



# CENTRAL INSTITUTE OF INDIAN LANGUAGES

DEPARTMENT OF HIGHER EDUCATION

Ministry of Human Resource Development, Government of India

Manasagangotri, Mysore - 570 006



## Linguistic Data Consortium for Indian Languages

### **Minutes of the Fifth Project Advisory Committee Meeting of the Linguistic Data Consortium for Indian Languages (LDC-IL) held on July 30, 2012 at CIIL, Mysore**

#### **Welcome**

Prof. S.N. Barman, Director, Central Institute of Indian Languages and Chairperson, Linguistic Data Consortium for Indian Languages (LDC-IL) welcomed the Members of the Fifth Project Advisory Committee meeting.

#### **Release of LDC-IL Publication:**

Shri P.K. Saha, Deputy Secretary, Language Division, Ministry of HRD, New Delhi released the LDC-IL Publication titled "POS Annotation for Indian Languages: Issues & Perspectives".

After the Welcome, the agenda items were taken up in the order.

#### **Agenda I: Confirmation of the Minutes of the Fourth PAC Meeting**

The Minutes of the Fourth Project Advisory Committee Meeting of the Linguistic Data Consortium for Indian Languages (LDC-IL) held on August 10, 2010 was confirmed with the following remarks.

- It was decided in the IV PAC to convene the next PAC meeting in Mysore within 3 months. But, due to some administrative problems like resolving licensing issues, pricing etc., and academic hurdles like ensuring quality data, the 5<sup>th</sup> PAC meeting could not be convened within the specified period. The same was explained to the members.

#### **Agenda II: Action Taken Report**

The actions taken on the recommendations of the Fourth Project Advisory Committee Meeting were explained item-wise.

#### **Agenda III: Achievements of LDC-IL in 11<sup>th</sup> Five Year Plan**

Dr. L. Ramamoorthy, Reader cum Research Officer & Head, LDC-IL made a brief presentation on achievements of LDC-IL Project in the last four years i.e., Financial Year 2008-12 and also progress of work in LDC-IL after 4<sup>th</sup> PAC Meeting i.e., from August 10, 2010 to July 30, 2012.

The PAC members appreciated the work done in LDC-IL for last four years and expressed their satisfaction on the progress of the work.

**Further, the following seven proposals under Agenda IV were placed before the PAC for discussion and consideration.**

**Proposal-1:**

**Annual Action Plan(2012-13) :** The following annual action plan for the current financial year 2012-13 was placed before the PAC for kind consideration.

<b>I</b>	<b>Speech Corpus:</b>	<b>Projected Targets</b>
1	Automatic Speech Recognition(ASR) - Read speech data	Annotation
2	Telephonic Data	Data collection & Setting Standards
3	Pronunciation Dictionary	Addition of sound file Phonetic transcription Phonetic variation
<b>II</b>	<b>Text Corpus:</b>	
1	Monolingual corpus	One million words (cleaned corpus for well resourced language)
2	Parallel Corpus	Data collection
3	Optical Character Recognition (OCR)	Data collection
4	Classical language corpus	Data collection
<b>III</b>	<b>Natural Language Processing</b>	
1	Morph Analysis	Morph dictionary entry (15K/ language)
2	Part of Speech Tagging (POS)	Part of Speech Tagging (50K per year)
3	Local word Grouping (LWG)	Local word Grouping (50K per year)
4	Shallow-parsing	Developing standards, annotation of 50K words/language per year
5	Indian Sign Language	Data Collection
6	Language Identifier	For Languages / Scripts which have not been taken up by other institutions

The above proposal was considered and approved by the PAC. Further, some suggestions were given on the following points;

**Telephonic data:** It was suggested that Corpus collection w.r.t. Telephonic data be in 16Khz instead of 8 Khz proposed by LDC-IL.

**Optical Character Recognition (OCR):** PAC members informed that Department of Information Technology, New Delhi has funded for OCR Consortium and already some corpus has been collected for this purpose and suggested that corpus collection for OCR may be suspended.

**Data base:** It was also suggested to focus on theme oriented data base for activities other than NLP such as language typology, genre based corpus etc.,

**Indian Sign Language:** It was suggested that KYNECT Camera which is available at reasonable price may be used for sign language corpus collection.

**Proposal-2**

**Tentative Targets of the 12<sup>th</sup> Five Year Plan:** The tentative targets set for the next five year plan i.e., 2012-2017 was placed before the PAC. The PAC suggested that, the proposed targets need to be placed before the expert committee.

Hence, it was decided to convene an expert committee to prepare a Road Map of LDC-IL and finalize the deliverable targets of 12<sup>th</sup> Five Year Plan and this will be provided to Evaluation Committee to be constituted by MHRD.

### Proposal-3

**Status of Project Fund Generation:** The PAC members were apprised of the status of fund generation and explained the reason for non-generation of funds i.e.,

1. Delay in ensuring Quality data
2. Non-availability of balanced corpus
3. Insufficient human resource
4. Delay in arriving at and setting standards
5. Delay in finalizing Costing & Licensing Policies

PAC members accepted the reasons for non-generation of funds and deliberated on the nature of complexities in languages, corpus and NLP activities in India.

PAC also appreciated the steps taken to create awareness among the researchers on NLP by LDC-IL.

Further, PAC felt that, self sufficiency of the LDC-IL Project is very difficult and would not be possible. After fixing the price also, the LDC-IL could generate certain funds and they further opined that attaining self sufficiency with full pledged revenue is highly impossible.

### Proposal-4

**LDC-IL Data Pricing:** A rough data pricing was done by the LDC-IL by taking the cost invested on data acquisition, corpus sanitization/segmentation and annotation of the total corpus in a particular language and the same was placed before the PAC for their consideration and kind suggestion.

Text Raw data	:	Rs.0.10 per token
Text annotated data	:	Rs. 1.20 per token
Speech Raw data	:	Rs.0.77 per word
Speech annotated data	:	Rs.217.60
Pronunciation Dictionary (5000 words)	:	Rs.2200.00

The major discussion was done about the data pricing and it was agreed to constitute a Licensing Committee to fix a price for LDC-IL data. Some of the following suggestions were given regarding pricing of the data.

**Membership Fees:** The membership fees needs to be re-debated by the Licensing Committee.

**Sample Data:** Sufficient sample data should be made available to get the feedback of the users before pricing.

**Non Disclosure Agreement (NDA):** In all cases of data licensing, the NDA has to be signed.

**Suggested price for Educational Institutions:** A reasonable price should be fixed for Educational Institutions.

**Data Package definition for members:** It was suggested that even if the price fixed for each word, it would be more for individual researchers/students. So, package data be introduced and it should be clearly defined what one gets if they become member.

**Copyright issues:** It was advised to include an expert in IPR matters as one of the member to the Licensing Committee to obtain suggestions while resolving the copyright issues.

**Preview data/Advance Copy:** PAC members should get advance copy of the data before it is taken up for releasing.

**Time-line:** It was strongly suggested to fix a time-line to prepare a Licensing Policy for data pricing.

However, all these points be taken up for elaborate discussion by the Licensing Committee.

**Constitution of Licensing Committee:** It was agreed to constitute a Licensing Committee with the following members to prepare Licensing Policy and also fix a price for the LDC-IL data. And also, it was proposed to hold a pre-meeting during next month.

**Licensing Committee:**

1. Dr. Hemant Darbari, Director General, C-DAC, Pune
2. Dr. Rajeev Sangal, Director, IIIT, Hyderabad
3. Dr. A.G. Ramakrishnan, IISc., Bangalore
4. Dr. Garg, Director, IPR Department, DIT, New Delhi
5. Dr. Ajai Kumar, Associate Director, C-DAC, Pune
6. One person from the Finance Division, C-DAC, Pune
7. Director (Languages), MHRD, New Delhi
8. Director, Finance Division, MHRD, New Delhi
9. Director, CIIL, Mysore
10. Dr. Ramamoorthy, Head, LDC-IL

**Proposal-5**

**Extension of the LDC-IL Project during 12<sup>th</sup> Five Year Plan:** A Proposal to extend the LDC-IL project for 12<sup>th</sup> five year plan was placed before the PAC.

The Director (Languages), MHRD, informed that before switching over from one plan to another plan, any scheme should be evaluated by an external agency appointed by the MHRD. After receiving the report from the appointed agency, decision will be taken by the Planning Commission about the extension of the scheme for the next five year plan. Regarding this LDC-IL Project, appointment of agency for evaluation is already initiated and will be finalized during next month.

**Proposal-6**

**Re-designation of the project positions:** It was observed by the audit that, the following regular designations should not be given to the project positions. Hence it was proposed to re-designate the positions as follows in the 12<sup>th</sup> plan period.

S/N	Existing Designation	Proposed to be Re-designated as
1	Reader/Research Officer	Assistant Project Director
2	Senior Lecturer/JRO	Senior Project Officer
3	Lecturer/RP	Junior Project Officer

LDC-IL was advised to place these before the evaluation committee to be constituted by the MHRD.

## **Proposal-7**

**a. Engagement of additional human resources:** The SFC of LDC-IL approved 49 project posts to be recruited on contract/deputation basis at the time of project implementation. The targets set for 12<sup>th</sup> plan needs more human resources to work on different aspects of Language technology in 22 languages i.e., Annotation of text data, collection of corpus for Text to Speech (telephonic data) & Annotation, Collection of corpus for Character Recognition & Sign Language need more specialists. Hence, it was proposed to create an additional 60 project posts by considering atleast 2-3 persons in each language, to fulfill the needs during 12<sup>th</sup> Five Year Plan.


PAC felt the need of engaging additional resource persons to work on different aspects of NLP in all Indian languages.

However, the Director (Languages) informed that, the above two proposal i.e., 6 & 7 would be taken into consideration once the evaluation of this project completes and based on the report submitted by the appointed agency.

**b. Annual increment:** During the 11<sup>th</sup> Five year plan, LDC-IL used to give an annual increment @8.5% to the project staff as mentioned in the LDC-IL SFC. In this 12<sup>th</sup> Five year plan, it was proposed to give yearly increment based on the DA rates of that period to sustain the trained personnel in the project for longer period.

PAC suggested that, since the mode of payment of annual increment is not yet decided for the 12<sup>th</sup> Five Year Plan, PAC expressed its consent to pay @8.5% increase on the present consolidated amount as approved in the LDC-IL SFC and as paid in the last five year plan.

The meeting ended with vote of thanks by the chairperson.

  
(S.N. BARMAN) 03.08.12  
Chairperson, Project Advisory Committee, LDC-IL &  
Director, Central Institute of Indian Languages, Mysore

## MEMBERS PRESENT

- |   |  |   |
|---|--|---|
| 1 | Prof. S.N. Barman<br>Director, CIIL, Mysore                | Chairperson                                     |
| 2 | Shri. P.K. Saha<br>Director (Languages)<br>MHRD, New Delhi | Member<br>Representing Language Bureau          |
| 3 | Prof. Rajeev Sangal<br>Director, IIIT, Hyderabad           | Member<br>Represented by Dr. Dipti Misra Sharma |
| 4 | Director, Indian Institute of Technology,<br>Bombay        | Member<br>Represented by Dr. Malhar Kulkarni    |

## SPECIAL INSTITUTIONAL INVITEES

- |   |                       |        |
|---|-----------------------|--------|
| 5 | Director, C-DAC, Pune | Member |
|---|-----------------------|--------|

## SPECIAL INDIVIDUAL INVITEES

- |    |   |                 |
|----|---|-----------------|
| 6  | Prof. K. Narayana Murthy<br>University of Hyderabad, Hyderabad    | Member          |
| 7  | Prof. A.G. Ramakrishnan<br>Indian Institute of Science, Bangalore | Member          |
| 8  | Prof. G Umamaheshwar Rao<br>University of Hyderabad, Hyderabad    | Member          |
| 9  | Prof. M. Ganesan<br>Annamalai University, Chidambaram             | Member          |
| 10 | Dr. Ajai Kumar<br>Associate Director, C-DAC, Pune                 | Member          |
| 11 | Dr. L. Ramamoorthy<br>RRO & Head, LDC-IL, CIIL, Mysore            | Member-Convener |

## MEMBERS ABSENT

1. Finance Division, MHRD, N. Delhi  
*(IFD has sent some suggestion on the agenda of V PAC (proposals: 6 & 7) which will be taken into consideration.)*
2. Director, Indian Institute of Technology, Madras
3. National Law School of India University, Bangalore

## SPECIAL INSTITUTIONAL INVITEES

- |    |   |
|----|---|
| 4. | Director, Indian Institute of Technology, Kharagpur |
|----|---|

## SPECIAL INDIVIDUAL INVITEES

- |    |                        |
|----|------------------------|
| 5. | Prof. Peri Bhaskararao |
|----|------------------------|

## SPECIAL INDUSTRY INVITEES

7. MICROSOFT, Bangalore
8. MOTOROLA, Bangalore