

A Procedure for Transliteration from Tamil to Sinhala

Gihan V Dias¹

G Balachandran²

Department of Computer Science and Engineering
University Of Moratuwa, Sri Lanka
¹gihan@uom.lk, ²balag@uom.lk

ABSTRACT

Transliteration is the process of representing text in one script by the characters of another script. It preserves spelling and pronunciation rather than meaning. Transliteration is the preferred method of converting data such as names, addresses, etc. in databases and applications. Both Tamil and Sinhala scripts belong to the Indic family. There is a correspondence of characters between the two scripts - although their visual appearance may differ - which facilitates transliteration. However, as Tamil has fewer letters than Sinhala, transliteration is more complicated than a one-to-one mapping.

Many government documents in Sri Lanka, for example National Identity Cards, are written in Sinhala. Many errors are now found in Tamil words manually transliterated into Sinhala in such documents. A standard method of transliteration from Tamil to Sinhala will obviate such errors.

The research developed a procedure for transliterating words from Tamil to Sinhala. This may be used as a framework for transliteration between other Indic scripts as well. The procedure contains five levels of rules. The first level maps irregular words which do not follow the other rules. The next two levels map letters using one-to-one and one-to-many rules, respectively. The fourth and fifth levels process the pronunciation information according to contemporary usage.

This procedure consistently and accurately maps names, addresses and other non-translated words from Tamil to Sinhala.

Keywords: Transliteration, Indic Scripts, Framework, Tamil, Sinhala

1.0 INTRODUCTION

In linguistics the *transliteration* is the process of representing graphic characters of a source script by the graphic characters of a target script [1]. Countries which have more than one official language will face the need of the transliterations regularly. In addition, transliteration should not be confused with translation, which involves a change in language while preserving meaning.

Indic scripts are phonetically-based *abugidas*. There is a clear correspondence of characters between two Indic scripts, although their visual appearance may differ. English uses an alphabetic script which does not correlate well with Indic scripts. Both Tamil and Sinhala use Indic scripts, which facilitates transliteration. However, as Tamil has fewer letters than Sinhala, transliteration is more complicated than a one-to-one mapping. Both spelling and pronunciation need to be considered in this transliteration.

A large number of government and private organizations in Sri Lanka deal with data acquired in the form of any of the Sri Lankan official language's scripts. Sinhala and Tamil scripts are main sources of input to most organizations where data is collected using forms such as registration forms, tax forms, visa forms and census forms. After the data is been collected transliterating the data into another script is a common task in many organizations for the reason of policies they follow. In many instances, data originally in Tamil is converted to Sinhala. Currently all the data needs to transliterated manually which is time consuming and error prone. These organizations would benefit greatly from an automated transliterated system.

There are few attempts done in this area related to Tamil and Sinhala Indic scripts. In November 2005, Balachandiran Balamurali in his masters' degree (University of Moratuwa) proposing an Algorithm for automated English-Tamil name Transliteration. Balachandiran Balamurali

developed a Java application to test the algorithm and concluding the research as “The results of the system are more than 98% accurate with the name base”. Moreover in January 2006, H.M.Weerasinghe from same university made research in “Transliteration of Names from English to Sinhala” for his masters’ degree. Both above researches were supervised by professor Gihan. V. Dias.

In addition during mid 2008, ICTA Sri Lanka was attempting (contracted to Science Land) to develop English-Sinhala-Tamil tri lingual transliteration software. Furthermore Daham Widurinda Jayatilake and K.M.J Karunaratne are contributed in this area.

This paper develops a procedure to consistently and accurately transliterate Tamil words to Sinhala and suggests a framework for transliterating among other Indic scripts.

2.0 INDIC SCRIPTS

Most languages spoken in India derive their orthography from the Brahmi script. Brahmic is a family of abugidas (writing systems) used in South Asia, Southeast Asia, Tibet, Mongolia, Manchuria. In an abugida, each character denotes a consonant accompanied by a specific vowel, and consonants accompanied by other vowels, or without a vowel, are denoted by a consistent modification of the consonant symbols [2].

Tamil script is mostly used to write Tamil language and Tamil language is belongs to Dravidian language family. It is an official language of Sri Lanka, which is used by the one of the minority ethnic group called Tamils and Tamil Muslims. Other languages that belong to this family are Telegu, Kannada and Malayalam.

Sinhala script is presently used to write Sinhala language and is an official language of Sri Lanka. It is used primarily by the Sinhalese who are the largest ethnic group in Sri Lanka. As mentioned by Fernando et al. [3], Sinhala is member of the Indo-Aryan family of languages. Other languages that belong to this family are Sanskrit, Hindi and Bengali.

3.0 PROBLEM DOMAIN

Practically any natural language will get polluted by other language words with the time. Moreover one community might be using more than one language words in their day to day life. When is come to person’s name and place’s name people

tend to use different language originated words. Sri Lankans use words originated from several languages. Major ones are Sinhala, Tamil, Sanskrit, Pali, Arabic, English, Dutch, Portuguese, and Malay. There could be others as well.

When we consider contemporary Tamil to Sinhala transliteration, according to the origin language of the Tamil word, the transliteration is done. For an example

கண்ணன் (Tamil language origin) is transliterated as කනේනන් and கங்கா (Sanskrit language origin) is transliterated as ගංගා in Sinhala. In Sinhala the two different character is been used for the single Tamil character க.

Similarly Tamil script written Arabic language name would be difficult to transliterate correctly into Sinhala script if there is no knowledge in Arabic language. Practically each script has to be handled by a person with the knowledge in the word originated languages. This process makes the automated transliteration into a complex procedure.

Analysis shows that Tamil to Sinhala transliteration rules are not followed well in Sri Lanka. Government documents, registrations certificates, even dictionaries are lack in following the rules. No proper standardization and less awareness are the major factors for this issue. Problems arises can be looked in following categories.

- Ambiguity In Letters
- Ambiguity In Sounds
- Incorrect Mapping

3.1. Ambiguities in Letters

In Tamil to Sinhala transliteration there are occasions which needs one-to-one mapping among the letters. Simple explanation will be a Tamil letter ஁ will always map to Sinhala letter ඉ, there is no other alternatives or ambiguity in this. Even though in many letters follow the one-to-one mapping during the transliteration, this is not implemented correctly because of the ambiguity in the pronunciations.

Tamil Words	Pronunciation
இலை	ல - The tongue should touch the teeth-ridge just above the base of the front upper teeth
இளை	ள - The tongue must be curled back and flapped forward
இழை	ழ - Pronounce as for an ordinary soft ř (r), but draw back the whole tongue making it spread the blade across the mouth. In addition should be turned back against the hard plate. The result should be slurred, obscure sound between ř (r) and ல (l).

Table 1: Tamil ல, ள and ழ

In Tamil where இலை, இளை and இழை has different in meanings and where might not followed strictly in the spoken. During manual transliteration process the Tamil letters ல, ள and ழ set mixed with ட and ட Sinhala letters set. Even though there are well defined transliteration rules exists for these letters it been mixed often. This problem exists for Tamil letters ந, ண, and ன in same manner; it mixed with ண and ன.

3.2. Ambiguities in Sounds

In phonetics (“Study of sound production”), an *allophone* is one of several similar speech sounds (phones) that belong to the same phoneme. In linguistics *phoneme* is a minimum unit of distinctive sound feature in the language. In Tamil, use of allophones is more compare with other Indic languages. This feature of the language causes problems in manual transliterations. Examples are given in appendix (Example 1 and 2)

3.3 Incorrect Mapping

For the reason that there are no standard transliteration rules exists people are not familiar with the situations. This will leads to the incorrect mappings of the letters. Example is given in appendix (Example 3)

Above mentioned issues were collated from various sources such as National identity cards, Birth certificates, Sri Lanka government gazettes, Government postal department guides and Dictionaries. It clearly shows that there are no

transliteration rules been followed within a unit of data processing. This without a doubt shows the need of this research and the needs of the implementation on the transliteration rules.

4.0 FRAMEWORK

A sequence of instructions used for do the transliteration in this research. Tamil to Sinhala transliteration done with a list of well-defined instructions, when given an initial state, proceed through a well-defined series of successive states, eventually terminating in an end-state. The transition from one state to the next is deterministic in this research.

This framework solution contains five levels of rules. The first level includes the words segments mapping for irregular words which do not obey the transliteration rules. The next two levels process the spelling mapping by one-to-one and one-to-many mapping rules. The fourth and fifth levels process the pronunciation information to go along with the contemporary usage.

4.1. Level 1 – Direct Word Mapping

In words derived from Sanskrit, or any other language, the letters may transliterate into original mapping into Sinhala language. Words such as நாகம், குரு, கணேசன் will not follow next levels of the framework. Exact words will be matched to transliterated values in a table.

4.2. Level 2 – One-to-One Mapping

In transliteration the writing conventions are primary importance than the pronunciation conventions. In this regards the elements of the Tamil language is been categorized as below rather than usual classification.

Category	Count
Vowels	12
Consonants	25
Consonant plus vowel அ (a) syllables	25
āytam	1
ஸ்ரீ (Sri)	1
Total	64

Table 2: Tamil language elements regards to transliteration

The Tamil and Sinhala have a phonetic based layout of the alphabet, i.e., the sequence and

layout of the standard presentations of the characters have a relation to their sound and place of origin in the human vocal system. There is a unique mapping between characters and sounds. This facilitates transliteration to some degree.

Since Tamil and Sinhala are rather close to the phonetics of their sounds, in this level 2 without considering the allophones, the direct phoneme mapping is considered. The ISO 15919 1st edition, “Information and documentation – Transliteration of Devanagari and related Indic scripts into Latin characters” standard is providing the basis mapping rule in well manner. This standard has been taken as the main source, for defining the one-to-one phoneme mapping in this research.

Table 3 in appendix summaries the one-to-one mapping of the Tamil 12 vowels and 21 (21 out of 25 from Table 2) consonants one-to-one mapping with Sinhala letters. The consonants are differentiated based on its articulator characteristics and distributional characteristics.

In the direction of the transliteration of Tamil to Sinhala, the vowels ඇ, ඇ, ඩා, ඩාා, ට, and ටා will not occur in the transliterated words since Tamil does not includes the above vowels. In addition the all Tamil vowels always follows one-to-one mapping with Sinhala vowels while writing.

In Tamil the consonants are pronounced with their original phoneme in every situation. Even though Sinhala language has 41 consonants above 21 Tamil consonants always phonetically map in one-to-one to 21 Sinhala consonants.

The analysis in the contemporary Sinhala writing shows that consonant ඩ is not been used when ඙ been transliterated. It will be always written using the anusvara “◌◌”. The ඡ consonant in Tamil represent a peculiar sound. Pronounce as for an ordinary soft ř (r), but draw back the whole tongue making it spread the blade across the mouth. In addition should be turned back against the hard plate. The result should be slurred, obscure sound between ř (r) and ல (l). Since Sinhala does not have an equal consonant, ඡ is used during transliteration. On the other hand to transliterate ள consonant ඡ is normally used. Tamil ற consonant pronounced as, applying the tip of the tongue to the ridge of the palate, and pronouncing a rough, vibrating ř (r). Sinhala does not have an equal consonant. During Tamil to Sinhala transliteration ට is used to represent the ற. On the other hand the syllables of ற are not

transliterated as ට syllables. Tamil ள consonant pronounced as, applying the tip of the tongue to the ridge of the palate, and pronouncing a distinct ற (n). Theoretically ள is a little harder than ற. Since Sinhala does not have an equal consonant, ඡ is used during transliteration. On the other hand to transliterate ற consonant ඡ is normally used. Sinhala has the conjunct concepts and the சஞ் consonant it will be transliterated into ங in Sinhala text.

Above explained ஡, ற, ள and சஞ் consonants gives the total of 25 consonants which been used in the contemporary Tamil.

When it comes to “Consonant plus vowel அ (a) syllables” category, even though consonant ற transliterated as ට, the syllables of ற are transliterated as ජ syllables, since Sinhala does not have an equal consonant for ற. Similarly anusvara (◌◌) used to transliterate the Tamil consonant ங, but ඩ consonant’s syllables are mapped to the ங consonant’s syllables. However grammatically anusvara will not form any syllables; e.g. அங்ஙனம் transliterated as අංඛනම්

The āytm is unique to the Tamil language. Even the ISO 15919 1st edition, standard does not provide any mapping to this among the other major Indic languages, such as Devanagari, Gujarati, Bengali, Oriya, Malayalam, Kannada, Telugu and Sinhala. Conversely the contemporary Sinhala map the visarga (◌:) to the ඌ; அஃறிணை transliterated as අඃරිඣන.

The conjunct syllable (ஶ்ரீ) Sri is a word and which has goodness Luxmi, gifted opportunity and few other meanings which is borrowed from Sanskrit. It also has the “respectable” meaning. When it appears in names it mostly comes with “respectable” meaning. ஶ்ரீ and ஶ்ரீ are two different shapes for this syllable and both have been used for more than one hundred years, and both are mapped to same ශ්ரී.

During level 2 one-to-one mapping Tamil language elements shown in Table 2 are transliterated without considering any other factors. As an example any syllable formed using ‘க’ will be transliterated into ‘ක’; கண்ணன் as කණ්ණන් and கங்கா as කංකා

4.3. Level 3 – One-to-Many Mapping

Changing a phoneme in a word can produce another word. Speakers of a particular language perceive a phoneme as a distinctive sound in that language. An allophone is not distinctive, but rather a variant of a phoneme; changing the allophone result may sound non-native, or be unintelligible. A good example will be phoneme ‘p’ producing two different meanings according to the pronunciation in word பாவம் [4]. பாவம் (pāvam): Accumulated result of sinful actions, பாவம்(bāvam) : Contemplation, Meditation

A comparison studies showed that, Tamil use the allophones concept more compare with other Indic languages, for the reason that, it has less consonants in the alphabet than the other Indic languages. In arranging the consonants, the Tamil and Sanskrit alphabet follows the same arrangement of “vargas” or rows. Sanskrit is the ancient sacred language of India; it is believed to be the oldest language of India. However, Tamil system has only the first and the last consonant of each row, omitting the altogether the intermediate consonants.

In the first or guttural row, Tamil alphabet has only the ‘k’, and its corresponding nasal ‘ñ’ and not the ‘kh’, ‘g’, and ‘gh’. This pattern follows to palatal, lingual / cerebral, dental and labial rows. Figure 1 in appendix shows a comparison between Sanskrit (via Grantha script) and Tamil

Even though these letters do not exist in the alphabet, the sounds of them are commonly used in Tamil language since the geographical reason. Analysis shows that sounds of voiced unaspirated consonants are been used in Tamil, i.e. g, j, ḍ, d, and b. In view of the fact that Tamil does not have the every consonants, laws of sounds requires making the same consonant to be pronounced the above voiced unaspirated consonants. In this regards the first consonant of each vargas used to represent the voiced unaspirated consonants with certain laws. Consequently the consonants ‘k’, ‘c’, ‘t’, ‘t’, and ‘p’ are will be used to represent more than one sound.

These rules are obtained by study of the language grammar. Some rules are listed below.

- In the beginning of the word, ‘ka’ will be ‘ka’ (க)
- ‘ka’ as ‘ha’(ஹ) sound when is single between two vowels sounds

நாகமணி => ந் + ஆ + க் + அ + ம் + அ + ண் + இ. i.e. : நாகமணி => நாஹணீ

- ‘ca’ after ஞ் it will be “aj” (அ)
- ‘ṭa’ after the ண் and ம் it will be ‘ḍa’ (ட)
- ‘ta’ after the ந் family letters it will be da (ட)
- ‘pa’ after க், and ண் it will be “pa” (ப)

Level 3 one-to-many rules will be applied to the one-to-one transliterated words and passed to the level 4 of the framework. As an example in level 2 கமலசேகரம் will be transliterated கமலசேகரம் and passed to level 3. At end of level 3, according to the one-to-many rules it will be கமலசேகரம்.

After the Level 3 stage it will be almost in completed stage. However the data analysis shows that there are few more rules are been followed during the transliteration. These rules are not following the phonemes mapping, and they do actually leave out the transliteration rules. It can be categorized as Contemporary Compulsory Rules and Contemporary Non-Compulsory Rules.

4.4. Level 4 – Contemporary Compulsory Rules

Ideal transliteration is loss-less, i.e., the informed reader should be able to reconstruct the original spelling of unknown transliterated words. In other words a Tamil name ராமலிங்கம் should be transliterated as රාමලිංගම according the rules. Complex fact is originally in Tamil, it can be written as ராமலிங்கம் or இராமலிங்கம் and both mean the exactly same, and it will be always transliterated as රාමලිංගම in Sinhala. Even though இராமலிங்கம் transliterated into රාමලිංගම and it loses the letters, this would be the expected way. Such behaviors identified and specified as level 4 rules; there might be more rules exists. Some rules are listed below.

- According to the Tamil grammar a word cannot start with “ர” row words. A borrowed word starting with ர, ரா, ரி, ரீ, ரெ, ரே, and ரை may be written with leading “இ”. A borrowed word starting with ரு, ரூ, ரொ, and ரோ may be written with leading “உ”. During transliteration it must be omitted

- According to the Tamil grammar a word cannot start with other than “யா” in the “ய” row words. A borrowed word starting with other than “யா” may be written with leading “இ”. During transliteration it must be omitted

- The Sanskrit ஸ்ரீ will be always mapped to ශ්‍රී. However attempts may be done to write the ஸ்ரீ in pure Tamil, and it may be சிறீ or சிரீ. These two forms must be transliterated as ශீරீ, and not ශීරී.

The table 4 shows an example action would take place in level 4.

Word	Level 2	Level 3	Level 4
இராமலிங்கம்	ஓராமலிங்கம்	ஓராமலிங்கம்	ராமலிங்கம்

Table 4: An Example from Level 2 to Level 4

4.5. Level 5 – Contemporary Non-Compulsory Rules

Every language does have its esthetical characteristics. A good example to describe this would be Tamil ங transliterated into ඩ, but the consonant ங is transliterated into anusvara “◌” and not consonant ன. Writing ன for ங is seen like out of the ordinary in the Sinhala language. See figure 2.

அங்கனம் (අංකනම) එසේ, එතුන

Figure 2: Tamil Sinhala Dictionary – (Participatory Development Forum) – Page 3
Such behaviors identified and specified as

level 5; there might be more rules exits. However these rules are categorized as not compulsory rules since it might not use in every situations. Some rules are listed below.

- If the form ந்திர appears in Tamil, it may be written as නේර form in Sinhala. In Tamil ந்திர is not a conjunct form therefore in Sinhala theoretically it will not transliterate into conjunct form. Since conjunct form exits in Sinhala it may be written in the conjunct form.

Tamil ந்த = Sinhala නේ

Tamil ந்த = Sinhala නේ (Conjunct form)

- Diphthong ஐ (ai) may get decomposed to අයි in Sinhala.

- If the ய appear in Tamil, it may be written as ය in Sinhala.

The table 5 shows an example action would take place in level 5.

Word	Level 2	Level 3	Level 4	Level 5
இரா மச்சந் திரன்	ஓராமலி வநீர ந்	ஓராமலி வநீர ந்	ராமலி நீர ந்	ராமலி நீர ந்

Table 4: An Example from Level 2 to Level 4

5.0 APPLICATION

Originally JSP, Servlet based Java application is developed to test the framework. It is been deployed in Tomcat server and developed using Unicode version 5.0. After successfully passing the test stage it been taken by Faculty of Information Technology of University of Moratuwa, undergraduate research students (2008) and converted into Web Service program.

6.0 CONCLUSION

We have tested the procedure on a sample of several thousand personal names and place names and find that it works very well on names of Tamil origin. Words of Arabic and Sanskrit origin, however, do not follow the rules exactly, and need to be handled by Level-1 rules.

We recommend that this procedure be adopted as the standard method of transliteration from Tamil to Sinhala.

7.0 ACKNOWLEDGEMENTS

Gratefully acknowledged the guidance and valuable comments provided by Professor Gihan V. Dias (University of Moratuwa), Dr. Sanath Jayasena (University of Moratuwa), Mr. Anura Tissera, former CIO, Associated Newspapers of Ceylon (ANCL), and Dr. Chathura De Silva, (University of Moratuwa) during the research reviewing.

Appreciations goes to the Faculty of Information Technology of University of Moratuwa, students (2008) N.A.C.Napagoda, A.N.N.Bandula, D.B.G.N.Bandara, N.W.E.G.E.Nanayakkara, V.T.Darmasiri and M.D.S Prasadini who has converted Java codes into Web Service program.

8.0 REFERENCES

[1] ISO 15919, "Information and documentation – Transliteration of Devanagari and related Indic scripts into Latin characters" *ISO*, 1st ed.

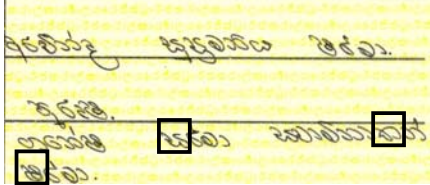
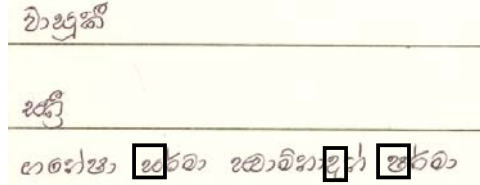
[2] Daniels P.T. and Bright W., *The world's writing systems*, 1st ed, New York: Oxford University Press, p.4 1996.

[3] Fernando, H. C., kodikara, N. D., Hewawitharana, S., 2003. A database for handwriting recognition research in Sinhala language, In: Seventh International Conference on Document Analysis and Recognition, August 03-06 2003 Edinburgh Scotland. 1262-1264.

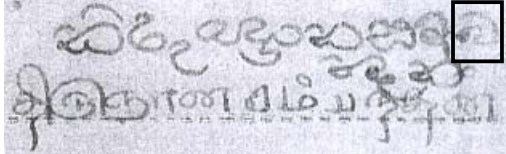
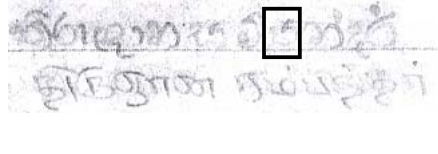
[4] University Of Madras, 1982, 'Tamil Lexicon Vol. 5', Reprinted, University Of Madras, India, pp 2633.

APPENDIX

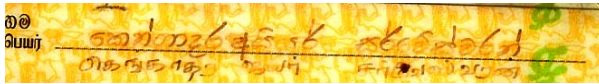
Example 1: Birth Certificates of 2 Children of a father (Sri Lanka)

1997 (Kandy)	2003 (Wellawatta)
	
<p>In Tamil, if த appears after ந or any letters of ந it will be always pronounced as Sinhala ජ not ஞ. In one child's birth certificate it is written with ஞ and ජ as another child's. In addition the same word சர்மா is written, as සර්මා and ඡර්මා.</p>	

Example 2: Sri Lanka National Identity Card

	
<p>In Tamil, if ப appears after ஞ it will be always pronounced as Sinhala ඞ not ඞ. In word திருஞானசம்பந்தர் it is written with ஞ and ජ with ඞ another occasion.</p>	

Example 3: Sri Lanka National Identity Card

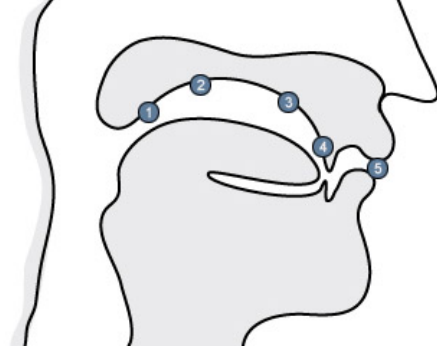


கெங்காதர written as කෙත^ඞදර

Vowels			Consonants							
Tamil	Latin	Sinhala			Velar/ Guttural	Palatal	Retroflex/ Lingual	Dental	Labial	
அ	a	අ	Plosives	Voiceless	Unaspirated	ක් ක්	ඡ් ඡ්	ඳ් ඳ්	ඤ් ඤ්	ඞ් ඞ්
					Aspirated					
Voiced	Unaspirated			ඣ් ඣ්						
	Aspirated									
				Nasals	ඟ් ඟ්	ච් ච්	ඡ් ඡ්	ඣ් ඣ්	ඞ් ඞ්	
ஊ	ū	ඌ	Fricatives	Voiceless		ඡ් ඡ්	ඣ් ඣ්	ඤ් ඤ්	ඞ් ඞ්	
				Voiced						
				Flapped & tapped sounds				ඣ් ඣ්		
Aspirate, semi-vowels and liquid					ඟ් ඟ්	ච් ච්	ඡ් ඡ්	ඣ් ඣ්	ඞ් ඞ්	
ஊ	ai	ඌ								
ஊ	o	ඌ								
ஊ	ō	ඌ								
ஊ	au	ඌ								

Table 3: Vowels and Consonants one-to-one mapping

Guttural	क व ग घ ङ k kh g gh ñ क ङ
Palatal	च छ ज झ ञ c ch j jh ñ च ञ
Cerebral / Lingual	ट ठ ड ढ ण ṭ ṭh ḍ ḍh ṇ ट ण
Dental	त थ द ध न t th d dh n त न
Labial	प फ ब भ म p ph b bh m प म



1. Guttural - Using the back of the tongue against the soft palate.
2. Palatal - Using the flat of the tongue against the back of the hard palate.
3. Cerebral - Using the tip of the tongue against the top of the hard palate.
4. Dental - Using the tip of the tongue against the top front teeth.
5. Labial - Using the lips.

Figure 1: Sanskrit Consonants (via Grantha script)