should create and make accessible free of charge, a database on current experiences, standards and technical recommendations, software products and on-going innovations bearing on computerization of languages.

Workshop on Lexical Resources for Natural Language Processing on 5-8 January 2001 at IIIT, Hyderabad

Lexical resources for Indian languages, for example, electronic dictionaries, thesauri, grammars, corpora etc. are far behind those for other languages of the world, even some smaller Asian languages. This in turn is affecting the development of applications. There is a need to make concerted and coordinated effort in building the lexical resources.

While developing a lexical resource, it is important to identify at least one application, which would benefit directly from the resource. Two important applications are: search engine for Indian languages, and machine translation. The following four LRs need to be built: (a) Transfer lexicon and grammar (between English & Indian languages), (b) Annotated corpus for each Indian language, (c) Bilingual dictionary of core meanings, (d) Word Net for Indian languages.

Details regarding how to build each of the above resources (except (d)) and a methodology was worked out. It would require a close monitoring of TIME and QUALITY. The distribution license for the LRs would be "free" software under GPL.

Symposium on Translation Support System (STRANS) on Feburary 15-17 2001 at IIT, Kanpur

The symposium STRANS2001 was first of its kind after CPAL-2 and SMATAC-96 to provide a platform for presenting and discussing all the issues related to machine aided translation focussed to Indian languages including English. Prof. R.M.K. Sinha, Program Chair for the STRANS Symposium, presented a brief review of the on-going activities in the area. The first key-note address delivered at the inaugural session was that from Prof. Harold Somers of UMIST, UK. Prof. Somers group has been working on South Asian languages, which are minority languages in United Kingdom. Prof. Somers presented his thoughts on "A corpus based approach to resource development for low density languages". Dr. Om Vikas, Ministry of Information Technology, Govt. of India, delivered another keynote address on issues related to Indian Language processing vis-a-vis the international scenario.

A special session was devoted to the presentation on TDIL VISION2010document and Standardization issues. There was a general agreement on the Vision 2010 document. It was agreed that standardization of lexical format, various romanization codings, transliteration, OCR data and speech data as contained in the document be created in a phased manner and be forwarded to BIS for making them a national standard. Some people expressed apprehension about their execution in the stipulated time frame and wanted Govt. to take initiatives to coordinate the efforts.

Some of the major recommendations of the panel discussions are summarized below: (1) Software and technology be developed to cater to all Indian Languages and scripts. However, priority be given to developing interfaces between English and Hindi, English and regional languages, and among Indian languages in that order. Funding agencies at the Central and State Govt. levels must exchange information to avoid duplication of efforts or re-inventing of wheel. (2) Lexware is a basic infrastructure for all works on language processing. Lexware work is a highly time consuming, laborious and requires a skill for which training is required. It cannot be created at a single centre or by a single group. Therefore all researchers and developers for easy sharing and portability. (3) Follow a common standard. The TDIL web site maintained by MIT should be the gateway for disseminating and exchanging ideas. All major developments and policy matters should be posted there. Industries and users including state govt. agencies should also be encouraged to post their problems, experiences and exchange solutions through this site. (4) STRANS be made a regular annual affair to provide a platform for researchers, developers, industries and users to report their work and exchange their ideas.

Universal Networking Language on April 10, 2001 at IIT, Bombay

Universal Networking Language is a recent Interlingua proposed by a group of researchers from 18 countries of the world. The United Nations University, Tokyo, leads the UNL project. The overall goal of the project is to enable natural languages on the Internet so that the access of information on the Internet can be done by a large section of humanity crossing the language barrier.

In India, the Indian Institute of Technology, Bombay, is handling the UNL project. The responsibility of IIT Bombay is to link Hindi with the rest of the languages over the Internet. Towards that goal a team of researchers led by Prof. Pushpak Bhattacharyya of the Computer Science and Engineering Department, IIT Bombay is working on conversions of English and Hindi to UNL and also the reverse problem of language generation from UNL.

Since UNL has generated lot of interest in the research community for natural Lang a means for exchange of information on the internet in natural languages. He acknowledged the key role played by three very special individuals in pioneering the UNL project- Dr. Hiroshi Uchida, Dr. Kazuhiko Nishi and Dr Della Senta. He explained the problems faced by the United Nations (UN) related to the "Language Barrier". There are 186 major languages spoken throughout the world out of which six are the official UN languages. The division of UN which is closest with respect to its functions to Information Technology is International Telecom Union (ITU). ITU spends 40% of its annual budget to ensure that its products are translated in all six official UN languages. To overcome this language barrier, Universal Networking Language (UNL) was conceived.

Gradually Hindi also was brought into the umbrella of this project, as India is a country with vast population and great language diversity. Prof Menon further added that, at present, there exist mainly three types of divides in this world. The first divide is the North-South divide, i.e., Rich and the poor divide. The second is the Digital Divide, i.e., the divide between people who have access to the Internet and those who do not. The third divide is the language divide due to the dominance of English language on the Internet. The challenge faced by UN today is how to ensure that these divides break down. Help is needed from commercial centers to take it down the line to the masses. The idea is to have it in public domain and not to make it a commercial property.

Dr. Uchida explained in detail the concept of the Universal Networking Language (UNL). The three pillars of this language are the Universal Words (standing for concepts), the Relation Labels (standing for relationships among the concepts as they appear in a specific sentence) and attribute labels (standing for speaker's attitude with respect to tense, number, mood etc.). The processes of "Deconversion", i.e., generation of natural language from UNL and "EnConversion", i.e., analysis of natural languages into UNL were explained . These two put together lead to (semi) automatic machine translation.

Prof. A. Sen and Prof. S. Sanyal described the ongoing work on speech and language technology in TIFR. They talked about the tools developed for speech processing at TIFR related to segmentation, speaker dependent phoneme identification, speech database etc. It was clear that today's speech research has to draw a lot from the natural language processing research.

Mrs. Tessie George, narrated what a professional translator expects from a machine translation software.

Dr. Mudur gave an overview of various activities like encoding techniques and standards, multilingual text processing and font design, OS level support for localization, Intelligent Information Retrieval and Machine Translation. Mr. Durgesh Rao, NCST, talked at length about Information Retrieval, Machine Translation and Language Engineering related work at NCST, particularly the English-Hindi translation tool, MaTra.

Dr. Sengupta described the ongoing research activities on language and speech technology at Tata Infotech Ltd. i.e creation of speech database was one of TIL's focus.

The last session include the presentations and demonstrations by the UNL Research group of IIT Bombay on construction of Hindi-UW dictionary, EnConversion (analysis) of English text to UNL , Hindi EnConversion process, Hindi DeConversion (generation), Marathi DeConversion, building the Indo Word net, Ontology building, UNL based inferencing, Automatic Hyper Linking using UNL, UNL-ML, i.e., XMLification of UNL. Each of these presentations was followed by small demonstrations.

Indo-UK workshop on Language Engineering for South Asian Languages (LESAL), on April 23-25, 2001 at NCST, Mumbai

An Indo-UK workshop on Language Engineering for South Asian Languages (LESAL), was organized at the National Centre for Software Technology, Mumbai, on April 23, 24 and 25, 2001. The aim of the workshop was to stimulate interaction among the Natural Language Processing communities in the UK and South Asia, for the mutual benefit of both communities, and for the state of the art in Language Engineering in the languages of South Asia.

The workshop covered a wide spectrum of NL activity, including corpus linguistics, lexicography, encoding standards, language engineering tools, and application areas such as machine translation, speech processing, OCR and spell checking.

The major suggestions / recommendations were as follows.

- Linguistic Data: There is scope for consolidation of current efforts, and joint work in building parallel corpora for English and Indian languages. Some of the possible collaborators and sources of parallel corpora include CIIL, CIEFL, the National Book Trust, the Sahitya Academy, news agencies, media and government bodies. This could aid further work in lexical resource building, for example, the LERIL initiative for development of Lexical Resources in Indian Languages. Speech corpora, and corpora for Indian English are other lacuna that need to be filled. Tribal languages may also benefit from such efforts to the extent possible.