Internship Course for Rajiv Gandhi University of Knowledge Technologies Students

(Linguistic Data Consortium for Indian languages, CIIL, Mysore)

3rd June 2014 to 4th July 2014

An Official Report

An Internship Course for Rajiv Gandhi University of Knowledge Technologies Students was organized by LDC-IL, CIIL, Mysore for a complete month. The program was well over and done with from the 3rd June 2014 to 4th July, 2012, at the LDC-IL Lecture Hall. There were 20 students, selected for the program through screening, and 22 Resource Persons engaged for presenting the talks and deliverance for the entire Internship Course.

The first day started off with an introduction gathering. Dr. L. Ramamoorthy, Head, LDC-IL, CIIL, Mysore initiated the course by delivering a talk on ***Introduction to Language and Technology***. This talk grounded the students about the themes and objectives of the course specified. Ms. Atreyee Sharma, Senior Lecturer/JRO, followed the session by giving a lecture on the ***Overview of LDC-IL***. She detailed the work and assignment of the project of LDC-IL in regard to the repository of linguistic resources in the form of text and speech for all Indian languages. On the second day, 4th June 2014, Mr. M. Md. Yoonus delivered a lecture on ***Introduction to NLP***, which covered various areas of both technical and linguistic perspectives involved in developing NLP tasks. Mr. Premkumar presented a talk on ***Linguistic Concepts***. It made the eyes open for the linguistic issues to discuss in the following days. On the third day, 5th June 2014, Ms. Sarbjeet delivered a presentation on ***Basic Phonology and Phonetics***. There was an extra hour for phonetic drill exercise. Ms. Vijayalaxmi, in the next session, talked about ***Morpholog***y, interacting with the students in identifying the morphemes out of a given text. The following **fourth day**, 6th June 2014 was kept for workshop for the students, particularly for corpus cleaning, database creation and the technology involved in developing the NLP tools. In the second week of internship and on the fifth day, 9th June 2014, Mr. N Rajesha gave a talk on ***Morphological Analyzer: Suffix Stripping***. There was a give and take between the students and the presenter on the technical issues and methodologies to be applied during the development of Morphological Analyzer tools. The next session was handed over to Ms. G. Manasa, who took the class of ***Morphological Analyzer for Apertium Model***. The whole day session was full of technical discussion. On the sixth day, 10th June 2014, Ms. K.S. Rejitha gave a presentation on ***Types of Word Formation***. It drew the attention of the students from the non-linguistic backgrounds. There were questions and answers in the hour. Mr. Premkumar continued the next session and presented for the second time a paper on ***Introduction to Speech Corpus***. There was a demonstration class to get the idea merged in the stream. On the seventh day, 11th June 2014, Ms. Sarbjeet, for the second time, gave a talk on ***Speech Segmentation and Annotation***. She explained the guidelined for speech segmentation and showed a demonstration with manual annatotation. Mr. A Vadivel followed the session and gave a presentation on ***Speech tools***. The presentation explored the Building systems/ helping systems building of LDC-IL. The students made queries into the technologies LDC-IL had been establishing and developing so far. It was a valuable and memorable minute. On the eight day, 12th June 2014, Mr. N. Rajesha, for the second time, talked about ***Font Encoding***. The students enjoyed the class since it reminded them of the basic character encoding which is used to represent a repertoire of characters by some kind of an encoding system. The next session was followed by the presentation on ***Development of IME***, by Mr. Rupesh Rai. He talked about Imput Method Editors in respect to keyboard layout. The following ninth day, the 13th June 2014, was kept for workshop. In the third

week of internship and on the ninth day, 16th June 2014, Ms. Atreyee Sharma gave a presentation on ***Text Corpus Collection***. She explained the essential methodology and lines of principles to be followed while collecting text corpus. Dr. Arunkumar and Ms. Bi Bi Mariyam gave a presentation on ***Corpus Cleaning***. The presenters explained to a great extent about the thumbs of rules to be adopted while corpus cleaning is going on. On the tenth day, 17th June 2014, Ms. Purva Dholkia gave a presentation on ***POS: Introductory and tagset***. She elucidated about the parts of speech and their levels to be assigned while tagging them in the text provided. She also discussed the standpoint of tagset to be standardized across the languages. Miss. Poonam Dhillon continued the session by delivering a lecture on ***POS Tagging and Issues***. She came forward with issues and put them in the eyes of the students for their linguistic assimilation towards the decision. It was an interactive session in a concrete way. On the eleventh day, 18th June 2014, Mr. M. Md. Yoonus, gave a lecture on ***Text Corpus*** particularly from the technical ground. Mr. Yoonus mainly discussed the progress of Text Corpus using tools developed by LDC-IL. Mr. Saurab Varik continued the session by delivering a lecture on Spoken Language Technology. Mr. Saurabh explained about TTS and ASR and also gave a demonstration for expounding the area under discussion. On *the twelfth day*, 19th June 2014, Mr. Pramod Kumar Raut and Md. Mansoor Khan gave a presentation on **Tokenization**. Both of them explicated the underlying nature of token required for splitting on the basis of word boundary since one word may correspond to a single entity or many entities. Mr. Santosh Kumar Mohanty followed the session and gave a lecturer on ***Sandhi***. He elucidated various morphophonemic rules in Indian languages. The following twentieth day, 20th June 2014 was kept for workshop for the students. In the fourth week of internship and on the thirteenth day, 23rd June 2014, Mr. Amom Nandaraj Meetei gave a lecture on ***Syntax: Karaka Theory***. Mr. Amom remarked that the need for dependency relations somewhere within syntax is generally agreed to. He explained that the Indian grammarian, more particularly Panini recognized the semantic relations that hold between the object denoted by NPs and the event or state denoted by the verb. This notion is then known as 'Karakas' or 'Causal relations'. Mr. Nandaraj Meetei continued the following session and gave another lecture on ***Syntax: Western Theory***. Mr. Amom explained about the western case theories and compared them with Indian Karaka Theory. He explicated that Panini was careful in describing the correspondences of different case forms of the same meaning. On the fourteenth day, 24th June 2014, Mr. Shahid Mustaq Bhat gave a lecture on ***Chunking***. Mr. Shahid described rule-based shallow parser that handles chunking of Indian sentences. It was an interactive moment among the students and the presenter, focusing more on the nature of linguistic properties. Mr. Shahid continued the following session by delivering another lecture on ***Tree Banking***. Mr. Shahid put emphasis on dependency tree banking good for dependency parsing. He also talked about close relation between grammar and Treebank guideline. On *the fifteenth day*, 25th June 2014, Mr. Rajesh Singh gave presentation on ***Indian Sign Language: Alphabets, Words and Story***. He explained about lexical signs which are used for alphabets and gave a demonstration in the form of story out of which he taught various forms of hand shapes used for words and interpretation. Ms. Atreyee Sharma followed the session by giving a lecture on ***Indian Sign Language Corpus Development***. Ms. Atreyee remarked that developing a corpus will directly lead to a better understanding of ISL structure and use. This information is imperative for the education of Deaf children, and for ISL teachers. On the seventeenth day, 26th June 2014, Mr.Amaresh Gopalakrishnan, delivered a lecture on Indian Sign Language: Segmentation and Annotation. Mr. Amaresh mainly focused on the annotation of Indian Sign Language by using a software tool called ELAN. He explained the four levels involved in performing annotation from glossing of conventional signs to the glossing of non-manual features to POS tagging to translation/interpretation. Mr. Umesh Chamling Rai followed the session and gave a lecture on ***Language Documentation***. Mr. Umesh explained what language documentation is and for what it is good. He emphasized that the real time for

preserving the linguistic primary data has arrived and documentation over the data has become a must. The following twenty-first day, 27th June 2014, was kept for workshop for the students. In the fifth week of internship and on the twenty second day, 30th June 2014, Ms. G. Manasa gave a lecture on ***Spell Checker***. Ms. manasa explained about how to automatic spell checker based on the dictionary and grammatical rules of the language concerned. Mr. A. Vadivel continued the following session and delivered a presentation on ***Introduction to Database Concept***. He detailed the perspectives of data management and warehouse procedures. Basic concepts of SQL server database were put into discussion. On the twenty-third day, 1st July 2014, Mr. Shahid Mustaq Bhat gave a lecture on ***Machine Translation***. Mr Shahid remarked that Machine translation, which is also known as Computer Aided Translation, have been specifically designed to translate both verbal and written texts from one language to another. He talked about the benefits to the translators from using machine translation and explained about rules-based and statistical based machine translation systems. Ms. G. Manasa continued the next session and delivered a lecture on ***Transliteration***. She provided a demonstration showing how transliteration from Indian Scripts to Roman Scripts and vice versa. On the twenty-fourth day, 2nd July 2014, Ms. Yumnam Premila Chanu delivered a lecture on ***Linguistics and Other Disciplines***. Ms. Premila stressed that language operates at many more levels and states and so the developments for linguistics and language related disciplines gives rise to an interaction not only for lingual subsystems but also how these as a consequence interact with social, psychological, and other cognitive systems. Mr. Shahnawaz Alam continued the following session and presented a lecture on ***Linguistic Structure Across Languages***. Mr. Shahnawaz talked about the typological dimension concerning not only for variation but also for the limitation on the degree of variation found in the languages of the Indian country, in particular. On the twenty-fifth day, 3nd July 2014, presentations from the students were made shown. The students were divided into team and already allotted assignment for GUI application development, particularly in relation to Language Identifier, Stemmer, Transliteration Tools, Simple Dictionary, Sentence Disambiguation, Metadata Retriever and N-grams etc. All these assignments had been loaded in response to the classes given to them. The valedictory function was chaired by Dr. L. Ramamoorthy, Head, LDC-IL, Reader-cum-Research Officer, CIIL, Mysore. Dr. Ramamoorthy delivered speech encouraging the spirit of the organizing program committee and the students for their valuable participation. Mr. A. Nandaraj Meetei, LDC-IL, and Mr. M. Md. Yoonus, Co-ordinators of the program, extended their thanks-giving speeches. For the comments and suggestions for more improving the programs to follow were from amongst the students and the presenters. The chair person of the closing day declared the ***Internship Course for Rajiv Gandhi University of Knowledge Technologies Students*** closed.