Language Specific Peculiarities Document for

Assamese as Spoken in Assam

1. Special handling of dialects

Dialects	Districts in Assam
Eastern Assamese	Sibsagar, Lakhimpur, Jorhat, Golaghat, Sonitpur, Karbi Anglong, Nagaoan, Dibrugarh, Dhemaji, Tinsukia
Central	Marigaon, (Eastern) Kamrup
Western	Goalpara, Dhubri, Bongaigaon, Kokrajhar, Barpeta, Nalbari, Darrang, (Western) Kamrup

Despite the language spoken in Sibsagar being considered the standard, and the literary dominance of Eastern Assamese since the 13th Century, the modern standard dialect has started assimilating to the dialect spoken in Guwahati (Eastern Kamrup in the table above), mirroring a recent shift of the cultural center from Sibsagar to Guwahati. Guwahati did not historically have a dialect of its own, as it was largely a place of military fortification.

2. Deviation from native-speaker principle

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

3. Special handling of spelling

There will be no particular special handling of spelling in this collection. English loan words will be spelled in the Assamese script rather than the Latin alphabet.

The *Hemkosh* (also known as the *Hema Kosha*) will be used as a reference for spelling.

4. Description of character set used for orthographic transcription

The Assamese script will be used for the orthographic transcription of Assamese.

The Unicode range for both Assamese and Bengali is U+0980-U+09FF. Presentation forms of these glyphs depend on the display font used, however, this does not affect the underlying Unicode. The Lohit Assamese font will be used. This Unicode-based font correctly renders all Assamese characters and can be downloaded from the following website: https://fedorahosted.org/lohit/.

Combined presentation glyphs are represented through the use of the *virama* (“*hashanta*” in Assamese) character (U+09CD), which suppresses the inherent vowel and dictates that the characters should render together.

In some cases, the characters should render separately, including the *virama* (*hashanta*) symbol, but this is generally a function of the font used to view the text. In some cases where separate rendering must be forced, such as for morphological boundaries or loan words, zero-width characters (U+200c and U+200d) may be used.

Certain words that would otherwise be homographs are distinguished by the use of the apostrophe (“*urdhokôma*” in Assamese; U+0027), as in ল'ৰা "boy" and লৰা "move”. This apostrophe is added to the inherent vowel and its initial variant “অ” to specify which of its pronunciations is being used.

5. Description of Romanization scheme

The following is Appen Butler Hill's Romanization scheme which is fully reversible. Appen Butler Hill’s Romanization schemes are being used for this project. These schemes are designed to be as similar in form as possible, but cannot be identical due to the different writing systems and spelling conventions in each language.

5.1. Assamese Romanization Scheme

Bengali script is used for Assamese.

UNICODE	ASSAMESE	ROMAN	DESCRIPTION
0x981	◌ঁ	M	BENGALI SIGN CANDRABINDU
0x982	◌ং	W	BENGALI SIGN ANUSVARA
0x983	◌ঃ	9	BENGALI SIGN VISARGA
0x985	অ	a	BENGALI LETTER A
0x986	আ	A	BENGALI LETTER AA
0x987	ই	I	BENGALI LETTER I
0x988	ঈ	i	BENGALI LETTER II
0x989	উ	U	BENGALI LETTER U
0x98a	ঊ	u	BENGALI LETTER UU
0x98b	ঋ	r[	BENGALI LETTER VOCALIC R.
0x98f	এ	e	BENGALI LETTER E

UNICODE	ASSAMESE	ROMAN	DESCRIPTION
0x990	ঐ	e3	BENGALI LETTER AI
0x993	ও	o	BENGALI LETTER O
0x994	ঔ	o3	BENGALI LETTER AU
0x995	ক	k	BENGALI LETTER KA
0x996	খ	K	BENGALI LETTER KHA
0x997	গ	g	BENGALI LETTER GA
0x998	ঘ	G	BENGALI LETTER GHA
0x999	ঙ	N	BENGALI LETTER NGA
0x99a	চ	c	BENGALI LETTER CA
0x99b	ছ	C	BENGALI LETTER CHA
0x99c	জ	j	BENGALI LETTER JA
0x99d	ঝ	Z	BENGALI LETTER JHA
0x99e	ঞ	J	BENGALI LETTER NYA
0x99f	ট	t`	BENGALI LETTER TTA
0x9a0	ঠ	T`	BENGALI LETTER TTHA
0x9a1	ড	d`	BENGALI LETTER DDA
0x9a2	ঢ	D`	BENGALI LETTER DDHA
0x9a3	ণ	n`	BENGALI LETTER NNA
0x9a4	ত	t	BENGALI LETTER TA
0x9a5	থ	T	BENGALI LETTER THA
0x9a6	দ	d	BENGALI LETTER DA

UNICODE	ASSAMESE	ROMAN	DESCRIPTION
0x9a7	ধ	D	BENGALI LETTER DHA
0x9a8	ন	n	BENGALI LETTER NA
0x9aa	প	p	BENGALI LETTER PA
0x9ab	ফ	P	BENGALI LETTER PHA
0x9ac	ব	b	BENGALI LETTER BA
0x9ad	ভ	B	BENGALI LETTER BHA
0x9ae	ম	m	BENGALI LETTER MA
0x9af	য	Y	BENGALI LETTER YA
0x9b2	ল	l	BENGALI LETTER LA
0x9b6	শ	S	BENGALI LETTER SHA
0x9b7	ষ	s`	BENGALI LETTER SSA
0x9b8	স	s	BENGALI LETTER SA
0x9b9	হ	h	BENGALI LETTER HA
0x9be	◌া	A2	BENGALI VOWEL SIGN AA
0x9bf	◌ি	I2	BENGALI VOWEL SIGN I
0x9c0	◌ী	i2	BENGALI VOWEL SIGN II
0x9c1	◌ু	U2	BENGALI VOWEL SIGN U
0x9c2	◌ূ	u2	BENGALI VOWEL SIGN UU
0x9c3	◌ৃ	r2	BENGALI VOWEL SIGN VOCALIC R
0x9c7	◌ে	e2	BENGALI VOWEL SIGN E
0x9c8	◌ৈ	e4	BENGALI VOWEL SIGN AI

UNICODE	ASSAMESE	ROMAN	DESCRIPTION
0x9cb	◌ো	o2	BENGALI VOWEL SIGN O
0x9cc	◌ৌ	o4	BENGALI VOWEL SIGN AU
0x9cd	◌্	+	BENGALI SIGN VIRAMA
0x9ce	ৎ	t2	BENGALI LETTER KHANDA TA
0x9dc	ড়	R	BENGALI LETTER RRA
0x9dd	ঢ়	r`	BENGALI LETTER RHA
0x9df	য়	y	BENGALI LETTER YYA
0x9f0	ৰ	r	BENGALI LETTER RA WITH MIDDLE DIAGONAL
0x9f1	ৱ	w	BENGALI LETTER RA WITH LOWER DIAGONAL
0x027	‘	`	APOSTROPHE

6. Description of method for word boundary detection

Word boundaries in the orthography are determined by localization of white spaces (blank, tab, etc).

In terms of word boundary issues, words in Assamese (such as compound words and stems with grammatical endings) often combine together to form one word. In such cases, words will be spelled without white spaces and will use the traditional spelling alterations associated with this phenomenon. Note however, that in some cases those words will appear next to each other with white space (this carries a different meaning).

Spelling of words as either a compound (without white spaces) or as separate words is checked and standardized throughout the transcription project by identifying and reviewing words which have been spelled both together and apart. Occurrences of words both with and without white spaces typically carry different meanings.

Hyphens are used in compounds to join components together when any of the components do not carry meanings on their own.

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 phones is 48. There are 30 consonants, 9 vowels (7 monophthongs and 2 diphthongs) and 9 nasal vowels (7 monophthongs and 2 diphthongs). 7 of these are foreign phones (/f, v, tS, dZ, Z, S, j/) which are not part of the native Assamese sound system but are commonly heard in English words. These phones are represented as being in allophonic variation with their equivalent pronunciations for a native speaker of the loan words. The orthography for these foreign phones would be the same as that given for the native equivalent phonemes.

Assamese Phone Chart

TYPICAL ASSAMESE CORRESPONDENCE	UNICODE	ROMAN	IPA	SAMPA	COMMENTS
			CONSONANTS		
প	0x9aa	p	p	p	
ফ	0x9ab	P	p^h	p_h	
			f	f	*allophone of /p_h/, mainly occurring in English words e.g. "**ph**one, **f**inal"*
ব	0x9ac	b	b	b	
ভ	0x9ad	B	b^h	b_h	
			v	v	*allophone of /b_h/, mainly occurring in English words e.g. "**v**ideo, **v**ery".*
ত	0x9a4	t	t	t	
ট	0x99f	t`			
ৎ	0x9ce	t2			
থ	0x9a5	T	t^h	t_h	
ঠ	0x9a0	T`			
দ	0x9a6	d	d	d	
ড	0x9a1	d`			
ধ	0x9a7	D	d^h	d_h	
ঢ	0x9a2	D`			
ক	0x995	k	k	k	
খ	0x996	K	k^h	k_h	
গ	0x997	g	g	g	
ঘ	0x998	G	g^h	g_h	
চ	0x99a	c	s	s	
ছ	0x99b	C			
			tʃ	tS	*allophone of /s/ in some English words, e.g. "spee**ch**"*

TYPICAL ASSAMESE CORRESPONDENCE	UNICODE	ROMAN	IPA	SAMPA	COMMENTS
জ	0x99c	j	z	z	
			ʒ	Z	*allophone of /z/ in some English words, e.g. "mea**s**ure, vi**s**ion"*
			dʒ	dZ	*allophone of /z/ in some other English words, e.g. "ju**dg**e"*
য	0x9af	Y	ɟ	j\	
ঝ	0x99d	Z	$ɟ^{h}$	j_h	
ম	0x9ae	m	m	m	
ন	0x9a8	n	n	n	
ণ	0x9a3	n`			
ঞ	0x99e	J			
ঙ	0x999	N	ŋ	N	
◌ং	0x982	W			
শ	0x9b6	S	x	x	
ষ	0x9b7	s`			
স	0x9b8	s			
			ʃ	S	*allophone of /x/ in some English words such as "**sh**eet"*
হ	0x9b9	h	h	h	
ৰ	0x9f0	r	ɾ, r	r	
ড়	0x9dc	R	ɽ	r`	
ঢ়	0x9dd	r`			
ল	0x9b2	l	l	l	
ৱ	0x9f1	w	w	w	
য়	0x9df	y	j	j	*occurs in English words, such as "**u**nite, **y**es". May be substituted with native vowel /i/ or /e/*

TYPICAL ASSAMESE CORRESPONDENCE	UNICODE	ROMAN	IPA	SAMPA	COMMENTS
ORAL VOWELS					
অ	0x985	a	ɔ	O	*may be pronounced /o/ in some environments; see notes below*
আ	0x986	A	a	a	
◌া	0x9be	A2			
ই	0x987	I	i	i	
◌ি	0x9bf	I2			
ঈ	0x988	i			
◌ী	0x9c0	i2			
উ	0x989	U	u	u	
◌ু	0x9c1	U2			
ঊ	0x98a	u			
◌ূ	0x9c2	u2			
এ	0x98f	e	e	e	
◌ে	0x9c7	e2			
এ	0x98f	e	ɛ	E	*may be pronounced /e/ in some environments; see notes below*
◌ে	0x9c7	e2			
য়া	0x9df 0x9be	yA2			
ও	0x993	o	ʊ	U	
◌ো	0x9cb	o2			
ঐ	0x990	e3	oi	oi	
◌ৈ	0x9c8	e4			
ঔ	0x994	o3	ou	ou	
◌ৌ	0x9cc	o4			
NASAL VOWELS					
অঁ	0x985 0x981	aM	ɔ̃	O~	

TYPICAL ASSAMESE CORRESPONDENCE	UNICODE	ROMAN	IPA	SAMPA	COMMENTS
আঁ	0x986 0x981	AM	ã	a~	
ইঁ	0x987 0x981	IM	ĩ	i~	
ঈঁ	0x988 0x981	iM			
উঁ	0x989 0x981	UM	ũ	u~	
ঊঁ	0x98a 0x981	uM			
এঁ	0x98f 0x981	eM	ẽ	e~	
এঁ	0x98f 0x981	eM	ɛ̃	E~	*may be rarer than other nasal vowels*
ওঁ	0x993 0x981	oM	ʊ̃	U~	
ঐঁ	0x990 0x981	e3M	oĩ	oi~	*may be rarer than other nasal vowels*
ঔঁ	0x994 0x981	o3M	oũ	ou~	

Notes

- Note that unlike in closely related Bengali, the Assamese consonants /t, d, t_h, d_h, n, l/ are alveolar, as indicated by the IPA symbols used above. Assamese has lost the distinction between dental and palatal consonants; both have merged to an alveolar place of articulation.
- The nasalised vowels are quite rare, but do occur in certain words and have phonemic status in Assamese. This phone set allows for a nasalised counterpart for all of the native vowels.
- The inherent vowel may be pronounced as /O/ or /o/. The two have a near-systematic pattern of distribution in Assamese (there are some exceptions). /O/ is pronounced as /o/ if there is a following /i/ or /u/. The same pattern of distribution applies to /E/ and /e/.
- /p/ and /b/ could be pronounced in their allophonic forms (IPA) [ɸ] and [β] between vowels and in word final positions. Depending on the variety spoken, however, this may not be a systematic occurrence.

8. Complete list of all rare phonemes

8.1. List of rare phonemes

We expect the following phonemes to be rare in the database:

ou
O~
a~
i~
u~
e~

E~
o~
oi~
ou~

8.2. List of foreign phonemes

The following phonemes are foreign (English):

j
v
f
tS
dZ
Z
S

9. Other language specific items

9.1. Table of digits

Digit	Digit Assamese	Assamese	Romanization
0	০	শূন্য	Su2n+Y
1	১	এক	ek
2	২	দুই	dU2I
3	৩	তিনি	tI2nI2
4	৪	চাৰি	cA2rI2
5	৫	পাঁচ	pA2Mc
6	৬	ছয়	Cy
7	৭	সাত	sA2t
8	৮	আঠ	AT`
9	৯	ন	n

9.2. Other Numbers

Number	Number Assamese	Assamese	Romanization
100	১০০	এশ	eS
10,000	১০,০০০	দহ হাজাৰ	dh hA2jA2r

Number	Number Assamese	Assamese	Romanization
100,000	১,০০,০০০	এক লাখ	ek lA2K
10 million	১,০০,০০,০০০	এক কৌটি, এক কোটি	ek ko4t`I2, ek ko2t`I2

10. References

Baruah, Sanjib and Masica, Colin P. (2001). “Assamese.” In J Garry & C. Rubino (Eds.), *Facts about the world’s languages: An encyclopedia of the world’s major languages, past and present* (pp. 43-47). New York: New England Publishing Associates.

Goswami, G.C. and Tamuli, Jyotiprakash. “Asamiya.” In Cardona, George and Jain, Dhanesh. 2003. The Indo-Aryan Languages. Routledge.

Immihelp.com, Telephone numbering scheme in India www.immihelp.com/nri/ phone-number-scheme-india.html.

Kloss, Heinz and McConnell, Grant D. The Written languages of the world: a survey of the degree and modes of use, Volume 2, Book 1., Université Laval. Centre international de recherches sur le bilinguisme Presses Université Laval, 1978.