Language Specific Peculiarities Document for
Tamil as spoken in India

1. Dialects

Tamil is the official language of the Indian state of Tamil Nadu and the Indian union territory of Puducherry. It is also a national language of Sri Lanka and an official language of Singapore. The collection will take place in Tamil Nadu. Five regional varieties have been identified within Tamil Nadu, based on four regional varieties of Northern, Central, Southern and Western[1] with a division of the Central dialect to account for the distinctness of Tamil spoken in Madurai. All five dialects are represented in the collection.

Region	Districts or Cities
Northern	Chennai, Kanchipuram, Tiruvannamalai and Vellore Districts
Central – Other	Tiruchirappalli, Thanjavur, Cuddalore and Villupuram Districts
Central – Madurai	Madurai
Southern	Ramanathapuram, Tirunelveli and Kanniyakumari Districts
Western	Salem, Dharmapuri, Coimbatore and Nilgiris Districts

Standard Spoken Tamil is the common standard dialect that is understood by most speakers. It is based on the everyday speech of higher-caste, educated, non-Brahmin Tamils in urban areas of Tamil Nadu, such as the Central and Southern districts of Tiruchirappalli, Thanjavur and Madurai. It is disseminated through media such as film and radio. It is understood everywhere Tamil is spoken including Sri Lanka and Singapore (Schiffman 1999).

Standard Spoken Tamil is an 'informal' register that exists in a diglossic situation with a formal or literary variety. Among speakers there is a range of competence in literary Tamil (LT). All speakers can understand LT to some degree, but the ability to speak or write LT is confined to educated speakers (Steever 1987; Britto 1986). LT is used in most writing, some broadcasting and in political oratory.[2]

2. Deviation from native-speaker principle

No special deviation – only non-Brahmin, Indian native speakers of Tamil, will be collected in this project. No Sri Lankan or Singaporean speakers will be represented.

1 http://www.lisindia.net/Tamil/Tamil_vari.html. Some sources (e.g. Steever 1990) also indicate 5 regional dialects – but with the addition of an Eastern dialect comprising Puttukottai and Ramanathapuram and the regrouping of Madurai into one Central dialect. However, the distinctness of Tamil in Madurai from other Central dialects, as well as from Puttukottai and Ramanathapuram suggests that it should be considered as a separate group.

2 In addition, spoken Tamil amongst Hindus also varies according to caste. In particular, there are identifiable Brahmin (priestly caste) and non-Brahmin (other castes) varieties of Tamil marked by differences particularly in phonology (McDonough & Johnson 1997) and lexicon, but also in grammar and even semantic structure (Bright 1968:457-458). Such differences in caste varieties may also be present in the Tamil varieties spoken by Christians where caste identities are still maintained (see, for example, Sherinan 2007).

3. Special handling of spelling

Standard Spoken Tamil is not usually a written language. In some cases there is an accepted spelling for the Standard Spoken forms and transcriptions will be standardized to these norms. In other cases an accepted form does not yet exist, and so Appen determined their own standard that was as systematic as possible, while also taking into account frequency of use and native speaker expert opinion.

Business names and loan words from other languages are spelled in Tamil script and standardized.

4. Description of character set used for orthographic transcription

Tamil script will be used for orthographic transcription. The Unicode range for Tamil is U+0B80-U+0BFF. The complete script consists of 12 independent vowels, 11 dependent vowels, 23 consonants, and the signs Pulli (which suppresses the inherent vowel) and Aytam (which indicates frication of /p/ and /dʒ/ for foreign sounds).

5. Description of Romanization scheme

The following is Appen's Romanization scheme, which is fully reversible. Transcription work is done by Tamil speakers working with the Tamil script and no Romanization; the Romanization scheme is primarily used as a reference for those unfamiliar with the Tamil script. Some of the symbols may appear arbitrary; however, they have been chosen to fit with similar schemes for other Indian languages.

UNICODE	TAMIL	ROMAN	DESCRIPTION
CONSONANTS			
0xb95	க	k	TAMIL LETTER KA
0xb99	ங	N	TAMIL LETTER NGA
0xb9a	ச	c	TAMIL LETTER CA
0xb9c	ஜ	j	TAMIL LETTER JA
0xb9e	ஞ	J	TAMIL LETTER NYA
0xb9f	ட	t`	TAMIL LETTER TTA
0xba3	ண	n`	TAMIL LETTER NNA
0xba4	த	t	TAMIL LETTER TA
0xba8	ந	n	TAMIL LETTER NA
0xba9	ன	n^	TAMIL LETTER NNNA
0xbaa	ப	p	TAMIL LETTER PA
0xbae	ம	m	TAMIL LETTER MA
0xbaf	ய	y	TAMIL LETTER YA
0xbb0	ர	r	TAMIL LETTER RA
0xbb1	ற	R	TAMIL LETTER RRA

UNICODE	TAMIL	ROMAN	DESCRIPTION
0xbb2	ல	l	TAMIL LETTER LA
0xbb3	ள	l`	TAMIL LETTER LLA
0xbb4	ழ	L	TAMIL LETTER LLLA
0xbb5	வ	w	TAMIL LETTER VA
0xbb6	ஶ	S	TAMIL LETTER SHA
0xbb7	ஷ	s`	TAMIL LETTER SSA
0xbb8	ஸ	s	TAMIL LETTER SA
0xbb9	ஹ	h	TAMIL LETTER HA
INDEPENDENT VOWELS			
0xb85	அ	a	TAMIL LETTER A
0xb86	ஆ	A	TAMIL LETTER AA
0xb87	இ	I	TAMIL LETTER I
0xb88	ஈ	i	TAMIL LETTER II
0xb89	உ	U	TAMIL LETTER U
0xb8a	ஊ	u	TAMIL LETTER UU
0xb8e	எ	E	TAMIL LETTER E
0xb8f	ஏ	e	TAMIL LETTER EE
0xb90	ஐ	e3	TAMIL LETTER AI
0xb92	ஒ	O	TAMIL LETTER O
0xb93	ஓ	o	TAMIL LETTER OO
0xb94	ஔ	o3	TAMIL LETTER AU
DEPENDENT VOWELS			
0xbbe	ா	A2	TAMIL VOWEL SIGN AA
0xbbf	ி	I2	TAMIL VOWEL SIGN I
0xbc0	ீ	i2	TAMIL VOWEL SIGN II
0xbc1	ு	U2	TAMIL VOWEL SIGN U
0xbc2	ூ	u2	TAMIL VOWEL SIGN UU
0xbc6	ெ	E2	TAMIL VOWEL SIGN E
0xbc7	ே	e2	TAMIL VOWEL SIGN EE
0xbc8	ை	e4	TAMIL VOWEL SIGN AI
0xbca	ொ	O2	TAMIL VOWEL SIGN O
0xbcb	ோ	o2	TAMIL VOWEL SIGN OO
0xbcc	ௌ	o4	TAMIL VOWEL SIGN AU

UNICODE	TAMIL	ROMAN	DESCRIPTION
OTHER SIGNS			
0xb83	ஃ	9	Aytam
0xbcd	◌்	+	Pulli

6. Description of method for word boundary detection

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

7. Table containing 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 34. There are 22 consonants and 12 vowels (10 monophthongs and 2 diphthongs). This number includes 5 consonant phonemes used only in borrowed words.

UNICODE	TAMIL	ROMAN	IPA	SAMPA	TAMIL	ROMANIZATION
CONSONANTS						
0xb95	க	k	k	k	கல்	kl+
0xb99	ங	N	ŋ	N	வங்கி	wN+kI2
0xb9a	ச	c	ʧ	tS	சரி	crI2
0xb9e	ஞ	J	ɲ	J	ஞாயிறு	JA2yI2RU2
0xb9f	ட	t`	ʈ	t`	பட்டு	pt`+t`U2
0xba3	ண	n`	ɳ	n`	பணம்	pn`m+
0xba4	த	t	t̪	t	பத்து	pt+tU2
0xba8	ந	n	n̪	n	நரி	nrI2
0xba9	ன	n^	n		மௌன	mo4n^
0xbaa	ப	p	p	p	பல்	pl+
0xbae	ம	m	m	m	மாலை	mA2le4
0xbaf	ய	y	j	j	யானை	yA2n^e4
0xbb0	ர	r	ɾ	4	மரம்	mrm+
0xbb1	ற	R	r	r[3]	கறி	kRI2
0xbb2	ல	l	l	l	ஒலி	OlI2
0xbb3	ள	l`	ɭ	l`	ஒளி	Ol`I2

[3] Most dialects pronounce both ர and ற as /4/. The exception is Southern, where R=/r/

UNICODE	TAMIL	ROMAN	IPA	SAMPA	TAMIL	ROMANIZATION
0xbb4	ழ	L	ɻ	r\`[4]	பழம்	pLm+
0xbb5	வ	w	ʋ	v\	கவி	kwI2
FOREIGN CONSONANTS						
0xbb6	ஶ	S	ʃ	S	ஶ்ரீ	S+ri2
0xbb7	ஷ	s`			வருஷம்	wrU2s`m+
0xb9c	ஜ	j	dʒ	dZ	ஜலம்	jlm+
0xb83 0xb9c	ஃஜ	9j			ஃஜிப்ட்	9jI2p+t`+
0xbb8	ஸ	s	s	s	ஸம்ஸ்க்ருதம்	sm+s+k+rU2tm+
0xbb9	ஹ	h	h	h	ஹரி	hrI2
0xb83 0xbaa	ஃப	9p	f	f	ஃபோன்	9po2n^+
VOWELS						
0xb85	அ	a	a	a	அடி	at`I2
0xb86	ஆ	A	a:	a:	ஆடி	At`I2
0xbbe	ா	A2			நாடி	nA2t`I2
0xb87	இ	I	i	i	இடி	It`I2
0xbbf	ி	I2			சிரி	cI2rI2
0xb88	ஈ	i	i:	i:	ஈ	i
0xbc0	ீ	i2			நீ	ni2
0xb89	உ	U	u	u	உரி	UrI2
0xbc1	ு	U2			சுட்டு	cU2t`+t`U2
0xb8a	ஊ	u	u:	u:	ஊதி	utI2
0xbc2	ூ	u2			மூடு	mu2t`U2
0xb8e	எ	E	e	e	எரி	ErI2
0xbc6	ெ	E2			பெண்	pE2n`+
0xb8f	ஏ	e	e:	e:	ஏறு	eRU2
0xbc7	ே	e2			பேனா	pe2n^A2
0xb92	ஒ	O	o	o	ஒரு	OrU2
0xbca	ொ	O2			பொரி	pO2rI2
0xb93	ஓ	o	o:	o:	ஓடு	ot`U2

[4] Not all dialects maintain a distinction between ழ /r`\/ and ள /l`/.

UNICODE	TAMIL	ROMAN	IPA	SAMPA	TAMIL	ROMANIZATION
0xbcb	ோ	o2			போ	po2
Diphthongs						
0xb94	ஔ	o3	au	au	ஔரங்கசீப்	o3rN+kci2p+
0xbcc	ௌ	o4			மௌன	mo4n^
0xb90	ஐ	e3	ai	ai	ஐப்பசி	e3p+pcI2
0xbc8	ை	e4			தை	te4

OTHER SYMBOLS	
.	syllable break
#	word boundary

Notes

- In words ending in a vowel + /n/ or /m/, the vowel may be nasalized and the consonant deleted. The nasal vowels do not have phonemic status in Tamil.
- Voiced stops can contrast with voiceless stops in initial position in foreign words. Elsewhere the contrast is neutralized.

7.1. List of rare phonemes

The following phonemes will be rare in Tamil but not in English borrowings.

IPA	SAMPA
au	au
ai	ai

7.2. List of foreign phones

The following phonemes are foreign. These phones can be assimilated to native equivalents in speech.

IPA	SAMPA
ʃ	S
dʒ	dZ
s	s
h	h
f	f

8. Other Language Specific Items

Some of the forms in the tables below are not often written down and were standardized by Appen.

8.1. Table of Digits

Numeral	Digit	Tamil	Romanization
0	-	சுழி பூஜ்ஜியம் ஜீரோ	cU2LI2 pu2j+jI2ym+ ji2ro2
1	க	ஒன்னு	On^+n^U2
2	உ	ரெண்டு	rE2n`+t`U2
3	ங	மூன்று	mu2n^+RU2
4	ச	நாலு	nA2lU2
5	ஞ	அஞ்சு	aJ+cU2
6	கூ	ஆறு	ARU2
7	எ	ஏழு	eLU2
8	அ	எட்டு	Et`+t`U2
9	கூ	ஒம்பது	Om+ptU2

8.2. Other Numbers

Numeral	Tamil	Romanization
10	பத்து	pt+tU2
100	நூறு	nu2RU2
1000	ஆயிரம்	AyI2rm+
10,000	பத்தாயிரம்	pt+tA2yI2rm+
100,000	லட்சம்	lt`+cm+
10 million	ஒரு கோடி	OrU2 ko2t`I2

9. References

Arden, A.H. (1954) *A Progressive Grammar of Common Tamil.* Madras: Christian Literature Society.

Agesthialingom, S. (1967) *A Generative Grammar of Tamil.* Annamalai University.

Bright, W. (1968). Social dialect and semantic structure in South Asia. In M. Singer & B. S. Cohn (Eds.), *Structure and change in Indian society* (pp. 455-460). New Jersey: Wenner Gren Foundation for Anthropological Research.

Britto, F. (1986) *Diglossia: A study of the theory with application to Tamil.* Washington: Georgetown University Press. 229-230

Lehmann, T. (1992) *A Grammar of Modern Tamil*. Pondicherry: PILC

McDonough, J., & Johnson, K. (1997). Tamil liquids: An investigation into the basis of the contrast among five liquids in a dialect of Tamil. *Journal of the International Phonetic Association, 27*(1-2), 1-26.

Pon.Kothandaraman. (1997) *A Grammar of Contemporary Literary Tamil*. Chennai: International Institute of Tamil Studies.

Ramanujan, A. K. (1968). The structure of variation: A study in caste dialects. In M. Singer & B. S. Cohn (Eds.), *Structure and change in Indian society* (pp. 461-474). New Jersey: Wenner Gren Foundation for Anthropological Research.

Schiffman, H. F. (1998). Standardization or restandardization: the case for "Standard" Spoken Tamil. *Language in society*, *27*(03), 359-385.

Schiffman, H. F. (1999). *A reference grammar of spoken Tamil.* Cambridge: Cambridge Univ Pr.

Sherinan, Z. C. (2007) Musical style and the changing social identity of Tamil Christians. Ethnomusicology, 51(2), 238-280

Steever, S. (1987) Tamil and the Dravidian Languages. In B. Comrie (ed) *The World's Major Languages,* Oxford University Press.

Language Variation in Tamil (http://www.lisindia.net/Tamil/Tamil_vari.html), Accessed 9/08/12.

UCLA Language materials project 'Tamil' (http://www.lmp.ucla.edu/Profile.aspx?LangID=99&menu=004), Accessed 9/08/12.

Encyclopedia Brittanica Entry on Dravidian Languages (http://ccat.sas.upenn.edu/~haroldfs/sars238/encybrit.html), Accessed 9/08/12.

Online Dictionaries/Corpora:

Online Tamil Dictionary (http://www.tamildict.com/), Accessed 9/08/12.

http://crea.in, Accessed 12/8/12

http://www.crea.in/corpus.html