TERMS OF REFERENCE FOR THE R&D PROJECT
(LITD 20, Indian Language technologies and products.)

1. Title of the Project: **Study & comparative analysis of Part-of-Speech (PoS) Tagsetting schemes for Northeast Indian Languages**

Duration of project: 04 months

2. Background:

a. Part-of-speech tagging, often abbreviated as POS tagging, is a process in natural language processing (NLP) that involves assigning grammatical parts of speech (such as nouns, verbs, adjectives, adverbs, etc.) to each word in a given text. The purpose of POS tagging is to analyze and understand the syntactic structure of a sentence, enabling computers to interpret the meaning of words in context.

b. In POS tagging, each word in a sentence is labeled with its corresponding part of speech.This information is valuable for various natural language processing tasks, such as text analysis, information retrieval, and machine translation, as it helps algorithms understand the grammatical structure and relationships between words in a sentence. POS tagging is typically part of the preprocessing steps in many NLP applications.

c. BIS has published Indian standard for PoS Tagset for Indian Languages (IS 17627:2021), covering a Superset of PoS applicable across the Indian Languages except North East (NE) Indian Languages (i.e., Kannada, Malayalam, Tamil, Telugu), Bangla, Gujarati, Hindi, Kashmiri, Konkani, Maithili, Marathi, Punjabi, and Urdu).

d. PoS tagging is critical for many Language Technology research and development works including Machine Translation and it is the first step in many Natural Language Processing (NLP) applications.

e. MeitY has already come up with a mega project on Indian Languages Multilingual Machine Translation Platform to be developed for bidirectional Machine Translation across the Indian Languages. Machine Translation requires linguistically rich parallel corpus, and one of the essential components is the Parts of Speech tagging. Thus, it is imperative to develop standardized POS Tag Setting schemes for Scheduled NE Indian languages.

3. Scope:

a) Study and analyse the lexical and syntactic behaviour of North East Scheduled Indian languages, i.e., Assamese, Bodo, Manipuri, and Nepali, in line with sequence to sequence tagging.

b) Analyze the gap of Indian Language PoS tagset and special requirements as per linguistics characteristics, for NE Indian Languages, with reference to IS 17627:2021.

c) Compile Language Specific Tagset for the four scheduled languages of North East India.
d) Conduct a preliminary study, i.e., requirements and gap analysis on the PoS behaviour of other non-scheduled languages like Khasi, Garo, Mizo, Kokborak, and Nagamese.

4. Deliverables:

a. Detailed study report for language specific PoS tag setting schemes specific to four scheduled languages of North East India- Assamese, Bodo, Manipuri, and Nepali.
b. Study and gap analysis document on language specific PoS tagset for non-scheduled languages of North East India, like Khasi, Garo, Mizo, Kokborak, and Nagamese. which may be the foundation for future revision of related standards.
c. Questionnaires, discussion & visit reports to be appended to the project report.

5. Research Methodology:

a. The project shall consist of detailed study and research on the NE Indian scheduled languages' specific behaviour, lexical and syntactic characteristics, and specifically the Lexical Categories, and their mapping to the Indian Standard of the PoS tagset as defined in the IS 17627:2021.
b. Review the literature in respect of areas covered in the scope.
c. Experts and researchers from the North Eastern region who are working in PoS related research and development activities in the Scheduled and non- scheduled languages shall be consulted for the studies and discussions, through various meetings, workshops, conferences. Being a region specific subject, to collect appropriate input for detailed study, linguistics groups working on NE Indian languages should be consulted alongwith with the relevant stakeholders to organize meetings and workshops for consultations and discussions.
d. Identified and likely stakeholders who may be consulted during the project period for evolving the NE Indian Languages language specific PoS tagset schemes:
    i. Stakeholders who are in Linguistics/NLP/Language Technologies in the concerned languages, primarily from the Universities/Institutes in the Northeast region.
    ii. Two Annotators and at least two linguists for each language who are currently working either in related projects, or working for related research.
e. The entire process shall also comprise of different modes of discussions and brainstorming sessions amongst focused groups for evolving appropriate subset PoS tag setting schemes for NE Indian languages.

6. Requirement for the CVs:

The individuals/organizations engaged in this project should have knowledge and experience in NLP & NE Indian Languages.

7. Timeline and Method of Progress Review:

a) Month 0-1: Study, and gap analysis.
- Interim Review of Progress of work through a meeting with the nodal officer before taking up Brainstorming, Discussions.

b) Month 1-3: Brainstorming, Discussions.
- Interim Review of Progress of work through a meeting with the nodal officer

c) Month 3-4: Drafting & Final report submission.

8. Support from BIS:

BIS will provide access to latest available editions of Indian standards and/ or international standards relevant to the project, on request

Contact Details:
Sh. Devansh Deolekar, Sc. D, LITD, litd20@bis.gov.in