



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 THIRD PROJECT ADVISORY COMMITTEE MEETING OF THE LINGUISTIC DATA CONSORTIUM FOR INDIAN LANGUAGES (LDC-IL) HELD ON DECEMBER 8, 2008 FROM 11.00 A.M. ONWARDS

I Welcome

Prof. Udaya Narayana Singh, Director, Central Institute of Indian Languages and Chairperson, Linguistic Data Consortium for Indian Languages (LDC-IL) welcomed the Members for the Third Project Advisory Committee meeting. After the Welcome, the agenda items were taken up in the order.

II Confirmation of the Minutes of the Second PAC

The Minutes of the Second Project Advisory Committee meeting of the Linguistic Data Consortium for Indian Languages (LDC-IL) held on **June 9, 2008** were discussed in brief and confirmed.

III Action Taken Report

1. The reports on action taken on the recommendations of the Second Project Advisory Committee were considered item-wise in detail and were accepted.

2. The PAC was informed about the composition of the Selection Committee for recruitment of academic, technical and administrative posts. It is as follows:

- The Director, Central Institute of Indian Languages and Chairman, LDC-IL - Chairman of the Selection Committee of all positions of the LDC-IL.
- There will be three Committees to consider the candidates who are to be called for tests/interviews: (a) Committee for Academic Staff, (b) Committee for Technical Staff and (c) Committee for Administrative Staff.
- The Liaison Officer for the SC/ST is being the Member of all the 3 Selection Committee.
- The Asst. Director (A) is the Selection Committee Member for the Administrative positions.
- One of the institute Technical staff - for the Technical positions.
- Apart from the above 6 External Experts were co-opted to the Selection Committee.

3. The proposal of LDC-IL to take interns to complete the short duration projects of the LDC-IL since these kinds of projects do not need regular faculty for longer duration was accepted.

4. Also, the LDC-IL was advised to give grants out of its GIA to such short term projects as and when necessary.

IV Presentation & Reviewing Progress Achieved During 2008-09 (Till November 2008)

1. Dr. B. Mallikarjun, Reader cum Research Officer & Head, LDC-IL made a presentation on the progress made in the work of LDC-IL from April 2008 to November 30, 2008.

2. The PAC expressed its satisfaction over the work done during the said period. The clarifications sought by the members of the PAC relating to the parallel Corpora and Speech data were provided.

V Proposals for the Year 2009-10

1. The proposals for the programs to be conducted during 2009-10 were agreed to. In addition to these, the PAC has recommended to sponsor IJCNLP proposed by Prof. Rajeev Sangal to be held before March 2010 in lieu of one international seminar proposed by the LDC-IL.

2. Instead of two Regional Seminars it was agreed that one National Seminar to be proposed by the **All India Institute of Speech and Hearing, Mysore** focusing on “**Sensitizing Linguistics Disorder of Speech and Hearing**” shall also be sponsored by the LDC-IL.

3. The LDC-IL was advised to conduct **awareness programmes** to create awareness among researchers about NLP and other related areas.

VI Launching of LDC-IL Website and Release of LDC-IL Publications

1. Smt. Rita Chatterjee, Director (L), Department of Higher Education, Ministry of Human Resource Development launched the LDC-IL Website and Dr. B. Mallikarjun provided the details of links in the web site by giving a demonstration.

2. Prof. Rajeev Sangal, Director, IIIT, Hyderabad released the LDC-IL publications of **Frequency Dictionaries in Bengali, Hindi and Kannada**. Prof. Peri Bhaskararao released the **Speech Data Set in Bengali, Hindi, Nepali and Tamil, Speech Data Collection & Annotation Manual, and Coding Convention for dot.net framework**.

VII Discussion on Roadmap of LDC-IL

The roadmap of the LDC-IL was taken for detailed discussion since it is going to set a tone of work for the LDC-IL for the next 5 years. In addition to acceptance of the roadmap the following was also suggested.

a. Indian Phonetic Alphabet

Need to evolve Indian phonetic alphabet was also felt. Set of parameters for sign symbol correspondences have to be identified and appropriate proposals as and when needed are to be sent to the International Phonetic Association.

b. Samvardhith Devanagari

Creation of Parivardith Devanagari for Indian language scripts ceased many decades ago. There is a need to initiate creation of Samvardith Devanagari to cater to the needs of all the Indian languages including the one which use tone etc., if necessary.

c. Existing Speech Data

Since Dr. Shobha Sathyanath informed that she has data recorded from some of the languages and she could be given a short term project to tag the sound data after digitization to give the same to the LDC-IL in DVD format as per the LDC-IL standards. Before that, she will provide the full details of the language, data quantity etc.

d. Parallel Corpora

The parallel corpora created by the LDC-IL could also be called Comparable Corpora.

e. Tools

(i) Under the Tools for Corpora Management and Analysis (7 d.) in the roadmap it is stated that “Speech Synthesizer for Indian languages that will facilitate to create speech corpora which acts as input for text corpora. That way parallel Text and Speech Corpora (Text Corpora parallel with respective Speech Corpora) could be created”. This may be deleted.

(ii) Wherever tools etc., are already available like **Sanchay** etc., could be tried out at the first instance. Many research groups have developed the NLP Tools. An attempt could

be made to procure them for refinement, testing etc., purposes and if appropriate LDC-IL can adopt them and use.

- (iii) In case of languages taken up for preparation of Computational Grammars, the tasks listed in the roadmap shall form the horizontals, and languages shall form the vertices. It is possible that many horizontals and the work on many languages continues simultaneously.
- (iv) The Pronunciation Dictionary shall have information like the word, standard transliteration, phonemic transliteration and phonetic transliteration in IPA. The Sound Dictionary shall have all these and also the sound file. This distinction was clarified to the PAC.
- (v) Before taking up the work of Indian Sign Language, issues relating to standardization have to be sorted out by the LDC-IL having a dialogue with the researchers who are working in the area.

f. Script Encoding Initiatives

1. Under the Text Corpora, LDC-IL is expected to create Historical/Inscriptional databases of Indian languages which is most important to trace not only as the living documents of Indian History but also historical linguistics of Indian languages.

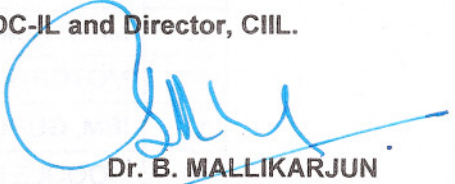
2. The PAC recommended the constitution of a Committee under the Chairmanship of Prof. Peri Bhaskararao, Tokyo University & Foreign Studies, Speech Sciences, ILCAA, Tokyo, Japan with a Working group to look into the history of Indian scripts to evolve super set of scripts to create master list of script symbols to go into the UNICODE for code points. This could be called **LDC-IL Script encoding initiative**. This group can have one person per language as language expert, a representative of TDIL, MCIT. A roadmap for the same shall be created by the working group and they can follow bottom up approach to create the super set of script.

VIII Consideration and Discussion of the Draft Minutes of the Licensing Group

The discussion relating to Licensing Policy covered all the issues listed in the document presented to the PAC by Dr. Hemant Darbari, C-DAC, Pune on behalf of the Licensing Group.

- a. It was felt that the Membership fee proposed to be fixed for software related industry is very high. In general it was felt that the membership fee needs to be re-debated by the Licensing Group and the membership policy and fee structure should be made reasonable and simple.
- b. Sample data of very limited quantity could be made available freely but not the whole data. The Licensee, individuals, institutions, industry etc., who use the data, should invariably acknowledge the source of the data.
- c. In case the LDC-IL has funded for the creation of data tools etc., one who has received the funds can use for himself/herself/itself only. But, he/she/they should not have any right to re-distribute the data either in the original or in the value added form to anybody.
- d. In all cases to data licensing, the Non Disclosure Agreement (NDA) has to be signed.
- e. The LDC-IL when it receives data, tools etc., it can redistribute with value addition with a revenue sharing model. This has to be worked out by the Licensing Group.
- f. Data acceptance format has to be decided by the LDC-IL.
- g. The list of persons who have received corpora till date for R & D purpose from CIIL with full address etc., have to be placed at the website.

IX The minutes have been approved by the Chairperson, LDC-IL and Director, CIIL.


Dr. B. MALLIKARJUN
Head, LDC-IL &
Reader-cum-Research Officer

MEMBERS PRESENT

1.	Prof. Udaya Narayana Singh, Director, CIIL, Mysore	Chairperson
2.	Mrs. Rita Chatterjee Director (L), MHRD, New Delhi	Member Representing Language Bureau
3.	Prof. Rajeev Sangal, Director, IIIT Hyderabad	Member
SPECIAL INSTITUTIONAL INVITEES		
4.	IIT, Kharagpur, KHARAGPUR	Member Represented by Prof. Anupam Basu
5.	C-DAC, Pune	Member Represented by Dr. Hemant Darbari
6.	AIISH, Mysore	Member Represented by Dr.K.S. Prema
SPECIAL INDIVIDUAL INVITEES		
7.	Prof. Shobha Satyanath, Delhi University, Delhi	Member
8.	Prof. A.G. Ramakrishnan, IISc. Bangalore	Member
9.	Prof. Peri Bhaskararao, Tokyo University, Japan	Member
10.	Ms. Rekha Sharma, Centre for Speech Sciences, CIIL	Member
SPECIAL INDUSTRY INVITEES		
11.	MICROSOFT, Haryana	Member Represented by Dr. Kalika Bali
12.	Dr. B. Mallikarjun, Head, LDC-IL, & Reader cum Research Officer CIIL, Mysore	Member-Convener

MEMBERS ABSENT

1.	Member from Finance Division, MHRD, New Delhi
2.	Prof. C.N. Krishnan, Director, AU-KBC Chennai
3.	Director, IIT Bombay, Mumbai
4.	Director, IIT Madras, Chennai
5.	Joint Secretary, Deptt. of IT, MCIT, New Delhi.
6.	Vice Chancellor & Professor of Law National Law School University, Bangalore - 560 072
7.	Director, IISc. Bangalore
8.	E-Governance, DIT, New Delhi
SPECIAL INDUSTRY INVITEES	
9.	HP LABS, Bangalore
10.	MAIT, New Delhi
11.	MOTOROLA, BANGALORE
12.	IBM, GURGAON, Haryana
13.	GOOGLE INDIA, Bangalore