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.

Vietnamese

Vietnamese is an Austro-Asiatic language spoken by about 68 million people of Vietnam. Vietnamese is the official language in Vietnam and is also spoken in some parts of Australia, Cambodia, Canada, China and Thailand [1, 2].

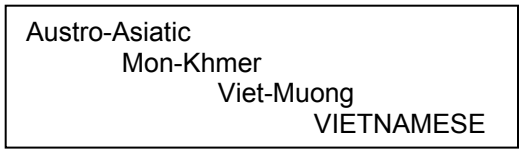


Figure 1: Language Family Tree of Vietnamese [1]

Vietnamese was initially written in classical Chinese writing Chữ-nho. This was later adapted in 10th century AD and called Chữ-nôm. In 17th centuary, western missionaries started developing a Latin based script, called Quốc Ngữ, which is now in widespread use [3]. Quốc Ngữ is the nationally adopted standard since 1945 [4]. Use of Chữ-nho and Chữ-nôm continued until about 1918. However, these are no longer in use (see [3] for more details).

Character Set and Encoding

Initially, due to absence of any standards, more than 30 different encodings were developed by different national and international vendors. These included single byte code table, e.g., 3C30 (3C Corporation), Cyrillix, VW1, VSCII, Daisy, 2font, Vietkey, ACAD, ABC, etc., and two byte code tables, e.g., 3C25, VW2, ATM2, VNij, VNI, etc. [4].

To resolve the situation, a task force was set up in 1991 by the Ministry of Science and Technology, which produced a draft for Vietnamese Standard Code Set for Information Interchange (VSCII) or TCVN 5712:1993, which was approved in 1993 as the first national standard in IT [4, 5]. This standard has been revised to TCVN 5712:1999 in 1999 [9]. Other encodings include Microsoft's cp 1258 (developed in 1996) and IBM's cp 1129 (developed in 1997), which are slightly different from each other [4].

Now Unicode is increasingly used. As Latin script is used, Vietnamese does not have a separate code table within Unicode, but characters within the Latin tables are used. These characters are spread across multiple tables within Unicode. For one byte, VSCII provides an alternative. However, there is a national standard for two-byte standard as well, TVCN 6909:2001.

There has also been work to revive the earlier writing systems. Relevant national standards include Chu Nom 16-bit character encoding standard TCVN 5773:1993 and Chu Nom Han 16-bit character encoding standard TCVN 6056:1995 [9].

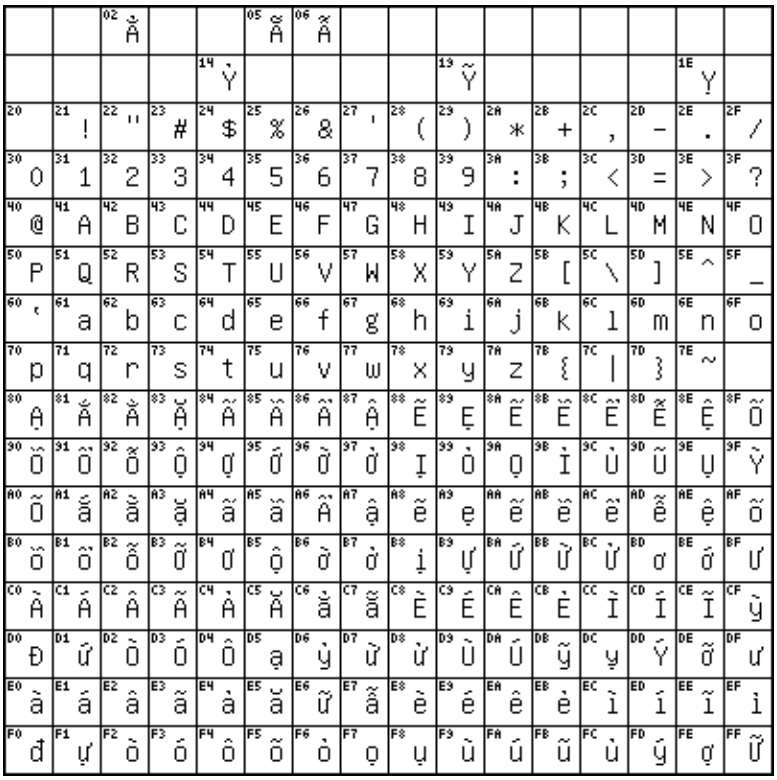


Figure 2: VSCII 1993 Character Set Encoding [5]

Text conversion utilities among various encodings are also available [10, 11]. Figure 3 shows one such utility.

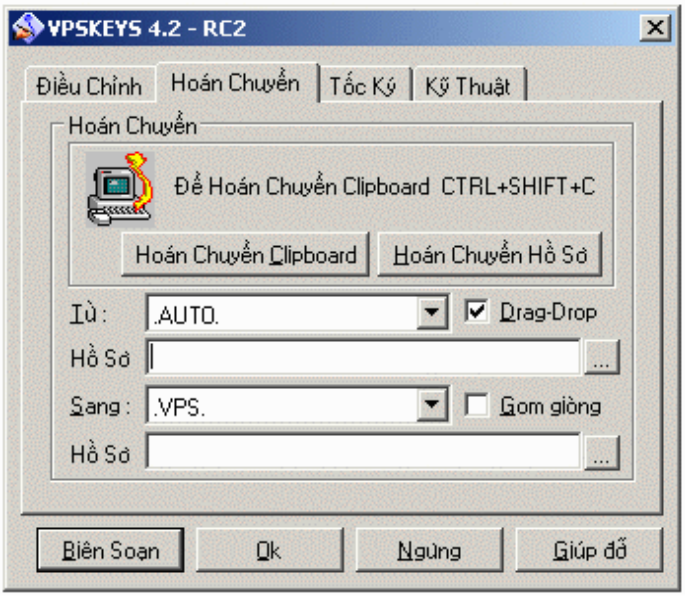


Figure 3: Encoding Conversion Utility for Vietnamese [10]

Fonts and Rendering

Latin script is supported by most platforms and many Vietnamese fonts are available for use. These fonts are based on a variety of encodings discussed above (e.g. see [6]).

Windows Platform

Microsoft XP does not include specific fonts for Vietnamese language but fonts for Unicode (e.g. Arial Unicode MS) cover the characters in Vietnamese as well. However many Vietnamese Unicode Open Type and True Type fonts have been developed by different research groups [4].

Linux Platform

On the Linux platform, fonts based on Unicode [6], TCVN, VNI and VPS [7] encodings can be adequately used to input Vietnamese text.

Keyboard

A national standard for Vietnamese keyboard TCVN 6064:1995 has been developed. It is based on international keyboard layout ISO 9995. TELEX and VNI are two other popularly used standards [4]. These keyboards are available, e.g. Unikey keyboard which allows user to configure any of these layouts on the keyboard for any of the encodings, as shown in Figure 4.

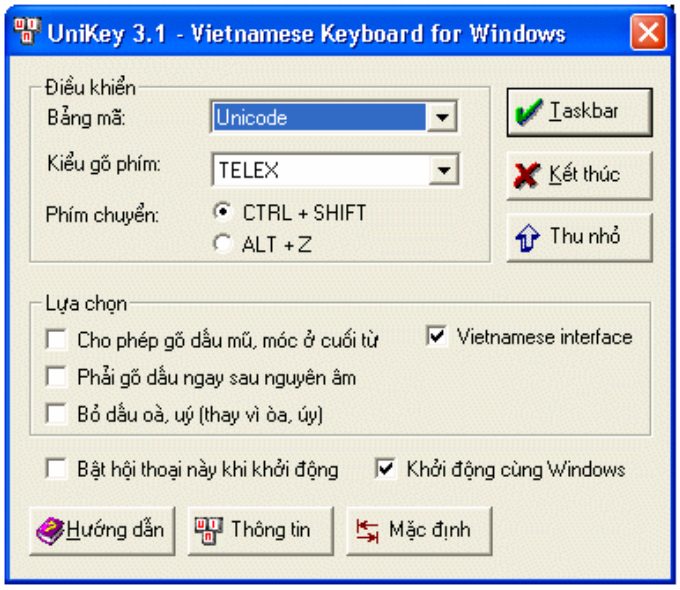


Figure 4: Unikey, Vietnamese Keyboard Layout Utility [11]

Keyboards use different schemes to generate the characters. Some generate pre-composed characters while others generate base character and diacritics separately. The variation in output has to be normalized before further processing. Normalization utilities are also available.

Microsoft Platform

Microsoft provides built-in Vietnamese keyboard. The basic Vietnamese characters are in similar position as in default QWERTY keyboard layout. Figure 5 below shows the normal state Vietnamese keyboard layout on MS platform.

Figure 5: Keyboard Layout for Vietnamese [8]

Linux Platform

XVNKB is a Vietnamese keyboard input for X-Window. It provides a useful way for editing Vietnamese on X-Windows environment with popular input methods and character sets. This software has been released under GPL license [12].

Figure 6: XVNKB Menus [12]

Collation

Collation order for Vietnamese written in Quoc Ngu has been developed as part of TCVN 5712:1993 standard [4] (also see [13, 14] for details of collation sequence). Sorting for Vietnamese has been enabled on Microsoft and Linux platforms, with minor problems still persisting [15].

Locale

Locale for Vietnamese (vi_VN) has been developed. TCVN/JTC1 and ITSC developed basic locale definition such as date, time, length, volume and weight measurements, and a conversion between the Gregorian (Western) and Lunar (Vietnamese) calendars. Standard locale for Vietnamese has also been included in IBM ICU and CLDR 1.3.

Microsoft Platform

Microsoft Windows XP provides support for Vietnamese locale. If system locale is switched to Vietnamese, changes may be monitored in date, time, and currency symbol within all application of Microsoft. The following figure shows Vietnamese locale on MS platform.

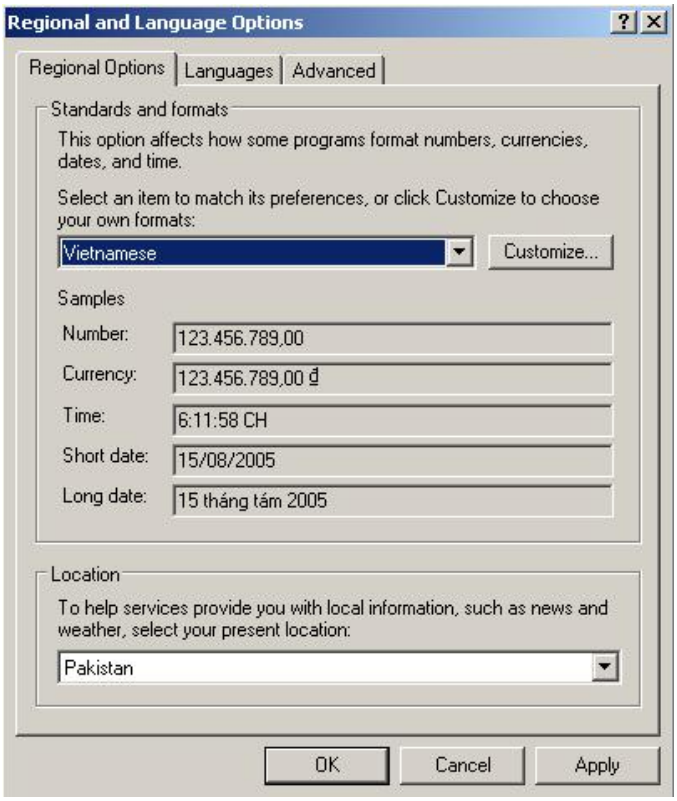


Figure 7: Vietnamese Locale on MS Platform

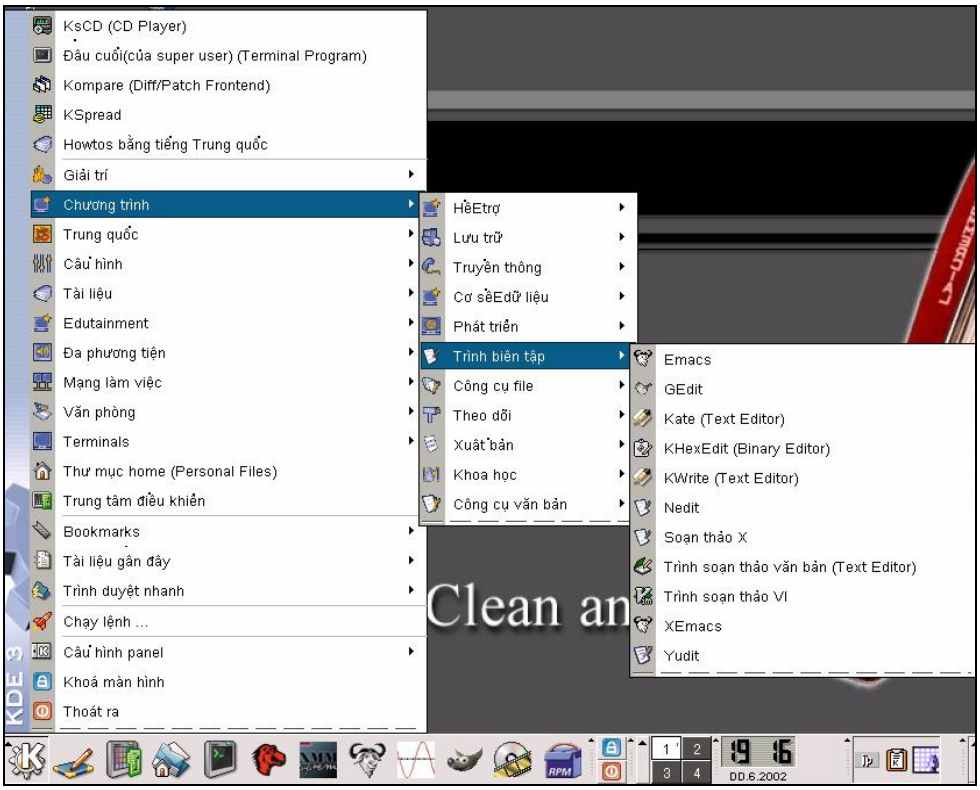
Interface Terminology Translation

TCVN/JTC1 also established glossary of Vietnamese and English terms for all the graphical user interface (GUI), icon names, dialogue boxes in 2000 through the standard TCVN 6695-1:2000.

Localized interface in Vietnamese is available on both Microsoft and Linux platforms. Microsoft support has been available for a long time, since Windows 95. Microsoft Language Interface Pack for Vietnamese is available [17].

Linux support has been developed later. The project on the development of VNLinux consists of two sub-modules: Development of a VNLinuxCD based on Mandriva Linux, designed for desktop use, and a VNLS server oriented distribution based on EnGarde Secure Linux [16]. A Vietnamese localization team has registered for GNOME glossary translation. As reported on the website about 96.41% of the GNOME glossary translation has been completed [18] while glossary translation of KDE has also been completed [19]. Figure 8 given below shows localized GNOME and KDE based Linux in Vietnamese.

(a)



(b)

Figure 8 (a) Konqueror KDE Browser, (b) Localized KDE Start-Up Menus

Status of Advanced Applications

Much work has been done on Vietnamese language processing. Vietnamese spell checkers are available for Microsoft [21] and Linux platforms, and through other vendors (e.g. [20]). There has also been considerable work on Vietnamese and Vietnamese-English lexicons. English-Vietnamese machine translation systems were available as early as early 1990's. Much more work has been done on it and on Vietnamese-French machine translation systems [22]. Vietnamese text to speech systems and speech recognition systems are also available through private vendors. Work has also started on Vietnamese hand writing recognition systems.

References

[1] http://www.ethnologue.com/show_language.asp?code=vie
[2] http://www.omniglot.com/writing/vietnamese.htm
[3] http://www.omniglot.com/writing/chunom.htm
[4] Chuong, T., Hoang, N., Nhan, N., Phuoc, D., Viet, D. "Current Status of Vietnamese Language Processing Multilingual Processing," in Proceeding of MLIT '97, Tokyo, Japan, 1997. Also available at http://www.informatik.uni-leipzig.de/~duc/software/misc/viet.html
[5] http://czyborra.com/charsets/vietnamese.html
[6] http://www.vietgate.net/fonts/
[7] http://www.fedu.uec.ac.jp/~vuhung/linux/font-rpms/
[8] http://www.microsoft.com/globaldev/keyboards/kbdvntc.htm
[9] "Human Resource Development Policies of Information Technology in Vietnam." http://www.cicc.or.jp/japanese/kouryu/pdf_ppt/vietnam.pdf
[10] http://www.vps.org/article.php3?id_article=274
[11] http://unikey.sourceforge.net/
[12] http://xvnkb.sourceforge.net/
[13] Lương, V. "Vietnamese Sorting Rules for Dictionary Entries." Vietnam Lexicography Center. http://vietunicode.sourceforge.net/charset/quytacABC_en.html
[14] http://vietunicode.sourceforge.net/charset/vietalphabet.html
[15] http://blogs.msdn.com/michkap/archive/2005/08/27/457224.aspx
[16] http://distrowatch.com/table.php?distribution=vnlinux
[17] http://www.microsoft.com/downloads/details.aspx?FamilyID=0db2e8f9-79c4-4625-a07a-0cc1b341be7c&DisplayLang=vi
[18] http://l10n-status.GNOME.org/GNOME-2.8/vi/index.html
[19] http://i18n.KDE.org/teams/index.php?a=i&t=vi
[20] http://www.vnisoft.com/?http://www.vnisoft.com/xpfeatures.html
[21] http://www.worldlanguage.com/ProductScreenShots/104793.htm
[22] Dien, D. and Kiem, H. "State of Art of Machine Translation in Vietnam," in AAMT Journal, Special Issue, September 2005, Thailand.