TELUGU HYPER GRAMMAR

Uma Maheshwar Rao[1], G., Santosh[6] Jena, Bharathi[4], D.V, Christopher Mala[2], Krupanandam[3], N., Srikanth[9], M., Bindu Madhavi,[5] B., Parameshwari[7], K. and Sreenivasulu[8], N.V.

Center for Applied Linguistics and Translation Studies
University of Hyderabad
Hyderabad, India
{[1]guraohyd, [3]nityakrupa}@yahoo.com
{[7]cuteparamesh, [2]efthachris, [5]madhavihcu, [8]nv.sreenivasulu, [6]santosh.jena, [9]mudhams, [4]vijaya.anhony}@gmail.com

Introduction:

Grammatical descriptions of human languages are the results of efforts in modelling of the design features and the internal organization of the structures and the mechanisms of language. 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.

Concept:

A Hyper grammar is a non-linearly organized dynamic grammar based on 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 Telugu, with hyper-links. It can be accessed and retrieved for various purposes involving language, to experience the effect of a native speaker 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 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 in the computational framework. The contents are being developed from both the published and unpublished sources carefully selected and rewritten in the hypertext format.

The Contents:

The content of Telugu Hyper grammar has two main components, viz. the description of grammar in hypertext format and the applicational aspect of the Telugu Language manager.

The Telugu Grammar:

The grammar part includes a number of comprehensive descriptive

notes on certain linguistic facts of Telugu Language. It is conceived in terms of a Computational Grammar. It deals with the Orthography, the design features of Telugu script, orthographic syllables, the information on the frequency distribution of written syllables etc.

As part of the Telugu morphology, we have information on Telugu categories 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 100 words, five thousand words and ten thousand words in terms of their frequencies, and communicative contribution to the coverage in Telugu Texts. As regards to the frequency of Telugu 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 Telugu-Hindi, Telugu-Kannada, Telugu-Telugu, Telugu-Oriya, Telugu-Marathi, Telugu-English and English-Telugu – are included. Originally these dictionaries are conceived as bilingual and bi-directional dictionaries initially created using the most frequently occurring words ensuring the coverage.

The Telugu language manager:

This is the most crucial component of Telugu 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 language. In order to make to sitimulate the grammar, it should involve a working analyzer, generator, parser and lexical accessor, etc. Currently the language manager includes a word form generator, a morphological analyzer and lexical accessor among others.

The Morphological Analyzer:

The word analyzer incorporated here is intended to analyze the Telugu 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 Telugu Morphological Analyzer comprises of inflectional i.e. paradigmatic data and root dictionary. These data comprise purely linguistic information of the language, which are processed subsequently to enable for 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 morphosyntactically

related word forms of a given lexeme. Based on the inflection, there are six distinct morphological categories are identified and the paradigms are created. It includes 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 morphosyntactic functions. The major word categories are listed as below,

–Nouns
–Verbs
–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 are indeclinables) 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 Telugu:

Input:

1 himAlayAlu
2 sahaja
3 sixXaMgA

4 erpaddAyi
5 .

Output:

1 himAlayAlu <fs af='himAlayaM,n,,pl,,d,0,0'>|<fs af='himAlayaM,n,,pl,,d,vu,vu'>
2 sahaja <fs af='sahajaM,n,,sg,,o,ti,ti'>|<fs af='sahajudu,n,,sg,,o,ti,ti'>
3 sixXaMgA <fs af='sixXaM,n,,sg,,,gA,gA'>
4 erpaddAyi <fs af='erpadu,v,n,pl,3,,A,A'>
5 . <fs af='.,punc,,,,,,'>

Word form Generator:

A Telugu word-form synthesizer enables a user to generate Telugu 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 Telugu as second language. Such uses can interactively generate the requested word in Telugu.

The Morphological Generator of Telugu 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,

vaccu, "v m sg 3 A"

The above is an example verb for generating third person singular masculine past tense verb form as such *vaccAdu* '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 process. It has four colums delimited by comma. For instance, to generate

'*puswakAlakosaM*'
puswakaM + lu +kosaM
Eng: book + plural+ purpose

the suffix information table consists,

"MasokalA,aM,puswakaM,89"

Whereas the first is an inversed suffix of 'AlakosaM' 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 Telugu, its category and the name of paradigm as such it behaves in its inflection. For instance,

'winu,v,koVnu'

Here winu, the verb behaves morpho-phonemically as koVnu.

Lexeme: ***winu,v.***

Paradigmatic Members/wordforms	*Common Maximal match*	*Functional/ Formative Element*	*Feature Values*
winnAdu	*win*	nAdu	Past-m-sg-3
wiMtAdu	*wi*	MtAdu	np-m-sg-3
winadu	*win*	adu	neg-m-sg-3
wini	*win*	i	nf-past
wine	*win*	e	nf-adjl-ppl
winu	*winu*	0	imp-sg

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

Output : a valid word form

Input and Output Specifications in Telugu:

Input:

1 himAlayaM <fs af='himAlayaM,n,,pl,,d,0,0'>
2 sahajaM <fs af='sahajaM,n,,sg,,o,ti,ti'>
3 sixXaM <fs af='sixXaM,n,,sg,,,gA,gA'>
4 erpadu <fs af='erpadu,v,n,pl,3,,A,A'>
5 . <fs af='.,punc,,,,,,'>

Output:

1 **himAlayAlu** <fs af='himAlayaM,n,,pl,,d,0,0'>
2 **sahaja** <fs af='sahajaM,n,,sg,,o,ti,ti'>
3 **sixXaMgA** <fs af='sixXaM,n,,sg,,,gA,gA'>
4 **erpaddAyi** <fs af='erpadu,v,n,pl,3,,A,A'>
5 . <fs af='.,punc,,,,,,'>

Dictionary :

The Telugu-Hindi bilingual dictionary is built based on the concepts of languages. It differs from the conventional dictionary with respect to the use of concept as bases and listing a series of words which indicate that concept. 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 to obtain a concise and effective lexicon which can be used in many NLP applications.

Ex:

ID :: 748
CAT :: NOUN
CONCEPT :: స్వచ్ఛంద కదలిక ఉన్న జీవం
EXAMPLE :: "భూమి మీద అనేక రకాల జంతువులను చూడవచ్చు"
SYNSET-TELUGU :: జంతువువు/HIN1, చతుష్పాదం/HIN6, మృగం/HIN5, పశువు/HIN2, గొడ్డు/HIN7, జీవి/HIN4, ప్రాణి/HIN3

Machine Translation System :

The development of Machine Translation is one of the most challenging tasks of Natural Language Processing Applications. The development of Machine Translation (MT) System which translates texts from Hindi or Tamil to Telugu and vice-versa (Bi-directional) are incorporated here. This MT system was developed as part of IL-ILMT consortium project funded by 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 (SL), Source language to Target language Transfer module (SL-TL) and Target language generation module (TL).

(i) Hindi-Telugu Machine Translation system:

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

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

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

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

(ii) Telugu-Tamil Machine Translation:

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

a. Source Language Analysis
   - Telugu Sandhi Splitter
   - Telugu Morphological Analyzer
   - Telugu POS Tagger
   - Telugu Chunker
   - Telugu NER (Named Entity Recognizer)
   - Telugu Parser

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

c. Target Language Analysis

Tamil Agreement Module
Tamil Word form Generator

The user may acess the Hyper Grammar any of these lexicon tools and applications as he/she wishes to do.

WX-Notation used in the Transcription of examples:

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 v S R s h

References :

Krishnamurti. Bh and J.P.L. Gwynn. 1985. *A Grammar of Modern Telugu*. New Delhi. Oxford University Press.
Uma Maheshwar Rao, G. 2002. A Computational Grammar of Telugu. (Mimeo) Hyderabad: University of Hyderabad.
Uma Maheshwar Rao, G. 2005. Telugu Hyper Grammar. (Mimeo and Electronic form) Hyderabad: University of Hyderabad.
Uma Maheswar Rao G, Amba P. Kulkarni and Christopher M. 2007. Functional Specifications of Morphology (mimeo). Hyderabad.
Uma Maheswar Rao G. and Christopher M. 2010. Word Synthesizer Engine. In Morphological Analyzer and Generators. Mona Parakh (ed.) Page 73-81. Mysore; CIIL.