Kashmiri Speech Corpus Annotation and Text Corpus Sanitation

21st Oct-06th Nov

Workshop Report

The fifteen day workshop on *Kashmiri Speech Corpus Annotation and Text Corpus Sanitation* was conducted in the Department of Linguistics, University of Kashmir, from 21st Oct to 06th Nov. The main objectives of the workshop were:

1. To train the selected participants (Research Scholars/RPs).
2. To get the annotation & sanitation work done from them.
3. To achieve the preset targets, both in speech as well as text corpora.

The expected targets were achieved with little difficulty. 6 hr 24 min 31 sec Speech Corpus has been annotated and 413835 words Text corpus has been sanitized in the workshop. The annotated data has been submitted to the server but sanitized text corpus is yet to be finalized. The same will be finalized soon.

However, more speech data could have been annotated if working hours per day would have been more and machines/laptops wouldn't have given any trouble. Further, being able to sanitize/clean text corpus online was achievement in itself as it was carried out on experimental basis. It was possible with technical support provided by Mr. Venkateshan and Rajesha N along with the Lab and internet facilities provided by the department. Finally, it is worth to mention that the most of participants were hard-working & have given their consent to work with us in future, whether in Mysore or Kashmir.

Shahid Mushtaq Bhat

Lecturer/RP

LDC-IL, CIIL Mysore