# POS Tagset Design, Tagging Status and Challenges Tagset and Tagging Nepali Corpus

Bal Krishna Bal

Project Manager

PAN Localization Project

Madan Puraskar Pustakalaya, Nepal

URL : www.madanpuraskar.org

Email: bal@mpp.org.np

# Contents

- Background to the development of the Part-of-Speech Tagset for Nepali
- Part-of-Speech Tagset for Nepali, an overview
- List of Part of Speech Tags for Nepali
- Tagging Nepali Corpus
- Tagging issues
- Conclusion

# Background to the development of the Part-of-Speech Tagset for Nepali

- Part-of-speech Tagset development – a crucial endeavor.
- Size plays a vital role, generally the smaller the Tagset, the greater the accuracy in the tagging process.
- However, compulsory categories evident in the language also should not be missed.
- Hence , a middle ground has been adopted for Nepali.
- The Nepali Part-of-Speech Tagset consists of 43 tags and covers almost all of the categories in the language.

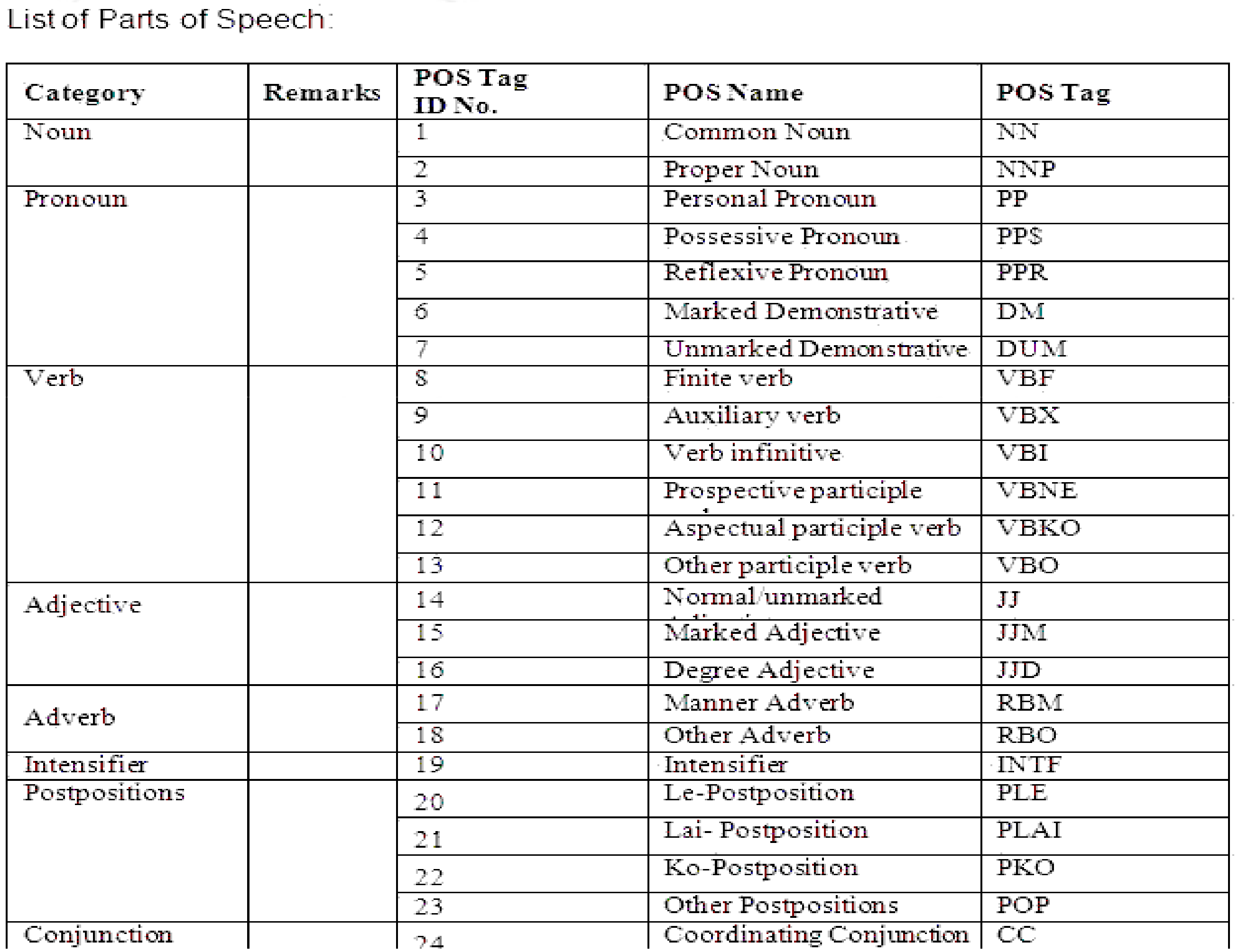
# Background to the development of the Part-of-Speech Tagset for Nepali …

- While developing the Tagset, the Penn Treebank Tagset has been consulted;
- Wherever applicable and practicable, the tag namings have been kept the same as that with the Penn Treebank Tagset;

# Part-of-Speech Tagset for Nepali, an overview

- Nepali – a morphologically rich language.
- Gender, Number and Person have a distinct role in pronouns, adjectival and verbal inflections;
- However, these distinctions have not been taken into consideration in the tagset;
- Honorificity, which is an important aspect of Nepali pronouns also has not been considered in the tagset;
- Verbal conjugation is also not taken into consideration in the tagset.

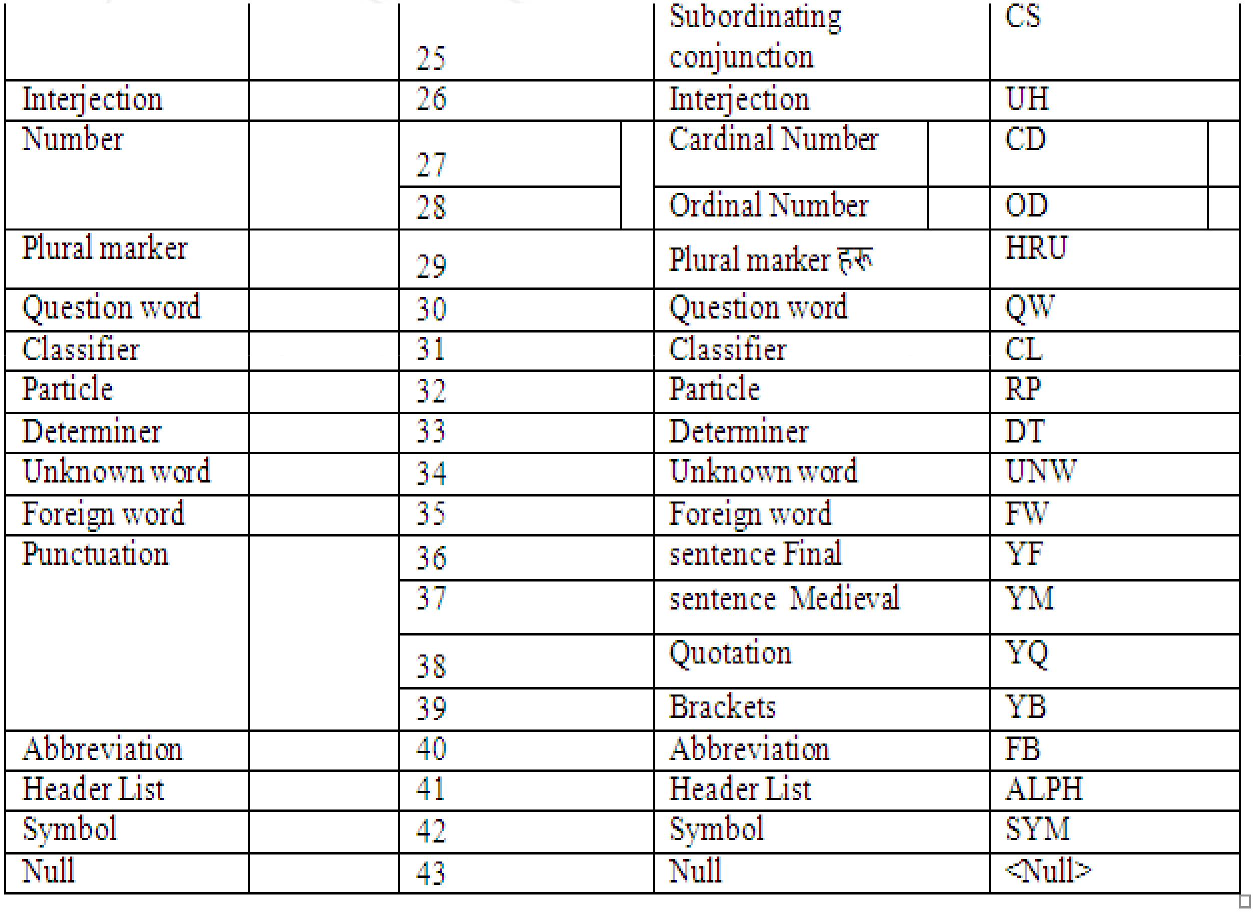
# List of part-of-speech tags for Nepali



List of Parts of Speech:

| Category | Remarks | POS Tag ID No. | POS Name | POS Tag |
|---|---|---|---|---|
| Noun | | 1 | Common Noun | NN |
| | | 2 | Proper Noun | NNP |
| Pronoun | | 3 | Personal Pronoun | PP |
| | | 4 | Possessive Pronoun | PP$ |
| | | 5 | Reflexive Pronoun | PPR |
| | | 6 | Marked Demonstrative | DM |
| | | 7 | Unmarked Demonstrative | DUM |
| Verb | | 8 | Finite verb | VBF |
| | | 9 | Auxiliary verb | VBX |
| | | 10 | Verb infinitive | VBI |
| | | 11 | Prospective participle | VBNE |
| | | 12 | Aspectual participle verb | VBKO |
| | | 13 | Other participle verb | VBO |
| Adjective | | 14 | Normal/unmarked | JJ |
| | | 15 | Marked Adjective | JJM |
| | | 16 | Degree Adjective | JJD |
| Adverb | | 17 | Manner Adverb | RBM |
| | | 18 | Other Adverb | RBO |
| Intensifier | | 19 | Intensifier | INTF |
| Postpositions | | 20 | Le-Postposition | PLE |
| | | 21 | Lai- Postposition | PLAI |
| | | 22 | Ko-Postposition | PKO |
| | | 23 | Other Postpositions | POP |
| Conjunction | | 24 | Coordinating Conjunction | CC |

# List of part-of-speech tags for Nepali…



| | | | | |
|---|---|---|---|---|
| | | 25 | Subordinating conjunction | CS |
| Interjection | | 26 | Interjection | UH |
| Number | | 27 | Cardinal Number | CD |
| | | 28 | Ordinal Number | OD |
| Plural marker | | 29 | Plural marker हरू | HRU |
| Question word | | 30 | Question word | QW |
| Classifier | | 31 | Classifier | CL |
| Particle | | 32 | Particle | RP |
| Determiner | | 33 | Determiner | DT |
| Unknown word | | 34 | Unknown word | UNW |
| Foreign word | | 35 | Foreign word | FW |
| Punctuation | | 36 | sentence Final | YF |
| | | 37 | sentence Medieval | YM |
| | | 38 | Quotation | YQ |
| | | 39 | Brackets | YB |
| Abbreviation | | 40 | Abbreviation | FB |
| Header List | | 41 | Header List | ALPH |
| Symbol | | 42 | Symbol | SYM |
| Null | | 43 | Null | <Null> |

# Tagging the Nepali Corpus

- The 100K English Nepali Parallel Corpus has been manually tagged.
- The Nepali tagset and the “Annotator” tool made available by the Regional Secretariat of PAN L10n was used.
- 2 linguists x 2 months was the human resource required for completing the work.

# Tagging issues

- English words simply transliterated into Devanagari are sometimes problematic to tag as their status is often not clear. For eg., जोन˜˛NNP)˜एण्ड˜(CC) कम्पनी (NN). Currently, such multiple words meaning a single entity have been tagged separately for each word constituent.
- Some words like पछि˜have multiple functions and hence quite problematic and confusing while tagging. In such cases, contextual occurrence has been taken into consideration. In घरपछि , पछि˜is a postposition (POP) and in चार˜वर्ष˜पछिfपछि˜is an adverb (RBO).˜
- Problems tagging the symbol “/”. Cases of १/२˜ and राम/श्याम
- Problems tagging the symbol “-”. For eg. घर-परिवार
- No provision for tagging negation words like न…न˜meaning “neither… nor”
- The post positions like लेfलाईfबाट˜etc. when comes in combination with other words is treated morphologically not leaving any spaces in between. For e.g. राम(NNP)ले(PLE) for रामले
- Compound verbs have been tagged as per the last verb component it contains. For e.g., खानुभयो<VBX>. Here the tagging is done taking into consideration भयो˜rather than खानु

# Conclusion

- This work is yet another step in the linguistic resource building for Nepali.
- The POS tagged text can have multiple applications like training data for stochastic POS Tagger as well as other Natural Language Processing Applications like the chunker.

# Acknowledgment

This work was carried out with the aid of a grant from the Language Resource Association (GSK) of Japan and International Development Research Centre (IDRC), Ottawa, Canada, administered through the Centre for Research in Urdu Language Processing (CRULP), National University of Computer and Emerging Sciences (NUCES), Pakistan.

# Thank You!!