Tamil Hyper Grammar

[1]*Uma Maheshwar Rao G,* [2]*Christopher M,* [3]*Parameswari K*
Center for Applied Linguistics and Translation Studies
University of Hyderabad, Hyderabad – 500046.
[1]*guraohyd@gmail.com,* [2]*efthachris@gmail.com,* [3]*parameshkrishnaa@gmail.com,*

Abstract: Grammatical descriptions of human languages are the results of efforts in modeling of the design features and the internal organization of the structures and the mechanisms. Therefore, Linguistics is about language modeling, designing and studying their theoretical and practical implications. However the activity of grammatical descriptions itself is molded by the specific needs of aims and the goals such as Teaching and Learning a language, investigating the issues related to the evolutionary biology with regard to discovering the universals of human language and development, philosophical and functional aspects of language and Linguistic Computing. Here, we would like to discuss certain issues towards building a Hyper grammar for a given language.

1. Concept: A Hyper grammar is a non-linearly organized dynamic grammar based on the hypertext format. It is intended to simulate certain functions of a native speaker. It can be used both as learning and teaching tool besides as a reference grammar.

It is comprised of a number of non-linearly arranged texts each with a comprehensive note on various grammatical facts of Tamil, with hyper-links. It can be accessed and retrieved for various purposes involving language, to experience the effect of a native speaker and hearer of the language. Functionally it serves better than any of the existing printed grammars, which are simply flat and linear. In a way the existing printed grammars are non-communicative i.e. passive, hence, they are monologues and do not participate or reciprocate to pass judgments about the linguistic facts of the respective languages. A grammar in order to reciprocate should have some of the computationally implemented tools like a morphological generator, analyzer, chunker, parser, lexical accessor etc.

The Hyper grammar is intended to be a reciprocative grammar, as it involves some of the properties like the native speaker's ability to make judgments on the grammaticality of the linguistic facts. This single feature makes it distinct from the printed grammars. Hyper grammars are extremely useful from the point of learning, teaching and as reference material.

The design features are borrowed from the hypertext format but conceived as a computationally cognitive model. The contents are being developed from both the published and unpublished sources carefully selected and rewritten in the hypertext format.

2. The Contents:

The content of Tamil Hyper grammar has two main components, viz. 1. the description of grammar in hypertext format and 2. the applicational aspect of the Tamil Language as a language manager.

2.1. The Tamil Grammar:

The grammar part includes a number of comprehensive descriptive notes on certain linguistic facts of the Tamil Language. It is conceived in terms of a Computational Grammar. It deals with the Orthography, the design features of Tamil script, the orthographic syllable, the information on the frequency distribution of written syllables etc.

As part of the Tamil morphology, we have information on Tamil categories viz. nouns, adjectives, verbs, adverbs, numerals, pronouns etc. In each of these, there is information regarding the setting up of paradigm types and a list of paradigmatic forms under each category. One can access information regarding the most frequent hundred words, five thousand words and ten thousand words in terms of their frequencies, and communicative contribution to the coverage in Tamil Texts. As regards to the frequency of Tamil characters and syllables as they occur in the 3 million-word corpus, one can find the relevant information. One of the most important and crucial is the lexical component. A number of bilingual dictionaries like Tamil-Hindi, Tamil-Kannada, Tamil-Tamil, Tamil-Oriya, Tamil-Marathi, Tamil-English and English-Tamil – are included. Originally these dictionaries are conceived as bilingual and bi-directional dictionaries initially created using the most frequently occurring words ensuring the coverage.

2.2. The Tamil language manager:

This is the most crucial component of Tamil Hyper grammar. It involves the actual functions of the practical aspect of the grammar outlined above. As said earlier, the grammatical description is only a statement about the competence of a native speaker – about his/her language. In order to make it to simulate the grammar, it should involve a working generator, analyzer, parser and lexical accessor, etc. Currently the Tamil language manager includes a word form generator, a morphological analyzer and lexical accessor among others.

a. The Morphological Analyzer:

The word analyzer incorporated here is intended to analyze the Tamil words in terms of the lexical root/stem, its category, the paradigm type and the inflectional or derivational affixes attached to it. A morphological analyzer (Morph) engine essentially learns from a morphological lexical database of a particular language. The functional coverage and efficacy of the engine is greatly dependent on the structure and the organization of the database. The database of Tamil Morphological Analyzer comprises of inflectional database and the root dictionary. These data comprise purely linguistic information of the language, which are processed subsequently to enable using it in morphological analysis. It uses the Word and Paradigm Model of analysis.

The Organization of the Linguistic data for Morph:

(i) The paradigmatic-data

The term Paradigm refers to an exhaustive set of morpho-syntactically related word forms of a given lexeme. Based on the inflection, six distinct morphological categories are identified and the paradigms are created. They include the major and minor categories of words.

(a) The major word classes which are productive and open class categories (new members are added from time to time) can inflect with distinct but characteristic suffixes which explicit morpho-syntactic functions. The major word categories are listed as below,

1.Nouns
2.Verbs
3.Adjectives

(b) The distinct minor categories which are productive but considered as closed class categories (no new members are added) are listed below,

4. Pronouns
5. Numerals
6. Locative Nouns

The other class of words which are not fallen under the above categories are a list of idiosyncratic word forms. They cannot inflect for any functional categories. They come under functional categories of language with defective morphology. The following words are usually known as indeclinable and have no morphology to process.

1.Postpositions
2.Adverbs
3.Conjunctions
4.Interjections
5.Particles

The above words are listed as 'Avy' (avyayas) in the dictionary.

(ii) Root Dictionary

Root Dictionary is a vast collection of lexemes which contains words, their categorical information and their suitable paradigms. It includes a certain number of minimally distinct words in the semantic system of a language. This is typically called as lexicon without semantics.

Input : a valid word form
Output : 1. Root
2. Lexical Category
3. Paradigm type
4. Morphological Category
(The output may be one or more analysis)

Input and Output Specifications in Tamil:

Input:

1	koyampuwwUr[1]
2	iraNtAvawu
3	mikappeVriya
4	wamilYYaka
5	mAnakarakam
6	Akum

Output:

1	koyampuwwUr	unk	<fs af='koyampuwwUr,n,n,sg,3,d,0,0'>
2	iraNtAvawu	unk	<fs af='iraNtu,num,,sg,,o,Avawu,Avawu'>
3	mikappeVriya	unk	<fs af='mikappeVriya,adj,any,any,any,,0,0'>
4	wamilYYaka	unk	<fs af='wamilYYakam,n,n,sg,3,o,0,0'>
5	mAnakarakam	unk	<fs af='mAnakaram,n,n,sg,3,d,0,0'>
6	Akum	unk	<fs af='Aku,v,n,sg,3,,pp,pp'>| <fs af='Aku,v,n,pl,3,,pp,pp'>

b. Word form Generator:

A Tamil word-form synthesizer enables a user to generate Tamil word forms. The user is prompted to select some choices leading to the generation of the desired word.

This is extremely useful to the learners of Tamil as second language. Such uses can interactively generate the requested word in Tamil.

The Morphological Generator of Tamil is based on Word and Paradigm Method. It is built using the feature values, suffix informations with add or delete rules and the root word dictionary with its category and paradigm. It uses the Machine Learning techniques to generate the word form from the given input.

The basic resources required for present word synthesizer:

1.**Feature Value:** It contains the category, its possible morpho-syntactic properties. It has five values, each viz., category, gender, number, person and the affix. For instance,

"v m sg 3 nw"

The above is an example for the verb for generating third person singular masculine past tense verb as such *vanwAnY* 'he came'.

2.**Suffix information and synthesis rule set:** This is generated from the paradigms and its feature values. It contains the rules for words based on their morpho-phonemic processes. It has four columns delimited by comma. For instance, to generate

'marafkalYE'
maram + kalY +E
Eng: tree + plural+ Accusative case

the suffix information table consists,
"Eylakf,m,maram,89"

Whereas the first is an inversed suffix of 'fkalYE' which is to be added, the second is the word which has to be deleted from root and the third is the name of the pardigm as such the word behaves in its morphophonemic process and finally the row number of the feature value file.

3.**Lexicon:** Lexicon consists of the root words of Tamil, its category and the name of the paradigm based on its phonological behaviour in its inflection. For instance,

'aNi,v,varE' Eng: put on, verb, draw

Here *aNi*, the verb act morpho-phonemically as *varE*.

aNi-nw-AnY	as in	*varE-nw-AnY* 'root-PST-3p.sg.m'
aNi-kirY-AlY	as in	*varE-kirY-AlY* 'root-PRS-3p.sg.f'
aNi-v-ArkalY	as in	*varE-v-ArkalY* 'root-FUT-3p.sg.m'

Input :
1. Root
2. Lexical Category
3. Morphological Category

Output : a valid word form

Input and Output Specifications in Tamil:

Input:

1	kampar	NNP	<fs af='kampar,n,m,sg,3,0,0,0'>
2	irAmAyaNam	NN	<fs af='irAmayaNam,n,n,sg,3,o,E,E'>
3	iyarYrYu	VM	<fs af='iyarYrYu,v,m,sgh,3,,nw,nw'>

Output:

1	**kampar**	NNP	<fs af='kampar,n,m,sg,3,0,0,0'>
2	**irAmAyaNawwE**	NN	<fs af='irAmayaNam,n,n,sg,3,o,E,E'>
3	**iyarYrYinYAr**	VM	<fs af='iyarYrYu,v,m,sgh,3,,nw,nw'>

c. Dictionary :

The Tamil-Telugu bilingual dictionary is built based on the concepts available in the language. It differs from the conventional dictionary which lists words but not concepts. Here, the lexeme(s) are related to each other on the basis of the concept i.e. the idea of ontological entity. The dictionary which is based on concepts is a better one for obtaining a concise and effective lexicon which can be used in many NLP applications.

d. Machine Translation System :

The development of Machine Translation is one of the most challenging tasks of the Natural Language Processing Applications. The development of Machine Translation (MT) System which translates texts from Telugu to Tamil and vice-versa (Bi-directional) are incorporated here. This MT system was developed as part of IL-ILMT consortium project funded by the Government of India at CALTS, University of Hyderabad. This Machine Translation system uses Transfer Based Approach. The System's Architecture is divided into three stages i.e. Source language Analysis module (SLA), Source language to Target language Transfer module (SL-TL) and Target language generation module (TLG).

(i) Telugu-Tamil Machine Translation system:

The *crucial* tools used in Telugu-Tamil Machine Translation system includes,

a. Source Language Analysis
1. Telugu Sandhi Splitter
2. Telugu Morphological Analyzer
3. Telugu POS Tagger
4. Telugu Chunker
5. Telugu NER (Named Entity Recognizer)
6. Telugu Parser

b. Source Language- Target Language Analysis
7. Telugu-Tamil Transfer Grammar Module
8. Telugu-Tamil Multi Word Expression Module
9. Telugu-Tamil Lexical Transfer Module

c. Target Language Analysis
10. Tamil Agreement Module
11. Tamil Word form Generator

(ii) Tamil-Telugu Machine Translation:

The *crucial* tools used in Tamil-Telugu Machine Translation system includes,

a. Source Language Analysis
1.Tamil Sandhi Splitter
2.Tamil Morphological Analyzer
3.Tamil POS Tagger
4.Tamil Chunker
5.Tamil NER (Named Entity Recognizer)
6.Tamil Parser

b. Source Language- Target Language Analysis
7.Tamil-Telugu Transfer Grammar Module
8.Tamil-Telugu Multi Word Expression Module
9.Tamil-Telugu Lexical Transfer Module

c. Target Language Analysis
10.Telugu Agreement Module
11.Telugu Word form Generator

3. Conclusion:

The Tamil Hyper Grammar thus described here is the most significant development in the recent applications of Natural language processing of the Tamil language to be used as teaching, learning as well as reference grammar for all kinds of language users.

[1] Transliteration Scheme using wx-notation:
Tamil Orthography :
a A i I u U eV e E oV o O H
k f c F t N w n p m y r l v lYY lY rY nY j s h R
Telugu Orthography :
a A i I u U q Q eV e E oV o O M H
k K g G f c C j J F t T d D N w W x X n p P b B m y r rY l lY lYY v S R s h

References :

[1] Arden A.H. 1891. A Progressive Grammar of the Tamil Language. Chennai: The Christian Literature Society.

[2] ILMT Consortium. 2007. ILMT SRS and Functional Specifications (mimeo). Hyderabad.

[3] Parameswari K. 2009. An Improvized Morphological Analyzer for Tamil: A case of Implementing an open source platform Apertium. Unpublished M.Phil. Thesis. Hyderabad: University of Hyderabad.

[4] Ramaswamy, Vaishnavi. 2003. A morphological Analyzer for Tamil. Unpublished Ph.D. Thesis. Hyderabad: University of Hyderabad.

[5]Uma Maheshwar Rao, G. 2002. A Computational Grammar of Telugu. (Mimeo) Hyderabad: University of Hyderabad.

[6] Uma Maheshwar Rao, G. 2005. Telugu Hyper Grammar. (Mimeo and Electronic form) Hyderabad: University of Hyderabad.

[7] Uma Maheswar Rao G, Amba P. Kulkarni and Christopher M. 2007. Functional Specifications of Morphology (mimeo). Hyderabad.

[8] Uma Maheswar Rao G. and Christopher M. 2010. Word Synthesizer Engine. In Morphological Analyzer and Generators. Mona Parakh (ed.) Page 73-81. Mysore; CIIL.

[9] Uma Maheswar Rao G. and Parameshwari K. 2010. On the Description of Morphological Data for Morphological Analysers and Generators: A case of Telugu, Tamil and Kannada. In Morphological Analyzer and Generators. Mona Parakh (ed.) Page 114-123. Mysore; CIIL.