

Report of the workshop on
STANDARDISATION OF THE DATA SET FOR INDIAN SIGN LANGUAGE CORPUS
at the Ali Yavar Jung National Institute for the Hearing Handicapped, Mumbai
from 10–14th September, 2012

The workshop on the standardization of the data set for Indian Sign Language Corpus was held in Ali Yavar Jung National Institute of Hearing Handicapped, Mumbai, from 10 – 14 September, 2012.

Inaugural programme:

The workshop started with a short inaugural session on Monday, 10th September, 2012. Mr. Amaresh Gopalakrishnan, Senior Resource Person, LDC-IL, gave the welcome address, followed by self-introduction of the experts and the participants. Ms. Atreyee Sharma, Senior Lecturer/JRO, LDC-IL, gave an introduction on the aims and objectives of CIIL and LDC-IL and elaborated on the work done by LDC-IL in the Indian Sign Language Corpus development, and also mentioned about the imminent need of LDC-IL about the finalization of the ISL dataset and the collection of ISL corpus. Prof. Rangasayee, Director, AYJNIHH, inaugurated the workshop by giving a short speech on the need of promotion of sign language studies in the national and international context. He stressed the need of a Universal Design and Barrier Free Environment which will be beneficial for the deaf community. The need of a collaborative, cooperative work by all institutes working for sign language was given focus in his speech. He praised the works done by LDC-IL in Indian Sign Language. He ended his speech with wishes to all the delegates and with the compliments that the workshop may yield fruitful results.

The inaugural session was followed by presentations by the LDC-IL staff. Ms. Atreyee Sharma highlighted the works carried out by LDC-IL in general, and on Indian Sign Language in particular, and the methodology adopted for Indian Sign Language corpus development in LDCIL. Mr. Aju Samuel Thomas gave an overview on the pilot study conducted by LDC-IL on Indian Sign Language Corpus Collection which concluded by equally stressing on both the merits and demerits of the pilot study. Mr. Rajesh Kumar Singh presented on annotating Indian Sign Language using the ELAN software. He also gave a short demo and explained the use of annotated corpus. Mr. Amaresh Gopalakrishnan gave a presentation stating the objectives of the workshop.

Objectives of the workshop:

- Presentation of the pilot study work done by LDC-IL in collaboration with Sri Ramakrishna Mission Vidyalaya, Coimbatore
- Presentation on the new methodology of work adopted by LDC-IL with regard to ISL corpus development, and to get a consensus from the experts on the same
- Presentation of the hand-shapes documented by the ISL corpus development team
- To list out a minimum of 10 sign-word under each hand-shape for data collection
- To list out at least 50 sentences for data collection

Synopsis of the pilot study:

LDCIL did a pilot study in collaboration with Faculty of Disability Management and Special Education (FDMSE), Ramakrishna Mission Vidyalaya, Coimbatore. Two expert committee meetings were held in this regard in which a data collection committee was constituted. In the second meeting, a data set consisting of 500 words and 200 sentences was finalized for data collection. The words were listed out from the Indian Sign Language dictionary documented by Sri Ramakrishna Mission, and the sentences were grouped under different sentence types. On the basis of the deliberations of the two meetings, data were collected and videos were recorded at FDMSE, Ramakrishna Vidyalaya, Coimbatore. The main aim of the pilot study is to list out any issues arising out of the video filming process that could be ironed out while the actual data collection work commences.

Methodology adopted by LDC-IL post pilot study:

After the pilot study was conducted and the issues listed out, LDC-IL decided to adopt a different approach to construct the data set for words and sentences. In this new approach, the data set for words will be constructed using the documented handshapes. A minimum of ten sign-words will be listed out under each handshape and that group of ten sign-words would ideally be able to elicit all the different parameters and features of sign language. The data set for sentences will be based on different emotions, namely, happy, sad, angry, etc. A minimum of ten sentences under each category will be listed down. These sentences would be the ones that are used in day-to-day conversation in the deaf community.

This approach to data collection was presented at the workshop and the sign language experts present in the workshop unanimously agreed to the methodology adopted. It was suggested at the workshop that LDC-IL has to collect as much of natural data as possible from the deaf community, as it might be difficult to elicit signs for controlled data.

Data set for words:

In this workshop, 54 handshapes documented by the LDC-IL team were shown to the sign language experts, out of which, 40 hand shapes were approved and the rest 14 have been rejected due to various reasons. More than 900 sign words listed out under those 40 hand shapes.

Data set for sentences:

The sign language experts divided into three groups and each group was asked to list as many sentences possible under different emotions like happy, sad, angry, etc. and also for questions. At the end of this exercise, more than 90 sign sentences were listed down.

The experts participated in the workshop agreed unanimously to the way and mode of work carried out by LDCIL for the development of Indian Sign Language Corpus. An email group was also created for further correspondences among the experts who participated in the workshop.
