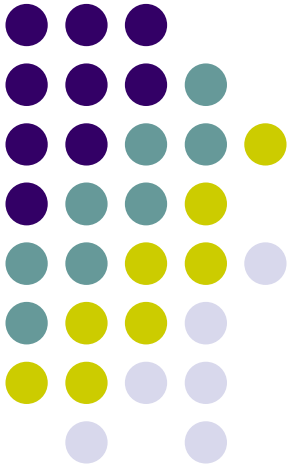
An Integrated approach for developing Morphological analyzer & Computational grammar using a POS tagger

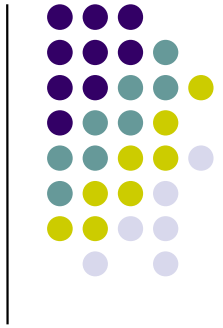


Presented By:

Pallav Kr Dutta

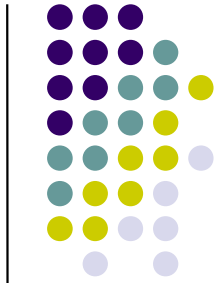
Indian Institute of Technology Guwahati

Topics



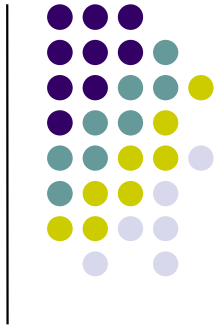
- Motivation
- Basic Morphology of the languages
- Our Approach
- Generation of Computational Grammar
- Future work
- References

Motivation



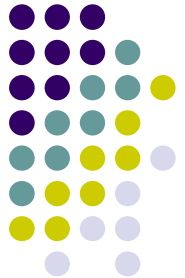
- Lack of Research in the field of Computational Linguistic for NE languages.
- Lack of significant tools for NE languages.
- Dependence of other NLP activities on POS tagging
- Lack of electronic resources.

Basic Morphology



- Assamese
  - Indo-Aryan group
  - 23 consonants, 8 vowels
  - Assamese script
- Bodo
  - Tibeto-Burman group
  - 16 consonants, 6 vowels
  - Devanagari script
- Basic POS :
  - Noun (বিশেষ্য)
  - Pronoun (সর্বনাম)
  - Verb (ক্রিয়াপদ)
  - Adverb (ক্রিয়া বিশেষণ)
  - Adjective (বিশেষণ)
  - Conjunctions (অব্যয় পদ)

Basic Morphology

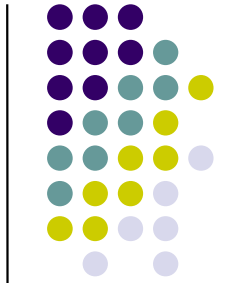


- Noun, Pronoun, Verb, Adverb, and Adjective are Inflected words and conjunction is Indeclinable.

- **Noun :** Noun is the important part of parts of speech. Maximum words are derived from this part. To inflect, all the base words of Nouns are dependent on case.

- **Verb :** Verbs are Inflects based on tense and person.

Basic Morphology

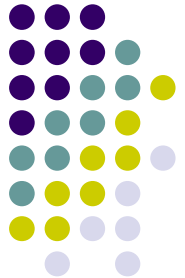


- **Adjective:** Assamese Adjective is basically not inflected. Bodo has derived adjectives

- **Adverb:** Assamese Adverbs are dependent on Adjective. When Adjective words take verbal form then the word becomes Adverb. Bodo Adverbs are derived from verb using suffix.

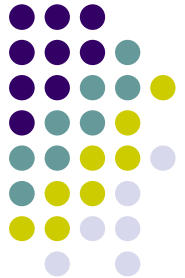
- **Conjunction:** Conjunction is indeclinable.

Basic Morphology
Number and Gender……



- Assamese noun morphology includes number and gender
- Singular and plural number
- Plurals are formed by using suffixes

Basic Morphology
Classifier and Postposition…

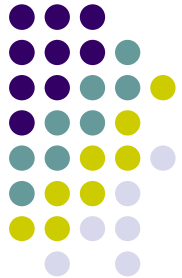


- Classifier is suffix to numerals
- Postpositions are used after nouns, pronouns and verbs

Basic Morphology
Case and Case markers

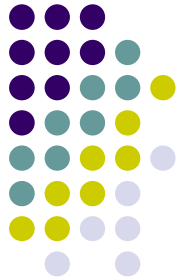
- Case is related to Noun.
- Nouns are inflected for case
- Each case has its own marker
  - **ACC**: ক /k/ **GEN**: ৰ /r/ **LOC**: ত /t/ **DAT**: লৈ/loi/ etc.
- Markers are always added at the end the word

Basic Morphology
Tense and Tense marker…



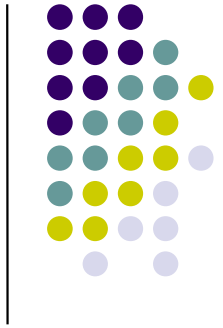
- Three tenses – Present , Past, Future
- Tense marker changes on person

Our Approach



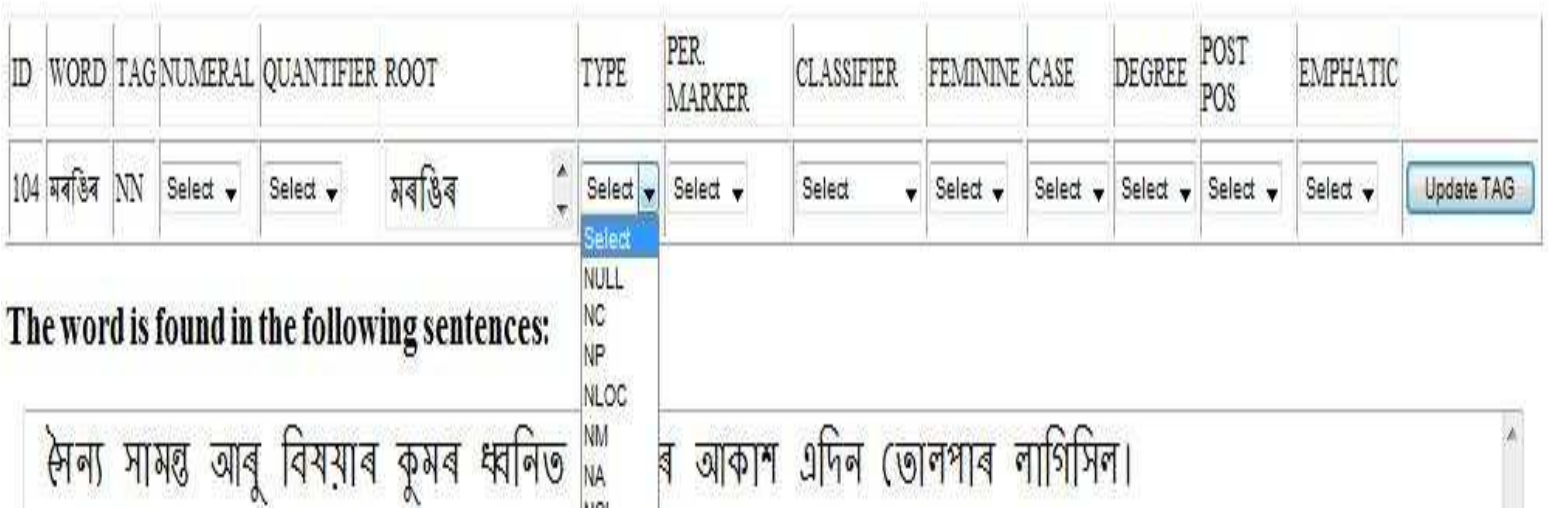
- Used the a-priori knowledge of Language experts.
- Morphological analysis of words in a base corpus of core set of words.
- Words in the corpus are analyzed, broken down and tagged by a team of experts.
- Database is used for storing tagged words.
- This database is used as a resource in the second stage where a large body of native speakers assists the system in tagging a larger corpus.

Our Approach



- One word has to be tagged once only.
- The lexical and grammatical rules are generated during tagging.
- Rely on pattern matching of new instances with what is already available with native speakers providing verification.
- Less involvement of Language Experts.
- Fast and less time and cost involvement .

Snapshot of POS tagger



Generation of Computational Grammar

- Used Bottom up approach
  - Taken POS tagger as the primary process of corpus generation
  - Starting from individual words to sentences
  - A-priori knowledge of the language

- Gathers the lexical and grammatical rules, e.g.
  - morphological composition of Noun & verb finite main (VFM)

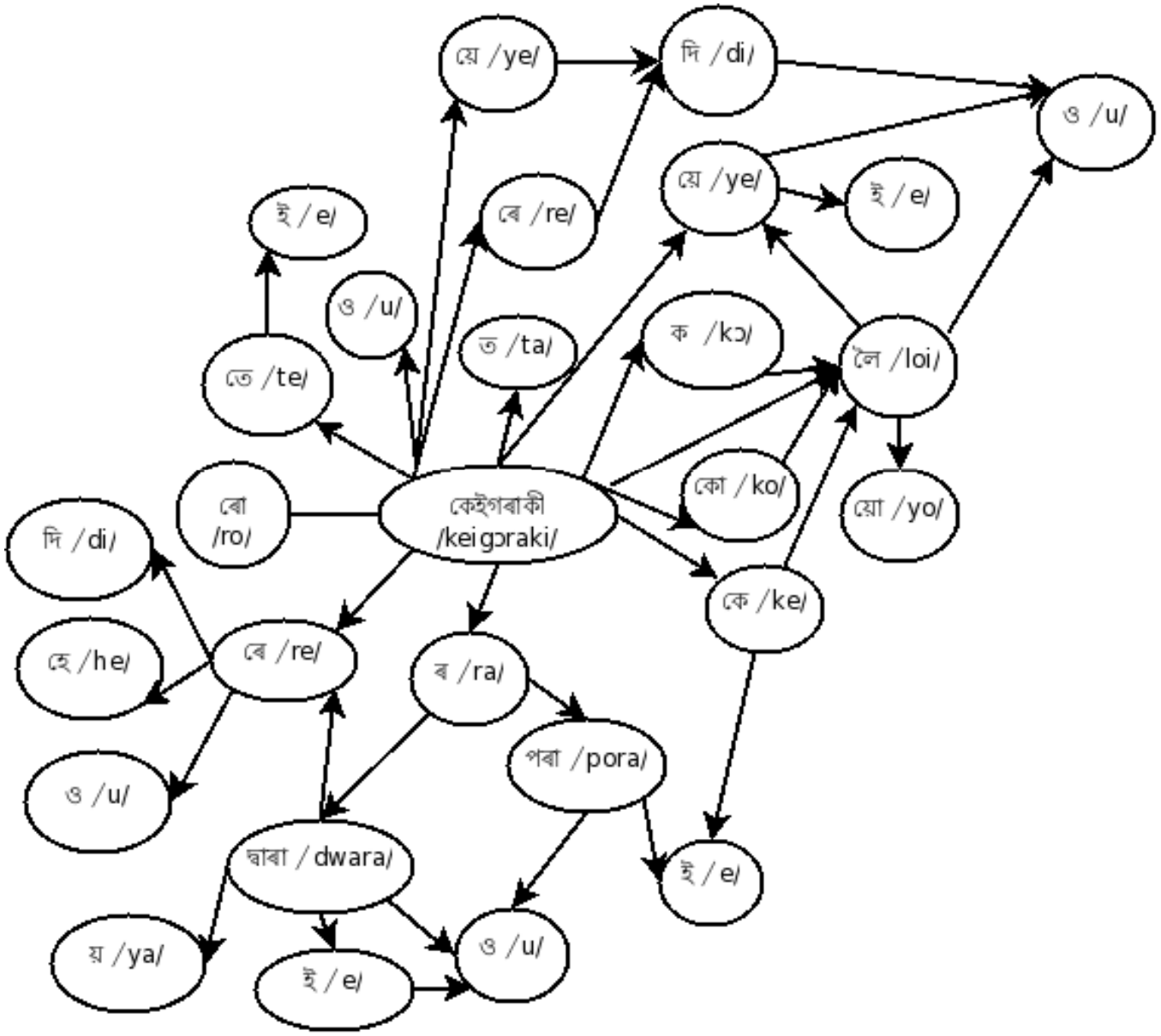
Numerical Prifix	Quantifier	Root Word	Noun Type	Personal Marker	Classifier Suffix	Feminine Suffix	Case Suffix	Degree Suffix	Post Position	Emphatic Suffix

Root Verb	Verb Type	Causative Suffix	Aspect	Tense	Person	Imperative	Mood

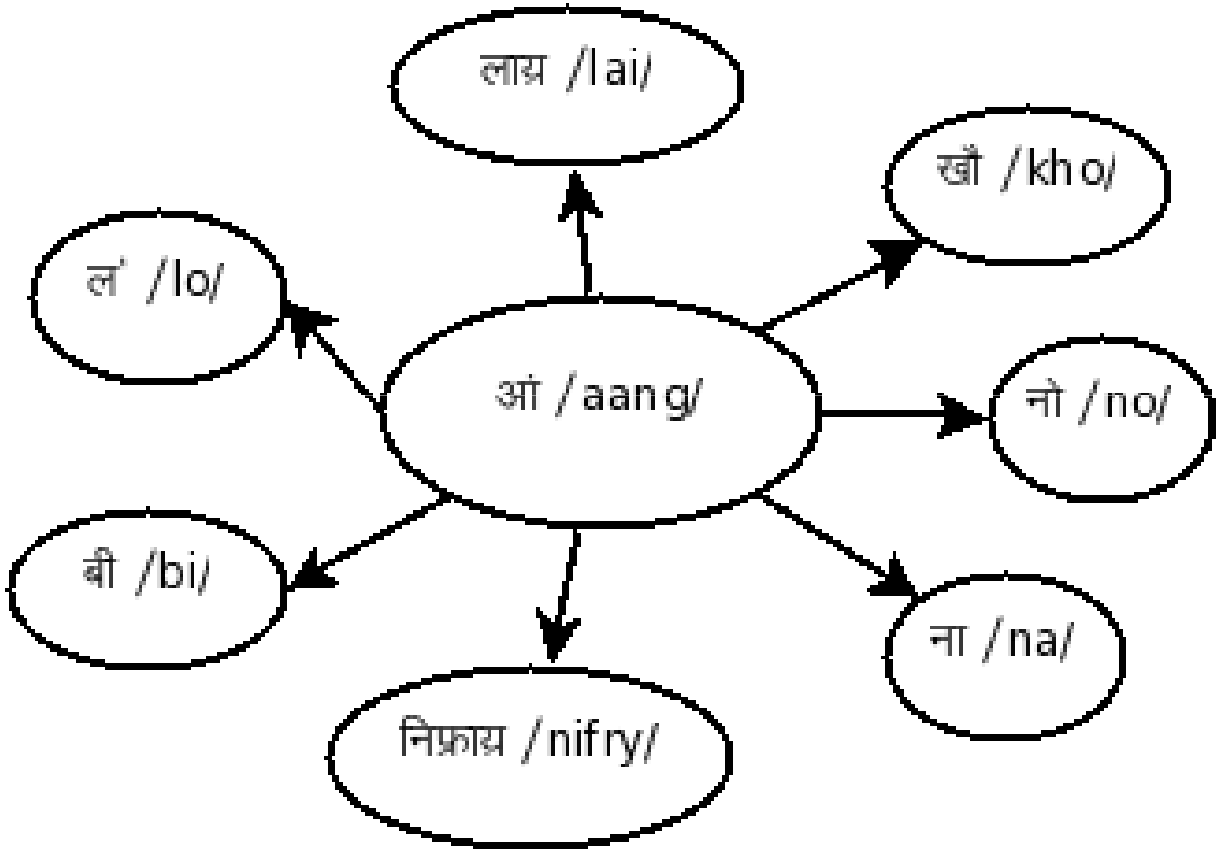
Generation of Computational Grammar

- Suffixation – a major morphological phenomenon.
- Derived, inflected and concatenative forms of suffixation are available.
- Suffixation is an iterative process.
- Uniqueness of Assamese – presence of personal marker ৰা, ra; eg. দেউতাৰা your father /deutara/).

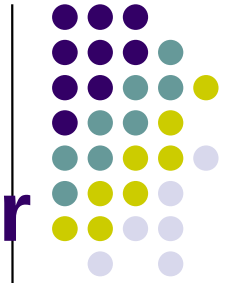
Assamese pattern of Suffixation



Bodo Pattern of Suffixation

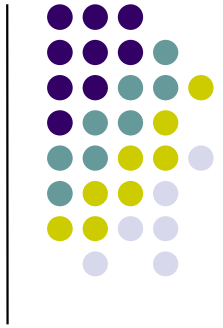


Generation of Computational Grammar



- Limited prefixes are available in Assamese.
- Negation of Verbs using a prefix ন /n/ and its variants like নো, না, নে, নি (/nɔ/, /na/, /ne/, /ni/)

Rules for word analysis



- W → PRS | PR|RS|R
- P → Numeral_prefix| Quantifier|উ /u/|অনা/ana/|আও /aau/| নি /ni/| 20 prefixes

  known to be originating from Sanskrit |eps
- S → sS|Ss| s| eps
- Numeral_prefix → এ /a/ |দু /du/ | esp
- Quantifier → ইমান /eman/ |esp

- Where W: word; P: prefix non-terminal; R: root word ; S: suffix non-terminal; s: suffix terminal; eps: epsilion;

Example of Suffixation

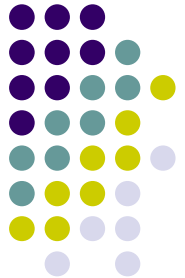
- মানুহকেইজনৰপৰাও [manuhkeijɔnɔrpɔrau]

মানুহ-কেইজন-ৰ-পৰা-ও [manuh-keijɔn-ɔr-pɔra-u]

Man-Pl.Classifier-Possessive-from(Post Position)-Emphatic

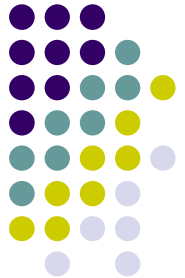
- W→ RS [মানুহকেইজনৰপৰাও] /manuhkeijɔnɔrpɔrau/
- W→ RsS [মানুহ-কেইজনৰপৰাও] /manuh-keijɔnɔrpɔrau/
- W→ RssS [মানুহকেইজন-ৰপৰাও] /manuhkeijɔn-ɔrpɔrau/
- W→ RsssS [মানুহকেইজনৰ-পৰাও] /manuhkeijɔnɔr-pɔrau/
- W→ RssssS [মানুহকেইজনৰপৰা-ও] /manuhkeijɔnɔrpɔra-u/
- W→ Rssssesp [মানুহকেইজনৰপৰাও] /manuhkeijɔnɔrpɔrau/
- W→ Rssss [মানুহকেইজনৰপৰাও] /manuhkeijɔnɔrpɔrau/

Example of Prefixing



- নাখাও [nakhau] { না - খা -ও [na-kha-u]}
- W→ PRS [নাখাও] /nakhau/
- W→ PRS [না-খাও] /na-khau/
- W→ PRS [নাখা-ও] /nakha-u/
- W→ PRsS [নাখা-ও] /nakha-u/
- W→ PRsesp [নাখাও] /nakhau/
- W→ PRs [নাখাও] /nakhau/

Grammar Rules



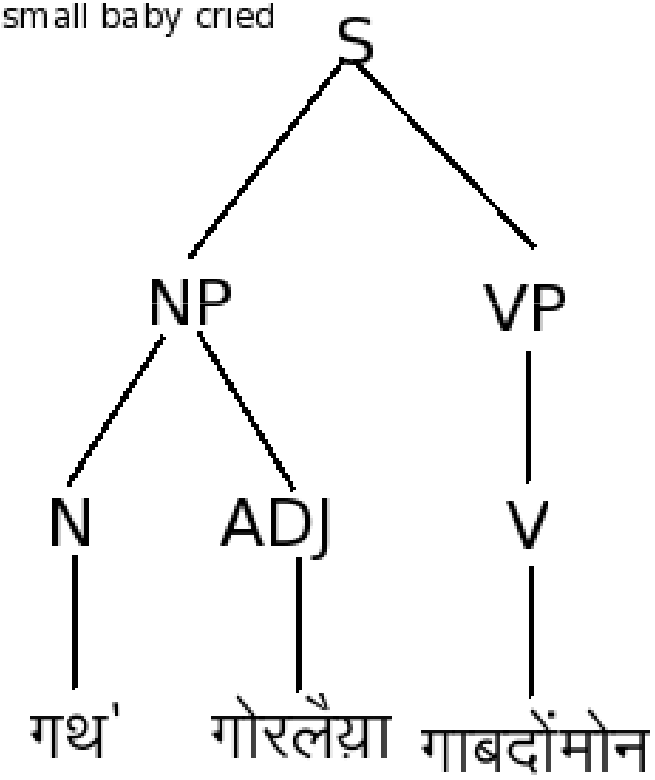
- S → S Conjunctions S | NP VP |NP|VP|
- NP → AdjP NP| NP Adj | Adj| Pronoun NP| NP Pronoun | Pronoun| Pronoun Adj| NP NP| NP
- VP → v NP |NP v| v Aux| v AdjP|Adj VP| Adv VP| AdvP VP|VP VP
- AdjP→ NP AdjP| VP Adj
- AdvP→ Adv Adv | Adj Adv

Example- 1

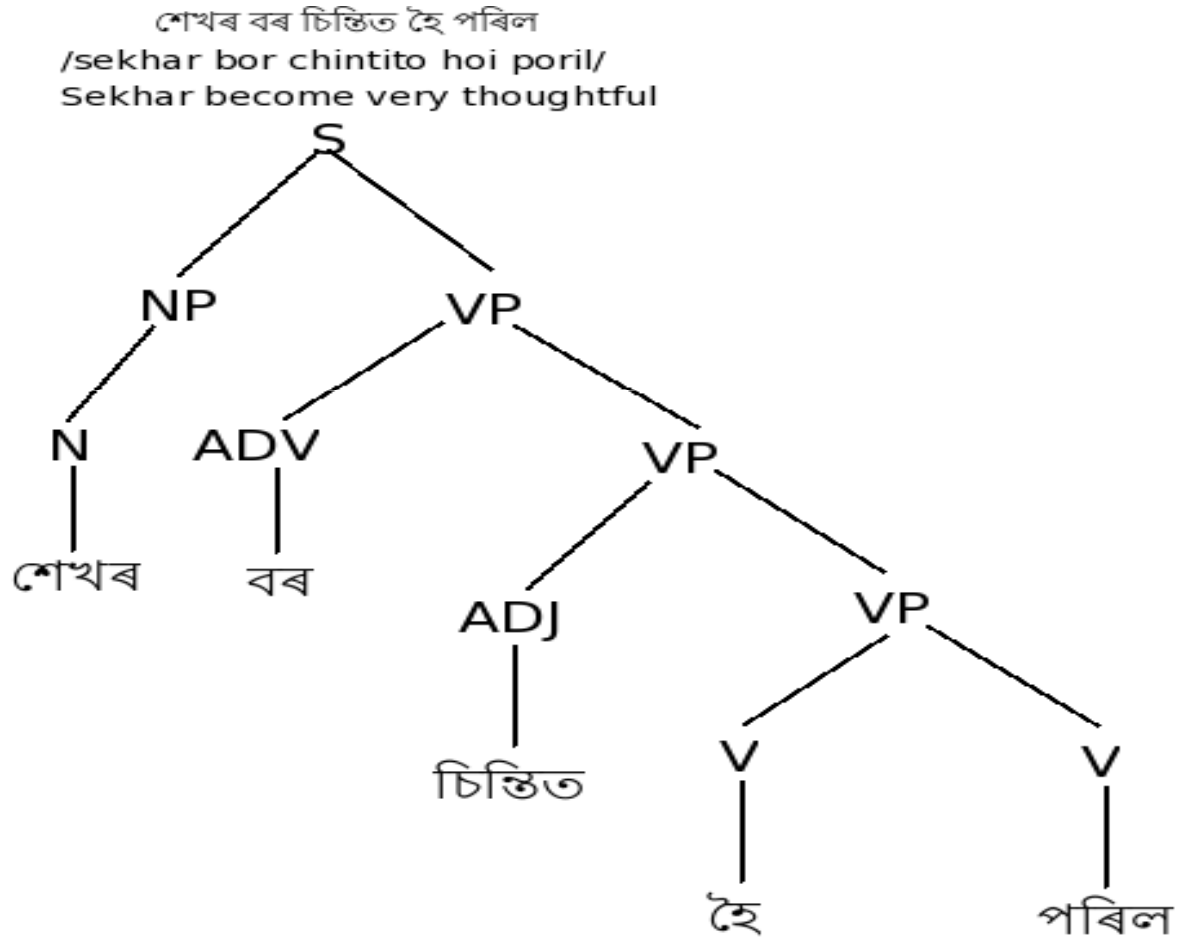
गथ' गोरलैंया गाबदोंमोन (गाब/Verb-दों/Aspect-मौन/suffix in past tense)

/Gotho gorlaiya gabdongmon/

The small baby cried

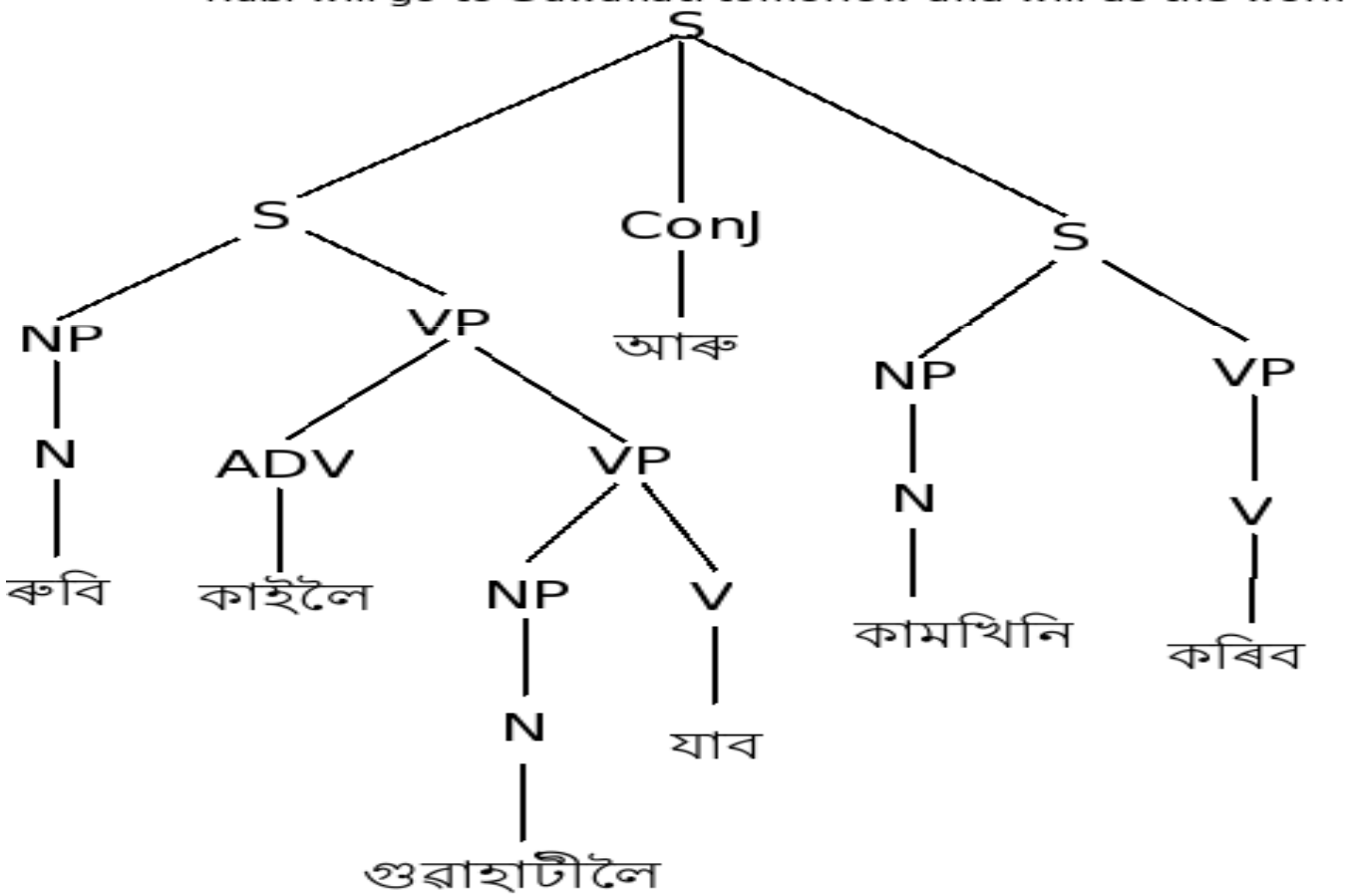


Example- 2

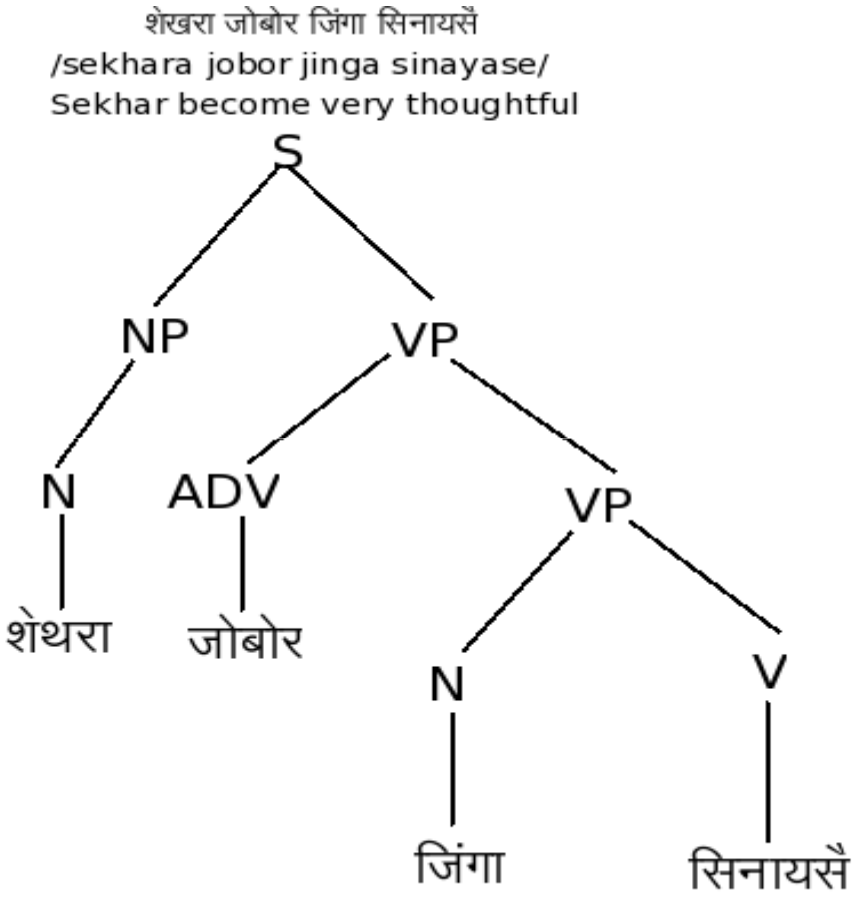

শেখৰ বৰ চিন্তিত হৈ পৰিল
/sekhar bor chintito hoi poril/
Sekhar become very thoughtful
S
NP
VP
N
ADV
VP
শেখৰ
বৰ
ADJ
VP
চিন্তিত
V
V
হৈ
পৰিল

Example- 3

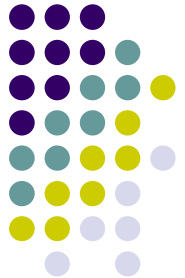
ৰুবি কাইলৈ গুৱাহাটীলৈ যাব আৰু কামখিনি কৰিব
/Rubi kailoi guwahatiloi jabo aru kamkhini koribo /
Rubi will go to Guwahati tomorrow and will do the work



Example- 4


शेखरा जोबोर जिंगा सिनायसैं
/sekhara jobor jinga sinayase/
Sekhar become very thoughtful
S
NP
VP
N
ADV
VP
शेथरा
जोबोर
N
V
जिंगा
सिनायसैं

Future Work



- Developing the complete grammar
- A common tag set to handle language specific issues

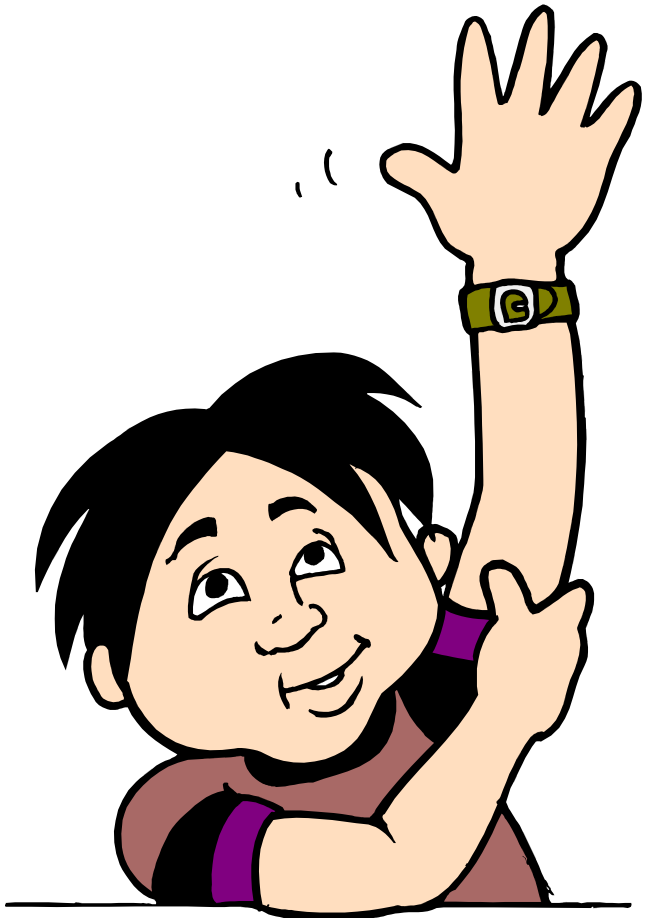
References:

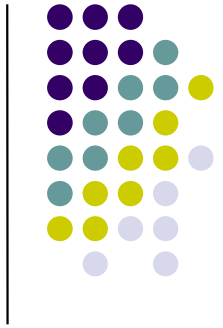
- Ms Lilabati Saikia Bora ,Asamiya Bhasar Ruptattva- First Edition, January 2006,published by M/s Banalata, Panbazar Guwahati-1. (ISBN: 81-7339-466-0)
- Satyanath Borah- Bahal Vyakaran, April 2006 Ed, published by B.C. Barua on behalf of Gopal Barua Agency, S.B Road, Guwahati-1
- Prof. G. C. Goswami, Asamiya Vyakaran Pravesh- Second Edition, April 2003, published by M/s Bina Library, Guwahati-1.
- M das, S Borgohain J Gogoi, 2002. “Design and Implementation of a Spell Checker for Assamese”, Proceedings of the Language Engineering Conference (LEC’02).
- Baskaran, S Et al., January 2008. "Designing a common POS-Tagset Framework for Indian Languages", Proceedings of the 6th Workshop on Asian Language Resources (ALR 6), Hyderabad., pp. 89

References contd…

- Shrivastava, Agrawal, Mohapatra, Singh and Battacharya, April 2005. "Morphology based Natural Language Processing tool for Indian languages", paper presented in the 4th Annual Inter Research Institute Student Seminar in Computer Science (IRISS05), IIT Kanpur.
- A. Dalal, K. Nagaraj, U. Swant, S. Shelke and P. Bhattacharyya, 2007. “ Building Feature Rich POS Tagger for Morphologically Rich Languages: Experience in Hindi “ . ICON.
- Sirajul Islam Choudhury, L. Sarbajit Singh, S. Borgohain and P. K. Das, 2004. "Morphological Analyzer for Manipuri: Design and Implementation", Applied Computing, pp. 123-129.
- T. N. Vikram and Shalini R. Urs. 2004. “Development of Prototype Morphological Analyzer for the South Indian Language of Kannada” ICADL 2007, LNCS 4822, pp 109-116, 2007
- http://www.iitg.ernet.in/rcilts/newassamesedesign.pdf

Questions ???





गोजोन्थों

ধন্যবাদ

Thanks