Minutes of the Seventh Project Advisory Committee Meeting

of the Linguistic Data Consortium for Indian Languages (LDC-IL)

held on June 27, 2018 at Conference Room, School of Sanskrit and Indic Studies, Jawaharlal Nehru University, New Delhi

The meeting was attended by the following members:

1	Prof. D.G. Rao	Director, CIIL & Chairperson
2	Dr. Pushpak Bhattacharya	Director, IIT Patna
3	Prof. Dipti Misra Sharma	IIIT, Hyderabad
4	Smt. Sangita Toppo	Under Secretary, Language Division, MHRD
5	Dr. Narayan Kumar Choudhary	Officer i/c, LDC-IL & Coordinator
	Special Institutional invitees:	
6	Dr. Ajai Kumar	C-DAC, Pune
7	Dr. Brajesh Priyadarshi	AIISH, Mysore
8	Ms. Swaran Lata	Head, TDIL, MeitY, New Delhi
9	Shri Vijay Kumar	TDIL, MeitY, New Delhi
	Special Individual invitees:	
10	Prof. Girish Nath Jha	JNU, New Delhi
11	Prof. Aadil Ahmed Kak	University of Kashmir, Srinagar
12	Dr. Kamal Kumar Choudhary	IIT Ropar, Punjab
	Special Industry invitees:	
13	Ms. Kalika Bali	Microsoft
14	Ms. Amrita Kamat	Google
15	Mr. Manish Chapla	Keypoint Technologies, Hyderabad
16	Mr. Vivekananda Pani	Reverie Technologies, Bangalore

In addition to the above, some members presented their views over email on the agenda papers. They are:

1	Prof. Rajeev Sangal	IIT-BHU, Varanasi
2	Dr. L. Ramamoorthy	Head, CCL, CIIL, Mysore

Welcome

Prof. D. G. Rao, Director, CIIL & Chairperson welcomed the members and expressed his happiness that the members have gathered to resolve the issues pending in the last PAC meeting that took place 3 months ago.

The members were briefed about the agenda items to be discussed on the day and the tasks at hand which needed a decision and consensus of the members.

Costing Plan for the Data Sets

The cost analysis document as finalized in the last meeting was shared with the public and other stakeholders for their feedback. Based on the feedbacks received, a cost analysis formula was derived for the following types of linguistic resources:

1. Raw Text Corpus
2. PoS Annotated Text Corpus
3. Chunk Labelled Text Corpus
4. Syntactic Parsing
5. Dependency Labelled Corpus
6. Raw Speech Corpus
7. Sentence Segmented Speech Corpus
8. Word Segmented Speech Corpus
9. Parallel Text Corpus
10. Scanned Image Corpus
11. Handwriting Image Corpus
12. Ontologies/Wordnet
13. Anaphora and Antecedent Annotated Text Corpora
14. Named Entity Annotated Text Corpora
15. Pronunciation Lexicon Dictionaries
16. Multi Word Expressions
17. Word Sense Disambiguation

The formula as envisaged the document prepared by Dr. Narayan Choudhary in consultation with the various stakeholders across the academia, industries and other stakeholders were finalized. All the caveats as discussed in the last PAC meeting were addressed the members agreed to the revised proposals of CIIL on this issue. Other similar resources may be priced based of above pricing models. TDIL, MeitY also helped by rallying inputs on various resources from NLP experts and accepts this finalized linguistic resources pricing model for adoption by MeitY for resources developed under TDIL Programme of MeitY. The document may be published as a policy document by CIIL. Updates may be done on a regular basis as and when the more feedback arrives.

Pricing Plans

It was decided that the data will be free for Academic (UGC recognized) and Research Organizations (Govt R&D organizations).

Calculating the Base Price

It was decided that a base price would be calculated by the controlling agency in consultation with the developing agency. An assumption may be made to divide the total/overall cost of a corpus by 10 which would work as the base price for a language resource.