# Computational Model for Oriya Morphological Analyser

*Prof. (Dr.) S.Mohanty*

*Dr. R. C. Balabantaray*

# *Overview*

- Introduction
- Oriya Morphology
- Nominal Inflection
- Verbal Inflection
- Computational Model
- Implementation & Output
- Conclusion

# *Introduction*

- NLP has been developed in 1960,as a sub-field of Artificial Intelligence and Linguistics.
- Aim of NLP is studying problems in the automatic generation and understanding of natural languages.
- Oriya or Odia (      ) is an <u>Indian language</u>, belonging to the Indo-Aryan branch of the Indo-European language family.
- It is mainly spoken in the Indian state of Orissa
- In India, the language is spoken by over 31 million people, and globally over 45 million speak Oriya.

- It is thought to be directly descended from the Prakrit known as Purva Magadhi that was spoken in eastern India over 1,500 years ago.

- It bears a very strong resemblance to the Bangla (Bengali), Maithili, and Ôxômiya (Assamese).

- Oriya appears to be the least influenced by Persian and Arabic.

- Morphological Analysis (MA) is essential for a language to analyse and generate different types of word forms.

- It plays an important role in application areas of Natural Language Processing like Spell Checking , WordNet , Information Retrieving System, Tagged Corpora Generator, Machine Translation etc.

- A morpheme is defined as the minimal meaning-bearing unit in a language

- It can be divided into three types such as prefix, suffix and infix.

# *Oriya Morphology*

- Three major classes of morpheme in Oriya such as :

    Pronoun Morphology (PM)

    Inflectional Morphology (IM)

    Derivational Morphology (DM)

- PM is the study on the grammatical classification of pronoun.

- The pronoun ଆମ୍ଭେମାନେ *(AmbhemAne)* (we) indicates that it is personal pronoun, 1st person and plural number.

- IM is the combination of root word with grammatical morphemes, usually resulting in a word of the same class as the original stem and failing some syntactic function like agreement.

- inflectional morphemes ମାନେ *(mAne)* for making plural form (number), third person (person), subject (*kAraka*) and 1$^{st}$ inflection (inflection) etc., in case of nominal word.

- DM is the combination of a word stem (root word) with a grammatical morpheme, usually results a word of different class, whose meaning is hard to predict.

- a verb word ଖୁଆ *(khuA)* (feed) can take the derivational suffix i.e. ଇବା *(ibA)* to produce a nominal word as ଖୁଆଇବା *(khuAibA)* (feeding).

# *Nominal Inflection*

- There are near about 60 morphemes for nominal words in Oriya language.

- Ex: the word କଲମଗୁଡିକ *(kalamguDika)* (pens) inflected the root word କଲମ *(kalam)*(pen) and suffix ଗୁଡିକ *(guDika)* (Suffix), which indicates that the word କଲମଗୁଡିକ *(kalamguDika)* (pens) is a Noun, Plural (number), Subject (Case-ending relation) and 1$^{st}$ inflection (inflection).

## Classes of Nominal Suffix in Oriya Language.

| ବିଭକ୍ତି (Inflection) | ଏକବଚନ (Singular) | ବହୁବଚନ (Plural) | କାରକ (Case-Relationship) | ଉପପଦ (Non-Case Relationship) |
|---|---|---|---|---|
| ପ୍ରଥମା (1st Inflection) | ଏ, ଟାଏ, ଟିଏ,କ୍ଷ ଟା, ଟି, ଟେ | ଏ,ମାନେ,ମାନ,ଗୁଡିକ, ଗୁଡାକ, ଗୁଡିଏ | କର୍ତ୍ତା (Subjective) | |
| ଦ୍ୱିତୀୟା (2nd Inflection) | କୁ,ଙ୍କୁ,କି,ଙ୍କି,ଟାକୁ, ଟିକୁ,ଟେ,ଠାକୁ,ଠିକୁ, ଠିକି | ଙ୍କୁ, ମାନଙ୍କୁ, ଗୁଡିକୁ, ଗୁଡାକୁ | କର୍ମ (Objective) | |
| ତୃତୀୟା (3rd Inflection) | ରେ,ଦ୍ୱାରା,ଦେଇ, ଙ୍କଦ୍ୱାରା | ଙ୍କଦ୍ୱାରା,ମାନଙ୍କଦ୍ୱାରା, ମାନଙ୍କଦେଇ,ମାନଙ୍କରେ | କରଣ (Instrumental) | |
| ଚତୁର୍ଥୀ (4th Inflection) | କୁ,ଙ୍କୁ,କି,ଙ୍କି,ପାଇଁ,ଲାଗି, ଙ୍କଲାଗି,ନିମନ୍ତେ, ଙ୍କନିମନ୍ତେ,ଙ୍କଯୋଗୁଁ, ସକାଶେ,ଙ୍କସକଶେ | ଙ୍କୁ, ମାନଙ୍କୁ, ଗୁଡିକୁ, ଗୁଡାକୁ, ମାନଙ୍କଲାଗି, ମାନଙ୍କନିମନ୍ତେ, ମାନଙ୍କଯୋଗୁଁ, ମାନଙ୍କସକାଶେ | ସମ୍ପ୍ରଦାନ (Dative) | |
| ପଞ୍ଚମୀ (5th Inflection) | ରୁ, ଠାରୁ, ଠୁଁ, ଠଁ, ଠଠୁଁ, ଙି, ଙ୍, ଠୁଁ | ଙ୍କଠାରୁ,ମାନଙ୍କଠାରୁ ମାନଙ୍କଠୁଁ,ଙ୍କଠାରୁ,ମାନଙ୍କଠଁ,ମାନଙ୍କଠୁ | ଅପାଦାନ (Ablative) | |
| ଷଷ୍ଠୀ (6th Inflection) | ର, ଙ୍କ, ଙ୍କର, ରି, ଙ୍କରି | ମାନଙ୍କ,ମାନଙ୍କର, ମାନଙ୍କରି, ଙ୍କରି | | ସମ୍ବନ୍ଧ (Genitive) |
| ସପ୍ତମୀ (7th Inflection) | ରେ, ଠାରେ, ଏ, ଠାଇଁ, ଠି, ଠେଇଁ, ଠଦ୍ୱି, ଠି | ମାନଙ୍କରେ, ମାନଙ୍କଠାରେ, ମାନଙ୍କଠାଇଁ, ମାନଙ୍କଠି, ମାନଙ୍କଠେଇଁ | ଅଧିକରଣ (Locative) | |

# *Verbal Inflection*

- There are near about 100 morphemes for verbal words in Oriya language.

- Most of the Oriya verbal word follow similar pattern of their suffixes except few one.

- For example the verbal word ଯାଇଥିଲି *(JAithiLi)* (had gone) inflected the root verb ଯା *(JA)* (go) and suffix ଇଥିଲି *(ithiLi),* which is indicated that the verbal word ଯାଇଥିଲ *(JAithiLi)* is past tense, 1$^{st}$ person and singular number.

Classes of Verbal Suffix in Oriya Language.

| Tense | Person (ପୁରୁଷ) | Singular Suffix (ଏକ ବଚନ) | Plural suffix (ବହୁ ବଚନ) |
|---|---|---|---|
| **Present Tense** (ବର୍ତ୍ତମାନ କାଳ) | 1st Person (ପ୍ରଥମ ପୁରୁଷ) | ଉଅଛି. ରଛି. ରଥାନ୍ତି. ଏ | ଉଅଛୁ. ରଛୁ. ଇଅଛୁ.ଇଛୁ.ରଥାନ୍ତୁ. ଉ |
| | 2nd Person (ଦ୍ୱିତୀୟ ପୁରୁଷ) | ଉଅଛୁ. ରଛୁ. ରଥାନ୍ତୁ. ଉ | ଉଅର. ରର. ଉଥାନ୍ତ |
| | 3rd Person (ତୃତୀୟ ପୁରୁଷ) | ଉଅଛି. ରଛି. ରଅଛନ୍ତି. ଉଥାନ୍ତା. ରଥାନ୍ତେ.ଉଥାଆନ୍ତେ.ନ୍ତି. ରଆନ୍ତା | ରଛନ୍ତି.ଉଅଛନ୍ତି. ରଥାଆନ୍ତି. ନ୍ତି |
| **Past Tense** (ଅତୀତ କାଳ) | 1st Person (ପ୍ରଥମ ପୁରୁଷ) | ଇଲି. ଇଛି. ଇଥିଲି. ରଥିଲି. ଇଥାନ୍ତି. ଇଅଛି | ଗଲୁ. ଇଥିଲୁ. ରଥିଲୁଁ. ଇଥାନ୍ତୁ |
| | 2nd Person (ଦ୍ୱିତୀୟ ପୁରୁଷ) | ଇଲୁ. ଇଛୁ. ଇଥିଲୁ. ରଥିଲୁ. ଇଥାନ୍ତୁ | ଇଲ. ଇର. ଇଅଲ.ଇଥିଲ. ରଥିଲ. ଇଥାନ୍ତ |
| | 3rd Person (ତୃତୀୟ ପୁରୁଷ) | ଇଲା. ଇଲେ. ଇଛି. ଇଛନ୍ତି. ଇଥିଲେ. ଇଥିଲା. ରଥିଲେ. ଇଥାଆନ୍ତା. ଇଥାଆନ୍ତେ | ଇଲେ. ଇଛନ୍ତି. ଇଥିଲେ. ରଥିଲେ. ଇଥାଆନ୍ତେ |
| **Future Tense** (ଭବିଷ୍ୟତ କାଳ) | 1st Person (ପ୍ରଥମ ପୁରୁଷ) | ଇବି. ଇଥିବ. ଥାନ୍ତି. ରଥିବି | ଇବୁ. ଇଥିବୁ. ଥାଆନ୍ତୁ. ରଥିବୁ.ଇବା |
| | 2nd Person (ଦ୍ୱିତୀୟ ପୁରୁଷ) | ଇବୁ. ଇଥିବୁ. ଥାଆନ୍ତୁ. ରଥିବୁ | ଇବ. ଇଥିବ.ଥାଆନ୍ତୁ.ରଥିବ |
| | 3rd Person (ତୃତୀୟ ପୁରୁଷ) | ଇବ. ଇବେ. ଇଥିବେ.ଇଥିବ.ଥାଆନ୍ତ.ଥାଆନ୍ତେ.ରଥିବ.ରଥିବେ | ଇବେ. ଇଥିବେ.ଥାଆନ୍ତେ.ଇଥିବେ |

# *Computational Model*

- The lexical resources implemented for the development of MA are as follows:

  - Development of reliable electronic dictionary (contains root words, their Parts-of-Speech and semantic information) with efficient searching technique.

  - Generation of nominal suffix table.

  - Collection of orthographic rules and grammar rules to model the changes that occur in a word.

  - Generation of syntax rules based on KARAKA theory.

- NFA For Nominal case can be defined as follows:

A Nondeterministic Finite State Automaton is a quintuple M= ( Q, $\Sigma^*$, $\delta$, q0,F ) Where

Q = Finite set of states { q0,q1,q2, .........,1s,2s......7s, 1p, 2p,........7p }

(1 .. 7 = Numbers are for inflection and s and p are for singular and plural respectively. )

$\Sigma^*$ = Finite set of input symbols { ଏ,ୟ, ଟାଏ, ଡିଏ, ଟ, ଠା, 6ଠ, ଠ̂ , ଗୁଡିଏ, ମାନ,ଦ୍ୱାରା, · · · · }

q0 = A member of Q which is the initial state.

F = A subset of Q , is the set of final states { 1s, .........7s, 1p............7p }

δ= The transition function which takes a state from Q and an input symbol from Σ and returns a set from Q again i.e a function from QxΣ* to Q.

- Oriya word is processed by the system taking characters from right to left.

◈ For example lets take ବାଳକଟିଏ (*bALakTie*)(a boy)/ ବାଳକଗୁଡିଏ (*bALakguDie*) (boys) for analysis.

$$(q0, \text{ବାଳକଟିଏ}) \vdash_M (1s, \text{ବାଳକଟି})$$

$$\vdash_M (1s, \text{ବାଳକ})$$

$$(q0, \text{ବାଳକଗୁଡିଏ}) \vdash_M (1s, \text{ବାଳକଗୁଡି})$$

$$\vdash_M (1p, \text{ବାଳକଗୁ})$$

$$\vdash_M (1p, \text{ବାଳକ})$$

The state transition table for FSA for ବାଳକଟିଏ (bALakTie) / ବାଳକଗୁଡିଏ(bALakguDie)

| Q | Σ | δ (Q, σ) |
|---|---|---|
| q0 | ଟ | q1 |
| q0 | ଏ | { 1s, 7s } |
| q0 | ଠ | q2 |
| q1 | ଟୋ | 1s |
| q2 | ଟ | 1s |
| 1s | ε | 1p |
| 1s | ଟା | 1s |
| 1s | ଟିଏ | 1s |
| 1s | ଗୁଡିଏ | 1p |

Output of the Oriya Morphological Analyser (OMA)

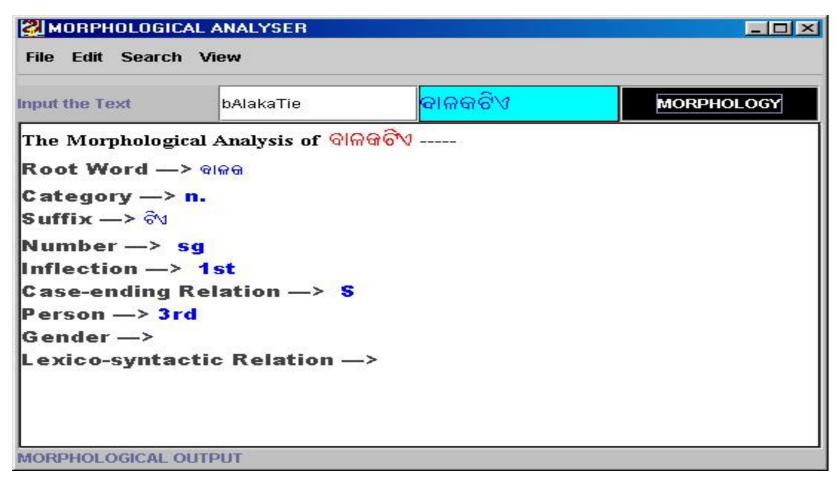| Input | Morphological Analyser Output |
|-------|-------------------------------|
| ଆଳକଟିଏ | ବାଳକ $+N+s+1^{st}+S+3P$ |
| ବାଳକଗୁଡିଏ | ବାଳକ $+N+p+1^{st}+S+3P$ |
| କଲମଗୁଡିକ | କଲମ $+N+p+1^{st}+S+3P$ |
| ବିଦ୍ୟାଳୟରେ | ବିଦ୍ୟାଳୟ $+N+s+1^{st}+I+3P$ (or) ବିଦ୍ୟାଳୟ $+N+s+7th+L+3P$ |
| ବାଳକମାନଙ୍କୁ | ବାଳକ $+N+p+2^{nd}+O+3P$ (or) ବାଳକ $+N+p+4^{th}+D+3P$ |

A portion of state diagram for the Nominal case in Oriya Morphology.

# *Output*

# *Conclusion*

- Development of Morphological Analyser for all types of word forms is a challenging task for agglutinative language like Oriya.

- The techniques developed for this OMA is based on syntactic rule of Sanskrit Language.

- The implementation is going ahead to incorporate more and more lexical information of Oriya language.

- Addition of semantic feature will solve the problems more effectively which is under implementation phase at this moment.

*Thank You*