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	**Development of Anusaraka Support Modules**
3	Duration of the project	2 years (2 – 5 years)
4	Total Cost	
5	Name address and phone numbers of PIs and Co-PI's	**PI: Dr. Anil Kumar Singh** Department of Computer Engineering IIT-BHU Email: nlprnd@gmail.com Mobile:08795240608 **Co-PI: Dr. Sanjukta Ghosh** Department of Linguistics Faculty of Arts Banaras Hindu University Email: mi_anil@yahoo.com Mobile: 9452070216

6. **General Description of the project:**

Anusaraka, one of the major machine translation system technologies developed in India for English-to-Indian languages, has been used in various languages pair environments. However, certain issues at the level of language complexities still require an intensive linguistic research to enable the system perform better. We propose to work on the development of four modules with a view to support the Anusaraka system. These modules are in the form of four short-term/long-term tasks briefly outlined as below.

A. WSD-MT Divergence Module (2-3 Years): Complexities related to MT divergences are the focus of this module. Divergences related to TAM, sentential adverbs, sentential discourse markers and preposition-postposition mappings are still major source of complexities that have to be resolved to better the system. The task will consist of working on the existing files in Anusaraka system for identifying the

complex issues that need to be resolved working towards formulating disambiguation and mapping rules for them.

B. **Parallel corpora creation and alignment (2 years)**: This module consists of the task of corpus creation and alignment, an important resource for the improvement of the Anusaraka system. The task will consist of exploring and identifying existing and usable parallel English-Hindi data. In case of such data to be insufficient for the purpose, focus will be to create sufficient data and then work for their alignment. Alignment will be initially sentence-level alignment and subsequently word-level alignment will be done.

C. Creating and Maintaining a Mirror for the Anusaaraka Project (3-5 years):
A website of Anusaaraka already exists at IIIT-Hyderbad. The aim of the proposed project is to create and maintain a mirror of this project at IIT (BHU). A copy of the Anusaaraka project will be kept at IIT (BHU) and it will be synchronized with the copy at IIIT-Hyderabad. There are also some speed related issues, particularly when using Java libraries in Anusaaraka. Another aim of the proposed project is to resolve these issues so that jar libraries can be kept in memory for the MT server in order to increase the speed of the system. Other speed related issues may be investigated and work can be done on them under the project to resolve them.

7. **General Description of experience/ expertise of team on such/ similar projects**
The project will require expertise mainly from language, linguistics and computer engineering. The faculty members from the department of linguistics who are in the team have worked on teaching and research in computational linguistics/language technology including corpus linguistics and syntactic analysis for last several years and have published in these areas. The faculty members from computer engineering have extensive experience in natural language processing and have been engaged in research and development in this area for the last several years. The team will be mentored by Prof. Rajeev Sangal with his enormous expertise in this field of research. Besides, some senior faculty members from LTRC, IIIT Hyderabad (Prof. V. Chaitanya, Prof. Dipti M Sharma, Dr. Soma Paul) are in the team who will guide the works with their wide expertise in the area.
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

MT System

(b) Process Prototype

Linguistics analysis/rules and resource

(c) Design/ Technical Document

Language technology tools/Computational grammar

(d) Software

MT system and some additions/changes to it

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

System, technical report/research document, documentation

(f) Any other

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

Work will start on every module from the beginning. In the first module, an extensive study will be done for obtaining different types of constructions involving lexical/TAM ambiguity in the context of English-to-Hindi language pair. This will be done by accessing both printed and online resources in the form of books, ebooks, edictionaries, corpus, etc. A deep and extensive linguistic analysis (syntactic, semantic) will be done on the obtained resources to identify senses and obtain the mapping rules. This will also include of visiting IIIT Hyderabad and working with the Anusaarak team.

The work on the second module will begin with exploring existing parallel corpus for English-Hindi. In case we have sufficient amount of corpus, we will start work on alignment. In case sufficient corpus is not available, we will build parallel corpus (say, one lakh sentences) and then we will work on their alignment.

In the third module, the work will begin by accessing Hindi texts and identifying in them the words rooted in Sanskrit. The will be a short-term (6 months) task.

In the fourth module, the PIs and project fellows will work to prepare the guidelines, implement them and maintain the mirror.

10. Time line / mile stones for achieving the deliverables.

A. First year: To obtain and study the resources/texts to understand the distribution of lexical ambiguities and their types.

Second year: To work on linguistic analysis towards obtaining mapping rules

Third year: To integrate the obtained knowledgebase into the anusaarak system and finally documentation of the study in the form of a technical report/research thesis

B. The first three months, a parallel corpus will be obtained from different sources.

In the following months, alignment task will be done on the obtained corpus.

C. First two months: preparing guidelines to be followed and preparing for implementation.

Rest of the period: creating and maintaining the mirror and creating the user manual. This will also involve creating the software.