International Conference on Universal Knowledge and Language (Goa, India from 25th to 29th Nov. 2002)

Introduction

The ***International Conference on Universal Knowledge and Language (ICUKL2002)*** was co-organized by the *Universal Networking Digital Language (UNDL) foundation Geneva, IIT Bombay* and *Transcultura Paris*.

The *universalizing* trend in the social, economic, political and cultural environment is compelling all nations to search for reciprocal knowledge and

transcultural understanding. These would hardly happen without bridging the gap between the *local knowledge* and *universal knowledge* and without crossing language barriers.

ICUKL2002 was unique in its theme for reflecting on the interdependence of knowledge, culture and language from philosophical, social and engineering approaches. It was also intended to be an opportunity for learning and discussing specific issues concerning the Universal Networking Language (UNL)–which is a language for knowledge access and distribution on the internet–from the point of view of multilingual infrastructure and related linguistic and technological aspects.

The Programme was organized under the broad and provocative theme *Universal Knowledge* and was addressed from four interrelated perspectives: Philosophical, Cultural, Linguistic and Engineering.

The conference was chaired by Prof. M.G.K. Menon. The Advisory committee co-chairs were Prof. Della Senta, Director, UNDL foundation and Prof. Ashok Misra, Director, IIT Bombay. The Program committee was co-chaired by Dr. Hiroshi Uchida of UNDL foundation and Prof. Le Pichon of Transcultura. The co-conveners were Prof. Pushpak Bhattacharyya and Prof. G.V. Parbhu-Gaunkar of IIT Bombay.

ICUKL2002 was organized in two parts :

- **Transcultura Symposium**
- **Knowledge and Language symposium**

The former had 15 invited speakers–all eminent scholars from Europe and India in various disciplines like philosophy of language, linguistics and culture. The theme of this track was the exploration of the idea of *Encyclopedia of Keywords.*

The latter had 6 invited speakers from major speech and language processing groups of this country including Ministry of IT, IITs at Chennai, Mumbai and Kanpur, NCST, IIIT Hyderabad *etc.* The *Technology Development in Indian Languages* initiative too was presented. There was a special session on *Konkani Languages.* In this session were discussed the History of Konkani Language and efforts on Konkani corpus creation. The **research tracks had 21 refereed papers** covering the following topics:

- UNL applications and tools.
- Machine translation and lexical resources.
- Information Retrieval *in parallel with* UNL enconverting workshop.
- Speech and visual information processing.

The last day also featured a **panel discussion titled *creating universal knowledge across language barriers***, chaired by Prof. M.G.K. Menon and participated in by Prof. P.V.S. Rao of TIL, India, Prof. J. P. Contzen of EU and Prof. Allan Le Pichon of Transcultura Paris.

Invited Speeches

Prof. M. G. K. Menon, spoke on the *Implications of Information Technology for Diversity of Languages and Cultures* touching upon a range of broad issues like the language barrier, the UNL project, the language policies and the deep impact of the internet on the languages. In his speech on *knowledge management and linguistic pluralism,* Shri Rajeeva Ratna Shah, Secretary, Department of Information Technology, addressed the problem of transferring linguistic information and then gave an overview of the major language technology initiatives in the country. Dr. Om Vikas from the Ministry of Information Technology- who is the prime force behind the Technology Development in Indian Languages projects- gave a talk titled *Annals of Indian Language Computing*, in which he traced the history of Indian language computing endeavors.

Dr. Hiroshi Uchida, the architect of the UNL system, and Ms Meying Zhu and Dr. Ronaldo Martins- researchers from the UNDL Foundation- gave an

exposition of the UNL system covering *the basic constructs of the UNL, the knowledge base* and *the universal word lexicon*. Prof. Gérard Chollet from the University of Paris talked about *the use of UNL in key words, key-images and key-concepts* in a transcultural setting. The speech on *Building Infrastructure for machine translation research: an Indian Perspective* by Prof. R.M.K. Sinha of IIT Kanpur dealt with the famous *Anglabharati* and *Anubharati* machine translation systems developed at IIT Kanpur. Prof. B. B. Choudhary of ISI Kolkata of Indian language OCR fame, described strategies for building *Spell Checkers in Indian Languages*. In his speech *a common architecture for machine translation for Indian Languages*, Prof. Rajeev Sangal, Director of IIIT Hyderabad, emphasized the need for building lexical resources, using machine learning techniques on corpora and touched upon the *Anusaaraka project* for MT among Indian Language and for between Indian Languages and English. Prof. B.Yegnanarayana of Indian Institute of Technology Madras delivered an exposition on *Speech technology in Indian context* narrating his many years of work on speech recognition and generation. Prof. Pushpak Bhattacharyya of IIT Bombay talked on *Interlingua Based Machine Translation and the Development of Lexical Resources* describing his group's research on tri-lingual MT system for English, Hindi and Marathi based on UNL and also the construction of the first wordnet in India for Hindi. Mr Walwalikar and Mrs. Madhavi Sardesia, linguist and littérateur of Goa, spoke on Konkani corpora development at *Asmitai*- an industry in Goa; the Konkani corpora is one of the best in the country today.

Prof. Subhadra Joshi of University of Mumbai spoke on the *Universality of key concepts in Vedic tradition.* Prof. Franson Manjali of JNU dwelt on *time, discourse and transculturality.* Mr. Uday Bhembre and Mrs Kiran Budkule, linguists and litterateurs of Goa, traced the *history of the Konkani language and culture*. The following speeches from Transcultura scholars were deeply appreciated by the participants:

- *Under the sky, a Chinese approach of "universality"*: Zhao Tingyang.
- *Strategy for Reciprocal Knowledge, Reciprocal anthropology and the question of universality or Key Words: Transcultural approach of universality and difference*: Alain Le Pichon.
- *Project of a Transcultural Encyclopedia of Key words. Words' meanings can change: an example about marriage in French language history - methodological proposals*: Pierre Varrod.
- *Zero as a key word for a transcultural approach to science*: Jacques Vauthiers.
- *Keywords and key concepts in economy: a transcultural approach of globalization:* Gustavo Martin Prada.
- *The principle of mutual recognition in the E.U. economic policy- about different possible understandings of a key-concept*: Fiorella Padoa.
- *Reciprocal Visions and Image Feedbacks: a visual experience of reciprocal anthropology in Amazonian cultures*: Patrick Deshayes.
- *Distance Learning: a EU-India Perspective* : G.V. Prabhu-Gaunkar and Allan Le Pichon.

Prof. P.N. Murthy, Tata Consultancy Services, Trivandrum spoke on *language learning- a method of imparting reading capability* prior to writing capability to an adult illiterate in the language spoken by him.

UNL Encoding Workshop

A highlight of the conference was a half-day workshop on encoding conventions in UNL, which Prof. Igor Boguslavsky of UNL-Russia Center led. Prof. Christian Boitet of UNL-France center, Prof. Irina Prodanoff of UNL-Italy center, Dr. Ronaldo Martins of UNDL Foundation and Mrs Shashi Palekar and Mr. Salil Barodekar of UNL-India center gave presentations on UNL representation, language independence, lexicon, language divergence and such other deep computational linguistics issues.

Conclusions

Several important conclusions were reached at the end of the conference. It was decided that

- To spread the awareness of the UNL system, a summer school will be held in IIT Bombay, in which NLP researchers from educational institutions and industries of India will participate. The lectures and the hand-on will be delivered by the researchers from UNDL foundation and IIT Bombay.
- The Konkani corpus will be stored in IBM PC format and will be processed for various frequency information. This is turn will be used for Konkani lexicon, Konkani morphology, Konkani wordnet and Konkani machine translation software.
- The next conference on universal knowledge and language will be held in Alexandria, Egypt.
- The Keyword Extraction project of Transcultura will explore the possibility of using UNL.
- The effort on creating lexical resources like machine readable dictionaries, wordnets and ontologies will be encouraged and intensified.

(Courtesy :Pushpak Bhattabaryya
Department of Computer Science & Engineering
Indian Institute of Technology Bombay
E-mail : pb@cse.iitb.ac.in
Website :http://www.cfilt.iitb.ac.in/icukl2002)

International Conference on Natural Language Processing (ICON) (NCST, Navi Mumbai, India from December 18 to 21, 2002)

Introduction

The International Conference on Natural Language Processing (ICON-2002) was co-organized by Language Technologies Research Centre, International Institute of Information Technology, Hyderabad and NCST.

Natural Language Processing (NLP) as a field has come of age with several practical mass applications varying from information retrieval to extraction and from machine translation to question answering. The availability of electronic texts in large amounts has brought about a major change. it has led to corpus-based study of actual language rather than idealized versions of it. The greatest change, however, is the recent phenomenal rise in the use of a combination of grammatical and statistical methods with an important role of machine learning. Machine learning works well when the information or rule to be learnt is relatively simple. This has put great demand on lexical resources such as annotated corpora annotated with the right information and grammar, dictionaries, collocations data bases etc, all of which need linguistic insight to be woven into machine learning systems. As a result, the recent machine learning trends demand a synergistic relationship with Linguists and Computer scientists. All this has brought about a significant change in the way research is done and the way it is reported, spurring the NLP research activity to new heights.

ICON-2002 was organized to start a high-quality annual conference on NLP in India with two main purposes. First, to create a platform where high-quality research on NLP gets reported and the researchers meet to exchange ideas and latest results. Second, to create a common forum for Linguists, Computer Scientists and others so that the researchers from different communities come together to learn from each other and collaborate with each other. The conference was organized along with Knowledge based Computing Systems (KBCS-2002) at the same venue which led to exchange of ideas between researchers across the two conferences.

Dr. P.V.S. Rao, Tata Infotech, Mumbai jointly inaugurated ICON-2002 and KBCS-2002 on December 19, 2002.

Key Note Address

Prof. Aravind Joshi, eminent Computational Linguist, University of Pennsylvania, Philadelphia, USA gave a joint keynote address to ICON-2002 and KBCS-2002 after the inauguration. He spoke on " Starting with Complex Primitives Pays Off: Complicate Locally, Simplify Globally" .The conventional approach in setting up formal system to specify a grammar formalism is to start with simple basic primitive structures and then introduce more complex structures. In his address, Prof. Joshi proposed an alternative approach of starting with complex primitives which directly capture some crucial linguistic properties and then introducing general operations for composing these complex structures. He described how this approach has led to some new insights into syntactic description, semantic composition, language generation, statistical processing and psycholinguistic properties. Prof. Joshi also discussed how the

approach is applicable to the description on some biological sequences.

Prof. Benjamin K. Tsou, City University of Hong Kong, a distinguished scientist working in NLP, gave an invited talk on "Natural Language Processing, Information Mining and the new Global Village" on December 20, 2002. He presented an in depth analysis on how the rapid developments in information technology have necessitated sophisticated efforts in information extraction due to major challenges in the identification of new or unknown words, particularly the proper nouns and named entities.

Panel Discussion

A Panel Discussion was arranged by Dr. Dipti Misra Sharma, LTRC, International Institute of Information Technology, Hyderabad on" Role of Linguists in Building Natural Language Processing Systems" Prof. Aravind Joshi chaired the panel and moderated the discussion. Distinguished scientists such as Dr Shivaji Bandyopadhyay, Jadhavpur University, Kolkatta, Dr. Pushpak Bhattacharya, IIT, Mumbai, Dr. Achala Misri Raina, IIT, Kanpur, Prof. Udaya Narayana Singh, CIIL, Mysore and Dr. Om Vikas, Ministry of Communications and Information Technology, New Delhi were invited for participation in the discussion. The discussion led to an agreement that the linguists would contribute towards computational grammar and building linguistic resources, particularly the tree-bank. It was proposed to start a new course on Computational Linguistics at various institutes. In addition, introduction of Computational Linguistics as a new subject in the existing syllabus was also suggested.

A Special Session on "Linguistics and Natural Language Processing was arranged on December 19, 2002. The session was chaired by Prof. U.N. Singh, Director, CIIL, Mysore. The eminent Linguists and Computer scientist who were invited to present their views were Dr. Om Vikas from Ministry of Communications and Information Technology who is the prime force behind the Technologies Development in Indian Languages projects, Dr. Lakshmi Bai, LTRC, International Institute of Information Technology, Hyderabad Prof. K. SubbaRao, University of Delhi, and Dr. Pramod Kumar, Kendriya Hindi Sansthan, New Delhi.

Paper Presentation

At ICON-2002 about 30 selected papers were presented in 9 sessions spread over three days. The papers covered a wide spectrum of NLP areas such as Morphology. Parsing, Semantics, Speech Processing, Statistical Language Modeling, Information Retrieval, machine Translation, Dialogue representing the current trends in NLP.

Pre-Conference Tutorials

The conference was preceded by four half-day Pre-Conference tutorials on December 18, 2002. The tutorials were in diverse NLP areas and received good response. The tutorial titles and presenters were:

- Automatic Text Summarization by Dr. Inderjeet Mani, The MITRE Corporation, Virginia, USA.
- Introduction to Computational Linguistics by Amba P Kulkarni and Dr. Dipti Misra Sharma, International Institute of Information Technology, Hyderabad.
- How to Sketch Words by Dr. Adam Kilgarriff, ITRL, University of Brighton, Brighton, UK.
- Interlingua based Information Processing in the Context of Indian Languages by Dr. Pushpak Bhattacharyya, IIT, Mumbai

ICON-2003

In the concluding session, which was common to both ICON-2002 and KBCS, Prof. Rajeev Sangal announced that ICON will be an annual feature. He also announced the formation of Natural Language Processing Association of India (NLPAI). ICON 2003 will be organized by NLPAI in 18-21, December 2003, at CIIL, Mysore (India).

(Courtesy : S.M. Bendre
Department of Mathematics and Statistics
University of Hyderabad, Hyderabad
E-mail : smbsm@uohyd.ernet.in
Website : http://www.iiit.net/conferences/icon2002.html)

Workshop on Spoken Language Processing (TIFR, Mumbai, India from January 9-11, 2003)

Introduction

A workshop on Spoken language Processing was organized by TIFR and was supported by the International Speech Communication Association (ISCA).

It is now widely accepted that in order to draw majority of the Indians to the ambit of the ongoing IT revolution, alternatives to English-type keyboard-monitor input/outputs are to be provided. Thanks to TDIL and similar efforts, Indian language keyboards, word processors and other utilities are now realities. But that still doesn't cover a sizeable number of illiterate and barely literate. Communication through speech is a simple way to realize the goal of interaction of masses with computers. It is not far from technological feasibility too at the current era.

Several organizations in India have been pursuing research and development activities in computer speech for decades. Many new efforts are also coming up in recent times. One of the problems faced by such R & D groups is their sub-critical size. This workshop aimed to provide a platform for such researchers to learn from each other and, hopefully, to evolve a common roadmap for future.

With that overall aim, the workshop started with tutorials on the basics of speech signal processing, speech recognition and speech synthesis. The tutorials were well attended, among others, by students and representatives from industries aspiring to expand to speech technology. The main workshop schedule included invited talks from experts, (contributed) oral as well as poster paper presentations and group discussions.

The proceedings of the workshop, containing the presented papers and some invited talks, is available at http://speech.tifr.res.in/wslp/proceedings in electronic form.

Invited Talks

Prof. Hiroya Fujisaki, Professor Emeritus, University of Tokyo, was the keynote speaker. In his address, he dealt, among other things, on **prosody**, currently a very relevant issue for text-to-speech synthesis in Indian languages. Backed up by his vast experience he grouped the information present in speech in relevant categories and, elucidating the underlying physical and physiological mechanism of speech production, specified ways to model each of them. Contrasting the standard empirical approach, he thus put prosody on a firm theoretical footing. This helps **across-the-language** study of prosody, very important for speech research in India.

In another special lecture, Prof. Fujisaki discussed **information retrieval through human-machine spoken dialogue**, another issue of great relevance in India. He described the ideas of (a) user and system modeling for efficient dialogue management (b) use of key concepts in information retrieval and (c) optimization of information retrieval through relevance estimation. He exemplified the concepts with the help of an academic information retrieval system by speech, developed in his laboratory.

Prof. P.V.S. Rao, Adviser, Tata Infotech, Mumbai, delved on speech recognition concepts at depth. Starting with **HMM**, the most popular methodology and then describing **HMM2**, its recent variant, he discussed the problems of speech recognition that could be successfully addressed by it and also the ones that could not be. He cited results from his vast experience and pointed to the enormity of the problems, the tasks still ahead and concluded with the suggestion of what may be termed as a 'holistic' approach for speech recognition.

Dr. S. S. Agrawal, Emeritus Scientist, CSIO, CSIR, narrated first-hand experience on applying KLSYN, currently the best formant synthesizer, for synthesizing speech in Indian languages. He also described the current research on gradually applying 'high level' articulatory features to KLSYN that will ultimately be able to rid formant synthesis of all heuristics and make it easy-to-implement. He presented many details of the R & D work done at the CEERI, New Delhi centre in collaboration with the Speech Communications laboratory, MIT, USA, for developing text-to-speech systems in Hindi and Bangla.

Prof. B. Yegnanarayana, IIT, Chennai, in his rather radical presentation, questioned the wisdom of over-dependence on very narrow and conventional signal analysis for interpreting speech. He cited the example of phase spectrum, ignored in conventional speech signal processing, that he had made very good use of in his research. He then elaborated the complexity of the complete process of speech understanding by human beings, of that signal processing forms merely a small part. He opined that success may elude us, unless we shed the narrow outlook and try to understand the marvel of cognition of speech in its profundity.

Oral Paper Sessions

There were five oral paper sessions on the following topics: (a) **speech processing** (b) **speech recognition** (c) **speech synthesis** (d) **hardware implementation**

and (e) **language identification**. While several presentations reflected quality research on basic issues, a distinct trend to link the research with tangible objectives was observed.

In **speech processing** session, **P. Patwardhan** and **Preeti Rao** proposed an innovative alternative to the standard Bark scale frequency warping, that may result in improvement of the quality of speech in low bit rate coding. **S.R.M. Prasanna, J.M. Zachariah** and **B. Yegnanarayana** proposed an improved method for end-point detection of speech, based on knowledge of vowel onset points, that was successfully employed to enhance performance of a speaker verification system. **G.V. Kiran** and **T.V. Sreenivas** presented a modified form of gammachirp filter bank for human auditory modeling that is expected to cover a wider range of phenomena and is also easier to implement. Presentation of **B. Prakash** is on the relevant acoustic measures that can differentially diagnose stuttering from normal non-fluency. The research is a useful step towards early diagnosis and treatment of children with stuttering.

In **speech recognition** session, the paper of **C. Chandra sekhar**, **W.F. Lee**, **K. Takeda** and **F. Itakura** compared the performance of different classification models for recognition of subword units of speech and shows that Support Vector Machine (SVM) based classifiers outperform the conventional HMM based ones. **R. Sinha** and **S. Umesh** presented a way to modify the standard Mel Frequency Cepstral Coefficient (MFCC) method with Weighted Overlapped Segment Averaging (WOSA) for better speech recognition performance. **D. Kanejiya**, **Arun Kumar** and **S. Prasad** presented a mathematical framework 'Syntactically Enhanced Latent Semantic Analysis' (SELSA) for language modeling for speech recognition.

In **speech synthesis** session, **A. Sen** presented the rule formulation methodology for an Indian English text processor that is the front-end of an 'Indian accented' English text-to-speech system. **P.K. Lehana** and **P.C. Pande** presented Harmonic plus Noise Model (HNM) for speech synthesis and applied it to analyze and synthesize speech sounds, including the ones typical for Indian languages. Paper of **A. Bandopadhyaya**, **B. Pal** and **S.K. Das Mondal** described a method of synthesizing intonated speech, using a prosody generation model, for a Bangla concatenation synthesizer.

The lone, but remarkable paper presented on the **hardware** session, by **R. Saini**, **S. Srivastava**, **A.S. Mandal**, **S. Kumar**, **R. Singh**, **A. Karmakar** and **Chandra Sekhar**, described the design of a special-purpose processor for hardware implementation of functions needed in a Klatt-type formant synthesizer. Its successful implementation will enable us to put a formant synthesizer into embedded systems.

In the **language identification** session, **T. Nagarajan** and **Hema Murthy** presented a new approach to language identification, using vector quantization and pairwise multiple codebook, that showed appreciable improvement in language identification accuracy with a standard international multi-language speech corpus. **V. Ramasubramanian**, **A.K.V. Sai Jayram** and **T.V. Sreenivas** extended the parallel phone recognition (PPR) concept for automatic language identification and after comparing various classifiers, concluded that the best performance is from maximum likelihood classifier.

Poster Paper Session

In the poster paper session, several interesting ideas were displayed. **M. Nagarajan** and **T.V. Sreenivas** proposed 'Product HMM', an integrated statistical model that provides a way of integrating different HMMs that model the sub-sequences of a complete vector sequence. **M. Vandana** proposed two non-linear models 'Quadratic Predictive Coding' and 'M-ary Predictive Coding', as against the standard linear prediction models. **S.R. Savitri** and **H. Rohini** presented the results of a study on the role of cerebellum in voicing perception for patients with cerebellar pathologies. **K. Samudravijaya** and **M. Barot** compared the speech recognition results with HTK and Sphinx, two freely downloadable software packages that help to build continuous speech recognition systems and concluded that while Sphinx gives better results, HTK is more user-friendly. **R. Verma** and **P. Chawla** presented results of their study on differentiating between dental and retroflexed consonants in Hindi that is very useful for Indian language speech synthesis. **S. Sethi** and **K.S.R. Anjaneyulu** outlined the scheme of their voice email (V-Mail) project, that proposes to use both speech recognition and synthesis. **A. Gupta** and **M. Sandeep** presented the results of a perception study of recognizing lexical and non-lexical words by a number of listeners. **S. Arora**, **K.K. Arora** and **S.S. Agrawal** presented a software tool 'Vishleshika' that can do various statistical analysis on Hindi text.

Discussion Sessions

(a) Informal discussion on development of Speech Databases in Indian Languages

During the workshop days, an informal meeting was held to discuss the issue of spoken language corpora in Indian Languages. The participants included Indian scientists with experience in development of speech databases as well as technology developers with an interest in databases for development of Indian

language applications. An informal list of Indian Language Speech Databases developed at various organizations was prepared and the possibility of sharing them within the Indian speech community was discussed. The gathering overwhelmingly felt the need for creation of larger and varied speech databases, e.g., with isolated words (application specific vocabulary) and also with continuous speech (read, spontaneous and dialogue modes); in different recording environments such as in sound-treated chamber, normal office environment and through telephone channel. Such databases need to be created in all Indian Languages.

(b) Panel Discussion

A panel discussion titled 'Speech science and technology for Indian languages - role and responsibility of academia and industry' was held as the concluding session of the workshop. Several representatives from academia and industry participated in the discussions. The following contains some of the opinions expressed by the delegates.

Prof. T.V.Sreenivas, I.I.Sc., the coordinator of the session, initiated the discussion by listing the salient points for discussion, e.g. long term research at academic institutes, support from government and industry for the same, roles and responsibilities of academia and industry in developing spoken language interfaces, the urgent need for creating speech databases and developing human resources.

Prof. B. Yegnanarayana, IITM, cautioned that a lot of thinking should be done prior to creation of spoken databases of Indian languages. He also suggested that the planners should have a vision of the types of database and the specific purposes for which the database will be used. The speech data need to be labeled by persons trained in the process. Courses on Speech Technology should be introduced as electives in all IITs and emphasis should be placed on basics.

Prof. Fujisaki, Japan said that governments should continue to support non-profit, long term, language independent research in speech area till speech technology matures. Not-so-good performance of recognition systems should not discourage developers. There are always applications that can exploit systems with even limited accuracies.

Prof. Hema Murthy, IITM suggested that one of the immediate applications in the Indian context is limited word (say 40 words) speech recognition systems that can be deployed in kiosks. In addition, these systems should have multi-modal interfaces.

Dr. Anjaneyulu, HP Labs pointed that collaborative efforts are needed to realize the dream of speech-to-speech translation. The speech database creation effort should be taken up on a priority basis. A 'special interest group' should be formed to expedite discussions, decisions and actions to develop speech databases for Indian languages on a sharable basis. It is also worth considering Sphinx4, an open source, java based, speech recognition system for software development effort.

Prof. Umesh, IITK said that attitude of industry also needs a change. Instead of expecting academia to provide fully developed systems, the industry should collaborate with academia in R & D efforts that may not always result in complete systems.

Prof. P.V.S Rao, TIL, Mumbai assured that Tata Infotech (TIL) would welcome collaborative effort between academia and industry and will accept even partially developed idea from academia for joint development into product. It will also consider collaborative projects primarily dealing with pursuit of promising concepts.

Dr. Nandini Bondale, TIFR suggested that it would be good to start a newsletter dealing with issues in spoken language technology. **Mr. A. Sen, TIFR**, added that a web site containing information about important references, source of software, tutorials should also be set up. **Dr. Dalvi, AYJNIHH, Mumbai**, said that it is desirable that interaction should extend to groups dealing with other areas of speech such as speech pathology data processing. **Dr. V. Ramasubramanian** said that the government and industry should support research on basic issues in speech science that are being conducted in academic institutions. The industry should look upon academia to provide solutions to core problems of speech technology; the industry should concentrate on language and application specific issues in order to develop products. **Mr. Sai Jairam, CDOT** stressed that the industry should motivate academics, through collaborative programmes, to take speech technology closer to the community. CDOT would welcome such programmes.

In summary, the following points were emphasized: **support of government and industry** for long-term research in speech technology; **collaborative programmes between academia and industry** for taking the speech technology closer to the community; **development of spoken language corpora in Indian languages**.

(Courtesy : K. Samudravijya A. Sen, School of Technology and Computer Science,TIFR
E-mail : asen@tifr.res.in
Website : http://speech.tifr.res.in/wslp proceedings)

Indo-Wordnet Workshop (CIIL, Mysore, India from 14-16 January, 2003)

Recognizing the immense importance of lexical resources, the Indian languages Wordnet Workshop was jointly organized by the Central Institute of Indian Languages (CIIL) Mysore and Indian Institute of Technology (IIT) Bombay from the 14th to the 16th of January, 2003.

The objective of the workshop was to explore methodologies for constructing the wordnets for Indian languages and then linking them internally to produce the Indo-wordnet which eventually would be linked to the English wordnet and the Euro-Wordnet (a conglomeration of European languages' wordnets).

In India, wordnet building activities are going on for Hindi and Marathi at IIT Bombay, Tamil at Anna University Knowledge Based Center (AU-KBC) Chennai and Tamil University Tanjavur, Gujarathi at MS University Baroda, Oriya at Utkal University Bhubaneswar and Bengali at IIT Kharagpur. The Hindi wordnet is at an advanced stage of development with about 11000 semantically linked synsets and with the associated software and the user interface.

On the first day, the Director of CIIL, Prof. Uday Narayan Singh welcomed the participants representing all the major languages of India. Prof. Singh stressed the need for utilizing the enormous amount of linguistic work in the country for the purpose of wordnet building and the need for setting up of a website where related information, software and resources will be kept in a browsable and freely downloadable form. Dr. Jayaram of CIIL and Dr. Pushpak Bhattacharyya of IIT Bombay explained the goal of the workshop and described the milestones that are expected to be achieved in a year's time.

A day-long tutorial was delivered by Dr. Bhattacharyya on the fundamentals, methodologies and the applications of the wordnet. The first wordnet of the country- the Hindi wordnet- is being built at IIT Bombay. The concepts of (i) Synsets, (ii) Semantic Relations and (iii) the Interface of the wordnet were explained. Since the synsets are the building blocks of the wordnet, considerable amount of time was spent on describing the structure, principle of creation and the associated parts of a synset. It was repeatedly stressed that the words may be polysemous, but when more than one synonymous word is put together, a unique meaning emerges. For example, the synset {ghar, griha and makaan} denotes the unique sense of "residence". This sense is attached as a "gloss" and is exemplified by a simple sentence. For example in the synset {ghar, griha, makaan, aalay, sadan}, 'manushya kaa aavaassthal', "raam kaa ghar mandir ke paas hai". -synset- -gloss- -example the gloss and the example are shown as above. The gloss plays a very important role in the wordnet since it is through this that the synsets are linked across wordnets. Thus, in the Indo-wordnet, the language specific wordnets are expected to have identical glosses and examples as far as possible. The advantage of this is the possibility of creating a multiway parallel corpora.

Dr. Bhattacharyya stressed that the synsets should be constructed abiding by the three principles of

- Minimality (the minimal set of words to make the concept unique)
- Coverage (The maximal set of words- ordered by frequency in the corpus- to include all possible words standing for the sense)
- Replacability (The example sentence should be such that the most frequent words in the synset can replace one another in the sentence without altering the sense)

In the above example, {ghar, makaan} is the minimal set, the rest of the words cover the concept and the words "ghar, griha" and "makaan" can replace one another in the example sentences with minor changes to the sentence structure.

The semantic relations of "hyperonymy/hyponymy", "meronymy/holonymy", "antonymy", "gradations", "entailment" and "troponymy" were explained with examples; so was the importance of cross parts of speech linkages and the connection between the wordnet and the ontology.

Next day, Mr. Nitin Verma of IIT Bombay demonstrated the application of the wordnet in automatically creating document specific dictionaries. The words obtain their disambiguators and semantic attributes from the wordnet. Dr. Rajendran of Tamil University, Tanjavur described their effort on the construction of the Tamil wordnet. Using an ontology- motivated by Nida's concept classification- the Tamil wordnet is created through the following steps: (i) extraction of words from the dictionary, (ii) grouping of words into domains and sub-domains and (iii) arranging the groups hierarchically. Dr. Sudeshna Sarkar of IIT Kharagpur described their work on the Bengali wordnet and placed it in the context of their other language processing activities.

After this, all the participating language groups exchanged notes on an exercise done on a 100 synset sample from the Hindi wordnet provided by IIT Bombay to all the language groups 3 months prior to the workshop. These 100 synsets cover all the parts of speech and were from major conceptual categories like natural object, action, quality etc. It was interesting to observe how words assume different shades across languages, how the glosses become tricky to create for commonly used terms, how words prefer collocations, how example sentences often are

directly adaptable with minimal changes from one language to another "close" language. While this exercise was going on it was realized that the following need discussions:

- The ontology behind the wordnet.
- Compositional approach to the construction of the wordnet.
- Culture specific considerations in the wordnet.
- Specialities of Indian verbs.

On the last day of the workshop, discussions continued around the sample synsets. Dr. Bhattacharyya emphasized that the glosses in the wordnet explicate the synset senses, but cannot really be encyclopedic, scientific or legal definitions. In explicating the senses, they are assisted by the members of the synset and also the accompanying example sentences. Since the gloss is used for linking and creating the synsets it was decided that

- the glosses will be short and simple.
- they will be expressed both in the specific language and in English.
- the example sentences also will be simple and precise; idiomatic and poetic expressions will be avoided.

Highlighting again was done of the fact that the glosses and the example sentences will give rise to multiway, parallel corpora.

Following this, Dr. Rajendran described their work on ontology. The top ontological categories are "things", "events", "abstracts" and "relationals" which correspond to "concrete nouns", "verbs", "adjectives and abstract nouns" and "postpositions and case markers" respectively. The details of the ontological categories at various levels were discussed. Dr. Uma Maheswar Rao described the componential approach to the wordnet creation. Introducing the interesting notion that "words are bundles of semantic features which are binary and parallel to those in phonetics", Dr. Rao proposed that a space of semantic features be designed and the words sharing ALL and ONLY a set of common semantic features be inserted into the same synset. He gave examples to illustrate this idea. Dr. Bhattacharyya suggested this highly interesting approach be worked out in detail especially for the "abstracts". He observed that the features- once detailed out- can be attached to the synsets of existing wordnets.

Verbs in Indian languages show some unique features like (i) conjunct verbs (ii) compound verbs (iii) causative formation {iv) pairings and (v) onomaetopia. Dr. J.C. Sharma of CIIL Mysore described the tests for conjunct verbs (nominal + verb) and compound verbs (verb + verb, with the second verb serving as the vector/explicator/intensifier). Not every nominal and verb combination qualifies as a conjunct verb; the whole unit must behave like a simple verb and the agreement must take place with entities outside the combination. Dr. Bhattacharyya brought up the computational issue of storing the verbs in the wordnet. It was decided that

- Conjunct verbs will be lexicalised in the wordnet.
- Compounds and all the other phenomena will be dealt with by a separate morphological module serving as the front end to the wordnet.

Dr Lalita Handoo showed with examples- especially from Kashmiri- how very culture specific concepts do not have their parallels in other languages. Their linkages with the synsets of other languages remains a question. A viable approach could be linking indirectly through the hyperonymy relation- suggested Dr. Bhattacharyya. Dr. Rajasri of CIIL said that the concepts could be classified as (i)universals across world's languages (ii)universals across Indian languages and (iii)those specific to individual languages. Initially the groups should concentrate on (i)and link with one another.

At the end of the workshop the following resolutions were adopted:

- By the end of 2003 each Indian language will create a wordnet of 5000 synsets. These will be for about 2000 most frequent content words in each language. Use will be made of the wordlist sorted by frequency- available with the CIIL.
- The language specific wordnets {are being}/{will be} developed by the following institutions-
  - CIIL Mysore: Kannada, Kashmiri, Punjabi, Urdu, Himachali
  - IIT Bombay: Hindi, Marathi and Konkani (in collaboration with the Goa research group for the last mentioned)
  - AU-KBC Chennai and Tamil University Tanjavur: Tamil and Malayalam
  - University of Hyderabad: Telegu
  - IIT Kharagpur: Bengali
  - University of Baroda: Gujarati
  - Utkal University Bhubaneswar: Oriya

Research groups have to be identified for building the wordnets of Assamese, Nepali and Languages of the North East.

(Courtsey : Pushpak Bhattacharyya, Indian Coordinator : Indian Wordnet Effort, Department of Computer Science & Engineering Indian Institute of Technology Bombay.
E-mail : pb@cse.iitb.ac.in
Website : http://www.cfilt.iitb.ac.in/wordnet/indowordnet.txt)