Morphological Generator for Tamil

P. Anandan, Dr. Ranjani Parthasarathy & Dr. T.V. Geetha

Resource Centre for Indian Language Technology Solution - Tamil,
School of Computer Science and Engineering,
Anna University, Chennai - 25
Email: ranjani-p@eth.net, anandan_p78@yahoo.com

ABSTRACT

Tamil is a relatively free word order language, the only constraint being that normally the verb comes at the end. This flexibility in word ordering is possible due to the morphologically rich nature of the language. Information such as case rules and auxiliary verbs indicating aspect, tense, mood are all conveyed through morphological attachments to the root noun or verb. This makes morphological generation of Tamil words a challenging task. Another issue is that unlike English where only limited number matching is necessary, in Tamil as in many other Indian languages person and gender matching between subject and verb is also necessary. A morphological generator designed for Tamil needs to tackle the different syntactic categories such as nouns, verbs, postpositions, adjectives, adverbs etc. separately, since the addition of morphological constituents to each of these syntactic categories depends on different types of information.

In this work a morphological generator has been designed for each of the syntactic categories and then combined to morphologically generate a complete sentence. The underlying morphological structure for a Tamil noun is as follows:

Noun stem
Or + [plural suffix] + [the euphonic suffix] + [the case suffix]
Oblique stem

While generating the noun derivatives from the roots, linguistic rules determining the form of a plural suffix has to be considered. The attachment of case suffixes to nouns is an important part of the morphological generator for nouns. In addition, this has to take into consideration the fact that certain nouns can take case suffixes only in oblique form. The euphonic suffix sometimes comes along with oblique suffixes or with plural suffixes. This has also to be considered. During the combination of the root noun with the above mentioned suffixes, "sandhi rules" have to be taken in to account.

Some nouns in Tamil take case only when it is converted to an oblique stem form. Oblique is a meaningless word stem which can come between two morphemes. An example of the rule for this conversion are root "maram" when converted to oblique becomes

"marattu". Noun stems ending with -"m" when converted to oblique would have "m' replaced be "ththu". Similarly root "vidu" when converted to oblique becomes "vitt". Hence separate rules to check endings of noun and converting them to oblique form if necessary is an important part of the morphological generator of nouns.

The Morphological structure of Tamil verb is quite complex since it caters to person, gender, and number markings and also combines with auxiliaries that indicate aspect, mood, causation, attitude etc. While morphologically generating the verb, the gender, number and person of the subject is necessary in order to select the appropriate suffix catering to the selected tense.

Verbs in Tamil are classified into more than 18 categories. They are categorized based on their tense markers. The past tense marker is used mainly to categorize the verbs. One of these categories is "kir-th-v", in which the "kir" indicates the present tense marker, "th" indicates the past and "v" indicates the future tenses. Considering "kir-th-v" pattern. For example take the verb செய for 3rd person singular it will be "cey-kir-aan" in present, "cey-th-aan" in past and "cey-v-aan" in future. Similarly for "kir-tt-v" take the verb "sappidu" it will take the form "sappidu-kir-aan" in present, "sappi-tt-aan" in past and sappidu-v-aan in future. In some cases past tense markers are same however future tense markers alone differ. Example let ஆnada", which comes the pattern "kir-nth-pp", "vizhu", which comes in the pattern "kir-nth-v" which has modification in their future tense markers. "nada" will take "nada-pp-aan" in future whereas "vizhu" will take "vizhu-v-aan" even though the past tense markers are same.

Apart from this, there is one more category with same tense markers, which vary in the root form changing. For example "vizhu" and "vaa" have the same tense markers "kir-nth-v". Here however their the root verb gets modified before taking the tense markers. "vizhu" will never change its root form while appending the tense markers to it. But "vaa" will change its root form when tense markers are appended to it. It can be shown by looking into the words "vizhu-kir-aan" "vizhu-nth-aan" "vizhu-v-aan" and "varu-kir-aan", "va-nth-aan", "varu-v-aan" in present, past and future tenses respectively.

The linguistic rules determining the combination of auxiliaries with the verb is also quite challenging since more than one auxiliary can attach itself to the verb as suffixes. The combination becomes exponentially large, hence we have used semantic based heuristics to prune the combinations. While using auxiliaries it is the last auxiliary that matches with the person, number, gender of the subject and not the root verb of the combination. The auxiliaries are categorized in more than six ways with respect to their occurrences of a verb or axillaries before them. The fact that the previous main verb/auxiliary (here "cey") will take a verbal participle/infinitive determined by the current auxiliary (here "koll"). The rules for changing root to verbal participle/infinitive determined by the word itself. Hence here "cey" the root changes to "cey-thu" when converted to verbal participle. If the verb "cey" and the auxillary verb "padu" come together then the previous comes in infinitive form. Infinitive is formed by word stem which can be "a" for weak verbs, "ka" for medium verbs and "kka" for strong verbs. These can be demonstrated with the words "cey- th-u-koll"

and "ati-kka-p-patu", "cey" root verb, "th-u" verbal participle and "koll" is the auxiliary. For the second "ati" is the , "kka" is the infinitive for the verb, "p" is sandhi and "patu" is the auxiliary.

Similar linguistic rules are used to generate derivatives for other syntactic categories also. Thus a Tamil morphological generator for different syntactic categories has been designed and implemented. This morphological generator can be used for suggestion list generation of a spell checker and as a basis for English to Tamil translation.