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.

Burmese

Burmese belongs to Tibeto-Burman language family and derives from Sino-Tibetan, as shown in Figure 1. It is the official language of Myanmar, where 32 million people speak it as their first language [1, 2]. Some people in China and India also speak Burmese.

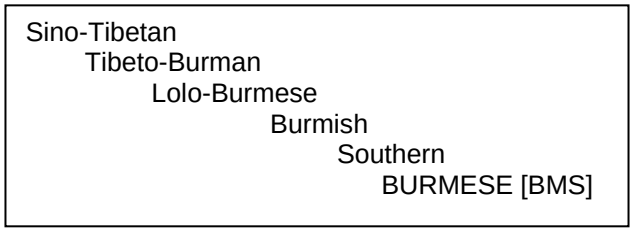

Figure 1: Language Family Tree of Burmese [2]

Myanmar or Burmese script is used to write Burmese language. The script has been developed from the Mon script, adapted from southern Indian Pali script. The earliest known inscriptions in Burmese script date back to 11th century [3].

Character Set and Encoding

Unicode code chart 1000-109F is the internationally standardized character set encoding for Myanmar script [4] but is not frequently used. Two other ad hoc character set encoding schemes, MyaZedi (developed by Solveware Solutions) and Win/CE/Geocomp, are more frequently used at the national level [5].

Fonts and Rendering

Microsoft's support for TTF and OTF fonts is able to render Myanmar fonts but fonts shipped by Microsoft do not support Myanmar. Many Myanmar Unicode fonts have been developed by local vendors and are available, e.g. MyaZedi [6], Pdadauk (Graphite) [7], MyaZedi M17N [8] and Myanmar_OTF [8]. Work is under progress to provide support in Pango rendering engine for GNOME. Mozilla (Firefox and Thunderbird) builds are also partially available in Myanmar [17].

Keyboard

Many keyboard layouts have been developed for Burmese character set. Popularly used keyboard layouts are Win/CE/Geocomp, SCIM-KMFL–Unicode, WIN Myanmar and MyaZedi. In terms of usage WIN/CE/Geocomp and MyaZedi are used widely by business, publishers, and government agencies [9]. SCIM-M17N developed by Myanmar Development Lab is the national standard, but it is not widely used [5]. A comprehensive list is given in [9]. Figure 2 shows the MyaZedi keyboard layout.

If Unicode is used, a more complex input mechanism is required. Unicode has released a short technical note which explains these issues [14].

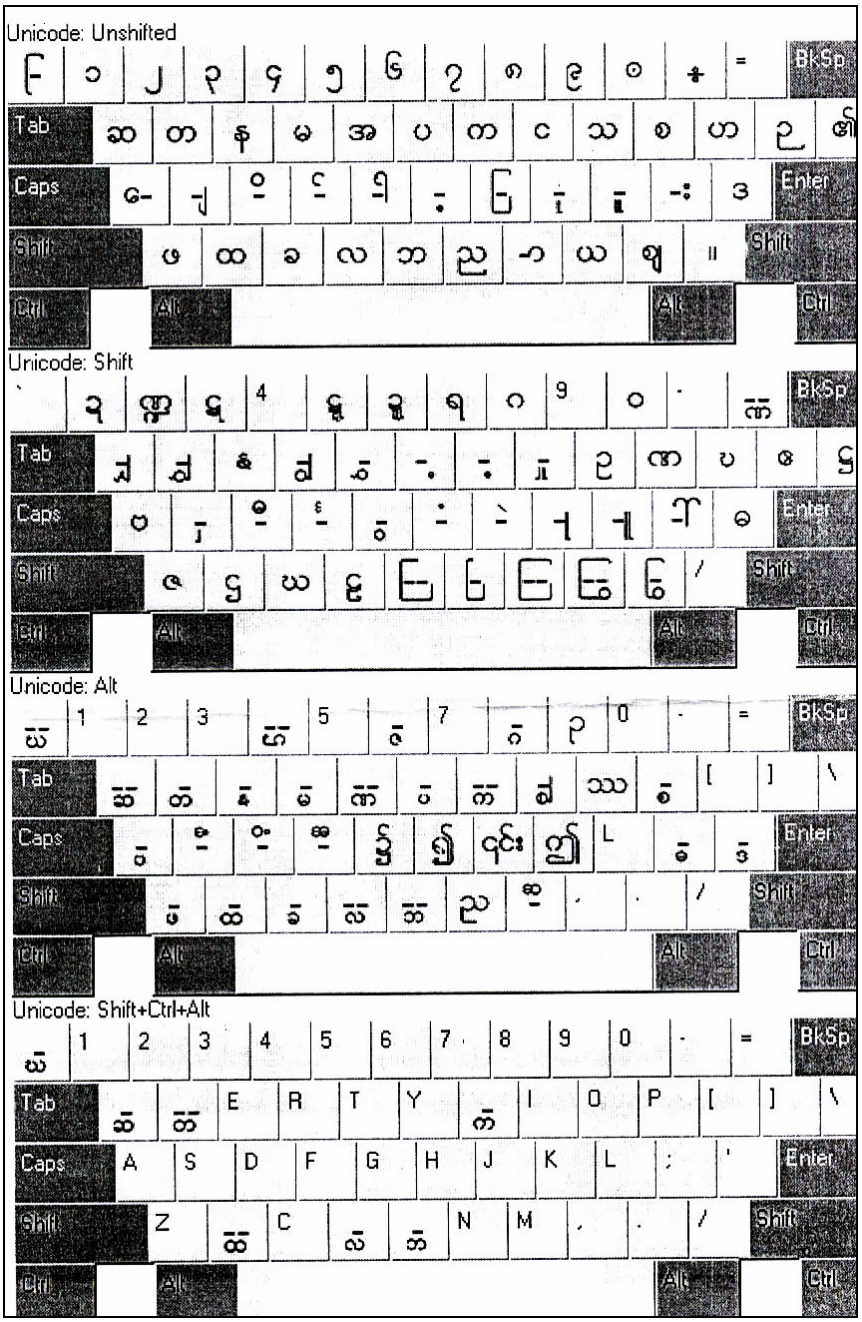

Figure 2: MyaZedi Keyboard Layout [10].

Microsoft Platform

Microsoft does not provide support for a Myanmar keyboard. However various keyboard layouts developed by local vendors can be used on Microsoft platform. A few of these keyboard layouts are based on layouts previously mentioned [9]. Following figures show the four different states of the SOAS Myanmar Keyboard layout, also available for Microsoft platform.

Figure 3: SOAS Myanmar Keyboard Layout [11]

Linux

SCIM-M17N keyboard (by Myanmar NLP Lab [8]) and Ava and traditional Win keyboards (by Myanmar Linux Users Group (LUG) [15]) are available for Linux but are rarely used [5].

Collation

There are two main collation sequences used for Myanmar, Pali order used for older dictionaries, and Spelling Book order used in modern dictionaries [16]. Non-Unicode fonts allow variable sequence of keystrokes to generate the same surface string, making it difficult to develop sorting sequences. However, Unicode enables a unique input sequence, on which collation can be built. Details of how to develop a collation sequence based on modern lexicographic order are

available in [16]. A Myanmar collation sequence developed by Myanmar NLP has been standardized nationally but is not widely known and used yet [5].

Microsoft platform does not provide collation support for Burmese. Myanmar NLP Research Center has developed a Myanmar sorter, which can sort Myanmar text in Unicode. GeoComp has also developed a sorting engine based on GeoComp Myanmar font encoding [12].

Myanmar collation is defined in IBM ICU and Glibc for open source platforms [18].

Locale

Burmese locale language name is “my” and country abbreviation is “mm” (earlier “bu” in ISO 3166). Myanmar locale data is not defined in the latest version of CLDR or IBM ICU. Microsoft does not provide support for Myanmar locale. Locale is being defined on Linux platform by Myanmar LUG and Myanmar NLP Research Center.

Interface Terminology Translation

Through the Myanmar Enabling Kit project [13], Myanmar LUG is developing interface terminology translations for the Linux based applications. Open Office 1.1.1 has already been released (with few reported errors) and work is in progress to translate applications on GNOME [17]. LIP for Myanmar on Microsoft platform has not been initiated but localized versions of Microsoft products are available from other vendors [18].

Status of Advanced Applications

Advanced local language applications for Myanmar language are being researched and developed by local vendors. Work has been done on defining line breaking [19] and developing utilities for its implementation (in Java [20]). Myanmar Spell Checker is also under development. Myanmar Unicode and NLP Research Center has released an initial version of Myanmar spell checker [5, 12]. Figure 4 shows the screen of this spell checker.

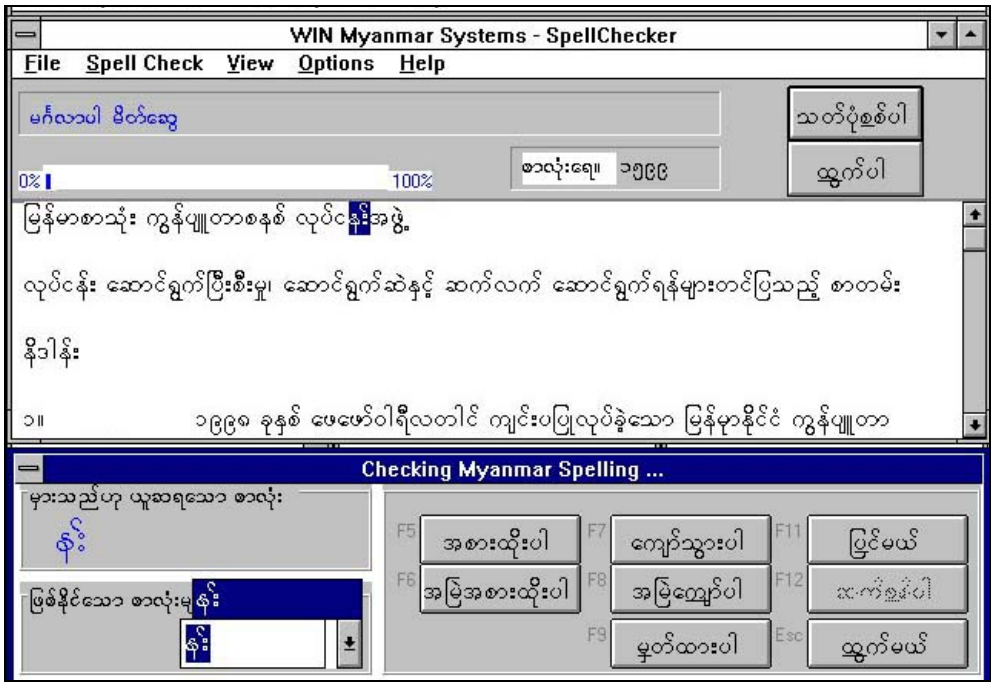

Figure 4: Myanmar Spell Checker [12]

Research on Myanmar OCR application has also been initiated. The current version of OCR, developed by Myanmar NLP Research Center, recognizes Myanmar digits. Work has also started on Myanmar speech recognition and text-to-speech systems [12].

References

[1] http://www.ethnologue.com/show_language.asp?code=mya
[2] http://en.wikipedia.org/wiki/Burmese_language
[3] http://www.omniglot.com/writing/burmese.htm
[4] http://www.unicode.org/charts/PDF/U1000.pdf
[5] Reported by PAN Localization project survey filled by Myanmar NLP Research Center.
[6] http://www.myazedi.com
[7] http://scripts.sil.org
[8] http://www.myanmars.net/unicode
[9] http://www.myanmars.net/unicode/keyboards/index.htm
[10] http://www.myanmars.net/unicode/doc/
[11] http://mercury.soas.ac.uk/wadict/burmese/SOASMyanmar_keyboard_and_font_user_manual.pdf
[12] http://www.myanmars.net/unicode/projects.htm
[13] http://www.iosn.net/country/myanmar/news/NLPTEAM
[14] Hosken, M and Tuntunlwin, M. “Representing Myanmar in Unicode: Details and Examples.” http://www.unicode.org/notes/tn11/myanmar_uni.pdf, 2004.
[15] http://www.thanlwinsoft.org
[16] Stribley, K. “Collation of Myanmar (Burmese) in Unicode: Sorting Myanmar in Unicode According to “Spelling Book Order” .” http://www.thanlwinsoft.org/ThanLWinSoft/MyanmarUnicode/Sorting/MyanmarCollation.pdf, 2005
[17] http://www.thanlwinsoft.org/ThanLWinSoft/MyanmarUnicode/Applications
[18] http://www.myanmars.net/winmyanmar/
[19] Stribley, K. “Syllable Based Dual Weight Algorithm for Line Breaking in Myanmar Unicode.” http://www.thanlwinsoft.org/ThanLWinSoft/MyanmarUnicode/Applications, 2005.
[20] http://www.thanlwinsoft.org/ThanLWinSoft/MyanmarUnicode/Parsing/MyanmarParser.java