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

1. **Title of the Project**: Study & comparative analysis of Tokenization schemes for North East (NE) Indian Languages
Duration of project: 04 months

2. **Background:**

   a. Tokenization is a primary step in many Natural Language Processing (NLP) tasks which is used to break down a given text into individual lexical units (most basic units). It can be sentence level or word level.
   b. This is a required pre-processing step for various NLP related tasks such as Part of Speech (PoS) tagging, Morphological Analysis, Chunking, Parsing, Named Entity Recognition, Spell Checking, sentiment analysis and Machine Translation etc. Without proper tokenization such tasks either cannot be achieved or achieved with incorrect/unacceptable/undesirable output.
   c. There are numbers of options for tokenization for English and other European languages, but none or a handful of options are available for Indic languages and no proper tokenization standard is found across the diverse set of languages from different language families found in Indic language context. Most of the Indic languages are morphologically rich, agglutinative and complex in terms of word formation as well as sentence structure. In this direction, it is again seen that languages from North East India fall short of such tokenizers. Though tokenization seems to be a very straightforward task such as separating words by space and punctuation marks and sentences by sentence end markers, it involves a lot more work than that. For example, how the tokenizer deals with numbers with decimal points, percentage symbol, various date format, special language specific characters etc., is very important for different tasks involved. So, it is of utmost importance to have a standard tokenizer for NE languages.

3. **Scope**:

   a. Study, analyse, and draft the scheme for Tokenization for the four scheduled languages of North East India- Assamese, Bodo, Manipuri, and Nepali
   b. Conduct a preliminary study, i.e., requirements and gap analysis on the tokenization of other non-scheduled languages like Khasi, Garo, Mizo, Kokborak, and Nagamese.

4. **Expected Deliverables**:
   a. Detailed study report for language specific tokenization schemes specific to four scheduled languages of North East India- Assamese, Bodo, Manipuri, and Nepali.
   b. Study and gap analysis document on language specific tokenization 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 tokenization schemes.

b. Review the literature in respect of areas covered in the scope.

c. Experts and researchers from the North Eastern region who are working intokenization related research and development activities in the Scheduled and non-scheduled languages shall be consulted for the studies and discussions, through variousmeetings, 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 tokenization schemes:

   a. Stakeholders who are in Linguistics/NLP/Language Technologies in the concerned languages, primarily from the Universities/Institutes in the Northeast region.

   b. 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 tokenization 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 BIS will provide:**

BIS will offer guidance and access to existing national & International standards and publications such as journals, magazines, research papers. BIS will also facilitate introducing the researcher to the relevant stakeholders.

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