3.0 Conference/Workshop Reports

3.1 SIMPLE' 04

Symposium on Indian Morphology, Phonology & Language Engineering

19th – 21st March 2004

Indian Institute of Technology, Kharagpur

The Communication Empowerment Laboratory and Department of Computer Science & Engineering, Indian Institute of Technology Kharagpur in collaboration with Central Institute of Indian Languages, Mysore organized a symposium on Indian language Natural Language Processing (NLP) with special focus on the languages of Eastern India. From 19th to 21st March 2004 and was attended by more than 100 delegates and several stalwarts in the fields of linguistics and NLP from all over India. SIMPLE'04 was sponsored by Ministry of Communication and Information Technology, Govt. of India, NLP Association of India, CIIL Mysore and Intel – Innovation in Education, besides IIT Kharagpur.

Western languages, like English, have long benefited from the use of computers with more than fifty years of NLP research. In India, several groups are engaged in research in Indian Language Speech processing and NLP. It has been identified that for multilingual India, the language barrier forms a major impediment towards alleviation of the communication bottleneck. The Communication Empowerment Laboratory (CEL) at IIT Kharagpur cherishes a dream — to develop and deploy 'appropriate' and 'affordable' information and communication technology and tools to empower the 'other billion' to a better life. They identify communication barrier to be one of the major hurdles in bridging the digital divide. The segment of people most adversely affected by the communication barrier includes rural *neo* and *pre* literates and people with special needs like the physically and mentally challenged.

To address these problems, the Communication Empowerment Laboratory has embarked on several projects, which are being deployed and used by the people. These include the Braille Transliteration System for the visually handicapped (being deployed by Webel Mediatronics and other agencies), Talking Communication System for the speech impaired and cerebral palsy affected children, Adult and Child literacy System, systems for converting Bengali and Hindi texts to speech output etc. These activities are being sponsored by agencies like Media Lab Asia- Ministry of Communication & Information Technology, Microsoft Research, and Texas Instruments etc. Many of these projects have significant NLP research components. The laboratory is seriously engaged in developing Indian language processing tools and systems, that include, morphology analyzer and synthesizer, grapheme to phoneme mapper, Part of Speech Tagger, Local Word Grouper etc. These tools are also being employed to develop English to Bangla translator and translators for Bangla to other Indian languages. The laboratory will also make available these tools through CIIL.

The effort towards organization of SIMPLE was a major step to address a commonly felt need — that of building and sharing of electronic resources such as dictionaries and corpora for Indian languages, and language processing tools that are being developed at different research laboratories in India.

The symposium was inaugurated by Prof S. K. Dube, director, IIT Kharagpur while Dr. R. K. Arora, Senior Director, MCIT graced the occasion as the chief guest. Prof. Pabitra Sarkar, the eminent linguist and educationist, delivered the keynote address. The major issues which were discussed during the symposium include electronic resource development, morphology (deriving the meaning of a word from its structure) and phonology (the pronunciation rules), machine translation and applications based on NLP such as spelling and grammar checkers, synthetic voice based assistive devices for the physically challenged and Indian language computer interfaces.

SIMPLE'04 was enriched by the presence of eminent scientists such as Prof Rajeev Sangal, Director, IIIT

Hyderabad, Prof B. B. Chaudhuri, ISI, Kolkata, Dr. Om Vikas, Senior Director, Ministry of Communication and Information Technology (MCIT), Govt. of India, Prof B. N. Patnaik, noted linguist and professor of IIT Kanpur, Prof G. Uma Maheshwar Rao of University of Hyderabad among others.

In order to facilitate the participation of young researchers in the area registration was made free while CIIL Mysore provided the travel support to the delegates. Around 35 papers were presented and several Indian language based NLP systems were demonstrated during the symposium.

For details about SIMPLE visit http://www.mla.iitkgp.ernet.in/~simple/

Courtesy: TDIL Programme,
Department of IT,
Ministry of Communication and IT,
New Delhi-110003
E-mail: tdilinfo@mit.gov.in

3.2 "Crossing the Digital Divide": The SCALLA working conference

The SCALLA project aimed to increase collaborations between European and South Asian workers in human language technologies and software localisation. The SCALLA working conference was organised in Kathmandu, Nepal in January 2004. This conference was the final conference of the SCALLA project, funded by the European Union under the Asia IT&C programme. The first conference was held in Bangalore in 2001, followed by many visits to Europe by researchers from South Asia, culminating in this final conference in Kathmandu in January

The aim of this final conference was to consolidate what had learn over the previous two and a half years, and to look to the future. As well as revisiting computational linguistics for South Asian languages, it considered in greater detail the localisation of software to those languages and the reasons why this is important. As well as technical support for languages, it also considered social and policy issues concerned with linguistic diversity. The conference aimed to identify gaps in current support, both technical and socio-political, contrasting the situation in South Asia with that in Europe. This conference highlighted, what advice and actions might be appropriate to help advance technical support for languages across South Asia, to help cross the digital divide. In this process it is also expected to identify advice and actions which in turn should be taken back to help the European Union in its own efforts to remove its own internal digital divide.

The conference was organised into a number of sessions as set out below. Each session began with a challenge of key questions, with many presentations and much discussion.

1. Inauguration

The conference was opened by welcoming addresses from

The Honourable Mr Kamal Thapa, Minister of Information and Communication, Nepal

Mr Sharad Chandra Shah, Vice Chairman of the High Level Commission for Information Technology, Nepal

Mr Lochan Lal Amatya, Computer Association of Nepal

Mr Rudiger Wank, the representative of the European Union in Nepal

All the dignitaries expressed their high level of commitment to information technology, in the development of Nepal and their appreciation at this conference being held in Kathmandu.

The keynote address for the conference was given by Mr. Kanak Mani Dixit, the member secretary of the Madan Puraskar Pustakalaya, the library of record of the Nepali speaking world. Mr. Dixit spoke of the need for full technological support for languages like Nepali, to enable citizens at large to benefit from the use of Information Technology. He illustrated his argument using Nepal, with its 23 million people most of whom are fluent in Nepali but have little English, who are barred access to computers not by cost since computers have become relatively cheap but by language. But he was optimistic, looking to the example set by the growing use of Radio, which shows such significant evidence of the creative application of technology within the Kingdom.

2. Language, culture, politics, and the digital divide.

The conference began by reviewing the state of languages in South Asia. Globally there are 6000 to 7000 languages, though the exact number cannot be determined due to the ambiguity concerning the difference between a language and a dialect. If two communities are unable to communicate with each other, is it due to a difference of dialect and accent, or a difference of language? This difficulty in identifying and cataloguing languages is very evident in South Asia, where the number of languages in the region has been estimated between the extremes of 500 and 2000 – the 'true' figure is probably around 1000.

The paper by Dr. B. Mallikarjun well illustrated this difficulty in counting languages. The 1991 census identified 1576 'mother tongues' which were subsequently rationalized into 216 and then 114 languages from five language families. In this rationalisation, linguistic communities of less than 10,000 people were dropped, and thus the number of languages is very much higher than that final 114. What is important to note is that there is a great variety of languages, written in a variety of writing systems. This variety and richness is repeated in other countries, and Dr. Tariq Rahman spoke about the richness of languages in Pakistan with around 70 distinct languages, and from Dr. Mark Turin about the 100 or so languages of Nepal. Only in Sri Lanka, presented by Mr. Vincent Halahokonege, is the variety limited, with just 2 languages. In Bangladesh there is the single dominant language, Bangla, spoken by maybe 96% of the population, though it was contested that Sylheti seen by some as a dialect of Bangla is in fact a distinct language – but there are also a number of small 'tribal' languages.

However not all languages are equal. While in India there are no prohibitions about the use of languages in the media, 22 languages are given the special status as official languages named in the constitution, favoured with financial support for technology investment. In Nepal all languages have been viewed equally since 1991, but the previous two centuries of favouring Ghorkali/Nepali has left a legacy of Nepali dominance, with other languages still lagging far behind in the resources to support them. In Pakistan language policy has been to favour Urdu over all other languages even though Urdu is a minority language. In Bangladesh it is Bangla that is favoured over all other languages.

It is not simply a matter of treatment under the law that determines a language's fate, there are significant social processes that are critical. The languages chosen by the powerful elite can bias people towards preferring those languages, and while this has usually meant favouring the official and national languages, is has often also meant a preference for English. One particularly surprising statistic from Pakistan is that Punjabi is by far the most used language, and yet it has low status and little support. While the people are drawn towards these languages of power, the elite themselves might use political and financial means to suppress other languages. This includes the level of

technology provision to support these languages - the powerful languages get the support, the others may not.

This situation was compared with that in Europe, where there are also many minority languages, both indigenous and migrant, that are also inadequately supported with technology. Describing this situation Prof. Jens Allwood spoke of English as a 'killer language' (quoting Skutnabb-Kangas) and the fear that Swedish may itself be endangered as people move to English. Allwood spoke of the thousands of years of conceptual development stored in a language that would be lost if the language was lost, outlining many arguments in favour of preserving languages, distinguishing between museum preservation and sustaining a language in use, and proposing 'language survival kits'.

The conference then turned to consider literacy and spoken versus written languages. Is literacy important, should all languages be given a written form? Spoken language is very different from written language in the way it works. Spoken language is interactive and multimodal with feedback and turn taking, the meaning being highly context dependent. It is very different from the written form of the language, and may even have a different syntax.

Contrasting the linguistic frameworks of South Asia and Euopre, it is noted that while there are not so many languages in Europe, both regions have a rich variety of languages with a relatively small number of dominant and more privileged languages whose dominant position is maintained a political, social, and economic forces.

3. Linguistics and Language Technology

The conference then turned to how much is known about these languages and what technology is available to support them.

Prof. B.N. Patnaik suggested that development of language technology to support tribal languages in order to overcome the 'English divide', the global dominance of English. Turning to models of language, he advocated the approaches of Panini and Tolkapiyam that had arisen in India, enriching them with western insights about language, but not to adopt western views of language. This moves away from universal views of language, a theme picked by Prof. Harold Somers in the context of Machine Translation which necessarily must use some common view of the two languages being translated - the conclusion from MT is that a universal interlingua is not possible, not even for closely related languages. As conjectured by Sapir and Whorf, languages reflect different conceptualisations of the world. However at the structural level there is much in common between languages, with the same grammatical phenomena occurring in many languages.

Prof. Rajeev Sangal spoke about the development of language resources, corpora and lexicons, for South Asian languages and in particular for Indian languages. Over the past few years there have been significant advances here. With the development of dictionaries and the collection of many speech and text corpora which have been tagged for parts of speech. Tags derived from established western systems work adequately. This has led to a number of machine translation systems, but in the long run to cover the range of languages needed, machine learning will be needed. Prof. Yogendra Yadava explained the situation in Nepal, with around 100 languages but where even the national language, Nepali, has very limited corpora and a dictionary that is in urgent need of updating and correcting and extending. Prof. Pushpak Battacharya then described a machine translation system based on the interlingual system UNL (Universal Network Language). A source sentence in one language is converted into UNL undergoing various normalisations in the process, and then the target sentence is generated by deconversion of the UNL. Studies are being made into English, Hindi, Marathi, and Bengali for application of this UNL-based approach to machine translation.

In discussions there was much concern that western and universal models of language might be inapplicable to South Asian languages, for both translation and tagging of corpora. BB Chaudhuri and BN Patnaik focused on examples of echo (eg guch-phuch) and reduplication (eg. the big big elephant) - in European languages reduplication denotes emphasis or de-emphasis, while in South Asian language the reduplication can take on new meanings.

The conference proceedings next moved on to statistical approaches to language engineering. Dr. Ruvan Weerasinghe reported an investigation into the use of statistical machine translation SMT) applied to the languages of Sri Lanka – English, Sinhala, and Tamil. SMT for European languages, such as English to French, works as well as the rule based systems like Systran. However when Weerasinghe applied SMT to English-Sinhala it did not work well, though it did do reasonably well for Sinhala-Tamil. Exactly why is not clear, perhaps because of the larger difference between Sinhala and English than between Sinhala and Tamil: Tamil is a Dravidian language while Sinhala is an Indo-Aryan language that has adopted many features from Dravidian languages. Dr. Peter Juel-Henrichsen then described his application of statistical methods to the description of unwritten unfamiliar languages: two different clustering methods were presented, which Battacharya conjectured resembled Wordnet similarity measures.

Dr. Roger Tucker and Dr. Ksenia Shalanova presented their work using text-to-speech (TTS) software to give speech access to information sources, motivated by the need to overcome literacy barriers. They provide a toolset based on diphones and a speech database, give advice on the use of this toolset, and support knowledge sharing among a network of groups using these tools. Collaborators include Rajeev Sangal and Dipti Sharma for Hindi and Ramakrishan for Tamil, with other sites in South Africa and in Nigeria, though they hope in the long run for global outreach. Dr. Hema Murthy described work at IIT Madras on multi-modal interfaces, and in particular speech recognition for Indian languages. Speech recognition has started relatively recently with speech corpora being built up covering a range of speakers, native and non-native, dialects, speaking environments. There is still much to be discovered about the spoken forms of Indian languages before quality recognition systems can be created. Early work on speech had also been done on ISI Kolkata for Bangla, and Dr. B.B. Chaudhuri reported on the promising results that they had with syllables rather than diphones. He reported that the best speech recognition system he had seen was a statistically based Hindi system at IBM Delhi. Discussion then moved on to agreement that the best speech generation would mix data driven approaches with rule based approaches to cover unusual words.

This then led onto an extended discussion about literacy: its many degrees and types of literacy, the way information is structured can itself provide a lot of guidance to understanding (eg railway timetables), that technology can impose particular language practices (eg spelling and grammar checkers), that it could be possible to support spoken language directly without any written form, that communication is multimodal, that standardisation is about power, that video conferencing is the most popular software in village kiosks, that a pure speech word processor Castor had been produced by Dr. Raman at Stanford.

The development of language technologies in South Asia is progressing well, but is dominated by western approaches to language engineering and the models of language that underlie these. There does seem to scope for building technology upon traditional South Asian views of language.

4. Software localisation and local software

Underlying all software that works in local languages are the computer encodings of the writing systems of those languages. In order to clarify the issues here, a short introduction into the history of the encoding of languages and writing system in the computer was given by Prof Pat Hall. There are three independent issues to be addressed in making computing technology work for a new writing system: the input methods including keyboard layout or other method; the internal code for the writing system, for example Unicode; and the output method, including fonts and rendering systems. It is the internal coding that is critical and this should capture the essence of the language regardless of the particularities of the writing systems. The ISCII encodings for the languages of India embody all the correct principles here. In discussion it was agreed that the situation in South Asia has been chaotic, with no acknowledgment of the need for standards, but that this is changing. Concern was also expressed about the lack of deep knowledge of Indian languages in Unicode, and thus the inclusion of inappropriate characters within Unicode, for example “leekar”.

The writing of Urdu using the Nastaleeq font was described by Dr. Sarmad Hussain. Urdu writing is cursive and multi-directional, with the 36 alphabetical characters changing into 15 to 20 different shapes as a function of context. Diacritical marks need to be moved around so as not to clash with each or the actual characters. So many open type font rules were required, well over a thousand, that the font designers ran out of space. While open type rendering works well for South Asian writing systems, technological limitations were encountered and need to be removed. Localised software needs to be used, and that requires both the marketing of the software and the training of people in its use.

The work on rendering Urdu in Pakistan is very recent, only possible now due to the development of Unicode and Open Type font rendering. Rendering for Indian writing systems came earlier, starting in the 1970s and 1980s, as described by Dr. S.P. Mudur. The need to separate input from output was known about then, but the limitations of computers at the time meant that Indic writing systems could not be rendered until the special GIST hardware had been produced. This GIST hardware and later software renderers were only possible because Indic writing follows very regular rules – however cursor positioning in composite characters (conjuncts) causes some problems. The renderers developed for Microsoft are also available for Linux.

In discussion two other OTF renderers – ICE and Pango – were described, together with issues around their use and the impact that these had on applications software. In India OTF fonts for Indian writings systems are being made freely available using public funds.

Keyboards were also raised in general discussion, and picked up by Prof B.B. Chaudhuri in the context of input devices in general. Keyboards should be reconfigurable so that a range of layouts can be used. OCR has been developed in Calcutta and is now being developed into a product at CDAC. Handwriting recognition and speech recognition are not well developed for Indian languages.

Dr. Shailey Minocha talked about human computer interaction and the way assumptions about users become embedded in the software and its interface. She illustrated this with banking web-sites in India and Taiwan, and explained the difficulty of getting evaluations of a site's usability in non-western cultures, and the need for usability laboratories. Evaluating computer systems then becomes like evaluating any product in a new market. This led on to discussion of the ethical basis for localisation of software – whether it was driven by commercial interests in finding global markets, whether interface design was more like advertising.

The social reasons for localisation were then picked up by Mr. Venky Hariharan, who pointed out that only some 10% of India's billion people were competent in English – which means that some 900 million people have no access to computers. A number of initiatives have distributed free fonts, and work is being done on localising Linux to 80 languages across Asia, which should be contrasted with MS Windows XP which has only been localised to 33 languages. These localised Linuxes add value to a billion dollar underlying open-source system, which are then distributed free. Harmonising terminologies is an important prerequisite, and the voluntary and distributed nature of open-source work makes this particularly challenging.

The contrasting situation in Nepal was described by Mr. Amar Gurung and Dr. Rhoddy Chalmers, with around 150,000 PCs mostly assembled locally and mostly without licensed software. Computer readable data has been converted to Unicode, keyboard drivers produced, a basic spell-checker and thesaurus produced, and Linux localised. The Linux localisation took around 18 person months of effort. All this had been achieved with much voluntary help and very little funding

In discussion Dr. Om Vikas emphasized the importance of standardisation, and the difficulties of multiple Linuxes – from IBM, Redhat, Mandrake and so on.

Localisation in Europe has become an industrialised process, with Ireland being the centre of this. Prof. Reinhard Schaeler explained that 95% of Irelands output was localised US software and digital content, with exports of software from Ireland exceeding the exports of software from the US. The localisation is done in Ireland by outsourcing, with nearly 7000

companies ('vendors') dominated by 3 very large localisation companies Bowne Global Solutions, Lionbridge, and SDL International. These are supported by the industry association GALA, and standards making associations Unicode, W3C, LISA, and the new OASIS who developed XLIFF and the Translation Web Services Standard. There are also a number of publications supporting localisation, and several conferences. The industry is driven by short term financial considerations leading to bland interfaces. Recently the ELECTOnline website http://www.electonline.org/index.php has been produced to share knowledge within the industry. In discussion the potential of India and China as centres of localisation was indicated, particularly based on open source software. There are many difficulties such as the smallness of the local market, the smallness of the localisation groups, and the lack of standardisation of terminologies.

Computers had been developed in the West to handle languages written using the Roman writing system, and developments to handle the richer writing system of Asia only appeared very recently based on Unicode and Open Type rendering. This has enabled software in South Asia to be localised for local languages, based on open source systems like Linux using voluntary labour. This contrasts sharply with the commercial drive for localisation in the West where localisation only happens where the market is large enough to promise a return on investment.

5. Linguistics and language technology futures

What is currently happening that could point the way forward in language engineering and linguistics for South Asia and South Asian languages? South Asian languages are important in Europe, and particularly in the United Kingdom, and two projects that point a way forward are the EMILLE and PLONES projects.

Language corpora are important for current linguistics, and Prof. Tony Mc Enery has considerable experience in this, from gathering material in English in the UK for the British National Corpus and on the EMILLE project gathering material in South Asian languages in both the UK and in South Asia. Practices for gathering spoken corpora which were acceptable for English were just not acceptable to the UK South Asian communities. Some locations for data gathering were not acceptable, formal copyright assignment was an alien practice and less formal measures had to be used. Mr. Vincent Halakonege recounted similar experience in Sri Lanka where all contact had to be personal and obtaining the support of the top person in an organisation was critical. While the volume of corpus data for South Asian languages is very significantly less that available for other world languages, the gathering of data has begun.

Prof. Harold Somers described his project to use language technology help health care patients with limited or no English (PLONEs), where all too often untrained people act as interpreters during medical consultations. The current language of study is Somali, with plans to move to Punjabi or Bangla, using speech and icons to interact with the patient. Often the language does not have the vocabulary to handle the health situations, and often people speak a dialect different from that described by linguists. They may expect a formal interaction with their doctor, expected in their birth culture but no longer practiced in the UK. Literacy and familiarity with computers may be an issue, and the project has looked for help from research work on support for disabled people. The technology must be shaped to meet the people's needs. Discussion led to consideration of language technology having a pedagogical role, to teach language and even the doctor-patient relationships of the UK. It was suggested that appropriate rather than high technology should be used as South Asia is multi-lingual rather than monolingual as in the UK or France, and that technical intervention should be accompanied by social and policy intervention.

In looking for ways forward, the PLONES project had looked outside linguistics to the assistive technologies developed for disabled people. Similarly translation theory might broaden our view of localisation and what language technologies might be appropriate. Prof. Anthony Pym gave us this translator's perspective. Localisation is a major employer of translators, but with the tools available to take care of equivalence, they become more concerned with controlling the source, with internationalisation. Human translators can do more

than maintain equivalence, they can change the discourse using their knowledge of the cultural values of the target readers. However it must be recognised that readers may wish to refer back to the original, particularly when that original is in an international language like English. There is a need to consider translation versus language learning, and if the project is long enough it may be better to teach people the target language rather than translate material. Discussion confirmed this complex multilingual situation in localisation and internationalisation.

It is observed that the basic to all language engineering must be acquiring more knowledge about the languages to be enabled, and that this needs the gathering of raw data in the form of language corpora. Methods for gathering this data need to be different for South Asian languages and in South Asian contexts. While localisation can be viewed as translation, there is a need to look more widely to contextualise what localisation is aiming to achieve, from assistive technologies to translation theory.

6. Localisation and local software futures

While linguistics has been developing over millenia, and language technology has been with us since the advent of computers fifty years ago, software localisation is relatively new, arising only some 25 years ago with limited aims. Localisation as currently practiced is ad hoc, using techniques and facilities developed for other purposes. How should software localisation be taken forward?

The Language Observatory project of Prof. Yoshiki Mikami was inspired by his travels around Asia and the many forms of typewriter he discovered. While it took 3 to 4 hundred years for printing to become localised from West to East, it took only 50 years for typewriters and 15 years for computers. This has been accompanied by a simplification of the writing of the languages, but also by a reduction in the range of languages covered so that today Windows XP only covers 50 languages of which only 2 are Asian. Discussion covered the systematic layout of keyboards and where the non-spacing vowels were placed relative to the consonants – but even where such designs take place and are demonstrated as giving benefits, people are very resistant to change.

Dr. S.P. Mudur described the work that NCST did for Microsoft in enabling the input and output of Indic Languages at the code level, and the sad experience that this the only worked in Notepad because other Office packages did their own input and output. There is a need to recognise different levels of localisation as a function of return on investment. He identified six levels – 1 no translation, 2 just user instructions, 3 enabled for local language input/output, 4 translated menus and dialogues, 5 translated online help and tutorials and redeems., 6 convert all cultural elements. During discussion this was related to levels in international marketing, and to the ignorance of software producers about actual language diversity.

Prof. Pat Hall then described the Glossasoft project, in which software had a high-level knowledge model about the software, and used language engineering methods to generate error and help messages, with an API that enabled language specific components to be replaced as required. Discussion focused on more general issues of what the basic languages components are in order to enable language survival, with suggestions ranging from basic speech corpora and dictionaries, to translation aids, to enabling the computer platform for the writing system, distinguishing language documentation from language survival from localisation needs.

The various localisation issues were then brought together by Dr. M. Sasikumar, distinguishing between localisation which enabled content creation by encoding the script and providing fonts, and changing the interface and look and feel. He argued that the first was the critical facility to enable information sharing, with users learning to use foreign keywords. The key interface localisation should aim at enhancing usability locally: for example NCST is enabling the searching of web content in Hindi, where the Hindi search terms are translated, the web is searched, and the results gist translated back into Hindi. He highlighted a number of issues: different Linux localisation may be incompatible with each other, computer terminologies used in translation need to be intuitive and standardised, usage of localised software by the intended users needs to be studied, and the potential of multi-modal interfaces. The key

lesson for localisation is that more collaboration and sharing will help progress.

7. Policy implications.

The final session was concerned with actions and policies to address the issues previously identified. The discussions began by addressing the criteria that should be used in selecting languages to work on and support. Mark Turin pointing out the topicality of this in Nepal which had moved to an educational policy which is inclusive and supports education in all of the countries 100 or more languages. Others highlighted the importance of the languages used in broadcast television and radio, the role of national rivalries wanting their country to match the resourcing of national and local languages of other countries, that supporting a minority language could further marginalise its speakers, that unwritten languages were more endangered but among those that with the largest number of speakers should be supported, that some languages have precious written records even if no longer spoken, that China requires all software sold in China to have been localised resulting in far higher PC usage in China relative to India, that China's Mandarin-only policy has disadvantaged all other languages in China such as Tibetan and Cantonese.

Dr. Om Vikas has been leading TDIL in India, coordinating technology and language engineering developments. He noted the increasing dominance of English as well as Spanish, French and German, measured by a range of criteria such as the proportion of Internet content or the volume of translation into those languages. He contrasted developed and developing countries in terms of: replacement of technology vs upgrades, interest in markets vs economic growth, IPR vs open source, affordable computers at $500 vs $40. India through TDIL has focused on adapting technology to meet core priorities, then developing its own technologies in collaborative programmes, and now focuses on creative technology within new sustainable structures. Areas of focus have been machine translation, OCR, text to speech generation and speech recognition, and many application areas like e-government and education. International collaborations are important. TDIL run an information and resource distribution service with a website and newsletter.

Dr. Khalid Choukri is the founding director of the European Language Resources Association (ELRA), which started in 1995. Europe is a multi-lingual multi-cultural region and faces many of the same problems as faced in South Asia. ELRA manages, evaluates, and distributes the full range of language resources of importance within the European Union. Before ELRA was founded in 1995 many language engineering projects had been funded, but their products had been lost. The languages covered include those of the enlarged European Union, plus many migrant languages now spoken within the union, as well as the languages of other regions like South America. The EU is now funding ambitious projects on speech-to-speech translation and multi-modal interfaces that will produce resources that will in due course be deposited with ELRA. There are still many gaps where no resources of a particular type are available for a language. Like TDIL, ELRA runs an information service with website and newsletter.

In discussion the idea of reusable generic application-neutral language resources was questioned, since theory-neutral language resources are not possible, questioning the whole idea of resource repositories like that managed for Europe by ELRA and being developed for India by TDIL, and suggested for South Asia. ELRA's resources are pre-competitive and not ready for commercial exploitation even though developed in part by industrial concerns using European Union funding. The irony is that the development of the 'knowledge society' has led to the loss of knowledge as languages are differentially advantaged and face extinction.

The final presentation to round off the conference was by Prof Ken Keniston, for has observed localisation developments in India for more than a decade. He had just spent a fruitful week at the WSIS meeting in Geneva. The tone of that meeting had been optimistic for the development benefits of ICTs with a focus upon local developments, but the absence of technology to support the local languages had not been recognised let alone seen as problematic. Investment in local languages is relatively small, even in India when the 5 million dollars invested so far is set beside the 7 billion dollars of revenue in 2003 from software services. Even though there have been

significant developments in South Asia of software that handles South Asian languages, there is still little software that is standards compliant or robust enough for sustained use, and projects to remedy this make promises but have yet to deliver. There seem to be cultural pressures against working on or in Indian vernacular languages, with English viewed as more prestigious. And for those who do work on local language software there is always the danger that efforts fragment as developers take pride in their own solutions and reject the solutions of others. Where commercial collaboration is needed, IPR interests can cause difficulties.

The solution throughout is standards and collaborations, with resources devoted to support these.

8. Position Papers

Available on the project website, http://www.elda.fr/proj/scalla.html

Allwood, Jens; *Linguistic Diversity and the Digital Divide*

Bhattacharyya, Pushpak; *Predicate Preserving Parsing* (with Jignashu Parikh, Yogadish Khot, and Shachi Dave)

Chalmers, Rhoderick and Amar Gurung; *Localising software: some experiences from Nepal*

Chaudhuri, B. B. *Input-Output Issues in Software Localisation of South Asian Language*. (Abstract)

Choukri, Khalid *European Language Resources Association, History and Recent developments*

Dash, Niladri Sekhar: *Language Corpora: Present Indian Need*. (not able to attend)

Hall, Patrick A.V.; *Linguistic Components to save Endangered Languages*

Halahakonege, Don Joseph Vincent; Sinhala, *Language Engineering and an Attempt to take Internet to the Masses*

Hariharan, Venkatesh (Venky); *Linux Localization in India* (abstract)

Juel-Henrichsen, Peter; *Working in Darkness, corpus based tools for low-budget field linguistics*.

Hussain, Sarmad; *Complexity of Asian Writing Systems: A Case Study of Nafees Nasta'leeq for Urdu*

Keniston, Kenneth; *The Politics of the Digital Divide* (extended abstract)

Mallikarjun, B; Indian Multilingualism, Language Policy and the Digital Divide

McEnery, Tony; *Constructing Corpora of South Asian Languages* (with Paul Baker, Andrew Hardie, and Sri B.D. Jayaram)

Mikami, Yoshiki; *The Language Observatory Project and its Experiment: Cyber Census Survey* (with Izumi Suzuki)

Minocha, Shailey; *Localisation Challenges in Usability and Customer Relationship Management of E-Commerce Environments. (with* Pat Hall and Lisa Dawson)

Mudur, S. P.; *A Reference Model for Software Localisation* (with Rekha Sharma)

Murthy, Hema; *A Multimodal Indian Language Interface to the Computer*. (with C.S.Ramalingam, Srinivas Chakravarthy and C Chandra Sekhar).

Patnaik; B. N. *On the Choice of an Appropriate Framework for Computational Linguistic Research on Indian Languages.*

Pym, Anthony; *Localization from the Perspective of Translation Studies: Overlaps in the Digital Divide?*

Rahman, Tariq; *Language Policy and Localization in Pakistan: Proposal for a Paradigmatic Shift* .

Sangal, Rajeev and Dipti Sharma; *Creating Language Resources for NLP in Indian Languages* (Dipti Sharma unable to attend)

Sasikumar, M. and Jayprasad J Hegde; *Software Localisation: Some Issues and Challenges*

Schäler, Reinhard ; *A Framework for Localisation*

Singh, Udaya Narayana: *Language Technology: a Road Map for South Asian Languages*. (unable to attend)

Somers, Harold; *Language Technology and Patients with Limited English*

Tucker, Roger and Ksenia Shalonova; *The Local Language Speech*

Technology Initiative – localisation of TTS for voice access to information

Turin, Mark; *Minority Language Politics in Nepal and the Himalayas*

Vikas, Om; *New Order of Knowledge-based Society: Rise, Raise & Race*

Weerasinghe, Ruvan; *A Statistical Machine Translation Approach to Sinhala-Tamil Language Translation*

Yogendra Yadava; Use of Software Support in Nepal's Languages.

Courtesy:
TDIL Programme,
Department of IT,
Ministry of Communication and IT,
New Delhi-110003
E-mail: tdilinfo@mit.gov.in

3.3 First Language Observatory Workshop: A report

What is "Language Observatory"?

- The Language Observatory surveys language activities in the virtual universe over the Internet.
- The Language Observatory tries to catch subtle messages of less spoken languages, as far as they appear in the virtual universe, and provides us such information like;
  - How many languages are found in the virtual universe?
  - Which language is missing in the virtual universe?
  - How many web pages are written by any given languages, say Tamil?
  - How many web pages are written by using Amharic script?
  - What kind of character encoding schemes (CESs) are employed to encode a given languages, say Khmer?
  - Is there any de facto CES standard for any given language?
  - How quickly ISO/IEC10646, or Unicode is spreading and is replacing conventional single byte encoding schemes on the net?
- In contrast to the existing multilingual search engines like Google, which can handle only limited number of languages, the Language Observatory is designed to handle every language. But please note that the Language Observatory is not a search engine, but an observation instrument.

FLOWS 2004- Objectives:

The First Language Observatory Workshop - Scanning Cyberspace for Languages (FLOWS 2004) was organised at Multimedia System Center, Nagaoka University of Technology, Nagaoka, JAPAN during February 20 - 21, 2004 with the following objectives:

- In recognizing the importance of the promotion and use of multilingualism and universal access to cyberspace;
- In cerebrating the International Mother Language Day;
- To initiate the Language Observatory Project

Mr. Ken-ichi Iwasaki, Director, Office of Research System Promotion, Research Institute of Science and Technology for Society (RISTEX), JST, Mr. Hiroshi Nagano, Secretary-General, Japanese National Commission for UNESCO, Director-General for International Affairs, Ministry of Education, Culture, Sports, Science and Technology (MEXT), Dr. Kozo Ishizaki, Vice President, Nagaoka University of Technology and Dr. Hajime Ohiwa, Professor, Keio University inaugurated the workshop.

In the first session Dr. Yoshiki Mikami, Leader, Language Observatory Project, JST, JAPAN Professor, Nagaoka University of Technology highlighted the programme background. He also presented a talk on "Chaos in character encoding schemes". Ms. Nagano Robin Lee, from Miskolc University, HUNGARY delivered he r presentation on "English as the de facto cyber mother tongue".

The theme of the second session was "Language Digital Divide & National/Regional/Global Initiatives". Dr. Om Vikas, Senior Director, Department of Information Technology, INDIA highlighted the various initiatives taken by Government of India to overcome the language barrier through his presentation "TDIL Mission: India's Initiative towards Bridging Digital Divide, Overcoming Language Barriers".

Dr. Ahmed Zaki Abu Bakar, Professor and Dean, Faculty of Computer Science and Information System, Universiti Teknologi Malaya, MALAYSIA presented "A Knowledge Management System for Arabic Character-Based Languages". Mr. Paul G. C. Hector, Assistant Programme Specialist, Information Society Division, UNESCO highlighted the UNESCO's Initiative in this direction.

The theme of the Session 3 was "How to Construct Language Observatory" Mr. Yoshihide Chubachi, Graduate School of Media and Governance Keio University, JAPAN talked about the Cyber Census 2002. Dr. Yoshiki Mikami, Project Leader, JST, JAPAN presented his talk on "Language Observatory - What it surveys, and how it works?". Mr. Izumi Suzuki, Assistant Project Leader, JST, JAPAN presented the paper "N-gram Statistics Approach to Automatic Language Identification". Mr. Ikechuku Anyanwu, Managing Director, Placade Japan Co., JAPAN demonstrated the "Test Crawler".

In the concluding session of the workshop Dr. Virach Sornlertlamvanich, Thai Computational Linguistics Laboratory, Communications Research Laboratory, THAILAND presented the "TCL's Computational Lexicon". Mr. Venkataraman Narayanan, Excel Solutions, SINGAPORE talked about the "Open Systems Approach to the Language Observatory".

Courtesy:
TDIL Programme,
Department of IT,
Ministry of Communication and IT,
New Delhi-110003

Email: tdilinfo@mit.gov.in

3.4 Speech Corpora Workshop – A report

A workshop on "Development of Annotated Speech corpora for Indian languages" on 17th – 18th July, 2003 was organized by CDAC-Noida.

The *annotated* speech corpora have been a critical component of research in the speech sciences. Various speech corpora by different groups have been made, but all of these are directly created & managed by the core group. The major issues involved are the propagation of repairs, consistency of references, and the ability to integrate annotations having different formats and levels of detail. This workshop forms framework, where these issues can be addressed successfully.

The workshop is focused on the following issues:

- To provide the awareness of recent developments in the representation and management of annotated speech corpora.
- To brainstorm the common structure of annotations and of annotated corpora for Indian languages so that a wide range of disciplines concerned with speech recognition & speech synthesis can be covered.
- To suggest the various techniques/ methodology for creating the database.

In the workshop, various experts from all over the country delivered lectures in their field of expertise.

After the inaugural speech by Sh. V.N. Shukla , Director (Sp.App.), CDAC, Noida , Dr.S.S. Aggarwal, Emeritus Scientist, CSIR & C-DAC (N) and Dr. Samudra Vijay, TIFR-Mumbai delivered the keynote addresses.

Sh. V N Shukla in his inaugural speech emphasized upon the need, importance and applications of speech databases.

Dr. SS Aggarwal in his Keynote address spoke about the Development of Speech Corpora for Spoken Language Processing Systems. He focused on Speech/ Language Processing Technology (Key Component in Next Century Communication), areas of speech processing e.g. Speech coding, Speech Synthesis, Speech Recognition, Speaker Recognition, Spoken Language Translation, Spoken Language Identification, knowledge sources (database) required for speech to text system, different parameters in speech processing, types of databases and important databases of the world.

Dr. Samudravijaya spoke about the spoken language technologies and need for annotated speech corpora. He said that different issues like type of data, methodology of data collection, standardization of annotation procedure, tools for annotation, training, validation, Organization and distribution should be addressed while creating the speech databases.

The first technical session was chaired by Dr. Samudravijaya. In the session, other main speakers included Dr. P.K. Chaturvedi, Dr. A.B. Saha and Prof. Ashok Dutta etc.

Dr. Chaturvedi spoke about the TDIL program activities on the Technology Development for Indian Languages.

Dr. A B Saha spoke about the annotated speech corpora development activities being carried out in Bengali and Assamese Language at Kolkata centre of CDAC Kolkata, the outcome of which were demonstrated as applications prototype by Sh. S Mandal.

Prof. Ashok Dutta raised following two questions:

1. Can we use expertise developed in connection with text corpora? And
2. What should be the content of speech corpora?

He said that Speech Corpora can be divided into 3 parts according to the need of Speech Research and Speech Technology Development.

1. Speech Research
2. Automatic Speech Recognition
3. Speaker Recognition

Then he presented a methodology of content selection for speech corpora development.

The second session was chaired by Dr. Om Vikas, Sr. director TDIL, DIT, MC&IT. In this session, the main speakers of the workshop included Ms. Sunita Arora,

Prof Pabitra Sarkar, Dr. Nagamma Reddy, Prof K K Goswami etc.

Ms Sunita Arora, CDAC-Noida described techniques and strategies about the Selection of Corpus & structure of speech corpus. She said that speech corpora should cover:

(a) Phonetically rich / phonemically balanced set of words / sentences

(b) Set of sentences covering variety of intonation patterns.

She proposed a directory structure for the storage. She also emphasized the need of training for the developers focusing on Language, Annotation Tools to be used and Corpus management & documentation.

Prof Pabitra Sarkar, Kolkata in his presentation expressed that in order to understand the basic things about speech recognition, speech analysis and other computer directed operations with natural language, the following relationships have to be explored

1. Verbal speech
2. writing (traditional)
3. Phonetic and other such representations
4. Acoustic representations by a Sona-graph or other such devices.

He also said that synthesized speech should sound as close as natural speech.

1. The computer should be enabled to 'represent' as well as 'identify' individual segments by using the representations it has created for individual contextual sounds, both their core and their margins.
2. It should be enabled to 'represent' and 'identify' words in isolation.
3. It should be enabled to 'represent' and 'identify' words in concatenation.
4. To do the same about phrases, clauses, sentences, and hopefully, discourses beyond sentences.
5. It should also be able to 'represent' and 'identify' stresses, pitch levels, and intonation patterns of sections bigger than words.

Dr. Nagamma Reddy, spoke about the Acoustic Features distinguishing voiced & voiceless plosives contrasts.

Prof K K Goswami presented some views on Linguistic aspects of speech corpora. He said that Speech corpora projects are largely concerned with language application. Linguistic part of the project can not be ignored hence computer scientist and linguist should join together to have best results.

The Other invitees of the workshop were Sh. Rajesh Verma, Sh. H.K. Sardana, Sh. Akhilesh Srivastava, Prof Kapil Kapoor, Prof Suraj Bhan Singh, Prof. Vaishna Narang, Prof. V.R.Jagannathan, Dr. R.C.Garg, Dr. Ashwani Kumar, Prof. Thakur Das, Prof. Ravindra Gargesh, Dr S K Chauhan, Sh. Arun Kumar, Sh. Deepak Kumar, Dr A M Ansari also presented their views during meeting.

Dr. Om Vikas in his concluding remarks summarized the need for development of speech database and efforts being made across the country. He emphasized upon following points for further development related activities

- Selection of sentences and other speech samples for creation the database
- Selection of appropriate tool for analysis and annotations
- Procedure for recording two channel database
- Storage of annotated database in the prescribed format.

Courtesy:
TDIL Programme,
Department of IT,
Ministry of Communication and IT,
New Delhi-110003
E-mail: tdilinfo@mit.gov.in

3.5 Universal Digital Library Programme

- A report of the Steering Committee Meeting at CMU, USA

The Steering Committee Meeting of the Universal Digital Library Programme was held during May 28-31, 2004 at Carnegie Mellon University, USA. Participants from China, India and USA were present during the three day steering committee meeting to discuss the progress of the UDL programme at the participating countries and curve out the Roadmap and strategy for the future.

Address by His excellency President of India:

His excellency President of India Dr. A.P.J. Abdul Kalam addressed the steering committee meting via video-conferencing link. In his address, he presented the idea of establishing "Village Knowledge Centre" throughout the country for widespread proliferation of the benefits of Information Technology to the masses. Some excerpts from speech of His Excellency President of India:

"… India has approximately 2.3 lakhs Village Panchayats. I visualize establishment of village knowledge centers in these Panchayats to empower the villagers with the knowledge and to act as a nodal center for knowledge connectivity for the villagers. Initially 100 remote Panchayats may be chosen for the Model Pilot project. Based on this experience, it will be possible to replicate Village Panchayat Knowledge Centers in other Panchayats.

This Centre will be equipped with 1000 books of different subjects. In addition it will have a computer terminal, Telephone connections, Modem, Printer, Photocopier, Camera, Scanner, Internet connection and other support facilities for functioning as a digital library. The knowledge center will provide hard copies of relevant pages to the members of the village, if required. Also the center can have facility for reading the text with an audio output so that people who are unable to read, can benefit from the library. The knowledge center can also be used for collection, digital storage and dissemination of village specific information pertaining to agriculture, craftsmanship, arts, artisan techniques, informal judicial system practiced in village based on values, local remedies for simple ailments, village stories with moral values, village history, village folk songs, village cultural traditions, traditional medicinal practices followed in villages and village marketing information and methods. Such information presently is being transmitted through word of mouth and with changing generation, it is being lost.

It is essential to preserve and carry forward this knowledge base for the benefit of future generations. This village panchayat knowledge centers will become a helpdesk and confidence builder for the villagers. In course of time, there could be inter panchayat information transfer leading to assured information flow to the needy. This Village Panchayat knowledge center will also act as an e-governance nodal point for the villagers. The cost of establishing and running the hundred remote panchayat knowledge centers for three years is estimated to be Rs.11 crores (approx.). Let us work together to establish this Village Panchayat knowledge centers in 100-village panchayats in India."…

Presentation by Prof. Raj Reddy:

Prof. Raj Reddy, Prof. Herbert A. Simon, CMU, USA & Coordinator of the Universal Digital Library (UDL) Programme presented a brief outline of the progress and international status of the UDL programme. He mentioned that China has 30,000 books, which are out of copyright. These books would be scanned and sent back to CMU , USA and after that these books would be available on-line . Prof. Raj Reddy mentioned about the enormous amount of task involved in the digitization process. He mentioned that Harvard University has a repository of 12 Million books; Library of Congress has 30 Million books & OCLC has a catalog of 42 Million books. In addition, there is a huge number of multilingual books. The universal digital library programme aims at digitizing 1.0 million books, which are only 1% of all books available in the world. Prof. Raj Reddy mentioned that presently 14 scanning centers are in operation in China and 1 Centre in Egypt. India has 21 scanning centers in operation and so far 78,000 books have been digitized.

Prof. Raj Reddy presented various issues and research challenges involved in the Digital Library Programme.

He mentioned that there is good scope of research in the domains of Universal Access, Design of distributed cached servers, Multilingual Information Retrieval, Machine Translation & Summarization technologies. He also stressed the need of renewed effort towards improvement of OCR technology. Prof Reddy touched upon various issues involved in the digital library programme such as copyright, metadata and scalability of architecture. Regarding the copyright issue he mentioned that 5 % of the books are out of copyright; 92 % of the books are out of print but they are in copyright period & 3 % of the books are in print and copyrighted. Prof Reddy suggested the idea of tax credit & usage based government funded compensation to resolve the issue of copyright. He mentioned and stressed the need of digital submission of all books "born-digital" to the national achieves.

Presentation by Prof. N. Balakrshnan

- Prof. N. Balakrishnan has presented the progress of scanning activity of digital library scanning centers in India. The statistics presented by Prof. Balakrishnan are as follows:

- Number of Scanners in operation:	100
- Total Number of books scanned:	70,294
- Total Number of pages scanned:	2,50,03,877
- Number of books available at web:	32,738
- Number of books in English	45,000 (10,000 books are from CMU)
- Number of books in Indian Languages:	25,000
- Abby Fine reader 6.0 is available to all centers	
- SCANFIX software is available to all centres.	

Prof Balakrishnan also presented the development of various software & tools at IISc, Bangalore, such as Universal Digital Dictionary (in association with CMU) , Machine Translation Engine (accuracy about 70 %) associated with the development of Digital Library. He also mentioned about the training programme pertaining to Digital Library Initiative in India.

Progress of UDL programme at various countries:

China:

Prof. Zhao Jihai , Professor, Zhejiang University Library presented the progress of Million Book Universal Digital Library project in China. The Project started in Dec 2000. The first funding at RMB 40 million yuan(US$4.84 million) from the Ministry of Education , China were allocated to Zhejiang University in Sept. 2003, for support of the staffing and operation of the 14 scanning centers and 2 technical centers. 14 Chinese universities have joined the project.

The first 6 Chinese partners are as follows:

- Chinese Academy of Sciences,
- Fudan University,
- Nanjing University,
- Peking University,
- Tsinghua University, and
- Zhejiang University

The second 8 Chinese partners are as follows:

- Beijing Normal University,
- Huazhong University of Science & Technology,
- Jilin University,
- Shanghai Jiaotong University
- Sichuan University,
- Wuhan University,
- Xi'anJiaotong University, and
- Zhongshan University (Sun Yat-SenUniversity).

Each center has been equipped with two AVA3+ scanners recently, while 9 Minolta scanners were provided in the first 6 scanning centers in 2002. Two technical support centers have been established in Zhejiang Univ. (China-US Million Book Digital Library Project.htm) and in CAS (The Universal Library -China Site.htm), to provide technical support for the scanning centers in southern and northern areas respectively.

Scanned centers	Scanned volumes	Scanned centers	Scanned volumes
CAS	791	Huazhong Univ of ST	1278
Fudan Univ	4218	Shanghai Jiaotong Univ	50
Nanjing Univ	8170	Sichuan Univ	600
Peking Univ	1673	Wuhuan Univ	800
Tsinghua Univ	5378	XI, an Jiaotong Univ	703
Zhejiang Univ	20941		
Beijing Normal Univ	98	Total	44700

Chinese Digital Library Website:

Future Plan:

- Chinese book selection and digitization.
- Over 40 thousand old Chinese books to be digitized in the coming months in the 14 scanning centers, and more non-copyrighted books and materials will be selected for scanning.
- English book selection in US and shipment to China for digitization.
- Establishment of Shenzhen scanning center for digitizing the English books.
- Improvement in the Chinese book digitization system.
- Increase in scanning and data process efficiency.

Progress of Digitization activity at Tsinghau University, Beijing, China:

Dr. Jiang Airong Tsinghau University Library, Beijing, China presented the digitization activity at this Digitization Centre.

Four Repositories (SiKu QuanShu)	222 578 pages	(276 volumes)
The ancient books	273 669 pages	(2937 volumes)
The Republic of China(Min Guo) books	436 499 pages	(2157 volumes)
Total is	932 746 pages	(5370 volumes)

Dr. Jiang mentioned about the future plan of this center:

Future Plan:

30 000 volumes of ancient books and the Republic of China(Min Guo) books would be digitized by the end of 2005.

India:

Arulmigu Kalasalingam College Engineering (AKCE):

K Sridharan, Secretary Arulmigu Kalasalingam College Engineering, presented the progress of Scanning activity of the Digital Library Centre. The Centre was launched May 25, 2001 and presently equipped with 2 Minolta PS 7000 heavy duty scanners, 2TB storage servers, 4 Workstations along with Abby FineReader6.0 & Scanfix Softwares. So far 5562 books (1.1 million pages) have been scanned and digitized.

The Goal of the center is to establish a universal access center for accessing books scanned under the UDL programme.

Tirumala Tirupati Devasthanam: (TTDS)

The progress of Shri Venkateswara Digital Library of Tirumala Tirupati Devasthanam has been presented. The center started operation in May 2002 with 4 scanners. Futher, it was annexed with Poorna Pragyna Vidya Peeta , Bangalore. It has been reported that , so far 10,216 books were scanned and digitized in various Indian Languages. The Statistics of the digitization activity is given below:

Language	No. of Books Scanned	No. of Pages Scanned
SANSKRIT	1,918	7,06,139
TELUGU	1,588	4,14,939
HINDI	336	24,320
TAMIL	586	1,02,196
KANNADA	1,242	3,68,335
URDU	3	826
MULTI	919	3,55,101
ENGLISH	3,515	19,59,474
TOTAL	10,216	39,31,330

Sanmugha Arts Science, Technology & Research Academy (SASTRA):

Dr. S. Vidyasubramanium , Dean , SASTRA presented the status of the Digital Library Scanning Centre at SASTRA. Presently, the scanning center has two scanners. It has been reported that so far , 2150 books in various languages such as Tamil, Sanskrit, Marathi, Hindi English & French have been scanned and cropped , out of which 2030 books have been OCRed and 1600 books have been hosted to the Web.

IIIT-Allahabad:

Dr. M.D. Tiwari presented an overview of the progress of scanning activity at IIIT Allahabad. He mentioned presently, 6000 books (~ 7 Million pages) have been scanned and digitized. Dr. Tiwari mentioned that IIIT Allahabad is presently collaborating with G.N. Jha Sanskrit Vidyapeeth which is having a rich collection of 80,000 books & rare manuscripts and Allahabad Pubic Library , which is having collection of another 10,000 rare books. He told that efforts are made towards digitizing these books and making them available on web. Dr. Tiwari stressed the need of augmentation of scanning capability & bandwidth at IIIT-Allahabad. He mentioned that IIIT-Allahabad requires 2 No of 1 Tb/s server, 10 AVA3 flat scanners & 8Mbps BW for enhancement of the connectivity.

Dr. Tiwari also mentioned that under the aegis of the mega-centre project IIIT-Allhabad would develop Cross-lingual Information retrieval, Speech to Text & Text to Speech in Hindi & Sanskrit & Multimedia authoring tools. He is hopeful that a version of these softwares & tools would be ready by December 2004

IISc. Bangalore:

Dr. Venkadesan , IISc. Bangalore presented the draft Meta-data standard for Digital Library Programme in India . A snapshot of the meta-data entry form is given below:

GUI of Meta-data

Rashtrapathi Bhawan, New Delhi:

Dr. V. Ponraj Director (TI) , President's Secretariat, Rashtrapati Bhavan, New Delhi, presented the progress of the project .

Presently, 4277 books (1700693 pages) were scanned out of which 1692 books OCRed , cropped and digitized. The target is 2.0 Million pages up-to March 31, 2005 out of the available 20.0 million pages.

Dr. Poonraj also mentioned the need of high-speed connectivity in order to have better access of the e-content available through the Digital Library of India.

USA

The Universal Digital Library Programme is coordinated by Carnegie Mellon University. USA.

Development of Hardware , Software & Network architecture:

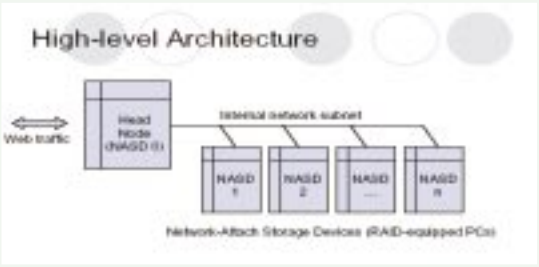
Dr. Eric Burns & Ed Walter presented the Design Challenges and possible solution about the Middleware platform for Universal Digital Library Programme. He mentioned that the main Design challenges are :

- Ease of installation/maintenance.
- Low cost / commodity hardware.
- Storage redundancy..
- Allowing customization but enforcing standards
- Simple software upgrade procedure.
- Simple storage addition procedure

Dr. Burns presented snapshot of the Universal Library Software and hardware platforms along with the configuration of the network architecture.

GUI of ULIB Software developed at CMU:

Schematic of the Network Architecture:



Dr. Eric Burns also proposed hardware & Software requirements for Digital Library :

Hardware platform:

- Low-end commodity PC
- Four removable 250GB IDE drives in software RAID-5 (data redundancy)
- Current usable space per unit: 750 GB
- Boots from CD-ROM or network

Software platform:

- Bare-bones store/search of books in MBP format (TIFF/HTML/TXT/[MARC|DC])
- Simple reading interface w/ TIFF => web-viewable GIF conversion
- Search features:
  - One query searches all bibliographic records and document full text
  - Full Google-style phrase and Boolean matching
  - Book reading interface supports searching and match highlighting within boo
  - ¸ ¸Management software automatically checks
  - ¸ ¸adding all books uploaded to serve "staging area
  - ¸ ¸Site-specific customization: all HTML ca be created by sit

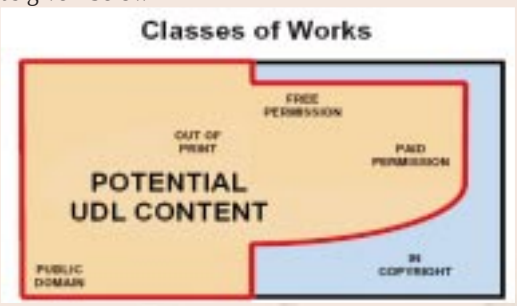
1 ¸Upcoming features

1 ¸Search by specific fields in bibliographi records

1 *Copyright Issues*

1 Dr. Micheal Shamos, Director, Universal Librar Programme presented copyright aspects involved i the process of accessing contents of Digital Library

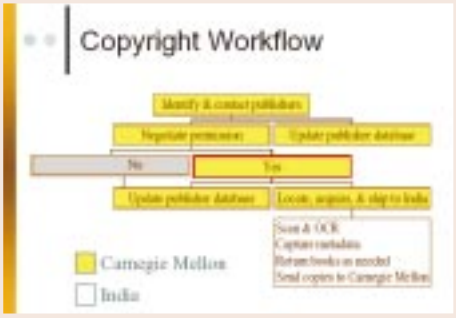
1 The potential UDL can be classified in to categorie as given below



According to Berne Convention, (1886) which is subscribed by 96 countries including India and China, every treaty country must regard non-nationals the same way it regards its own nationals

Dr. Shamos mentioned briefly about various acts (Public Domain Enhancement Act, Digital Era Copyright enhancement act)and their implications in the perspective of Digital Library Programme. Dr Shamos stressed the importance of public lending systems and its usability in the Digital Library System. The mode of access to content is also necessary for sustainability of DL systems. Out of various models such as Government-paid (e.g. public lending right), Subscription model (e.g. HBO) , Metered use (e.g. page views), Free-to-read, pay-to-print , Compulsory licensing etc an optimized model needs to be evolved for wider penetration of the UDL programme.

Dr. Gloriana St. Clairs , CMU presented a model copyright workflow to be followed by the participating countries for the Universal Digital Library Programme and is given below:



Courtesy : Mr. Somnath Chandra,
TDIL Programme, Department of IT,
Ministry of Communication and IT,
New Delhi-110003
Email: tdilinfo@mit.gov.in