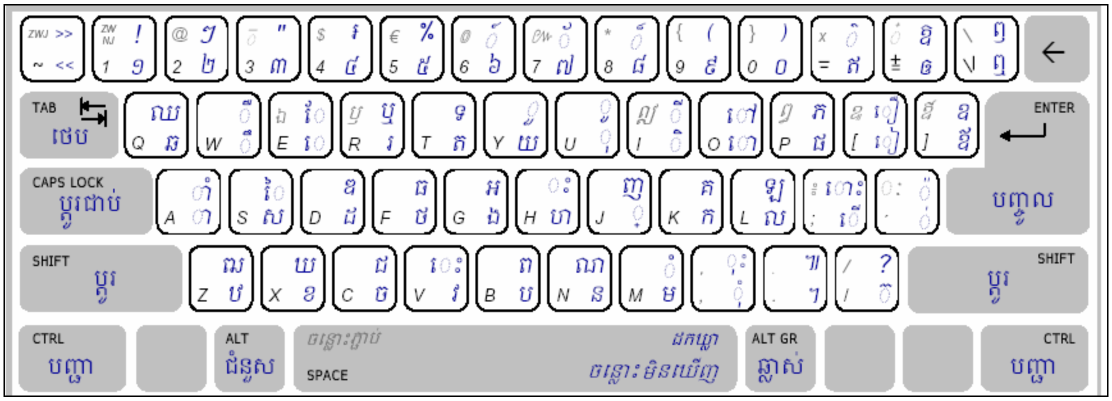


Figure 2: AZERTY French Khmer Keyboard [5]

Additional keyboard layouts are being proposed by various individuals and organizations, especially to cater to Unicode standard, e.g. by National Committee for Standardization Khmer Script in Computers (NCSKSC) and KhmerOS [5] (see Figure 3). Keyboard by NCSKSC has been submitted for standardization.



(a)



(b)

Figure 3: Unicode Based Khmer Keyboards by (a) NCSKSC and (b) KhmerOS [5]

Microsoft Platform

Microsoft does not provide built-in keyboard support for Khmer. However keyboard setups based on MSKLC have been developed (e.g. see Figure 3).

Linux Platform

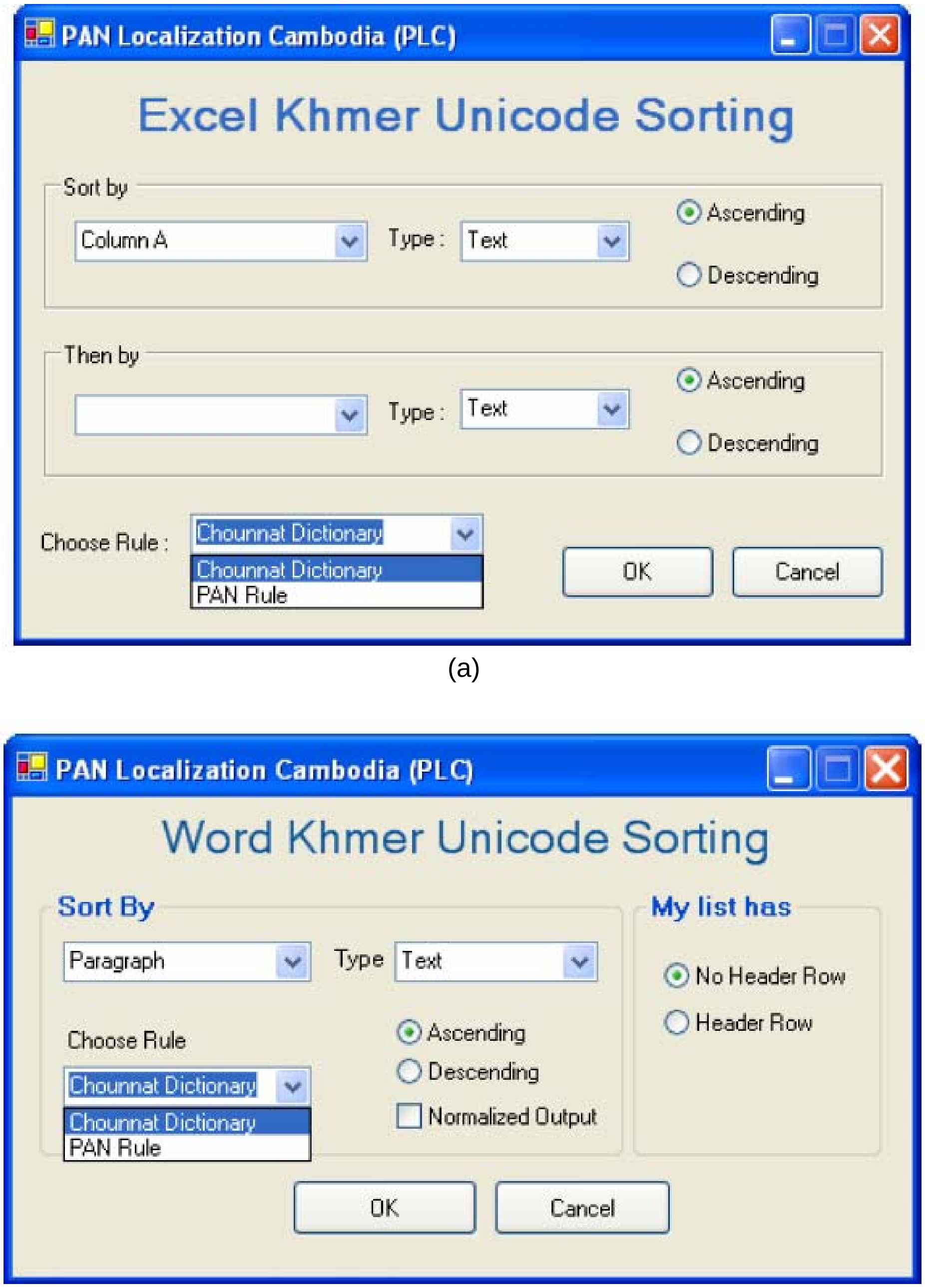
Keyboard driver for Khmer has also been developed. Both AZERTY and KhmerOS versions are available [5].

Collation

Standardization of a single linguistic Khmer sorting sequence is currently under review by national as well as international organizations. This collation is based on the Chuonnat dictionary, the only official Khmer dictionary. A different collation based on Headley's Khmer-English Dictionary is also possible [5].

Microsoft Platform

Microsoft does not support Khmer sorting according to Chuonnat dictionary. However, sorting utility has been developed by PAN Localization Cambodian component [4], which can sort data in Microsoft Office applications. New words, not in Chuonnat dictionary, are sorted through a phonetic mechanism. Screenshots of this application for Word and Excel are given in Figure 4 below.



PAN Localization Cambodia (PLC)

Excel Khmer Unicode Sorting

Sort by: Column A, Type : Text, ● Ascending ○ Descending

Then by: , Type : Text, ● Ascending ○ Descending

Choose Rule : Chounnat Dictionary

Chounnat Dictionary
PAN Rule

(a)

PAN Localization Cambodia (PLC)

Word Khmer Unicode Sorting

Sort By: Paragraph, Type: Text

Choose Rule: Chounnat Dictionary

Chounnat Dictionary
PAN Rule

● Ascending
○ Descending
□ Normalized Output

My list has: ● No Header Row ○ Header Row

(b)

Figure 4: Khmer Sorting Utility for (a) Excel, and (b) Word for Microsoft Office 2003 by PAN Localization Project

Linux Platform

KhmerOS has also developed collation algorithms based on Headly's and Chuonnat dictionaries which can be compiled on Linux [5] and on other platforms.

Locale

Khmer locale is defined in CLDR 1.3 [8]. It has been incorporated in some Linux platforms [5]. However, there is still no support for Khmer locale in Microsoft.

Interface Terminology Translation

No standard terminology translations exist for Khmer. There is no support on Microsoft platform. However, significant work is being done for open source applications through KhmerOS initiative for Microsoft and Linux platforms. Localization of many open source applications has now been initiated. For example a completely localized beta version of Open Office 2.0 is available in Khmer language. Localized Firefox and Thunderbird, Internet and email clients based on Mozilla, are also available.

Status of Advanced Applications

Work is under progress on development of line breaking algorithm by both PAN Localization and KhmerOS projects. Both are lexically based. Algorithm being developed through PAN Localization is also statistical and based on Khmer corpus.

Khmer corpus and lexicon are also being developed by PAN Localization Cambodian component. The lexicon contains lexemes and additional information, e.g. parts-of-speech. In addition, work is also under progress within this project to develop Khmer a spell checker [4]. .

References

[1] http://www.ethnologue.com/show_language.asp?code=khm
[2] http://www.omniglot.com/writing/khmer.htm
[3] http://www.unicode.org/charts/PDF/U1780.pdf
[4] http://www.PANL10n.net
[5] http://www.khmerOS.info/khmerOS_download.html
[6] http://www.microsoft.com/typography/OpenType%20Dev/Khmer/intro.mspx
[7] http://www.travelphrases.info/gallery/fonts_Khmer.html
[8] http://www.unicode.org/cldr/version/1.3.html