TelMore: Morphological Generator for Telugu Nouns and Verbs

Madhavi Ganapathiraju and Lori Levin

Language Technologies Institute, Carnegie Mellon University, Pittsburgh, PA 15213, USA
madhavi@cs.cmu.edu, lsl@cs.cmu.edu

***Abstract* — Telugu is an Indian language spoken by over 50 million people in the country. The language is rich in literature and has been studied by native and foreign linguists significantly, yet it has not benefited significantly from the recent advances in computational approaches for linguistic or statistical processing of natural language texts. However with the recent progress in standardization of machine representation of text, applications like machine translation and information retrieval are beginning to surface, for example with the collaborative efforts under the umbrella of Digital Library of India (DLI). There is a need for a morphological generator for Telugu that forms an integral part of applications like machine translation and universal dictionary.**

Here we present the development of a tool, called TelMore, that can generate morphological forms of nouns and verbs of Telugu. It has been developed based on the previously established linguistic analyses of Telugu by C.P. Brown and H. Krishnamurthy. The tool is developed in Perl® and is made available with a web interface and in source code for use and further development at

http://www.cs.cmu.edu/~madhavi/TelMore/.

1. Introduction

Telugu is an Indian language spoken by over 50 million native speakers. It ranks between 13-17 largest spoken language in the world alongside of Korean, Vietnamese, Tamil and Marathi. The distribution of spoken language in India is geographic, and each of the different states of the country usually speak a different language (apart from a large number of Hindi-speaking states). Andhra Pradesh state, where Telugu is spoken, shares borders with 5 different states which speak Tamil, Kannada, Marathi, Hindi and Oriya. Thus, in regions along the borders with these states, the dialect of Telugu is different, although the script and formal (written) language are the same. It has recorded origins around 7th century AD and became a literary language around 11th century AD.

Computational approaches to linguistic analysis of Indian languages have so far been hindered due non availability of a standardized digital representation and in turn by the non-availability of large amounts of text data. This scenario has recently started to change, primarily through the projects that are being carried out under the umbrella of the Digital Library of India (DLI) (www.dli.ernet.in). Multilingual applications such as machine translation, multilingual book reader with transliteration and translation capabilities, and Indian language search engine have all been built over the transliteration tools and Indian language data generated by DLI [1-3]. A foremost contribution of DLI towards language technologies has been the creation of the transliteration scheme called *Om*, which allows representation of Indian language alphabet with English transliteration [4]. Om transliteration scheme has been shown to be useful for reading Indian language text with or without native font rendering [5] and for development of language processing tools, for example, *good-enough* translation for Indian languages [3], *OmSE* Tamil information retrieval [5], multilingual book reader with transliteration and translation capabilities [3]. Thousands of literary books that are being scanned under DLI should soon be available as plain text with the optical character recognition (OCR) technology for Indian scripts currently at greater than 95% accuracy for clean images. Om, and other DLI activities are bridging the gap between Indian languages and the computational language technologies.

An integral part of many of these applications based on natural language processing is the morphological analyzer/generator. Although Telugu language has ancient origins, and is today spoken by such a large number of people, and is spoken in a State with significant advancements in the information technology (www.aponline.gov.in/), few advances have been made in computational linguistic processing or computational natural language processing in this language. Of the few tools that have been developed elsewhere earlier is a morphological analyzer [6]. This tool is non-descriptive, that is, it is based on a dictionary and not rule-based, and is limited in its span of the lexicon. For a large number of words it fails to give morphological analysis. Further, it does not generate related morphological forms of the given word. In the work presented here, we developed a computational morphological generator for nouns and verbs of Telugu. In the rest of the paper, we describe this tool.

2. TelMore

Telugu Morphological generator for nouns and verbs which we henceforth refer to as *TelMore*, is built over the linguistic rules described in previous literature [7, 8]. It accepts a noun and a predefined lexical class as input, and generates all the applicable noun forms, namely nominative, genitive, accusative, dative, vocative and instrumental forms for masculine, feminine or neutral genders and for singular and plural numbers. For the verb types, TelMore accepts the infinitive *t’a* form (e.g. pan’put’a, cheyyut’a, etc) and generates, present, past and future tenses, and the aorist affirmative, aorist negative, imperative and prohibitive forms of the verb, for all genders and numbers.

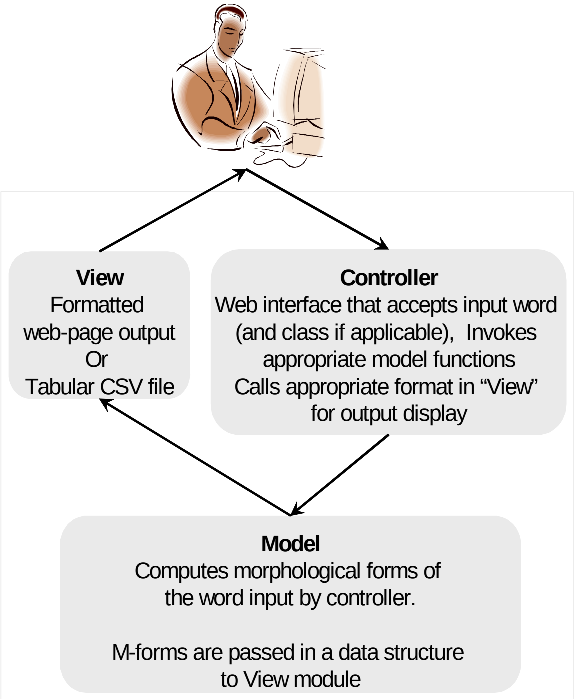
View
Formatted
web-page output
Or
Tabular CSV file
Controller
Web interface that accepts input word
(and class if applicable), Invokes
appropriate model functions
Calls appropriate format in “View”
for output display
Model
Computes morphological forms of
the word input by controller.
M-forms are passed in a data structure
to View module

Fig.1 Model-view-controller architecture of TelMore. The three modules are separated from each other in architecture and interact only through the passing of parameters between them, allowing each of the modules to *evolve* independent of the rest of the application. Model component consists of the morphological generator functions, View consists of the output formatting functions and the Controller consists of the user inputs and activation of appropriate model functions.

A. *Software Architecture*

The software architecture of TelMore follows the standard architectural pattern of the model-view-controller (MVC) (see Fig. 1). The controller unit allows the user to choose the input and output formats, which can be one of (1) comma-separated flat files (2) user interactive mode on terminal or (3) formatted HTML pages for display or (4) stream to and from another application. The output or display formats themselves form the visualization unit. There are multiple print functions that form the core of this visualization module and can be updated without affecting other modules. The interaction of this module with other modules is only through function arguments and return values. The actual generation of the morphological forms of nouns and verbs is done by a set of library functions of the model module. This module takes in a word and generates the different morphological forms. There are many functions corresponding to the different lexical forms of the input word, and can be up corrected or updated as found necessary. The uniformity of interaction of this module with other modules allows the extension of the functionality for other grammatical classes of words, for example adjectives, adverbs and so on.

The, separation of the three modules, namely models, view and control, allow the package to evolve to support more lexical classes and for development of other applications. For example, addition of the feedback module to perform evaluations has been added at a later stage to the print module, without in any way affecting the rest of the program. The MVC architecture is best suited for this collaborative development setting, wherein different applications and data generation processes are continuously added to the system by different groups at different geographical locations. The web interface module, for example, has been generated at Indian Institute of Science by a project student, with the least burden of understanding the underlying the tools.

The tool is expected to form an integral part of other systems, specifically the example based machine translation and universal dictionary generation for the Universal Library (http://ul.cs.cmu.edu) and the Digital Library of India (http://www.dli.ernet.in and http://dli.iiit.net).

In all of this paper, Telugu text is written as per Om transliteration [4], and is shown in fixed-width font.

3. Nouns

A. *Plural Formation of Telugu Nouns*

Last syllable for plural nouns is always the *plural suffix* lu, although plural formation happens in a number of different ways. The regular way of forming the *nominative plural* of a common noun is to add the plural suffix lu to the *nominative singular*. For example:

aavu → aavulu
anna → annalu
kurchii → kurchiilu
pet’t’e → pet’t’elu

However, on addition of the plural suffix, there may be a *san’dhi* (conjugation) formation, because of which lu becomes l’u or l’l’u or n’d’lu. There is also a list of nouns that do not form plural according to these rules. Plurals sometimes have variant forms in use (for example kannulu or kal’l’u).

Rules of san’dhi formation

Let nominative singular be referred to as *stem* in the context of plural formation.

1. If stem final is: [t’/n’t’/n’d’] + [i/u], then the final vowel [i/u] is lost before the plural suffix lu.

   [t’/n’t’/n’d’] + [i/u] → [t’/n’t’/n’d’] + lu
   t’i → t’lu
   n’t’i → n’t’lu
   n’d’i → n’d’lu
   t’u → t’lu
   n’t’u → n’t’lu
   n’d’u → n’d’lu
   n’d’lu freely becomes l’l’u

2. If stem final is: [d’i/d’u/lu/ru] or if stem is more than 2 syllables and ends in [li/ri], the final syllable becomes [l’] before adding [l’u].

   Exception: Masculine nouns of Sanskrit origin ending in d’u replace d’u by lu
   Example: sneihitud’u → sneihitulu

3. Stem final [t’t’/d’d’]+[i/u] becomes [t’/d’]+l’u

4. Stem final [llu/nnu] becomes n’d’lu or l’l’u.

   Exception: Following stems add lu to the basic stem to form plural: Pannu, vennu, ponnu, jannu, tannu, t’annu.

5. Stem final [an’/aan’] is replaced by aa and stem final [en’] by e’ before plural suffix lu.

6. Stem final aayi, having more than 2 syllables add lu.

7. Stem final [y/yy]+i is replaced by tulu. Only 3 nouns in this class: cheyyi, goyyi, neyyi.

8. If above rules do not apply and stem ends in i, then

9. If stem is 2 syllables, or if 3 syllables and middle vowel is other than i, then i changes to u before lu. If middle syllable is i, then that also changes to u, unless the noun is of Sanskrit origin. (atithi, parithi, samiti).

Exceptions to plural formation:

The following nouns to do not conform to plural formation rules given above:

raayi → raal’lu
poyyi → poyyilu
pen’d’li → pen’d’in’dd’lu → pel’l’il’l’u
vari → vad’lu
+gaaru → +gaarlu
eddu → eddulu → ed’lu
veyyi → veilu
+saari → +saarlu
cheinu → cheilu
peinu → peilu
eid’u → ein’d’lu → ei’l’lu
+gaad’u → +gaal’l’u
allud’u → allun’d’lu → allul’l’u
manamaraalu → manamaraan’dlu → manamaraal’l’u

Plural generation in TelMore is performed both based on Brown’s informal description and Krishnamurthy’s linguistic formalism described in [7, 8]. However, the latter generation although very reliable in most classes, is erroneous for some (one or two) exceptions.

B. *Morphology of Telugu Nouns*

Morphology of nouns given here is as described by Brown [8]; they have three declensions and two numbers, namely singular and plural. Case markings found are nominative, genitive, dative, accusative, vocative, instrumental and locative. Case markings of singular nouns in general are given in Table 1, and are described in detail for each declension separately.

CASE	MORPHOLOGICAL FORM
Nominative	
Genitive	Inflection, +yokka
Dative	{+ki, +ku}
Accusative	{+ni, +nu}
Vocative	{oo + , +finalVowel}
Instrumental	{+cheita, +too}
Locative	+na

Table 1. Case markings of Telugu nouns

First Declension: masculine nouns ending in d’u

All nouns that are masculine and end in d’u, such as raamud’u, allud’u are placed in this declension. Many Sanskrit nouns are also placed in this declension after adding the suffix –ud’u, for example, braahmand-ud’u, vartakud’u, deivud’u. The inflection is formed by changing the -ud’u or d’u of the nominative singular to ni. Thus, tammud’u (nominative singular) forms tammuni (genitive). The morphological forms are derived exactly as described in Table 1.

1ST DECLENSION	2ND DECLENSION	3RD DECLENSION REGULAR
Raamud’u	Varamu	Puli
Deivud’u	Penamu	Anna
Viirud’u	Veidamu	Karra
Nat’ud’u	Pan’demu	Strii

Table 2: Examples of words in the first and second declensions and regular class of third declension

Second Declension: neutral nouns ending in mu

This declension contains neutral nouns of more than two syllables, ending in mu, aamu and emu. Many nouns use the nominative form instead of any inflection. Some genitives are alternatively formed by in pu, for example N: gurramu, G: gurrapu.

Third Declension—Regular Class

All regular nouns having no inflection in the singular are placed in this class. For example, anna, is morphed as anna, annayokka (G), annaku (D), annanu (A), annaa (V), annatoo (I) and annaloo, annayan’du (L). In plural, the locative suffix -laloo is frequently contracted as -lloo. Plural is formed by adding -lu to the nominative singular, and nouns ending in -i change -i to -ulu to form the plural.

Example words of first, second and regular classes of third declension are given in Table 2.

Third Declension—Irregular classes: There are eight irregular classes of Telugu words, and another class in which are placed all the foreign words. Examples of all irregular classes of nouns in this third declension are given in.

1. Regular in singular, irregular in plural.
2. Inflection singular is formed by changing last syllable to t’i and plural nominative into l’l’u or n’d’lu.
3. Last syllable of the nominative singular into n’t’i to form singular inflection and into n’d’ulu or n’d’lu or n’l’l’u to form plural nominative.
   Kannu, mannu, channu, minnu, villu, mullu, illu, pallu.
4. Use nominative singular as inflection OR change last syllable to t’i. Plural nominative is formed by adding lu or changing final syllable to l’l’u or n’d’lu.
5. Last syllable of nominative singular changed to ti to form inflection and tulu to form plural.
6. u of nominative singular changed to i to form inflection in singular. Nominative plural is formed by adding lu or l’l’u or n’d’lu. *Some change inflection irregularly.*
7. Words ending in rru form inflection in singular as rti. (strange words).
8. Some nouns form inflection singular in lu and nominative plural in l’l’u, n’d’lu and n’d’ru..
9. *Foreign words:* These words are quite often adopted into the language, such as doctor, glass, naukar (usually all end with a -u suffix). These words do not undergo any inflection, except in plural where inflection happens occasionally.

1ST CLASS	2ND CLASS	3RD CLASS	4TH CLASS	5TH CLASS	6TH CLASS	7TH CLASS	8TH CLASS	FOREIGN WORDS
Chiit’ii	Biid’u	Kannu	Choot’u	Vaayi	Cheinu	*	Kaalu	Naukaru
Paat’u	Taad’u	Mannu	Jood’u	Raayi	Guunu		Veilu	Vakiilu
Pod’i	Eiru	Villu	Modalu	Neyyi	Nuulu		Kood’alu	Munasabu
Kood’i	Doosili	Illu	Niiru	Goyyi	Uuru		Maradalu	
Ut’t’i	Nuduru	Minnu	Vennela		Kuuturu			
Ban’d’i	Kood’u	Mullu	Netturu					
Aavu	Nooru		Aakali					
Toolu	Chekkili		Kod’avali					

Table 3. Examples of words in the irregular classes of the third declension

*The seventh irregular class consists of some words that are unheard of in common language, it is not clear how they should be morphed, and hence morphological forms of this rare class are not generated.

4. Morphology of Telugu Verbs

Telugu verbs fall into three conjugations based on their morphology. The first and second conjugations mostly contain words of Telugu origin, and borrowed words from other languages fall into the third conjugation. Verbs are described based on their infinitive form ending in `t'a` in this tool. ***Voice****:* There are two voices namely affirmative and negative. Passive voice is compounded with `-pad'ut'a`, middle voice with `-konut'a` and causal voice with insertion of `-inchu`. ***Tense****:* Tenses of the verbs can be present, past, future, aorist and imperative. ***Numbers****:* Numbers are singular and plural. ***Persons****:* First, second and third. ***Gender****:* In singular third person, feminine has neutral termination, but masculine termination in plural.

Verbs are divided into 3 conjugations based on the ending letters of the root, as described below. Examples of the three classes of the verbs are given in Table 4

1ST CONJUGATION	2ND CONJUGATION	3RD CONJUGATION
Vinut'a	Cheiyut'a	Vachchut'a
Konut'a	Pooyut'a	Karachut'a
Chaduvut'a	Kosut'a	Viruchut'a
Pan'put'a	Neiyut'a	Gichchut'a

Table 4. Examples of words in the first, second and third conjugations of verbs

First Conjugation: all verbs other than described in second and third conjugations

All verbs that do not belong to second and third conjugations appear in this first conjugation. A majority of verbs fall into this conjugation.

Second Conjugation: verbs that end in `yut'a`*,* `yyut'a`*,* `sut'a` *and* `ssut'a`

Third Conjugation: verbs that end in `chut'a`*,* `chchut'a`

A. *Morphological forms of verbs of the three conjugations*

The basic suffix for all the conjugations is the same. The only difference between the three conjugations is the *san'dhi* formation due to the end-syllable of the stem. Some words however take irregular morphological forms. An example is the verb `vachchut'a` which means *coming*. In the prohibitive case the verb takes on an altogether different stem `raa` (`raanu`, `raaku`). The general suffixes for the three conjugations are given in Table 6 (at the end of the report).

5. Data

1st Declension		247
2nd Declension		539
3rd Declension	Regular	29
	Irregular-1	8
	Irregular-2	18
	Irregular-3	6
	Irregular-4	14
	Irregular-5	6
	Irregular-6	5
	Irregular-7	0
	Irregular-8	2
1st Conjugation		62
2nd Conjugation		5
3rd Conjugation		55

Table 5. Data entry: Number of words in each lexical class

A data set of nouns and verbs has been created by native Telugu speakers for testing the morphological generator. Where required, the lexical class (declension, conjugation) of the root noun is specified, and the verb is entered in the required infinitive `t'a` form. The number of words available in each class is as given in Table 5. (Note that these words are only provided as examples, and do not contribute to the morphological generation process. Morphological generation follows generalized rules, and is not derived from the input data).

6. Results and Discussion

A random selection of words from the word lists was used in generating morphological forms, and the results are presented to native speakers for evaluation. A second iteration after corrections to program based on errors in first iteration produced accurate results. Automatic generation of plurals is accurate for most nouns, but is erroneous for nouns with 2 forms of end syllables. This error has not yet been corrected.

There is a large amount of 'raw text' available in the form of online news paper articles. Although all of these newspapers adopt their own non-standard fonts for representation, some tools that convert these texts into the standard Om representation have been developed by DLI, which make the newspaper text parsable by traditional tools. Lexicons created from these newspapers can be semi-automatically divided into the input classes for TelMore, and their morphological variations can be generated, which prove extremely useful in machine translation, information retrieval and related tasks. Example based machine translation system is in use under digital library of

SINGULARS			
Masculine	Feminine	Neutral	
Present Tense			
-unnaanu, -utaanu			1st person
-unnaavu, -aavu			2nd person
-unnaad’u	-unnadi, -un’di		3rd person
Past tense			
-ini, inaanu			1st person
-itivi, inaavu			2nd person
-enu, -inaad’u	-inadi, in’di		3rd person
Future Tense			
-edanu			1st person
-edavu			2nd person
-ed’ini			3rd person
Aorist			
-edunu			1st person
-eduvu			2nd person
-unu			3rd person
Imperative			
-u, -umu, -umaa			2nd person
Prohibitive			
-aku			2nd person
PLURALS			
Masculine	Feminine	Neutral	
Present tense			
-unnaamu, -aamu			1st person
-unnaaru, -utaaru			2nd person
-utunaaru, -utaaru		-tunnavi	3rd person
Past tense			
-imi, -inaamu, -aamu			1st person
-itiri, inaaru, aaru			2nd person
-iri, -inaaru, -aaru		-enu, -inavi	3rd person
Future tense			
-edamu, -eimu			1st person
-edaru, -eiru			2nd person
-edaru, -eiru		-ed’ini	3rd person
Aorist			
-udumu			1st person
-uduru			2nd person
-uduru		-unu	3rd person
Imperative			
-udamu, -udaamu			1st person
-an’d’i			2nd person

Table 6: General Suffixes for Morphological Forms of Verbs.

India and relies on a set of rules for translation from English to Telugu. By integrating a Parts-of-speech tagger on the source language side and the integration of TelMore on the target language side, the machine translation would yield much more accurate results, even without any further adaptations.

7. AVAILABILITY

TelMore is available as an online resource with Open Source at http://www.cs.cmu.edu/~madhavi/TelMore/ and http://ashwini.dli.ernet.in/morph/ free for use and further development. The online versions work with a web interface and also allow users to give feedback to each of the generated forms, thus allowing developers to keep track of any possible errors that may come in to notice by extended use.

8. CONCLUSION AND FUTURE WORK

TelMore generates morphological forms of the two main lexical classes: noun and verbs of Telugu language. Except for a few irregular classes of nouns, some of which have only a few words per class, the process is automatic. For these exceptions, the user needs to identify the class to which the noun belongs, based on what form it takes for genitive case or based on how its plural is formed. Plural generation is done in two ways: based on informal rules after identifying the class of the noun, or through formal descriptive rules without the requirement for identification of class in which case no user interaction is required. For the latter however, there are 2 syllable endings, for which the generated plural is incorrect. Verb generation is fully automatic and is correct for all words. There are some irregular verbs which are erroneously generated, but this is what makes them irregular words. Current version of the toolkit is available in Open Source for review and enhancement by the World Wide Web community.

For a full fledged deployment of the tool, other forms, namely the pronouns, adverbs and adjectives are to be supported. Further, compound verbs and other complex forms are to be added into the system.

ACKNOWLEDGEMENTS

Authors would like to thank Prof. Raj Reddy and Prof. N. Balakrishnan for valuable comments. Authors would also like to thank R. Harsha who developed the web interface, Naveena Yanamala who configured the web server locally, and V. Mythili Shyam, G. Padmasree, P. Nava Mallika, V. Abhinay and B. V. Prashanth, who contributed towards the creation of the data and evaluation of the output words.

REFERENCES

1. Balakrishnan, N., et al. *Million Books to Web: Technological Challenges and Research Issues*. in *Proc. Tamil Internet conference*. 2004. Singapore.

2. Balakrishnan, N., et al., *Digital Library of India: A testbed for Indian Language Research.* IEEE Technical Committee on Digital Libraries Bulletin: Special Issue on Asian Digital Library Research, 2005. In press.

3. Balajapally, P., et al. *Multilingual book reader interface: transliteration and translation*. in *VALA*. 2006. Melbourne, Australia.

4. Ganapathiraju, M., et al. *OM: "One Tool for Many (Indian) Languages"*. in *ICUDL: International Conference on Universial Digital Library*. 2005. Hangzhou.

5. Jayaraman, A., et al. *OmSE: Tamil Search Engine*. in *Proc. Tamil Internet conference*. 2004. Singapore.

6. Analyzer, I.M., *http://www.iiit.net/ltrc/morph/morph_analyser.html*.

7. Krishnamurti, B., *A grammar of modern Telugu*. 1985, Delhi ; New York: Oxford University Press.

8. Brown, C.P., *The Grammar of the Telugu Language*. 1991, New Delhi: Laurier Books Ltd.