Sandhi Decoder for Modern Tamil texts

Dr.K.Umaraj
Assistant Professor, Department of Linguistics
Madurai Kamaraj University, Madurai
umarajk@gmail.com, www.umarajk.in , + 91 9487223316

The Morphophonemic rule or the Sandhi rule is a key factor for designing an automatic sandhi decoder for Tamil. The traditional Grammarians definitions are not sufficient enough to solve the problem of decoding sandhi where as even the definition of modern linguist's are not also serving the requirement as they consider the natural grammar but for computerization we need format grammar.

For example if we consider the situation of the plural suffix / kal /- and the accusative case or the second case marking suffix -ai results different outputs.

i) / maram / + / kaḷ / - maraṅkaḷ 'tree+plural '

ii) / maram / + / - ai / - marattai 'tree+ second case'

In both cases same root word derives differential output. The linguists classify this with the following rule.

-m --> ṅ

-m → tt

This is also not sufficient enough to decode the sandhi . Apart from the phonological information, the grammatical information also required for decoding the sandhi . So a new design was formulated for sandhi decoding.

Design:

1) Text without sandhi marking
2) Checking for the words in the Lexicon
3) Checking for POS information in the Lexicon
4) Checking for the Phonetic and phonological information of the proceeding and following words in the Lexicon
5) Apply sandhi rules
6)Text with sandhi marking.

Uses:

Automatic sandhi encoder and decoder will be useful for the correcting the typing errors due to misplacement of sandhi.

Sandhi Rules:

The following are the sandhi rules for developing a sandhi encoder(generator) and decoder(parser)

1)In casal construction, If the proceeding accusative or dative noun ends with a vowel and the following past tense verb starts with an voiceless plosive then plosive of the following verb is doubled.

rāman-ai + ka ṇ ṭān --→ rāmanik ka ṇ ṭān ‘ He saw Raman’

ramanukku + koṭuttān -- > ramanukkukkoṭuttān ‘He gave to Raman ’

2) If the proceeding negative relative participle of ceyyā type ends with an vowel and the following noun starts with a voiceless plosive then the plosive of the following noun will be doubled.
Vst –ā + p -(Vst-ā + PT)

kā ṇa+ kātci -- kāṇakkāṭci ‘the scene that which is not seen ’

ōṭa + ku l̲antai - ōṭakku l̲antai ‘ the child that which do not Run’

3) If the proceeding ceyya verbal participal ends with a vowel and the following past tense verb starts with a voiceless plosive then plosive of the following verb will be doubled.

Vst + a + PP	Vst +a + PP
Vara + ka ṇ ṭān	varakka ṇ ṭān ‘ He saw his arrival’
Ceyya + co n̲ n̲ār	ceyyacco n̲ n̲ār ‘He said to do ’

4)If the proceeding ceytu verbal participles or partciples ends with ‘i’ , ‘y’, ‘ppu’ and the following past tense verb starts with a voiceless plosive then the plosive of the following past tense verb will be doubled.
virumpi + pārttān - virumpippārttān ‘having liked he saw ’

p ōy + pārttān - p ōy ppārttān ‘ having gone and saw’

paṭi-ttu + k ūri ṉāṉ -(paṭi-ttuk k ūri ṉāṉ ‘he read and said ’

5) If the proceeding demonstrative and interrogative noun base ends with a vowel and the following noun starts with a voiceless plosive then the plosive of the following noun will be doubled.

a + kutirai -- akkutirai ‘that horse’

i + pa ṭam - ippa ṭ ṭam ‘this picture ’

e+talaivar - ettalaivar ‘which leader ’

6) If the proceeding manneral or time adverb ends with a vowel and following past tense verb starts with a voiceless plosive, then the plosive of the following past tense verb will be doubled.

appa ṭ i + ceytān - appa ṭ ic ceytān ‘He did in that manner ’

ippa ṭ i + ceytān ippa ṭ ic ceytān ‘He did in this manner ’

eppa ṭ i + kantān eppa ṭ ik kan ṭ ān ‘How did he see ’

a ṅ ku + cenrān a ṅ kuc cenrān ‘How went there ’

i ṅ ku + pārttān i ṅ ku pāettān ‘He saw here ’

e ṅ ku + kan ṭ ān e ṅ ku kan ṭ ān ‘Where did he see ’

Compound nouns: (All compound nouns except unmaittokai and vinayttokai have sandhi compulsory)

7) If the proceeding adjective or noun ends with **long vowel or geminate plosive plus u** and the following noun starts with a **voiceless plosive** then **voiceless plosive** of following noun became geminated.

Mā + pa ṭ ṭ a → māppa ṭ ṭ a 'an animal fell into ' (kuru:171:3)

Pulattu + pu → pulattu-ppu ' bloom in the back yard ' (Kuru:323:4)

t ī + ku ṇ am (t īkku ṇam ' bad habits'

Exception:

The following word is uyartinai then there is no doubling of plosives

Examples: pulavar ka ṇ ṇ ir 'Poet's war'

8) If the proceeding noun ends with **voiced retroflex nasal ṇ** and the following noun starts with a **voiced dental nasal n** then voiced dental nasal of following noun will be converted into **voiced retroflex nasal ṇ**

ka ṇ + nir -→ ka ṇ- ṇ ī r 'tears' (kuru:4:2)

ma ṇ + nilai -→ ma ṇ ṇilai ' ' (Kuru:1114)

9) If the proceeding noun ends with a kurriyalukaram(u) and the following noun is a semivowel (y,v) then u of the proceeding noun becomes i.

Maruppu + yānay > maruppiyānay 'tusked elephant' (kuru: 215:4)

10) If the proceeding noun ends with a kurriyalukaram(u) and the following locative noun starts with a vowel then the kurriyalukarm u of the preceeding noun simply drops

Kokku +in → kokkin ' the crabs' (kuru:117:1)

11) If the proceeding noun ends with a kurriyalukaram(u) and the following noun starts with a vowel, then u of the proceeding noun simply drops and the terminal consonant geminates.

ē l̲utu + ellām - ē l̲uttellām

12)If the proceeding noun ends with voiced bilabial nasal m and the following noun starts with a voiced dental nasal n, then voiced bilabial nasal m converts into voiced dental nasal n

Cem + n ā	cen-n ā	‘reddish tongue’ (kuru:14:1)
y ām + nā ṇukam	yān-nāṇukam	‘we (royal) shall blush’(kuru:14:6)

13) If the proceeding noun ends with voiced bilabial nasal m and the following noun starts a voiced nasal ñ , then voiced bilabial nasal m of the proceeding noun converts into voiced nasal ñ and the dental n of the following noun was dropped

y āyum + nāyum → yāyu- ñ āyum ‘my mother and your mother ’ (kuru:40:1)

atavam + ttu -(atavan-ttu “the fig tree s” (kuru: 24:3)

14) In non casal construction, If the proceeding noun ends with voiced bilabial nasal m and the following noun starts with a voiceless velar plosive k then bilabial nasal m of proceeding noun converts into voiced velar nasal n

Cem + k ō ṭ u → ‘ce ṅ -k ō ṭ u ‘reddened tusk’ (kuru:1:2)

Maram + ceti → ‘ maranceti ‘tree and creeping ’

Pa ṇ am to ṭ u → ‘ pa ṇ anto ṭ u ‘ear rings made out of palmyra leaves ’

15) If the proceeding noun ends with voiced post-**dental lateral l̲** and the following noun starts with a **voiced nasal m,** then **voiced post-dental lateral l̲** of proceeding noun converts **into voiced post-dental nasal n̲**

Ci **l̲** + mo**l̲**i → ci **n̲** -mo**l̲**i ‘few words’ (kuru:14:2)

Na l̲ + nāṭu → na n̲ -nāṭu ‘good country ’ (kuru:11:6)

Ko l̲ + ntu → ko n̲- n ṭu ‘having killed ’ (kuru:1:1)

Na l̲ + nirai na n̲ n irai ‘ good ’ (Thiru: 1111)

16) If the proceeding noun ends with voiced retroflex lateral ḷ and the following noun starts with a voiced nasal m, then voiced retroflex lateral ḷ of the proceeding noun converts into voiced Retroflex nasal n

Ko ḷ + mār ko ṇmār ‘to have ’ (kuru : 16:2)

Cāra ḷ + nā ṭu cāra ḷ nāṭu ‘a region with mountain slopes ’(kuru: 18:2)

Iru ḷ + nā ṭanāl Iru ḷ nā ṭa nāl ‘dark midnight’ (Kuru: 141:7)

e ḷ + ney eṇṇey ‘oil’

17) If the proceeding noun ends with voiced retroflex lateral ḷ and the following noun starts with a voiceless plosive p,t,c,k then voiced retroflex lateral ḷ of the proceeding noun converts into voiced retroflex plosive ṭ . Geminated plosive becomes simple plosives when it is

preceded by a nasal or plosive. This is to avoid three consonant cluster of this type.

U ḷ + kkay u ṭ kay ‘palm’ (kuru:60:3)

Po ṭ iyil +t t ō ṉṟu potiyit- tō ṉṟu ‘be seen in the common place’ (kuru:15:2)

Na ḷ +cce ṭ i na ṟcce ṭ i ‘Good news’ (Thiru)

u ḷ + k ō ṭ ṭ a m u ṭ k ō ṭ ṭ a m (Thiru:119)

18) If the proceeding noun ends with voiced retroflex lateral ḷ or voiced alveolar l and the following noun starts with a retroflex or alveolar plosive t then a lateral of alveolar or retroflex variety of proceeding noun becomes a larygnal sound (ie, the āytam in Tamil) when it is followed by the alveolar or retroflex plosive respectively.

Ka l̲ al + t ō ṭ i -(kala(aytam marker)-t ō ṭ i (kuru: 1:3)