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.

Cost should not be prohibitive to the buyers. Therefore, 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. The developing agencies are however free to make the division as which ensures that the prices are reasonable and within the reach of the target group.

Further to this, there would be some price tiers as denoted in the table below. The discounts/subsidies in the price have been proposed to the different tiers of commercial users which are more likely to contribute to the promotion of Indian languages and language technologies in Indian languages.

MNCs and Foreign Entities	Base Price
Non-MNC Indian Company	80%
MSME/Entities from SARC Countries	60%
Startups/MSMEs with a turnover of less than 5 crores	20%

The prices are subject to revision by the developing agencies with the approval of the HoDs/competent/controlling authorities of the agencies involved.

Promoting Shared Task Competitions

Limited versions of datasets may be shared for shared task competitions with the non-commercial license. The participants, be it commercial or non-commercial, may be given the limited dataset for test, development and evaluation. The distributing agency in consultation with the developing agency may decide upon the quantum of the limited data set.

Privacy Policy

The template of the privacy policy proposed before the committee was approved. However, it was noted that going forward, the data collection processes online or offline may also address the concerns of GDPR guidelines which is prevalent now in the European Union countries and elsewhere.

Undertaking/Terms & Conditions

The template of the undertaking for non-commercial users was approved. Additional changes with regard to the rights of the purchasers (commercial or non-commercial) were delineated as follows:

- The non-commercial users would be allowed to download the dataset after giving a proper justification for their request. The justification provided by them would be subject to verification by the competent authority. LDCIL reserves the right to deny anyone the license without giving any reason.
- A download request made at the portal must be processed by the distributing agency within a period of 7 days.
- A licensee would be free to make changes in the dataset as per their requirement however they should take utmost care that the datasets are not distributed further (except their internal

organization or collaborators or subsidiaries) and the same is not made public and follows the practices of extant fair copyright laws.

- Even a modified version of the datasets should not be made public by any of the users.
- Users would be free to derive applications such as a machine learning models, APIs, other web services etc. out of it or with the help of it. However, non-commercial users should not use such applications for any commercial purposes. Commercial users may make use of these derived items for commercial purposes.

Additionally, it was suggested that LDCIL may think devising a mechanism for detecting violation of the terms of use of the data sets.

Data Distribution Mechanism

It was decided that the datasets may be distributed through the data portal of LDCIL which is http://data.ldcil.org. A demo of this portal was also given by the LDCIL team. The following decisions were made in this regard:

- The datasets would be released online.
- Users would be able to browse through the portal and make download/purchase request only.
- It will be the onus of the user making the download request to prove that they belong to the category of the user as claimed by them.
- The users would be able to download the data online only. Request for providing the data in any other format may be entertained at the discretion of the distributing agency.
- The distributing agency will have the rights to verify the identity of the individual making the download request.
- Wherever necessary, checks and balances will take place offline. However, it should be ensured that the process is smooth, quick and users get what they want seamlessly.

Copyright Issues

It was found that for the textual data sets, there are some copyright permissions required before it can be released for commercial purposes. It was found that LDCIL has already started the process of getting the consent for the extracts used in the text datasets. However, the process has not reached a releasable status for most of the languages. It was decided that the wherever the copyright issues are, it should be resolved on an urgent basis so that the textual datasets are released at the earliest.

At the end everyone hoped that that the datasets are released at the earliest as taking more time will make many of the existing datasets irrelevant for use in current language technology practices.

The meeting ended with the vote of thanks to the chairperson.

Prof. D. G. Rao
Chairperson