Language Specific Peculiarities Document for Vietnamese as Spoken in Vietnam

1. Special handling of dialects

Vietnamese (tiếng Việt, or less commonly Việt ngữ) is the national and official language of Vietnam. As stated in Wikipedia, it is the mother tongue of 86% of Vietnam's 85 million inhabitants, and of about three million overseas Vietnamese. It is also spoken as a second language by many ethnic minorities in Vietnam. The dialects or regional variants of Vietnamese can be categorized according to their geographical location within Vietnam. Four dialect regions have been identified by Wikipedia, based on differences in their sound systems and vocabulary:

Region	Main cities in each dialectal region
Northern	Ha Noi (the standard variety)
North-central	Vinh (Nghệ An Province)
Central	Hue (Thừa Thiên – Huế Province)
Southern	Ho Chi Minh City (Saigon)

The Northern dialect is the standard variety. The Northern-central dialect is the most conservative phonologically, preserving features such as consonant clusters and non-diphthongized vowels. Regional variations in pronunciation are listed in the table below, with the IPA in square brackets and the SAMPA following:

Syllable position	Orthography	Northern	North-central	Central	Southern
syllable-initial	x	[s] s	[s] s	[s] s	[s] s
	s		[ʂ] s`	[ʂ] s`	[ʂ] `
	ch	[tɕ] ts\	[tɕ] ts\	[tɕ] ts\	[tɕ] ts\
	tr		[tʂ] ts`	[tʂ] ts`	[tʂ] ts`
	r	[z] z	[ɻ] r\	[ɻ] r\	[ɻ] r\
	d		[ʝ] J\	[j] j	[j] j
	gi		[z] z		
	v	[v] v	[v] v		
syllable-final	c	[k] k	[k] k	[k] k	[k] k
	t	[t] t	[t] t		
	t (e_)			[k, t] k,t	
	t (ê_)			[t] t	[k, t] k,t
	t (i_)				[t] t
	ch	[c] c	[c] c		
	ng	[ŋ] N	[ŋ] N	[ŋ] N	[ŋ] N
	n	[n] n	[n] n		
	n (i_, ê_)			[n] n	[n] n
	nh	[ɲ] J	[ɲ] J		

While “l” and “n” are distinct graphemes in Vietnamese orthography, their pronunciation is merged in some rural dialects. In these dialects [l] becomes [n][1].Section 7 contains the complete table and discussion of Vietnamese orthography, pronunciation, and the interpretation of diacritics as used in this document.

2. Deviation from native-speaker principle

No special deviation – only native speakers of Vietnamese, born in Vietnam will be collected in this project.

3. Special handling of spelling

There will be no particular special handling of spelling for native words in this collection. The standard Vietnamese writing system in use today represents all native phonemes, with additional diacritics for tones. Business names and loan words from other languages will be spelled in the adapted Roman script with necessary variation and standardization wherever possible. Foreign names and loan words can be spelled according to either their nativized pronunciation or the source writing system (especially for Roman-alphabet languages). Native speakers will be asked to follow the orthographic conventions found in official publications (e.g. dictionaries) and social media (newspapers, well-established websites). Orthographic variants of the same words created by different transcribers will be standardized during data post-processing.

For example, a foreign company name like “Hyundai” will be spelled using its source writing system as “Hyundai” because this is how this name is presented in the media. However, a loanword like “acid” will be spelled as “a-xít” because this format can be found in textbooks and published dictionaries. Variants of “acid” such as “a-xít” and “axít” will be standardized during ABH’s data post-processing to produce a single consistent version.

The Từ điển Tiếng Việt Phổ Thông (Dictionary of the Vietnamese Language), by the Vietnamese Institute of Linguistics, 2008, will be used as a reference.

4. Description of character set used for orthographic transcription

The Vietnamese script will be used for the orthographic transcription of Vietnamese. There are two methods for entering vowels plus their tonal diacritics for Vietnamese. We will use the method of entering the vowel and the combining diacritic as a single character, rather than having them as individual characters. The complete unicode character set for Vietnamese is given here.

[1] Because the table reflects only regional variation, orthographic “l” does not appear.

Vowels

Tone 1 no diacritic		Tone 2 acute		Tone 3 grave		Tone 4 hook above		Tone 5 tilde		Tone 6 dot below	
	Unicode		Unicode		Unicode		Unicode		Unicode		Unicode
a	0061	á	00e1	à	00e0	ả	1ea3	ã	00e3	ạ	1ea1
ă	0103	ắ	1eaf	ằ	1eb1	ẳ	1eb3	ẵ	1eb5	ặ	1eb7
â	00e2	ấ	1ea5	ầ	1ea7	ẩ	1ea9	ẫ	1eab	ậ	1ead
e	0065	é	00e9	è	00e8	ẻ	1ebb	ẽ	1ebd	ẹ	1eb9
ê	00ea	ế	1ebf	ề	1ec1	ể	1ec3	ễ	1ec5	ệ	1ec7
i	0069	í	00ed	ì	00ec	ỉ	1ec9	ĩ	0129	ị	1ecb
o	006f	ó	00f3	ò	00f2	ỏ	1ecf	õ	00f5	ọ	1ecd
ơ	01a1	ớ	1edb	ờ	1edd	ở	1edf	ỡ	1ee1	ợ	1ee3
ô	00f4	ố	1ed1	ồ	1ed3	ổ	1ed5	ỗ	1ed7	ộ	1ed9
u	0075	ú	00fa	ù	00f9	ủ	1ee7	ũ	0169	ụ	1ee5
ư	01b0	ứ	1ee9	ừ	1eeb	ử	1eed	ữ	1eef	ự	1ef1
y	0079	ý	00fd	ỳ	1ef3	ỷ	1ef7	ỹ	1ef9	ỵ	1ef5
A	0041	Á	00c1	À	00c0	Ả	1ea2	Ã	00c3	Ạ	1ea0
Ă	0102	Ắ	1eae	Ằ	1eb0	Ẳ	1eb2	Ẵ	1eb4	Ặ	1eb6
Â	00c2	Ấ	1ea4	Ầ	1ea6	Ẩ	1ea8	Ẫ	1eea	Ậ	1eac
E	0045	É	00c9	È	00c8	Ẻ	1eba	Ẽ	1ebc	Ẹ	1eb8
Ê	00ca	Ế	1ebe	Ề	1ec0	Ể	1ec2	Ễ	1ec4	Ệ	1ec6
I	0049	Í	00cd	Ì	00cc	Ỉ	1ec8	Ĩ	0128	Ị	1eca
O	004f	Ó	00d3	Ò	00d2	Ỏ	1ece	Õ	00d5	Ọ	1ecc
Ơ	01a0	Ớ	1eda	Ờ	1edc	Ở	1ede	Ỡ	1ee0	Ợ	1ee2
Ô	00d4	Ố	1ed0	Ồ	1ed2	Ổ	1ed4	Ỗ	1ed6	Ộ	1ed8
U	0055	Ú	00da	Ù	00d9	Ủ	1ee6	Ũ	0168	Ụ	1ee4
Ư	01af	Ứ	1ee8	Ừ	1eea	Ử	1eec	Ữ	1eee	Ự	1ef0
Y	0059	Ý	00dd	Ỳ	1ef2	Ỷ	1ef6	Ỹ	1ef8	Ỵ	1ef4

Consonants

	Unicode		Unicode
b	0062	B	0042
c	0063	C	0043
d	0064	D	0044
đ	0111	Đ	0110
f	0066	F	0046
g	0067	G	0047
h	0068	H	0048
j	006A	J	004A
k	006b	K	004b
l	006c	L	004c
m	006d	M	004d
n	006e	N	004e
p	0070	P	0050
q	0071	Q	0051
r	0072	R	0052
s	0073	S	0053
t	0074	T	0054
v	0076	V	0056
w	0077	W	0057
x	0078	X	0058
y	0079	Y	0059

5. Description of Romanization Scheme

None provided.

6. Description of method for word boundary detection

While word boundaries in transcription are typically represented by a white space, for Vietnamese we tokenize on syllables/morphemes. This approach is justified, for several reasons. Firstly, this approach uses standard and intuitive spelling conventions, reducing the risk of introducing inconsistencies and errors at the transcription phase. Secondly, by not grouping syllables/morphemes together into words, we will improve the lexicon quality without loss of information, as there are no phonological processes that cross morpheme/syllable boundaries, such as consonantal assimilation, tone sandhi or word-level stress. This second point in particular means that tokenizing on white space for Vietnamese will create a more efficient lexicon than would be achieved by tokenizing on lexemes.

7. All phonemes in the stipulated notation

The phonemic transcription of the words in this database uses X-SAMPA symbols, which can be found at http://www.phon.ucl.ac.uk/home/sampa/x-sampa.htm. The total number of phonemes is 54. There are 25 consonants and 45 vowels (12 monophthongs, 25 diphthongs and 8 triphthongs). The chart below includes the geographical variations as listed in section 1. Our preliminary research has shown that there is sometimes more than one orthographic representation for a particular phoneme in a particular variety, since the orthography was designed to accommodate the main regional variants

found in the language. This is reflected in the table below. The letter 'y' can be either a vowel or a consonant. A tilde ~ is used to indicate variant pronunciations. It is both used to mark variances across dialects such as in orthography "ch" and to present variances of the phoneme between the beginning and ending positions. A comma separates variant spellings.

Monophthongs			Diphthongs			Consonants		
Orth	IPA	SAMPA	Orth	IPA	SAMPA	Orth	IPA	SAMPA
a	aː	a:	ai	aːɪ̯	a:I	p	p	p
ă	a	a	ao	aːʊ̯	a:U	b	ɓ	b_<
â	ə	@	au	aʊ̯	aU	ph	f	f
e	ɛ	E	âu	əʊ̯	@U	v	v~j	v~j
ê	e	e	ay	aɪ̯	aI	m	m	m
i	i	i	ây	əɪ̯	@I	u, o, qu, w	w	w
o	ɔ	O	eo	ɛʊ̯	EU	t	t~k	t~k
ơ	əː	@:	êu	eʊ̯	eU	th	tʰ	t_h
ô	o	o	ia, iê, yê	iə̯	i@	đ	ɗ	d_<
u	u	u	iu	iʊ̯	iU	x	s	s
ư	ɨ	1	oa	ɔaː	Oa:	gi	z~j	z~j
y	i	i	oă	ɔa	Oa	n	n~ŋ	n~N
Triphthongs			oe	ɔɛ	OE	l	l	l
Orth	IPA	SAMPA	oi	ɔɪ̯	OI	tr	tɕ~tʂ	ts\~ts`
iêu	iə̯ʊ̯	i@U	ôi	oɪ̯	oI	s	s~ʂ	s~s`
oai	oaɪ̯	oaI	ơi	əːɪ̯	@:I	r	z~ɹ	z~r\
oay	oaɪ̯ː	oaI:	ua, uô	uə̯	u@	ch	c~t	c~t
uôi	uə̯ɪ̯	u@I	uâ	uə̯	u@	d	z~ɟ~j	z~J\~j
uyê	uɪə̯	uI@	ưa, ươ	ɨə̯	1@	nh	ɲ~n	J~n
ươi	ɨə̯ɪ̯	1@I	uê	u̯e	ue	y	j~z	j~z
ươu	ɨə̯ʊ̯	1@U	ui	uɪ̯	uI	c, k, q	k	k
yêu	iə̯ʊ̯	i@U	ưi	ɨɪ̯	1I	kh	x	x
			uơ	uə̯ː	u@:	g, gh	ɣ	G
			ưu	ɨʊ̯	1U	ng, ngh	ŋ	N
			uy	uɪ̯ː	ui:	h	h	h

In addition, a glottal stop is normally added before a vowel-initial segment.

The 6 tones in Vietnamese are represented by numbers from 1 to 6 in SAMPA

Tone	Diacritic	SAMPA
Ngang	(no mark)	_1
Sắc	◌́ (acute accent)	_2
Huyền	◌̀ (grave accent)	_3
Hỏi	◌̉ (hook)	_4
Ngã	◌̃ (tilde)	_5
Nặng	◌̣ (dot below)	_6

Note that every full syllable containing a vowel must appear with a tone. Partial syllables, without a vowel, do not receive a tone.

8. Complete list of all rare phonemes

Rare phonemes not provided.

9. Other language specific items

9.1 Table of digits

0	không
1	một
2	hai
3	ba
4	bốn
5	năm
6	sáu
7	bảy
8	tám
9	chín

9.2 Other numbers

10	mười
100	một trăm
10,000	mười ngàn/nghìn, một vạn
100,000	một trăm ngàn/nghìn
10 million	mười triệu

10. References

Thompson, Laurence C. 1991. A Vietnamese reference grammar. Seattle: University of Washington Press. Honolulu: University of Hawaii Press. (Original work published 1965). (Online version: www.sealang.net/archives/mks/THOMPSONLaurenceC.htm.)

Vietnamese Institute of Linguistics. 2008. *Từ điển Tiếng Việt Phổ Thông* 'Dictionary of the Vietnamese Language'.

"Vietnamese language." Wikipedia. Wikimedia Foundation, Inc.. 10 Oct. 2011.