Language Specific Peculiarities Document for

TELUGU as Spoken in INDIA

Telugu is the most widely spoken language of the Dravidian family. It is one of the five officially designated classical languages of India.

1. Dialects

Telugu is spoken in Andhra Pradesh state, which consists of 23 districts. These districts can be divided into three regions: Coastal Andhra (9 districts), Rayalaseema (4 districts) and Telangana (10 districts). There are four major dialect areas spread over these regions. Northern dialects are spoken in Telangana; Southern dialects are spoken in Rayalaseema and the two southernmost districts of Coastal Andhra; and the Eastern dialects are spoken in the three northern districts of Coastal Andhra. A set of Central dialects is spoken in the four middle districts of Coastal Andhra, at the meeting place of the three regions. The table below lists the major dialects of Andhra Pradesh based on geographical distribution.

Dialect region	Districts or Cities
Central (Standard)	Guntur, Krishna, East Godavari and West Godavari
East	Visakhapatnam, Vijayanagaram, Srikakulam
South	Nellore, Prakasham, Cuddapah (Kadapa), Kurnool, Chittoor, Anantapur
North	Ten different Telangana districts

Standard Telugu is spoken in Central Andhra (Guntur, Krishna, East Godavari and West Godavari). Unlike some other Indian languages, Telugu does not have a sharp diglossia between its spoken form and its formal written variety. The standard written form is a close match with the standard central dialects.

Appen collected data from each of the four dialect regions listed in the table above: Central, East, South, and North. The dialects of Srikakulam (which borders the region of Orissa) and Adilabad (one of the Telangana dialects, bordering the region of Maharashtra) are considered to be mutually unintelligible with the standard dialect and were not collected. There are also Telugu speaking populations in other neighboring states such as Karnataka and Tamil Nadu. These populations are not included in this collection.

2. Deviation from native-speaker principle

For this collection, only native speakers of Telugu are used.

3. Special handling of spelling

A major source of spelling variation is loan words from Hindi or English. These loan words can be spelled in many ways, depending on how the speaker pronounces the word and which Telugu character is chosen. For this collection, ABH spelled all loan words using the Telugu writing system (not English letters), and standardized to one spelling per loan word wherever possible. Note that some loan words may be given native or non-native affixes, and so may be given different spellings for one meaning, e.g., *servers* - సర్వర్లు/సర్వర్స్ "sarvarlu/sarvars".

Sanskrit loan words are also common in Telugu; these loan words are found in a dictionary and have a standard accepted spelling.

Variable word boundary decisions (morpheme joining or non-joining) are also expected to contribute to spelling variation in this language. ABH standardized to a single unjoined form wherever possible. Most compound words can be found in reference works, but can also be spelled separated by a whitespace. Since both forms are correct and commonly attested, ABH preferred to standardize to the separated spelling in order to preserve consistency in unique word lists and to prevent possible spelling variation where a term is found on its own and also as an element of a compound. By contrast, where a morpheme would not constitute a valid dictionary entry on its own, it is spelled together with its associated stem.

For the most complete guide to spelling in Telugu, please refer to the J.P.L. Gwynn Telugu-English dictionary or this online dictionary: http://www.andhrabharati.com/dictionary/.

4. Description of character set used for orthographic transcription

Only characters from the Telugu alphabet will be used for transcription of words. Underscores may be used to join elements of initialisms (such as "ABH"), which are occasionally borrowed from English. The Unicode range for Telugu is U+0C00 – U+0C7F.

Unicode specifies some characters in the Telugu range as exact equivalents. These consist of a single codepoint 0C48 (ై) which can also be represented as two separate codepoints 0C46 + 0C56 (ై). Not all fonts and programs can handle this equivalence; however, for this collection only the combined codepoint 0C48 is used.

Some characters are specified in the Telugu Unicode range, but are no longer a part of standard Telugu spelling. These characters have been excluded from the Romanization scheme. They include the letters RRA (0C31 ఱ), vocalic R (0C0B ఋ, 0C43 ృ), vocalic RR (0C60 ౠ, 0C44 ౄ), vocalic L (0C0C ఌ, 0C62 ౢ), vocalic LL (0C61 ౡ, 0C63 ౣ), TSA (0C58

ౘ), and DZA (0C59 ౙ), as well as the historical length markers (0C55 ౕ, 0C56 ౖ). Other characters do not have a unique pronunciation, but are nevertheless required elements of standard Telugu spelling. These characters have been included in the Romanization scheme. They include the diacritic symbols arasunna (0C01 ఁ), sunna (0C02 ం), and visarga (0C03 ః).

Note that NGA (0C19 ఙ) is a rarely used character and does not occur in the database.

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 all Indian languages within this program. 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.

Transcription work is done by Telugu speakers working with the Telugu script. The Romanization scheme is intended for use only as a reference for those unfamiliar with the Telugu script. It can also be of assistance to find and remove duplicated symbols that may not be visible when viewing the Telugu script alone in a text editor.

5.1 TELUGU Romanization Scheme

UNICODE	LANGUAGE	ROMAN	DESCRIPTION
		CONSONANTS	
0C15	క	k	TELUGU LETTER KA
0C16	ఖ	K	TELUGU LETTER KHA
0C17	గ	g	TELUGU LETTER GA
0C18	ఘ	G	TELUGU LETTER GHA
0C19	ఙ	N	TELUGU LETTER NGA
0C1A	చ	c	TELUGU LETTER CA
0C1B	ఛ	C	TELUGU LETTER CHA
0C1C	జ	j	TELUGU LETTER JA
0C1D	ఝ	Z	TELUGU LETTER JHA
0C1E	ఞ	J	TELUGU LETTER NYA
0C1F	ట	t`	TELUGU LETTER TTA

UNICODE	LANGUAGE	ROMAN	DESCRIPTION
0C20	ఠ	T`	TELUGU LETTER TTHA
0C21	డ	d`	TELUGU LETTER DDA
0C22	ఢ	D`	TELUGU LETTER DDHA
0C23	ణ	n`	TELUGU LETTER NNA
0C24	త	t	TELUGU LETTER TA
0C25	థ	T	TELUGU LETTER THA
0C26	ద	d	TELUGU LETTER DA
0C27	ధ	D	TELUGU LETTER DHA
0C28	న	n	TELUGU LETTER NA
0C2A	ప	p	TELUGU LETTER PA
0C2B	ఫ	P	TELUGU LETTER PHA
0C2C	బ	b	TELUGU LETTER BA
0C2D	భ	B	TELUGU LETTER BHA
0C2E	మ	m	TELUGU LETTER MA
0C2F	య	y	TELUGU LETTER YA
0C30	ర	r	TELUGU LETTER RA
0C32	ల	l	TELUGU LETTER LA
0C33	ళ	l`	TELUGU LETTER LLA
0C35	వ	w	TELUGU LETTER VA
0C36	శ	S	TELUGU LETTER SHA
0C37	ష	s`	TELUGU LETTER SSA
0C38	స	s	TELUGU LETTER SA
0C39	హ	h	TELUGU LETTER HA

UNICODE	LANGUAGE	ROMAN	DESCRIPTION
INDEPENDENT VOWELS			
0C05	అ	a	TELUGU LETTER A
0C06	ఆ	A	TELUGU LETTER AA
0C07	ఇ	I	TELUGU LETTER I
0C08	ఈ	i	TELUGU LETTER II
0C09	ఉ	U	TELUGU LETTER U
0C0A	ఊ	u	TELUGU LETTER UU
0C0E	ఎ	E	TELUGU LETTER E
0C0F	ఏ	e	TELUGU LETTER EE
0C10	ఐ	E3	TELUGU LETTER AI
0C12	ఒ	O	TELUGU LETTER O
0C13	ఓ	o	TELUGU LETTER OO
0C14	ఔ	O3	TELUGU LETTER AU
DEPENDENT VOWELS			
0C3E	ా	A2	TELUGU VOWEL SIGN AA
0C3F	ి	I2	TELUGU VOWEL SIGN I
0C40	ీ	i2	TELUGU VOWEL SIGN II
0C41	ు	U2	TELUGU VOWEL SIGN U
0C42	ూ	u2	TELUGU VOWEL SIGN UU
0C46	ె	E2	TELUGU VOWEL SIGN E
0C47	ే	e2	TELUGU VOWEL SIGN EE
0C48	ై	E4	TELUGU VOWEL SIGN AI
0C4A	ొ	O2	TELUGU VOWEL SIGN O
0C4B	ో	o2	TELUGU VOWEL SIGN OO

UNICODE	LANGUAGE	ROMAN	DESCRIPTION
0C4C	ౌ	O4	TELUGU VOWEL SIGN AU
OTHER SYMBOLS			
0C01	ఁ	M	TELUGU SIGN CANDRABINDU
0C02	ం	W	TELUGU SIGN ANUSVARA
0C03	ః	9	TELUGU SIGN VISARGA
0C4D	్	+	TELUGU SIGN VIRAMA
200C		|	ZERO WIDTH NON-JOINER
005F	_	_	UNDERSCORE
002D	-	-	HYPHEN

6. Description of method for word boundary detection

Generally in Telugu space characters are used as word boundaries. Compounding is common in Telugu, and these words are spelled together without hyphenation or other joining characters. As noted above, there is some variation in compounding, and variations in compound forms will be targeted for standardization.

7. Table containing all phones 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 34 consonants (14 of which are rare and/or foreign, found in borrowings from Sanskrit, English and Arabic), 2 semi-vowels, and 12 vowels (10 monophthongs and 2 diphthongs).

TELUGU PHONE CHART
(Krishnamurti 1985, 2003)

TYPICAL TELUGU CORRESPONDENCE	UNICODE	ROMAN	IPA	SAMPA	COMMENTS
	CONSONANTS				
క	0C15	k	k	k	
ఖ	0C16	K	kʰ	k_h	rare, allophone of /k/
గ	0C17	g	g	g	
ఘ	0C18	G	gʰ	g_h	rare, allophone of /g/
ఙ	0C19	N	ŋ	N	rare, allophone of /n/
చ	0C1A	c	t͡ʃ	tS	
ఛ	0C1B	C	t͡ʃʰ	tS_h	rare, allophone of /tS/
జ	0C1C	j	d͡ʒ	dZ	
			z	z	found predominantly in loan words, /dZ/
ఝ	0C1D	Z	d͡ʒʰ	dZ_h	rare, allophone of /dZ/
ఞ	0C1E	J	ɲ	J	rare, allophone of /n/
ట	0C1F	t`	ʈ	t`	
ఠ	0C20	T`	ʈʰ	t`_h	rare, allophone of /t`/
డ	0C21	d`	ɖ	d`	
ఢ	0C22	D`	ɖʰ	d`_h	rare, allophone of /d`/
ణ	0C23	n`	ɳ	n`	
త	0C24	t	t	t	
థ	0C25	T	tʰ	t_h	rare, allophone of /t/
ద	0C26	d	d	d	
ధ	0xc27	D	dʰ	d_h	rare, allophone of /d/

TYPICAL TELUGU CORRESPONDENCE	UNICODE	ROMAN	IPA	SAMPA	COMMENTS
న	0xc28	n	n	n	
ప	0xc2a	p	p	p	
ఫ	0xc2b	P	f	f	found in English and Arabic loanwo
			pʰ	p_h	rare, allophone of /f/
బ	0xc2c	b	b	b	
భ	0xc2d	B	bʰ	b_h	rare, allophone of /b/
మ	0xc2e	m	m	m	
య	0xc2f	y	j	j	
ర	0xc30	r	ɹ	r	
ల	0xc32	l	l	l	
ళ	0xc33	l`	ɭ	l`	
వ	0xc35	w	ʋ	v\	
శ	0xc36	S	ʃ	S	
ష	0xc37	s`	ʂ	s`	
స	0xc38	s	s	s	
హ	0xc39	h	h	h	
	VOWELS				
అ	0xc05	a	a	a	
ఆ	0xc06	A	aː	aː	
ా	0xc3e	A2			
ఇ	0xc07	I	i	i	
ి	0xc3f	I2			
ఈ	0xc08	i	iː	iː	
ీ	0xc40	i2			

TYPICAL TELUGU CORRESPONDENCE	UNICODE	ROMAN	IPA	SAMPA	COMMENTS
ఉ	0xc09	U	u	u	
ు	0xc41	U2			
ఊ	0xc0a	u	uː	u:	
ూ	0xc42	u2			
ఎ	0xc0e	E	e	e	
ె	0xc46	E2			
ఏ	0xc0f	e	eː	e:	
ే	0xc47	e2			
ఆ	0xc06	A	æ	{	Central/Eastern dialectal allophone of /a/ in certain word-final grammatical contexts.
ా	0xc3e	A2			
ఏ	0xc0f	e	æː	{:	Central/Eastern dialectal allophone of /eː/
ే	0xc47	e2			
ఐ	0xc10	E3	ai	ai	
ై	0xc48	E4			
ఒ	0xc12	O	o	o	
ొ	0xc4a	O2			
ఓ	0xc13	o	oː	o:	
ో	0xc4b	o2			
ఔ	0xc14	O3	au	au	
ౌ	0xc4c	O4			

OTHER SYMBOLS	
.	syllable break
#	word boundary

7.1 List of rare phones

The aspirated stops and the velar and palatal nasal consonants are chiefly historical and less commonly heard in modern Telugu. They are still used by some speakers for some words in modern Telugu, therefore where applicable these allophones are included in a dispreferred variant entry in the pronunciation lexicon, with unaspirated stops or alveolar nasals being given as the preferred variant.

Note that the velar nasal may occur more frequently than the others in this set due to its presence in loan words. The palatal nasal is expected to occur very rarely, if at all.

The list of potentially rare phones is as follows:

- k_h
- g_h
- tS_h
- dZ_h
- t`_h
- d`_h
- t_h
- d_h
- p_h
- b_h
- N
- J

7.2 List of foreign phones

- /z/ for some loan words
- /f/ for some loan words

7.3 Loan word behaviour in Telugu

As Telugu does not feature word initial consonant clusters or clusters other than geminates or homorganic nasal + stop, it is only in loan words (and thus predominantly in the speech of educated/multilingual speakers) we are likely to find these phonemic patterns. Likewise, Telugu words can only end in short vowels or /m n w j /, so a word with any other word-final phoneme can be expected to be a loan word.

8. Other Language Specific Items

8.1 Table of Digits

Roman digit	Telugu Digit	Telugu word	Romanization
0	౦	సున్నా	sU2n+nA2
1	౧	ఒకటి	Ok t`I2
2	౨	రెండు	rE2W D`U2
3	౩	మూడు	mu2D`U2
4	౪	నాలుగు	nA2lU2gU2
5	౫	ఐదు	E3dU2
		అయిదు	ayI2dU2
6	౬	ఆరు	ArU2
7	౭	ఏడు	eD`U2
8	౮	ఎనిమిది	EnI2mI2dI2
9	౯	తొమ్మిది	tO2m+mI2dI2

8.2 Other Numbers

Roman digit	Telugu Digit	Telugu word	Romanization
10	౧౦	పది	pdI2
100	౧౦౦	వంద	wWd
1000	౧౦౦౦	వెయ్యి	wE2y+yI2
10,000	౧౦,౦౦౦	పది వేలు	pdI2 we2lU2
100,000	౧౦౦,౦౦౦	లక్ష	lk+s`
10 million	౧౦,౦౦౦,౦౦౦	కోటి/పది మిలియన్లు	ko2t'I2/pdI2 mI2lI2yn+lU2

9. References

http://www.unicode.org/charts/PDF/U0C00.pdf

http://www.lmp.ucla.edu/Profile.aspx?menu=004&LangID=100

http://en.wikipedia.org/wiki/Telugu_language

Gwynn, J.P.L, 1991, *A Telugu-English Dictionary*, Oxford: Oxford University Press

Krishnamurti, B.,

1985, *A Grammar of Modern Telugu*, Oxford: Oxford University Press.

1998, “Telugu”, by Bh. Krishnamurti, in Sanford Steever (Ed.) *The Dravidian Languages*, 202-240. ed. by Sanford Steever (Routledge.

2003, *The Dravidian Languages*, Cambridge: Cambridge University Press, 1998; pp. 202-240)