**EXTRACTION OF NOUN’S FROM TELUGU CORPUS DATA USING PERL PROGRAM**

**AIM:**

Trying to get the nouns and their suffixes in Telugu language by writing a program using PERL programming language.

**TELUGU NOUN SUFFIXES:**

|  |  |
| --- | --- |
| SUFFIXES | SUFFIXES CALLED |
| -lo- | Locative sig,plu |
| -lu- | Nominative plural |
| -to- | Instrumental-sociative sig,plu |
| -ku- | Dative sig,plu |
| -ki- | Dative sig,plu |

Some real time examples representation table with suffixes mentioned above:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **S.NO** | **ROOT** | **SUFFIXES** | **GRAMMATICAL DESCRIPTION** | **EXAMPLES** |
| A | Doll  /bomma/  /bomma-lu/ | -lu- | NOM,Sig  NOM,plu | Bommalu konnaru |
| B | Doll  /bomma-to/  /bommal-a-to/ | -to-  -a-to- | Inst,sig  Inst,plu | Papa bommal-a-to aduthundi |
| C | Doll  /bomma-lo/  /bommal-(a)-lo/ | -lo-  -(a)-lo- | Locative,sig  Locative,plu | Bomma-lo cotton vunti |
| D | Doll  /bomma-ku/  /bommal-a-ku/ | -ku-  -a-ku- | Dative,sig  Dative,plu | Bommala-ku rangulu veyandi |
| E | House  /inti-ki/  /ill-a-ki/ | -ki-  -a-ki- | Dative,sig  Dative,plu | Illa-ku rungulu veyandi |

PROCEDURE:

First I collected some short stories in Telugu Ramayanam saved them in a folder called corpus data. And merge them into one file and named it as “rawcorpus.txt”, which I used this file as input file to run all the programs.

After I created a folder named as wordsplit-tokenizer-sorting-Harsha which contains three small files (code,input,output) in code folder I write and saved the wordsplit,tokenizer,sorting perl program, and in input folder I saved the rawcorpus.txt(text file) which is used as the input for the program. In the output folder I saved the output that I got after running the programs.

**WORDSPLIT,TOKENIZATION AND SORTING:**

I run the wordsplit,tokenizer,sorting program in perl command by giving the command as “perl wordsplit,tokenizer,sorting.pl rawcorpus.txt >wordsplit,tokenizer,sorting-output.txt” this program removes all the punctuation marks and sort the list and saved as wordsplit,tokenizer,sorting-output.txt this sorted file is used as the input file in the string extraction and root extraction program which I saved it in the output folder.

**STRING EXTRACTION AND ROOT EXTRACTION:**

String\_and\_root\_extraction\_dp folder also contains the same three folders (code, input, output). This string and root extraction program is used to extract the suffixes from the corpus. For running my string extraction program I used output of the wordsplit,tokenizer,sorting-output.txt as input file.

And I put all my possible suffixes for example (lo,ki,lu,to,ku)in this one program and at the same time I extracted all the root words from which the final output contains all the words with their suffixes and with the possible root words.

**RESULTS AND ISSUES:**

1. In Telugu language the suffixes of nouns, pronouns and some other categories also having the same suffixes. So in the result the words that are having the same suffixes as nouns also there in the result.

2. While sorting the tokenized-text some of the punctuation marks are not removing which is done manually.

3. In the result of root extraction some words are becoming meaningless without the suffixes.

4. Only proper nouns are shown correct in the result of root extraction.

**SOURCES:**

* [www.valmikiramayan.net/bala\_kanda1\_contents.html](http://www.valmikiramayan.net/bala_kanda1_contents.html)
* [www.valmikiramayan.net/bala\_kanda2\_contents.html](http://www.valmikiramayan.net/bala_kanda2_contents.html)
* [www.valmikiramayan.net/bala\_kanda3\_contents.html](http://www.valmikiramayan.net/bala_kanda3_contents.html)
* [www.valmikiramayan.net/bala\_kanda4\_contents.html](http://www.valmikiramayan.net/bala_kanda4_contents.html)

SUBMITTED BY

M.Durga Prasad