Tamil Hyphenator

P. David Prabhakar
Department of Tamil,
Madras Christian College, Chennai-600 059, India
E-mail:davidprabhakar@rediffmail.com

Abstract

Hyphenation is an integral part of text processing for any language. Hyphenation splits word into sub-strings and it is used to breakup words at the end of a text line.

The Hyphenation principles differ from language to language. Some of these differences are related to differences in the nature of syllables. Other differences raise the unifying nature of compound words. This paper aims to analyze the existing principles of manual hyphenation in Tamil and tries to suggest a model for developing a 'Tamil Hyphenator' can be used by word processors.

Since modern Tamil prose has more than 400 years of history, many of the existing hyphenation procedures are based on poetry, which has *acai* and *ciir* as sub-strings in a line (*aTi*). It is believed that the hyphenation should be meaningful and parallel to the pronunciation.

As Tamil is inflectionally rich, using a hyphenated dictionary is not a viable solution. Building a Tamil hyphenator is the ideal solution for Tamil. In this paper, an attempt is made to formulate hyphenation rules based on the ***acai*** patterns (*neer* and *niRai*) and it also recommends further morphological processing to incorporate the concepts like sandhi rules and structure of compound forms.

Introduction

Hyphenation is an integral part of text processing for any language. The hyphenation rule declares allowed hyphenation points, separated by spaces in which each hyphenation point is indicated by a '-' character and it is used to break up words at the end of text line.

The Greek word *huphen* means literally `under one'. It is derived from *hupo* `under' and *hen,* the neuter accusative case of heis `a mark'. Hyphenation principles differ from language to language. Some of these differences are related to differences in the nature of syllables other differences raise the unifying nature of compound words.

This paper tries to exemplify the existing principles of manual hyphenation in Tamil and also suggests the methods and rules for developing a Tamil hyphenator, which can be used in Word Processors. The hyphen exists to help fulfill the true purpose of the printed pages, its readability. In composing a column of justified type, it is sometimes necessary to divide a word so as to align the right-hand margin without allowing either too much or too little white space between the words. The hyphenation suggests the hyphen positions, so that it can be used with other applications which adjust the inter word spacing.

Objectives

The aim of this paper is two fold:

1. To understand the existing hyphenation
2. To suggest methods for Tamil hyphenation.

Conventional hyphenation principles in Tamil

Tamil was written continuously without word spaces, punctuation or word-divisions in the first books printed (16thCent. A.D) in Tamil. Similarly, we cannot find even word boundaries in the inscriptions, copper plates, palm leafs, bond-paper, and dairies in Tamil. It is also interesting to note the early writings of monks (5th Cent. A.D) who wrote their Latin without any space between words, and without a formal procedure for word division.

Later, we find inter, intra word spacing depending upon the intuition of printer or typist. We can trace at least two principles found in the Tamil hyphenation practice. They are as follows:

1. Single characters are avoided at the end of text line and at the initial position of text lines. (e.g. *po, ka*, etc.)

2. Dotted consonants are avoided in the initial position of text lines. . (e.g. *p, k, t*, etc.,)

One can assume that, these two principles exist due to the influence of Tamil prosody. Because,

1. Monosyllabic words do not exist in Tamil prosody.
2. Single character has *acai* status only in a very few occurrences (e.g. when CV pattern occurs as a last syllable in a *ciir*.
3. Dotted consonants (consonants without vowel) did not occur in the initial position of a word. And it has no value in deciding the *acai* type.

Along with the existing hyphenation principles, we need to elaborate and formulate acceptable hyphenation principles for computational implementation.

Basis of Tamil hyphenation

There are dictionaries like 'Oxford dictionary of spelling' (1986), which also show word-breaks. 'Chicago manual of style' contains a huge chart listing various sorts of phrases that are or are not to be hyphenated. We have no such dictionaries in Tamil. 'Tamil Style Manual' (2001) published by Mozhi Trust, 'A Handbook for Journalists' published by Si. Paa. Adhithanar, are notable style manuals in Tamil. There is no suggestion on hyphenation in the above mentioned manuals.

Lexicographer's approach to hyphenation is not suitable for Tamil. It is possible to have a dictionary with large lists of words, with every approved hyphen-point marked can be stored on a hard disk. But such dictionary requires more memory capacity than most of the DTP publishers require for a total system.

As Tamil is inflectionally a rich language, a single verb stem in Tamil can be inflected for more than 2000 different forms by affixing various auxiliaries, tense markers, PNG markers, clitics etc. Similarly, noun forms can be generated by affixing plural marker, *caariyai*, case markers, affixes, emphatic markers, clitics etc. Both the verb and noun form constructions can be split meaningfully so as to meet the needs of hyphenation. Prabhakar (2001) has developed an algorithm to split and label constituents of verb forms. A pattern for the description of noun inflections is found in Kothandaraman (1981). These two references can be used for building the hyphenator whenever a morphological analysis is needed.

Hence it is true that, Dictionary based solution is not possible for Tamil hyphenation. Unfortunately, no such dictionaries exist for total solution to any language. And, hyphenation rules are just guidelines. We may think of it as only pertaining to grammar, but hyphenation operates on an aesthetic level, too. An appropriate method can be suggested for Tamil hyphenation. It has two levels. They are:

1. Acai processing
2. Morphological processing.

Before we go into the details of the above, let us have a look at a few observations regarding hyphenation.

One can find variations in the hyphen placement within a language. Lexicographer may differ from typographer's approach. American practice may differ from British practice. US usage is to break words according to pronunciation while UK usage is to break them by etymology. British practice tends to favour morphological while, Americans favour syllabic breaks.

According to Prof. Walter W. Skeat, editor of Oxford's etymological dictionary, 'nothing is gained by pretending to keep the root intact, when the spoken utterance does nothing of the kind'. His advice was to divide all words by sound and pronunciation, but not to expect to establish rigid and invariable rules (as quoted in McIntosh & Fawthrop, 2000). Based on these observations, we suggest a syllable based hyphenation for Tamil. If needed, morphological word divisions can be taken into account.

A sample of 1500 words (continuous text from a Tamil daily) was taken for this study. Based on the *acai* patterns, hyphenation positions were marked. It shows that, 93-94% of words' splits are parallel to the manual hyphenation. Further processing with a sub-set of rules handles another 2-3% of words properly. It is felt that, morphological processing is necessary for the remaining words.

Rules for Tamil hyphenation

1. Avoid divisions wherever possible
2. Single characters at the beginning or end of word do not improve readability of the text, hence should be avoided
3. Dotted consonants at the beginning of any split should be avoided as in conventional practice. A word which has 5 characters with 3 consecutive consonants cannot be split.
4. Hyphenation positions have to be marked based on *acai* patterns .When single character gets *acai* status at the initial position of a word, the succeeding *acai* should be added to the preceding *acai.*
5. Morphological process can be admitted for further processing when sub-set of rules fail to overcome the difficulties in the *acai* based splits.

Acai processing

Acai processing is a kind of pattern matching. It is to recognize all possible character sequences which occur within a word. *acai* in Tamil prosody is different from the syllable. Kothandaraman (1976:58) says that, a*cai* is an intermediary node between a foot and a syllable. This paper suggests that, Tamil hyphenation can be handled on the basis of *acai* patterns found in the Tamil prosody. In Tamil, there are two types of acai.

1. *neer acai*
2. *niRai acai*

Both the forms of *acai* can be predicted on the basis of patterns. *neer* and *niRai* patterns are given below. Note that *niRai* pattern has two vowels while *neer* pattern has one.

1. *niRai* patterns

   (C) VCV: (C)
   (C) VCV (C)

2. *neer* patterns
   (C) V: (C)
   (C) V (C)

Sub rules to *acai* processing

The following sub rules can be deployed simultaneously.

1. A character (composed of consonants and a short U) can be added to preceding *acai.*
   e.g. veeNTu/menRu

These are accepted patterns in Tamil prosody as *neerpu* and *niRaipu.*

2. Semi vowels *y* or *v* can be added to following *acai*
   e.g. maaRi/viTu/kin/Rana

Morphological processing

Morphological processing is necessary in a very few cases, where certain affixes are added to noun or verb. These affixes can be short listed split can be made accordingly

To conclude, in many respects it seems more convincing to regard the *acai* pattern as basis for Tamil hyphenation.

Bibliography

Prabhakar, David. 2001. **Analysis of Tamil verbs: Computational approach.** Unpublished PhD thesis, Chennai: University of Madras.

-----. 2002. Hyphenation in Tamil, **4th International Conference on south Asian Languages**, Annamalai Nagar: Annamalai University.

Kothandaraman, Pon. 1976. **Modern Studies in Tamil.** Chennai: Tamil Nulagam.

-----. 1981.**IlakkaNa ulakil putiya paarvai (Vol.2).** Chennai: Tamil Nulagam**.**

McIntosh, Ronald and Fawthrop, David. 2000. **A discussion of the changing principles of word division, now implemented by computers, in British and American English, with notes on hyphenation of 39 other languages**, Computer Hyphenation Ltd, Halifax, UK, (Electronic version).