AMRITA MORPH ANALYZER AND GENERATOR FOR TAMIL: A RULE BASED APPROACH

Dr. A.G. Menon, Amrita University and Leiden (Netherland),
S. Saravanan, R. Loganathan and Dr. K. Soman, Amrita University, Coimbatore, India.

menon.govindankutty@gmail.com sarwanster@gmail.com loganathn@gmail.com
kp_soman@amrita.edu

THE CONTEXT

From 2006 CEN (Centre of Excellence for Engineering and Networking) of the AMRITA University, Ettimadai, Coimbatore under the guidance of Prof. K. Soman, is engaged in research and development in the field of Natural Language Processing (NLP). It is a young and dynamic university. AMRITA is part of a consortium of six IITs and two IIITs and CDAC, Pune, which are involved in the research and development of tools for the translation of English to Indian Languages, which is funded by DIT. A new project on Machine Translation is started recently (May 2009) with the funding of the MHRD for developing linguistic resources and machine translation tools. AMRITA is also developing its own engine for Machine Translation. The present Amrita Morph Analyzer and Generator (AMAG) for Tamil is an independent project carried out in CEN.

THE NEED

More than a dozen Tamil Morphological Analyzers and Generators are announced through the Internet and websites of many renowned institutions. The only DEMO version available is ATCHARAM displayed on the website of the IT Ministry, Resource Centre for Indian Language Technological Solutions – Tamil. However, none is available for our research and development from the open source. This deplorable situation has compelled us to build our own MAG for developing a system for the MT and other NLP applications.

MORPHOLOGY FOR COMPUTER

Morphology deals, primarily, with the structure of words. Morphological analysis detects, identifies and describes the meaningful constituent morphs in a word, which function as building blocks of a word. The densely agglutinative Dravidian languages such as Tamil, Malayalam, Telugu and Kannada display a unique structural formation of words by the

addition of suffixes representing various senses or grammatical categories, after the roots or stems. The senses such as person, number, gender and case are linked to a Noun stem in an orderly formation. Verbal categories such as transitive, causative, tense and person, number and gender are added to a verbal root or stem. The morphs representing these categories have their own slots behind the roots or stems. The highly complicated nominal and verbal morphology do not stand alone. It regulates the direct syntactic agreement between the subject and the predicate. Another important aspect of the addition of morphs is the change which often takes place in the space between these morphs and within a stem. A Morphological Analyzer and Generator (MAG) should take care of these changes while assigning a suitable morph to the correct slot to generate a word. The combination of sense and form in a morph and the possibility to identify the governing rules are the incentives to attempt to build an engine which can automatically analyse and generate the same processes taking place in the brain of a native speaker.

CHALLENGES IN BUILDING A MORPH ANALYZER AND GENERATOR FOR TAMIL

The slots behind the root/stem can be filled by many morphs. The rules governing the order of the morphs in a word and the selection of the correct morph for the correct slot should be formulated for analysis and synthesis. The inflections and derivations are not the same for all the nouns and verbs. The biggest challenge is the grouping of nouns and verbs in such a way that the members of the same group have similar inflections and derivations. Otherwise one has to make rules for each noun and verb, which is not feasible. The most difficult slot in a verb is the one which follows the verb root/stem. This position is occupied by the suffixes belonging to the category transitive. The elusive behaviour of these suffixes poses many problems, and most of the earlier Morphological Analyzers did not handle this problem adequately. Our system, as mentioned earlier, works on rules and these rules are capable of solving this clumsiness in an elegant manner.

Many changes take place at the boundaries of morphs and words. Identifying the rules that govern these changes is a challenge because dissimilar changes take place in similar contexts. In such cases it is necessary to look into the phonological as well as morphological factors which induce such changes.

The system we design involves building an exhaustive lexicon for noun, verb and other categories. The performance is directly related to this exhaustiveness. It is a laborious task.

STRUCTURE OF A DRAVIDIAN VERB

<table>
<tr><td colspan="5">Structure of a Dravidian Verb</td></tr>
<tr><td>1</td><td>2</td><td>3</td><td>4</td><td>5</td></tr>
<tr><td rowspan="3">Root/Stem</td><td rowspan="2">Intransitive</td><td>Personal object base</td><td rowspan="2">tense/mode</td><td rowspan="3">Personal endings (person, number and gender)</td></tr>
<tr><td>plural action base</td></tr>
<tr><td>Transitive</td><td>motion base</td><td>negative</td></tr>
</table>

The third position is not relevant for Tamil verbs.

The structure of a Tamil verb is given below:

The finite verb: Root/Stem + Transitive + Causative + Tense / Negative + Empty + PNG

Clitics can be added after the Person Number and Gender (PNG) marker.

Non -Finite Verb:

Root/Stem + Transitive + Causative + Negative + Infinitive / Conditional infinitive suffix

Root/Stem + Transitive + Causative + Tense / Negative + Relative Participle / Verbal Participle / Conditional Verbal Participle

Root/Stem + Transitive + Causative + Negative + Verbal Noun suffix

The above descriptions mention only the slots on the right side of the root/stem. Apart from this, many non-finite verbs occur on the left side of the root/stem and form complex verb structures such as main + auxiliary verbs.

PREDICTABILITY OF THE VERB SUFFIXES

One of the challenges is the predictability of the suffixes which fill the three slots after the verb root/stem: transitive, causative and tense. There are two types of verbs: verbs which have intransitive and transitive contrast such as *tā□ntā□* as against *tā□tti□ā□* and such as *cirittā□* without such contrast. We can divide the verbs broadly into three groups on the basis of the past tense suffixes *–nt-* , *-i□-* and *–t-*. They can be further divided into eight groups taking into account the first three positions after the root/stem. The fillers of each position are

determined by the verb root/stem. As far as the non-finite forms are concerned, the predictability of the verbal noun suffixes is an important task of MAG.

STRUCTURE OF A NOUN:

Stem

Stem + Formative / Oblique suffix

Stem + Formative / Oblique suffix + Case marker

Stem + Formative / Oblique suffix + Empty suffix + Case marker

Stem + Formative suffix + Plural + Case marker

Stem + Formative /Oblique suffix + Pronominal suffix

The closing slot can be followed by a clitic such as *–um* or interrogative morph such as *–ā*, *-ē* or *-ō*.

HANDLING THE NOUN SUFFIXES

The suffixes occupying the slots after the stem do not vary according to the stem. There is no direct relationship between them unlike the verb stems. However, the noun stems themselves vary before they take suffixes. This phenomenon is limited to a small number of groups of nouns. We have mentioned above that some of the stems take an oblique suffix before the addition of case markers. Since they are identifiable on the basis of their endings or specific phonological features of the stems, it is also easier to make rules for the changes which take place within the stem. For example, forms like *maram* 'tree' become *maratt*- and *nā□u* becomes *nā□□*-. The first and second person pronouns have also two different forms along with the third person neuter plural pronoun. However, when nouns like *kal* are preceded by another noun, problems arise in handling the *sandhi* rules because these rules are based on the phonemic makeup of the final noun alone. Creating rules governing the distribution of case suffixes is an important step towards the building of a MAG. The noun morphology is relatively less complex than the verb morphology.

TECHNOLOGY

Finite State Transducer (FST) is used for morphological analyzer and generator. We have used AT &T Finite State Machine to build this tool. FST maps between two sets of symbols.

It can be used as a transducer that accepts the input string if it is in the language and generates another string on its output. The system is based on lexicon and orthographic rules from a two level morphological system. For the Morphological generator, if the string which has the root word and its morphemic information is accepted by the automaton, then it generates the corresponding root word and morpheme units in the first level (Fig 1). The output of the first level becomes the input of the second level where the orthographic rules are handled (Fig 3), and if it gets accepted then it generates the inflected word.

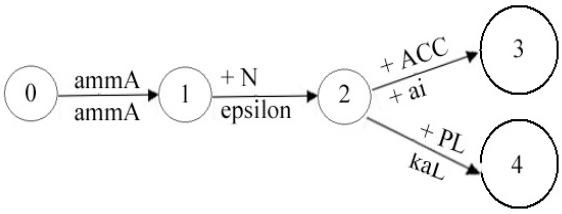

Fig 1: Morphotactics Rule

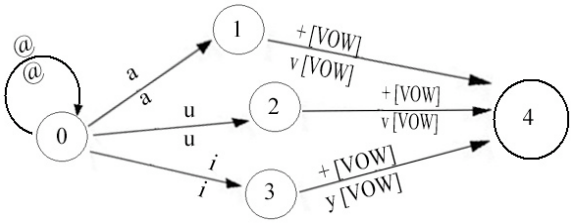

Fig 2: Sandhi Rule

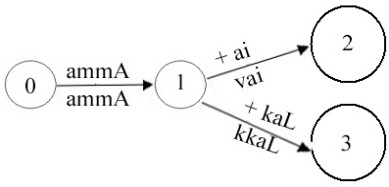

Fig 3: Application of Sandhi Rule

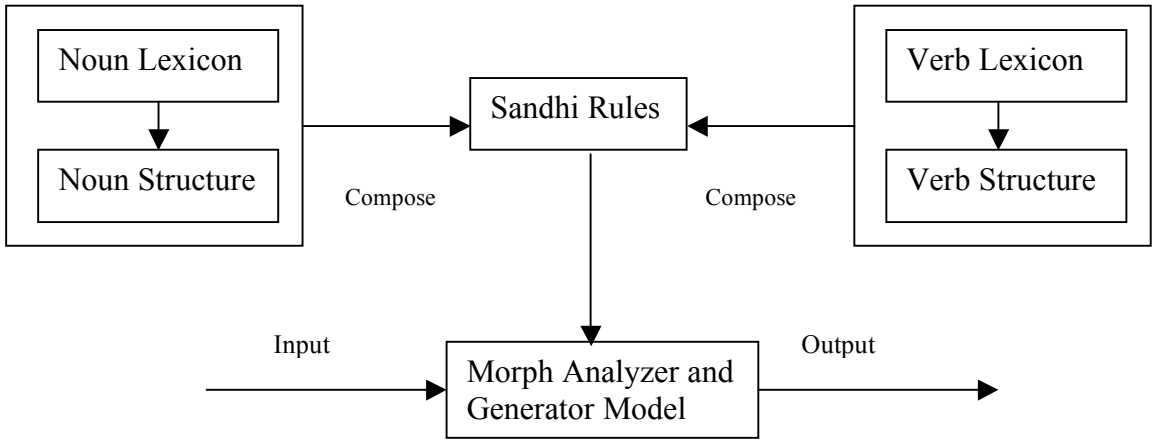

Fig 4: Model Morph Analyzer and Generator

CONCLUSION

At present we started with a list of fifty thousand nouns, around three thousand verbs and a relatively smaller list of adjectives. Our MAG is capable of analysing and generating more grammatical categories than ATCHARAM. In the future we are planning to expand our

lexicons for more exhaustiveness. The uniqueness of our MAG is its capacity to generate and analyse transitive, causative and tense forms apart from the passive constructions, auxiliaries and verbal nouns. A demo version of AMAG will be soon uploaded for testing.

BIBLIOGRAPHY

Anandan, P., Ranjani Parthasarathy & Geetha, T.V., 2001. "Morphological Generator for Tamil", Tamil Internet 2001 Conference, Kuala Lumpur, Malaysia.

Anandan, P., Ranjani Parthasarathy & Geetha, T.V., 2001. "Morphological Analyser for Tamil", *ICON 2002*, RCILTS-Tamil, Anna University, India.

Beesley, Kenneth R., 1996. "Arabic Finite-State Morphological Analysis and generation", *Proceedings of the 16th International Conference on Computational Linguistics*, Vol. 1.Copenhagen, Denmark. pp. 89-94.

Beesley, Kenneth R. & Karttunen, Lauri, 2003. *Finite State Morphology*, Stanford, CA: CSLI Publications.

Koskenniemi, Kimmo., 1984. "General Computational Model for Word-Form Recognition and Production", *COLING* 84. pp. 178-181.

Lakshmana Pandian, S and T.V. Geetha, 2008. "Morpheme based Language Model for Parts-of-Speech Tagging", *POLIBITS – Research Journal on Computer Science and Computer Engineering with applications,* Volume 38 (July-December 2008), Mexico. pp. 19-25.

Menon, A.G., 1976. "Tamil Verb Classification", *Actes du XXIXe Congres international des Orientalistes* 1973. Inde Ancienne, Vol. II.3. Etudes Dravidiennes. Paris: L'Asiatheque pp. 136-148.

Menon, A.G., 1988. "Tamil Verb Stem Formation", *International Journal of Dravidian Linguistics*, Vol. 27, part 1. Trivandrum: International Association of Dravidian Linguistics. pp. 13-40.

Menon, A.G. & Schokker, G.H., 1990. "Linguistic Convergence: the Tamil-Hindi auxiliaries", *Bulletin of the School of Oriental and African Studies*. London: University of London. pp. 266-282.

Renganathan, Vasu, 2001. "Development of Morphological Tagger for Tamil", *Tamil Internet 2001 Conference*, Kuala Lumpur, Malaysia (26-28 August 2001).