Format for submitting projects under Project Varanasi. (Do not change the serial no and headings of the items. If the headings are not relevant to your project then say so. The proposal document must not exceed 5 pages.)

1	Project type (Strike off those not applicable, refer to the policy document for project types)	~~(i) Technology Development or Prototype Development.~~ ~~(ii) Faculty Projects (Innovation and application Projects)~~ ~~(iii) Project (Student Nurturing)~~
2	Title of the Project	**Resources and Tools for Bhojpuri, Maithili and Magahi Machine Translation System – Phase II**
3	Duration of the project	**6 months**
4	Total Cost	
5	Name address and phone numbers of PIs and Co-PI's	**PI-- Dr. Anil Kumar Singh** Deptt. of Computer Science & Engineering, IIT(BHU), Varanasi-221005, Mobile: 8795240608 Email: aksingh.cse@iitbhu.ac.in **CO-PI – Dr. Swasti Mishra** Deptt. of Computer Science & Engineering, IIT(BHU), Varanasi-5, Mobile: 9389156777 Email: swasti.linguist@gmail.com

6. General Description of the project:

The institute (IIT-BHU) is planning to launch an initial version of a machine translation system between Bhojpuri/Maithili/Magahi and Hindi (preferably both directions). Efforts have been going on in this direction and considerable work has already been done. Three of the key tools required for building this system are POS (part-of-speech) tagger, chunker and morph analyzer/generator. Significant work has already been done on each of these in the phase I of this project. Language-wise, the work on Bhojpuri has almost reached a satisfactory level as per the goals of the project. The proposed Phase II of the project on **Resources and Tools for Bhojpuri, Maithili and Magahi Machine Translation System** will focus more on Maithili and Magahi, on which a lot of work is still needed. For more details on the previous project, please refer to the proposal on the phase I of the project titled above.

Since most of the planned work on Bhojpuri is complete, we will mainly focus on Maithili and Magahi. The ultimate goal (which is beyond the scope of this project, but towards which this project is directed) is to build machine translation systems for these three languages.

7. General Description of experience/ expertise of team on such/ similar projects:

Since this is Phase II of an ongoing project, all the required expertise is in place. We will, however, need to employ some fellows as mentioned under the manpower section. The PI and the co-PIs remain the same.

8. Deliverables (The deliverables are to be described in each section. If there is no deliverable in a particular section then say the same clearly.):
(a) Prototype -nil-

(b) Process Prototype -nil-

(c) Design/ Technical Document **-yes-**

(d) Software **-yes-**

(e) Document (audio, visual, write ups web sites etc) -nil-

(f) Any other -nil-

9. **Method/ Technology to reach the deliverable. (A detailed description of method or technology may be described):**

The method/technology remains the same as for the Phase I of the project and is partly repeated below for easy reference.

POS (part-of-speech) tagger is a tool for automatically marking whether a given word in a sentence is a noun, verb, etc. A chunker is a tool that groups POS tagged data into Local Word Groups or chunks. These are important parts of a machine translation system. A morph analyzer is a tool that takes an inflected form of a word and output the root form as well as the morphological features (such as gender, number, person, tense, aspect modality etc.). A morphological generator is the opposite of this and generates the inflected form given the root form and the features. These tools are crucial parts of a machine translation system. A Word Transducer takes one word in the source language as the input and gives an acceptable equivalent form in the target language (if applicable). This can increase the coverage of the system.

For these tool to be created, tagged or analyzed, the data has to be manually created first by language experts and/or linguists. Such data can then be used by computational tools for the tasks mentioned above.

Such POS tagged, chunked and morph data each will be created by a team of people who have the necessary competence under the supervision of the project PI and the co-Pi. A few people may also be assigned to preparing the data for the Word Transducer.

10. **Time line / mile stones for, achieving the deliverables:- 6 to 8 months.**

11. **Budget details:**

(i) **Manpower:** 23 project fellows

Honorarium to the project fellows:--

20 (full time) @

3 (Part-time) @

(ii) Equipment: Nil

(iii) Contingencies (should not normally exceed 10 % of the total cost):

Rs.

(iv) Travel (should not normally exceed 10 % of the total cost):

Rs

Total = Rs -

Note: Since this is a project that requires advanced skills from the manpower, which may or may not be available as expected or planned, we request some flexibility with regard to the number and mode of project fellows.

Anil Kumar Singh
(PI)

Swasti Mishra
(Co-PI)