Network Working Group Vietnamese Standardization Working Group
Request for Comments: 1456 May 1993

Conventions for Encoding the Vietnamese Language
VISCII: VIetnamese Standard Code for Information Interchange
VIQR: VIetnamese Quoted-Readable Specification
Revision 1.1

Status of this Memo

This memo provides information for the Internet community. It does not specify an Internet standard. Distribution of this memo is unlimited.

Abstract

This document provides information to the Internet community on the currently used conventions for encoding Vietnamese characters into 7-bit US ASCII and in an 8-bit form. These conventions are widely used by the overseas Vietnamese who are on the Internet and are active in USENET. This document only provides information and specifies no level of standard.

1. Introduction

In this paper we describe two conventions for representing Vietnamese characters. VISCII (pronounced "visky") is an 8-bit character encoding that is similar to that used with ISO-8859. VIQR (pronounced "vicker") is a mnemonic encoding of Vietnamese characters into US ASCII for use on 7-bit systems. There is substantial existing online freely distributable software that implements these conventions for UNIX and personal computers. These encodings enable Vietnamese-language users to take full advantage of powerful tools already developed for the English-speaking world, eliminating unnecessary reinvention. This paper describes these conventions in part so that MIME-compliant software might also support the Vietnamese language.

NOTE: The accented Vietnamese letters are herein represented by their VIQR equivalents, offset by enclosing angle brackets. For example, the single letter "a acute" is written as <a’>, where the apostrophe is the mnemonic symbol for the acute.

2. LINGUISTIC OVERVIEW

As a romanized language, Vietnamese appears to lend itself readily to integration into existing English-based systems. To cite a simple

example, consider implementing support for French in such systems. One can allocate code positions in the 8-bit space necessary for accented letters such as <e^> or <e'>, then provide a means for users to access these codes through the keyboard. The required number of "extra" code positions is small (see, e.g., ISO-8859/Latin-1 [1]), and the relatively low frequency of occurrence of accented letters does not place heavy demand on efficient keyboard input schemes. The same things cannot be said for Vietnamese, where both the number and occurrence frequency of accented letters are large. Apart from the alphabetics already available in ASCII, Vietnamese requires an additional 134 combinations of a letter and diacritical symbols.

Note that one can resort to a composite encoding scheme to reduce this requirement, but that would mean giving up on integration into today's computing platforms which for the most part do not support such schemes. In addition, the heavy use of diacritical marks in Vietnamese text calls for a keyboard input scheme that does not require extra keystrokes such as a special "compose" key to generate accented letters. Because of the large number of possible combinations, the scheme should also be easily learned and memorized.

Finally, to integrate Vietnamese into current electronic mail systems which are still limited to 7 bits, there should be a representation for Vietnamese text that is readily readable in its 7-bit form.

The Viet-Std group, an electronic standardization roundtable, has worked over the past few years to draft proposals addressing these issues. This has culminated in the conventions to be described briefly in the next two sections. The detailed technical considerations have been reported elsewhere [2]. In this memo we give a brief outline of the working standards and describe supporting software availability.

3. SPECIFICATION OF VISCII

VISCII stands for VIetnamese Standard Code for Information Interchange, an 8-bit encoding specification. Its salient features are:

1. Encoding of all Vietnamese letters as single units rather than separating base vowels and diacritical marks.

2. Retention of the complete ASCII graphics repertoire in order to facilitate integration.

3. Encoding the 6 least-often-used upper-case letters into 6 least problematic C0 (control) characters.

4. Character placement have been designed with consideration for Unix/X integration, ISO-8859/Latin-1 compatibility, coexistence with a wide array of existing software, including provisions for single- and double-line drawing characters in the IBM graphic character set.

The 8-bit VISCII encoding is shown below. Because of the limitations of the 7-bit US ASCII character set, here we use the mnemonic form to represent Vietnamese glyphs. See the VIQR specification below for clarification of how diacritical marks are applied. The online PostScript version of reference [2] may also be useful as it does display each character correctly.

Table 1. VISCII 8-bit Encoding Table (v1.1)

	0x	1x	2x	3x	4x	5x	6x	7x	8x	9x	Ax	Bx	Cx	Dx	Ex	Fx
x0	nul	dle	sp	0	@	P	‘	p	A.	O^‘	O˜	o^‘	A‘	DD	a‘	dd
x1	soh	dc1	!	1	A	Q	a	q	A(’	O^?	a(’	o^?	A’	u+’	a’	u+.
x2	A(?	dc2	"	2	B	R	b	r	A(‘	O^˜	a(‘	o^˜	A^	O‘	a^	o‘
x3	etx	dc3	#	3	C	S	c	s	A(.	O^.	a(.	O+˜	A˜	O’	a˜	o’
x4	eot	Y?	$	4	D	T	d	t	A^’	O+.	a^’	O+	A?	O^	a?	o^
x5	A(˜	nak	%	5	E	U	e	u	A^‘	O+’	a^‘	o^.	A(	a.	a(	o˜
x6	A^˜	syn	&	6	F	V	f	v	A^?	O+‘	a^?	o+‘	a(?	y?	u+˜	o?
x7	bel	etb	’	7	G	W	g	w	A^.	O+?	a^.	o+?	a(˜	u+‘	a^˜	o.
x8	bs	can	(	8	H	X	h	x	E˜	I.	e˜	i.	E‘	u+?	e‘	u.
x9	ht	Y˜	)	9	I	Y	i	y	E.	O?	e.	U+.	E’	U‘	e’	u‘
xA	lf	sub	*	:	J	Z	j	z	E^’	O.	e^’	U+’	E^	U’	e^	u’
xB	vt	esc	+	;	K	[	k	{	E^‘	I?	e^‘	U+‘	E?	y˜	e?	u˜
xC	ff	fs	,	<	L	\	l	|	E^?	U?	e^?	U+?	I‘	y.	i‘	u?
xD	cr	gs	-	=	M	]	m	}	E^˜	U˜	e^˜	o+	I’	Y’	i’	y’
xE	so	Y.	.	>	N	^	n	˜	E^.	U.	e^.	o+’	I˜	o+˜	i˜	o+.
xF	si	us	/	?	O	_	o	DEL	O^’	Y‘	o^’	U+	y‘	u+	i?	U+˜

4. SPECIFICATION OF VIQR MNEMONICS

VIQR, VIetnamese Quoted-Readable specification, is not an encoding convention but is rather a convention for typing, reading, and transferring Vietnamese data using only the 7-bit ASCII character set. With VIQR, accented Vietnamese letters are represented by the vowel followed by ASCII characters whose appearances resemble those of the corresponding Vietnamese diacritical marks. For example, the phrase "N<u+><o+’>c Vi<e^.>t Nam" is represented in 7-bits by "Nu+o+’c Vie^.t Nam". The complete list of diacritical mark equivalents is given in Table 2. There is also provision in the VIQR specification to prevent undesirable composition, for example, to

avoid getting "How are you?" composed into "How are yo<u?>". For details, please see [2]. VIQR therefore serves the following purposes:

1. It provides for a mnemonic, readable representation of Vietnamese in 7-bit form, which makes it easy to transfer Vietnamese electronic mail without special conversion. The originator and recipient can communicate in Vietnamese without the need for an 8-bit environment at any point in the data chain.

2. It provides a bridge for translation between 7- and 8-bit environments. In this context, typing in both 7-bit and 8-bit systems requires exactly the same keystrokes, the only difference is that the 8-bit user gets to see actual Vietnamese on-screen, whereas the 7-bit user sees a mnemonic representation thereof. The same options are available for the 7-bit and 8-bit recipients of Vietnamese text.

Because of its mnemonic nature, the VIQR typing method is easy to learn and remember. In pure 8-bit environments, special-purpose software developers may wish to devise more efficient input schemes, but the intent is for all Vietnamese keyboard software to support the basic VIQR method to minimize learning time for Vietnamese who will already be familiar with the mnemonic method described here.

Table 2. VIQR Mnemonics for Vietnamese Diacritics

Diacritic	Char	ASCII Code	D<a^'>u
breve	(	0x28, left paren	tr<a(>ng
circumflex	^	0x5E, caret	m<u~>
horn	+	0x2B, plus sign	m<o'>c
acute	'	0x27, apostrophe	s<a('>c
grave	`	0x60, backquote	huy<e^`>n
hook above	?	0x3F, question	h<o?>i
tilde	~	0x7E, tilde	ng<a~>
dot below	.	0x2E, period	n<a(.>ng
d bar	dd	(repeated d)	<dd>
D bar	DD	(repeated D)	<DD>

5. SUPPORTING SOFTWARE

VISCII & VIQR have been successfully implemented on various platforms. The work has been carried out primarily by the TriChlor software group, a non-profit spin-off from Viet-Std. Software by other individuals and groups have also been developed. In addition, commercial software entities have indicated that they would support the standards in the form of VISCII-compliant keyboards and fonts.

The current software selection from the TriChlor group enables users to use Vietnamese on existing Unix, MS-DOS, and Windows systems, including such operations as Vietnamese file naming, Vietnamese keyboarding within any application, electronic mail and news filters for Unix, printing to various printer languages, incorporating Vietnamese in such document preparation systems as TeX, Word for Windows, WordPerfect, using Vietnamese in databases (e.g., Paradox) and spreadsheets (e.g., SC on Unix or Excel in Windows). Vietnamese-specific applications are also available and include a large song lyric database, several poetry collections in hypertext format, a Windows-based fortune teller, a text-based multiple-choice test program in Vietnamese, etc. In short, software exists that supports thorough integration of Vietnamese into existing platforms, allowing Vietnamese users to take advantage of all the powerful tools already available in English-only environments.

Translation between 8-bit VISCII 1.1 and other character sets, particularly ISO-10646/Unicode 1.1, has been included in the Plan 9 operating systems' tcs utility that has been made available by Andrew Hume of AT&T Bell Laboratories.

6. MIME CONSIDERATIONS

For use with MIME-compliant software, the value "VISCII" has been registered as a charset with the Internet Assigned Numbers Authority for the VISCII encoding convention described above, and the value "VIQR" has been registered with the Internet Assigned Numbers Authority as a charset for the VIQR mnemonic encoding convention described above. Implementation of support for these two MIME character set types is not mandatory to comply with RFC-1341. If the encoding conventions described above are used in MIME email or news, the appropriate MIME character set type value should be used to label the body-part containing such text.

7. SECURITY CONSIDERATIONS

Security issues are not discussed in this memo.

REFERENCES

[1] International Organization for Standardization. ISO 8859/x: 8-bit International Code Sets. ISO, 1977.

[2] Viet-Std, "A Unified Framework for Vietnamese Information Processing-v1.1," published on the Internet, available for FTP from Sonygate.Sony.COM:tin/viet-std, September 1992.

AUTHORS' ADDRESSES

Cuong T. Nguyen
Center for Integrated Systems
CIS 062--MC 4070
Stanford, CA 94305-4070

Phone: (415) 725-3721
Email: cuong@haydn.Stanford.EDU

Hoc D. Ngo
Vista Research, Inc.
100 View St, Suite 200
P.O. Box 998
Mountain View, CA 94042

Phone: (415) 966-1171
Email: uunet!vri280!hoc

Cuong M. Bui
National Semiconductor Corp.
3388 Burgundy Dr.
San Jose, CA 95132

Phone: (408) 721-6873
Email: bui@berlioz.nsc.com

Thanh van Nguyen
Roche Image Analysis Systems
95 First Str Suite 110
Los Altos, CA 94022

Phone: 415-917-2022
Fax: 415-917-2025
Email: thanh@rias.com

For more information, please contact the authors at:
viet-std@haydn.stanford.edu