PARTS-OF-SPEECH TAGGING FOR KANNADA

Vijayalaxmi F. Patil, Shahid M. Bhat
[jeeveshwarip@gmail.com]
[shahid.bhat3@gmail.com]
LDC-IL, CIIL Mysore

CONTENTS

INTRODUCTION

- Kannada is the official language of the state Karnataka. Kannada is one of the Dravidian languages with SOV word order.

- It is very important language as it is not only one of the 22 scheduled languages of India but also one of the classical languages.

- It is spoken in Karnataka and its neighboring states like Maharashtra, Tamil Nadu, Andhra, Goa, etc by about 35 million speakers (Wikipedia).

CONT.....

- It is morphologically rich and agglutinative in its nature.
- It shares many morphological features with other Dravidian languages like defective verbs, like alla not and illa (no), some particles like inclusive particle kUDa (also) and some auxiliaries like –koLL (reflexive), paDu (passive) etc which are considered to be one of the type of an auxiliaries.

CONT.....

- Parts-of-Speech tagging refers to the process of assigning a POS tag to the words of a text.

- In other words we can say that it is a process of marking the words in corpus which corresponds to particular parts of speech based on both its definition and the context.

- POS tagging is one of the important level and the ground work for other higher level stages in NLP.

KANNADA & AVAILABLE LANGUAGE RESOURCES

- Like many of the Indian languages, very little work has been done in the area of NLP for Kannada. It is a resource-poor language. Even if resources exist somewhere, they exist without public access (Murthy 2000).

- CIIL has developed a corpus of about 3 million words for Kannada under a project funded by Department of Information Technology (DIT). Further,

- POS tagger and morphological analyzer have been developed for Kannada under ILMT consortium project. From last few years LDCIL is engaged in creating language resources for Kannada on large scale.

CORPUS USED IN THIS WORK

- In the current work we are concerned with POS tagging of text corpus. Text Corpus is a machine readable collection of the text which is generally used as a raw data for various NLP.

- We have used 10,000 words of Kannada corpus from a single domain (Aesthetics).

CORPUS USED IN THIS WORK

Category
Aesthetics

Literature-Short Stories
Number of words
5654
Aesthetics
Literature-Children's Literature
780
Aesthetics
Literature- Autobiographies
856
Aesthetics
Literature-Essays
6572
Aesthetics
Literature-Biographies
2407

CONT.....

- Kannada corpus is not directly used for POS tagging because of various problems that need to be settled before actual tagging.
- Whatever we do with corpus to make it fit for tagging is generally called preprocessing. It involves the following two subtasks.

a. Corpus Cleaning

- Corpus usually contains some extra symbols, Sanskrit shlokas and some stanzas of poems. we have removed such elements.

- we corrected spelling mistakes, added some missing words and sentences, removed some extra words, sentences and paragraphs according to the text available in the hard copies of the corpus.

- We remain faithful to the text and keep some spelling variations as such which would be considered wrong spellings otherwise.

b. Corpus normalization

- Normalization is sort of tokenization. Since Kannada is highly agglutinative language (with severe fusion of grammatical categories), we need to tokenize corpus so that we can assign POS tags easily.

- In corpus normalization, we tokenize corpus properly by separating punctuations from preceding tokens and by splitting sentences or phrases into their constituent tokens.

For Example:

We segment “hELikoLLuttiruttidda” (had-been-speaking-himself) into ‘hELi’ (having spoken), ‘koLLutta’ (himself), ‘irutta’ (been), and ‘idda’ (had).

Similarly;

- NOUN: mAtinallIga (now-in-speech) = mAtin-alli (speech-in) + Iga (now)
- PRONOUN: adariMdEnu (with it-what) = adariMda (with it) + Enu (what)
- PRONOUN: nimagArigU (to-you-anyone) = nimage (to-you) + yArigU (anyone)

TAG SET USED IN THIS WORK

In order to assign a tag to a token we must have a tag set according to which we will assign tag to a token. In this work we are using BIS Dravidian tag set. Which has a 11 categories and 35 sub categories. The tag set is summarizes below with examples.

CONT.....

- NOUN: In Kannada we have 3 types of noun namely Common (NN), Proper (NNP) and Nloc (NST).
- Common noun includes words like mara (tree), giDa (plant), manuSya (human) etc. Proper noun includes words like name of places, Institutions, companies like Mysore, CIIL, Bajaj etc. NLOC includes the locations, adverb what we call traditionally adverb of time and place for example mEle (on), keLage(down), Iga (now)

- PRONOUN: In Kannada we have 4 type of pronoun namely Personal (PRP), Reflexive (PRF), Reciprocal (PRC), and Wh-word (PRQ)
- Pronominal includes nAnu (I), nInu (you), avanu (he) etc, Reflexive includes words like tAnu (self), swataha (self)) etc, Reciprocal includes paraspara (each other), obbarigobbaru (each other) etc and Wh-word includes words like yAru (who), Enu (what) etc

CONT.....

- DEMONSTRATIVE: In Kannada we have two types of Demonstratives like Deicic and Wh-word. Deictic includes A (that) and I (this), Wh-word includes yAva (which).
- VERB: In verb we have 2 types of verb like main and auxiliary, main verb has 5 types namely Finite, Non-finite, Infinitive, Gerund and finally participle noun. Auxiliary verb has 3 types namely Finite, non-finite and Infinitive.
- Finite main verb includes words like baMdanu (has come), hOdanuhas gone) etc, Non-finite forms includes baMdu (having come), hOgi(having gone) etc, Infinitive includes baralu (to come), hOgalu (to go) etc, Gerunds includes baruvudu (coming), hOguvudu(going) etc, Perticiple noun includes baruvavanu (coming person), hOguvavanu (going person) etc. Auxiliary verbs in Dravidian languages occur with the main verb usually. For example bahudu (may), bEku (should) etc.

CONT.....

- ADJECTIVE: Adjectives has no sub-type and it includes words like suMdaravAda (beautiful), kliSTha (difficult) etc.
- ADVERB: it includes nidhAnavAgi (slowely), jOrAgi (fastly) etc.
- POSTPOSITIONS: it includes locations like mEle (on), keLage (down), hinde (back), muMde (front) etc.

CONT.....

- CONJUNCTIONS: this is divided into 3 three namely co-ordinator, subordinator and quotative etc.
- Co-ordinator includes words like mattu (and), hAgU (and) etc. Subordinator includes words like AddariMda (therefore), hAgAgi (therefore) etc and quotatives are eMdu (that), anta (that) etc.
- PARTICLES: three sub categories in this section and they are Default, Interjection and Intensifier.
- Default includes kUda (also) etc. Interjections like ayyO, oh etc and Intensifier tuMba (very), bahaLa (many) etc.

CONT.....

- QUANTIFIERS: we have 3 types in this namely General, Cardinal and Ordinal.

  General includes ella (all), bahaLa (many) etc, Cardinal includes oMdu(one), eraDu(two) etc and ordinals includes oMdaneya (first), eraDaneya (second) etc.

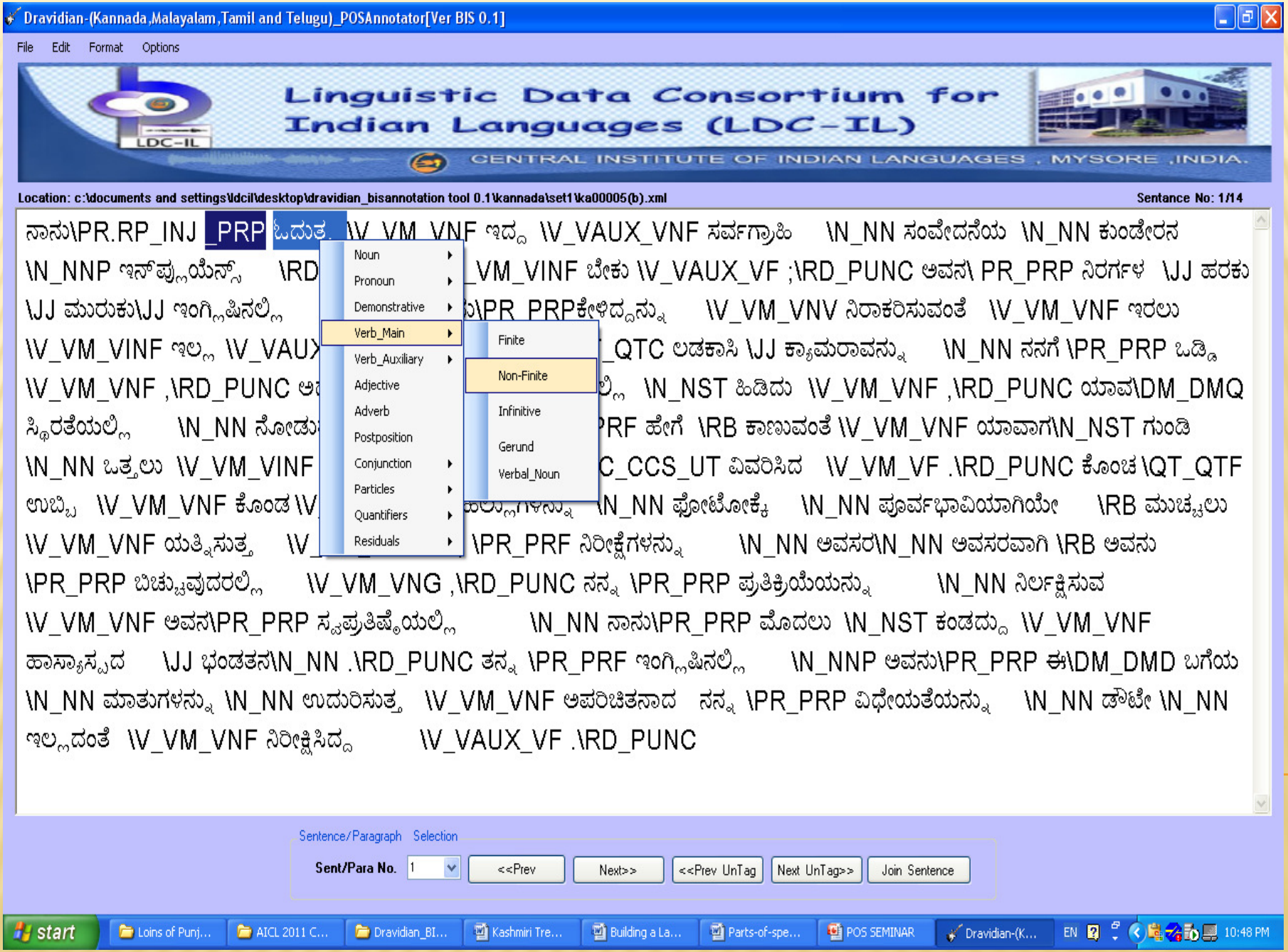
- RESIDUALS: it includes Foreign, Symbol, Punctuation, Unknown and Echowords.

  Foreign words usually includes book etc, symbol includes @,& etc. Punctuations like ?, < etc. Unknown includes ——— and finally Echowards includes mane gine (houses and other buildings), huli gili (tiger and others) etc.

KANNADA POS TAGGING

LDC-IL has developed annotation tool for POS tagging. It is a customizable manual tool that can be used to implement any tag set.

We have used this customized tool for implementing BIS Dravidian tag set for Kannada. In this work, we have used this tool for tagging the above mentioned preprocessed corpus.


Dravidian-(Kannada,Malayalam,Tamil and Telugu)_POSAnnotator[Ver BIS 0.1]
File Edit Format Options
Linguistic Data Consortium for Indian Languages (LDC-IL)
CENTRAL INSTITUTE OF INDIAN LANGUAGES , MYSORE ,INDIA.
Location: c:\documents and settings\ldcil\desktop\dravidian_bisannotation tool 0.1\kannada\set1\ka00005(b).xml
Sentance No: 1/14
ನಾನು\PR.RP_INJ _PRP ಓದುತ್ತ \V_VM_VNF ಇದ್ದ \V_VAUX_VNF ಸರ್ವಗ್ರಾಹಿ \N_NN ಸಂವೇದನೆಯ \N_NN ಕುಂಡೇರನ
\N_NNP ಇನ್‌ಪುಲ್ಸಯೆನ್ಸ್ \RD \V_VM_VINF ಬೇಕು \V_VAUX_VF ;\RD_PUNC ಅವನ\ PR_PRP ನಿರ್ಗಳ \JJ ಹರಕು
\JJ ಮುರುಕು\JJ ಇಂಗ್ಲಿಷಿನಲ್ಲಿ \PR_PRPಕೇಳಿದ್ದನ್ನು \V_VM_VNV ನಿರಾಕರಿಸುವಂತೆ \V_VM_VNF ಇರಲು
\V_VM_VINF ಇಲ್ಲ \V_VAUX QTC ಲಡಕಾಸಿ \JJ ಕ್ಯಾಮರಾವನ್ನು \N_NN ನನಗೆ \PR_PRP ಒಡ್ಡಿ
\V_VM_VNF ,\RD_PUNC ಅ ಲ್ಲಿ \N_NST ಹಿಡಿದು \V_VM_VNF ,\RD_PUNC ಯಾವ\DM_DMQ
ಸ್ಥಿರತೆಯಲ್ಲಿ \N_NN ನೋಡು RF ಹೇಗೆ \RB ಕಾಣುವಂತೆ \V_VM_VNF ಯಾವಾಗ\N_NST ಗುಂಡಿ
\N_NN ಒತ್ತಲು \V_VM_VINF C_CCS_UT ವಿವರಿಸಿದ \V_VM_VF .\RD_PUNC ಕೊಂಚ \QT_QTF
ಉಬ್ಬಿ \V_VM_VNF ಕೊಂಡ \V ಹಲ್ಲುಗಳನ್ನು \N_NN ಫೋಟೋಕ್ಕೆ \N_NN ಪೂರ್ವಭಾವಿಯಾಗಿಯೇ \RB ಮುಚ್ಚಲು
\V_VM_VNF ಯತ್ನಿಸುತ್ತ \V \PR_PRF ನಿರೀಕ್ಷೆಗಳನ್ನು \N_NN ಅವಸರ\N_NN ಅವಸರವಾಗಿ \RB ಅವನು
\PR_PRP ಬಿಚ್ಚುವುದರಲ್ಲಿ \V_VM_VNG ,\RD_PUNC ನನ್ನ \PR_PRP ಪ್ರತಿಕ್ರಿಯೆಯನ್ನು \N_NN ನಿರ್ಲಕ್ಷಿಸುವ
\V_VM_VNF ಅವನ\PR_PRP ಸ್ವಪ್ರತಿಷ್ಠೆಯಲ್ಲಿ \N_NN ನಾನು\PR_PRP ಮೊದಲು \N_NST ಕಂಡದ್ದು \V_VM_VNF
ಹಾಸ್ಯಾಸ್ಪದ \JJ ಭಂಡತನ\N_NN .\RD_PUNC ತನ್ನ \PR_PRF ಇಂಗ್ಲಿಷಿನಲ್ಲಿ \N_NNP ಅವನು\PR_PRP ಈ\DM_DMD ಬಗೆಯ
\N_NN ಮಾತುಗಳನ್ನು \N_NN ಉದುರಿಸುತ್ತ \V_VM_VNF ಅಪರಿಚಿತನಾದ ನನ್ನ \PR_PRP ವಿಧೇಯತೆಯನ್ನು \N_NN ಡೌಟೇ \N_NN
ಇಲ್ಲದಂತೆ \V_VM_VNF ನಿರೀಕ್ಷಿಸಿದ್ದ \V_VAUX_VF .\RD_PUNC
Noun
Pronoun
Demonstrative
Verb_Main
Verb_Auxiliary
Adjective
Adverb
Postposition
Conjunction
Particles
Quantifiers
Residuals
Finite
Non-Finite
Infinitive
Gerund
Verbal_Noun
Sentence/Paragraph Selection
Sent/Para No. 1
<<Prev
Next>>
<<Prev UnTag
Next UnTag>>
Join Sentence
start
Loins of Punj...
AICL 2011 C...
Dravidian_BI...
Kashmiri Tre...
Building a La...
Parts-of-spe...
POS SEMINAR
Dravidian-(K...
EN
10:48 PM

POS TAGGING ISSUES

1. Kannada has adverbial suffix which is responsible to make any word into adverb. For example: hasanAgi (cleanly), sukhavAgi (happily), nishcitavAgi (surely), AtmIyavAgi (closely), but there are other cases in Kannada where this Agi can come with the Proper noun like- rudranannu shivanannAgi kANuva kathegaLive (there are some stories where rudra is seen as shivA). If we tag it as adverb the important information like proper noun will be missed out in POS tagging.

CONT.....

2. baMdiruvavanannu (the person who has come +accusative case), baMdu + iruvavanannu here baMdu is main participle noun but iruvavanannu is an auxiliary with png and case marker. We are considering it as auxiliary. When we split this word into baMdu + iruvavanannu, this iru with png and case marker should be an auxiliary participle noun but this auxiliary participle noun is not mentioned in the Dravidian BIS tag set. Similarly the word like baMdiruvudannu we split this word into two tokens to give auxiliary information i.e iruvudannu this token has auxiliary verb with gerundial marker and case marker which is an auxiliary gerund but auxiliary gerund is not there in the Dravidian BIS tag set.

CONCLUSION

In this work we have summarized our experience of POS tagging of 10,000 words of Kannada corpus according to BIS standards. Moreover we have highlighted the problems which Dravidian languages face in general but Kannada in particular at the level of POS tagging because of their agglutinative nature.

REFERENCES

Schiffman, H. Sptember1979.Areference grammar of spoken Kannada.

Shridhar, S.N. Kannada (Descriptive Grammars).

T. N. Vikram and Shalini R. Urs. 2007. Development of prototype morphological analyzer for the south Indian language of kannada. In Proceedings of the 10th international conference on Asian digital libraries: looking back 10 years and forging newfrontiers, ICADL'07, pages 109–116, Berlin, Heidelberg. Springer-Verlag.

B. R Shambhavi, P Ramakanth Kumar, K Srividya, B J Jyothi, Spoorti Kundargi, and G Varsha Shastri. 2011. Kannada morphological analyser and generator using trie. IJCSNS International Journal of Computer Science and Network Security, 11(1), January.

Kavi Narayana Murthy. 2000. Computer processing of kannada language. Technical report, Computer and Kannada Development, Kannada University, Hampi.

P. V. S. Avinesh and G. Karthik. 2007. Part-Of-Speech Tagging and Chunking using Conditional Random Fields and Transformation-Based Learning. In Proceedings of the IJCAI and the Workshop On Shallow Parsing for South Asian Languages (SPSAL), pages 21–24.

P.J. Antony, M Anand Kumar, and K.P. Soman. 2010. Paradigm based morphological analyzer for kannada language using machine learning approach. Advances in Computational Sciences and Technology (ACST), 3(4).

P.J. Antony and K.P. Soman. 2010. Kernel based part of speech tagger for kannada. In Machine Learning and Cybernetics (ICMLC), 2010 International Conference on, volume 4, pages 2139 –2144, July.

THANK YOU