3.2 Sanskrit Informatics Workshop

Venue: Indira Gandhi National Centre for Arts(IGNCA), 3 Rajendra Prasad Road New Delhi -110001

A Report on the Proceedings

Day 1 : August 7, 2003 (Thursday)

Session 1 : Inagural Session

The following participants addressed the morning session

> Dr. G.C.Tripathi, IGNCA
>
> Dr Unruh Gerd
>
> Prof. R.K.Saxena, Rector, JNU
>
> Dr. Om Vikas, Sr. Director, DIT, MC&IT

Prof. G.V.Singh (gvs10@hotmail.com)

The session started with a welcome address by Prof. G.V.Singh. Prof. Singh stressed the fact that Sanskrit is not just one of the languages but *the language* that linguists and we computer scientists are talking about. It is structure of this language that interests we all.

Prof. G.C. Tripathi

In his keynote address, Prof. Tripathi said that India is a treasure of time proven scientific knowledge. All this knowledge is available in the form of Sanskrit literature which is not accessible to the scholars due to various reasons. So we need to morph this knowledge so that it is available to users in localized forms. We are moving towards a knowledge society and need to develop tools and technologies that have knowledge processing capabilities.

Prof. Gerd Unruh

Prof. Gerd said that in order to develop high-end technologies we have to combine and speed up effort to justify expenditure. We need to market these knowledge systems and tools to countries in Europe and America. We need to integrate the traditional and holistic knowledge of the east with the modern and fragmented knowledge of the west.

Prof. R.K. Saxena

Prof. R.K.Saxena, Rector JNU stressed the need for such initiatives and also lauded the efforts of Prof G.V. Singh and his team in this direction. He also emphasized on 'marrying the nascent Sanskrit department in JNU to the School of Computer and System sciences.

Dr. Om Vikas (omvikas@mit.gov.in)

Dr. Om Vikas said that it heartening to know that Sanskrit Informatics is no more a national cause but it has become an international cause. More and more of international scholars from France, Germany, USA, are showing interest in our heritage knowledge. He stressed that our heritage knowledge has a holistic and integrated approach. Sanskrit language has a mathematical framework. Therefore, abstraction of this knowledge is possible. He also talked about the new initiative taken by DIT, Govt. of India, i.e., the KUNDALINI project.

Session 2 : Development of Basic Information Processing Kit (BIPK)

Chair Dr. V.V.S. Sarma (vvs@csa.iisc.ernet.in)

Speakers Mr. M.D. Kulkarni, Mr. P.Ramanujam, and Prof. G.V.Singh

Mr. M.D. Kulkarni from CDAC Pune introduced various linguistic and computional resources that are being developed at C-DAC Pune. He talked about the limitations that are coming in the way of tools and technologies being developed. He talked about the requirement of a **Basic Information Processing Kit for Sanskrit (Vedic)**. After giving the background information of the project, he listed the requirements of a **Vedic Sanskrit** BPIK.

Mr Ramanujam of C-DAC Banglore, started with a need for '**Heritage Computing**'. He showed presentation of **Desika** system. He underlined the several problems that are being faced in converting *Desika* from a DOS based system to a Windows based system. He also presented his '**Rg Veda Ratnakar**' and '**Gita Reader**' projects

Prof G.V. Singh of JNU, presented his project "**Language Learning System for Sanskrit**". He also informed about the various Language Tools and Language Resources that are being developed at RCILTS, JNU. His group was able to demonstrate the work that they have done on Sanskrit lexicon (specially the Sanskrit-English dictionary) and exercise modules with static and dynamic exercises. They have also done significant work on Sanskrit grammar viz., *Dhaturoop, Sabdroop,* and *Sandhi*. The team also presented portions of *Siddhant Kaumudi* that have been brought on the web. The lexicon developed at RCILTS, JNU has provisions to make it multilingual lexicon.

Prof. V.V.S.Sarma concluded the session with the remark that we have seen the ages of Data Processing and Information Processing in that order. Now we are in the age of Knowledge Processing. So we need to do Information Mining, Information Extraction, Information Summarization, etc. For all these we need to have language based tools and resources.

Session 3 : Optical Character Recognition & Text to Speech Systems

Chair Prof. V.N.Jha (vnjha@vsnl.net)

Speakers Mr. V.N.Shukla, Mr. M.D.Kulkarni, and Mr. Vinamra Agarwal

Mr V.N. Shukla of C-DAC NOIDA, talked about creating a parallel language corpora for 12 Indian languages. He emphasized the need to productize the language tools and technologies developed at various research centers in the country. In his presentation of OCR technology developed at C-DAC NOIDA, he stressed the need for inclusion of the training module in the various OCR Systems. With this, the accuracy of present OCR systems with domain specific training modules will improve. He talked about a new problem that was regarding skew in absence of *Siro Rekha*.

Mr. Kulkarni's presentation on OCR stressed the need for addressing new requirements of the present OCR systems. The new requirements that needs to be incorporated or improved further are

Removal of noise

Multilingual Support

Font Diversity

Siro Rekha problem

Feature Extraction

Post Processing

Mr Vinamra Aggarwal—from Prologix Software, Lucknow, presented a Text-to-speech (TTS) system for Indian languages called *Vaachak*. He talked about various design approaches for building a text-to-speech engine. He briefly explained the classical phoneme based approach. He then explained the unit selection and concatenation approach. In case of Indian languages, he asserted that this new approach gives better quality.

Session 4 : Machine Aided Translation Systems & Language Resources(1430-1545 hrs)

Chair : Dr. Mukul K. Sinha)expert@vsnl.com)

Speakers : Prof. R.M.K.Sinha, Prof. Sanghmitra Mohanty

The session opened with a detailed presentation of a Machine Aided Translation (MAT) System, **Angla-Bharati** developed at RCILTS, IIT, Kanpur by Prof. R.M.K. Sinha. He explained the various approaches of MAT. He explained at large the following approaches of MAT:

Rule Based Machine Translation (RBMT)

Knowledge Based Machine Translation (KBMT)

Example Based Machine Translation (EBMT)

Hybrid Approach of RBMT & EBMT

Statistical Approach to Machine Translation

He talked about the pseudo inter-lingua that is being developed at IITK which can become the base for machine translation from one Indian language to another.

Prof. Sanghamitra Mohanty of Utkal University, presented 'Sanskrit Wordnet: A Pivot in Indian Language Machine Translation (ILMT)'. The presentation talked about the nature of Sanskrit language, online lexical resources and the Sanskrit Wordnet. Her discussion on the wordnet included the following:

*nymy (category, synonyms, antonyms....)

IWL, CNL

demo of Sanskrit Wordnet

She also talked of true inter-lingua of Sanskrit that shall be the base of machine translation from one Indian language to another.

She also had a brief presentation on Oriya OCR.

Session 5 : Standardization Issues (1545-1630 hrs)

Chair Dr. Om Vikas (omvikas@mit.gov.in)

Speakers Dr. R.K.Joshi, Mr. M.D.Kulkarni

Dr. R.K.Joshi enumerated the basic linguistic needs – input, internal coding, output and exchange of data. He demonstrated the many *styles* and *rich expressions* found in the fonts available in age old printed documents. He noted that there are almost 12,000 Glyphs possible for conjuncts. He stressed for the urgent need of first level Vedic Unicode standardization.

Mr. M.D.Kulkarni stressed the need for immediate standardization as the absence of standard severely restricts the proliferation of IT to the masses. He also talked about the problems of sorting that needs to be resolved. He suggested an approach of – *evolve, implement* and *don't tamper* the standards.

Session 6 : Sanskrit on Linux & Windows(1645-1800 hrs)

Chair Dr. R.K.Joshi

Speaker Prof. Jitendra N. Shah and Dr. Alka Irani

Prof. Jitendra Shah of VJTI, Mumbai, started with the agenda that if you want to solve peoples problem then you need to talk to them. The 98% of the population don't talk English. Commercial efforts have not been able to bridge the digital divide. He stressed the need (for children) to learn with the new technologies rather becoming Macaulay's clerks. He demonstrated the GNUBharati initiative that he has been pursuing for quite some time. He illustrated some of the applications that have been localized in Hindi, Marathi, and Gujarati. He proposed to take forward the GNUBharati initiative further with availability of Indian language Operating system with system level support.

Dr. Alka Irani also shared her experiences of the past and suggested how standardization issues should be handled, so that they become a facilitator rather than impedance. She suggested that as technology developers we should keep on adopting and embracing new technologies rather than get stuck with something that we had acquired in the past. She talked about the Roopantar transliteration scheme.

Day 2 August 8, 2003 (Friday)

Session 7 : Intelligent Cognitive System-KUNDALINI (Knowledge UNDerstanding, Acquisition of Language, INferencing and Interpretation)

Chair Dr. M.A. Lakshmithathachar

Speakers Dr. V.V.S Sarma and Dr. V.N.Jha

Prof. V.V.S.Sarma described in detail -Indian Knowledge Resources, Indian Cognitive Systems and-IT tools for handling and manipulating them. He emphasized the importance of these systems and tools in the following areas:

administration

business

education

language technology

KR cognitive systems

contextualization

curriculum and training

To achieve the desired results he defined the broad areas for manageable development as- LT(Language Technology), further subdividing into- LT1 LT2, TLT3; KT(Knowledge Technology) further subdividing into KT1 , KT2, KT3; CT(Cognitive Technology), further subdividing into CT1,CT2,CT3, HT(Heritage Resource Technology)), further subdividing intoHT1HT2,HT3 and HR(Human Resource Centres)), further subdividing into HR1,HR3,HR2

Prof. V.N.Jha's presentation talked about the following aspects of KUNDALINI project -

knowledge created thorough ages

recognize this knowledge and context and apply

Relevance of knowledge

understanding it and convince others through experiments

Need language for coding the knowledge

He stressed that the Navya Nyaya identifies the elements of interpretation (relate language to reality). He described the process:

Knowledge -> KR -> Knowledge Application. The development path according to him is:

Content creation activity from primary groups

Intermediate language (data + operation, data structure + algorithm)

Retrieval technology for easy access

Dissemination for common man

Mr Ramanujam's presentation was centered on setting up a Digital Library of Indian heritage for Vedas. He said that this is partly developed – annotation, tags, hyperlinks, indices etc are built. Additional grammar rules are required, and speech synthesis is to be done. He also highlighted the need for people to work together

Session 8 : Panel Discussion: Technology Roadmap For Sanskrit Informatics (1430-1700 hrs)

(The next session on "Joint Indo-European cooperation in the field of Sanskrit Informatics" was merged with this session)

Chair: Prof. Gerd Unruh

Panelists: Prof. V.V.S Sarma, Prof. V.N.Jha, Prof. R.M.K. Sinha,

Mr. M.D. Kulkarni, Dr. Mukul Sinha, Prof. N.J. Rao, Dr. R.K. Joshi, Prof. Jitendra Shah, Dr. P. Ramanujam

Prof. Gred was of the view that very little progress has been made to develop sustainable and robust software products. Prof R.M.K.Sinha agreed with objectives proposed by V.V.S. Sarma (in his KUNDALINI presentation). Prof. Sinha opinied that for creating interdisciplinary expertise two types of course contents have to be developed by experts

1. Contents for B.Tech students

2. Contents for Sanskrit students.

Prof. Sinha also emphasized the need for an Indian Networking Language – a concept based language.

Prof. Joshi was of the opinion that the technology is for human and must be based on the following principles of Indian Philosophy:

sah karma yoga

vividh gyan shakha

He stressed need for Panchasutri Sanskrit Technology development with

1. Bilingual Information

2. Indian Phonetic Code

3. Vedic Code Chart

4. Phonetic translation of Sanskrit documents

5. Phonetic equivalence to IPA Phono fonts, considering calligraphic aspects

Dr. G.C. Tripathi stressed the need for inter code

conversion of information to facilitate the reusability and interoperability of technology and tools developed by various agencies.

Mr. M.D.Kulkarni said that the standards are available, such as, ISCII but ISO standards are to be decided for establishing the compatibility.

Dr.Mukul Sinha said that what is to be done has already been spelled out by other panelists. But how it is be done is the question. There are difficulties which people face. He proposed that all government purchasing must make **inscript** keyboards mandatory (English and Hindi, etc.). Browsers have default encoding ASCII & UTF is optional whereas it should be other way round. He suggested that all RCs should put their work on the net using UNICODE. All the tools that are being developed by C-DAC and other research centers should have an option to save the content in UNICODE. He also informed that since October 2002, the Google search is possible now Hindi. He also informed the house about the computer Ganak Bharati proposed by him.

Prof. N.J. Rao put language technology related scholars in to following five categories:

1. traditional

2. university scholars

3. linguistics, AI, ??

4. working outside India

5. others

He emphasized that the scholars of all categories be involved in the development work. But initially we can start with a core professional group.

Prof. V.V.S.Sarma was more particular about the component re-usability in the processes of development.

Dr. Om Vikas saw development of language technologies in general and Sanskrit Informatics in particular falling in to three phases.

Phase I

Phase II

Phase III

His opinion is that Sanskrit is loved by everybody and the effort should be made to preserve the wealth of knowledge available in Sanskrit manuscripts. He informed that the manuscript project of IGNCA is worth Rs. 35 crores.

He talked about the various initiatives of DIT for language technology development, namely, –

1. Language/knowledge, lexicons, wordnets ontology

2. Corpora - 3 million

3. Knowledge tools (re-usable tools)

4. Morph analyzers, text editors (basic information tools)

5. Text summarizing and inference rules

6. Fonts (open type fonts) IPA

7. INL (CML)

8. Knowledge Representation

9. OCR (Human Machine Interface)

10. TTS – Speech corpora proposal

The immediate attention needs to be paid to issues related to:

Machine Translation

Standardization

HRD issues mismatch – Coputational Linguistics, engineering courses for IT and Sanskrit people

He also spoke on government policy on

Digital library

International cooperation

Europe-India joint research projects

GNU-Bbharti OS

And emphasized on early initiatives to achieve closer cooperation between groups and countries.

He informed that at the moment the government has no plans for major international cooperation on the subject.

Prof. Gerd emphasized that specific responsibilities be assigned to individuals or groups for the following

Re-usability responsibility to be given to Prof N.J. Rao

Lexicon everyone needs

Check the lexicon against the corpora anyone needs

Fonts anyone needs

According to Prof. N.J. Rao, European participation should be

Providing external reference

Machine oriented

Provide Multilinguability

Use Internet media

Prof. Gerd stressed the need for

Establishing a center

Establishing a site

All supporting Lexicons and fonts

According to Prof. R.M.K.Sinha the European standards should support ISCII.

Dr. Om Vikas stressed on the need for the following:

Prof. Gerd to identify controls

A feedback workshop on

Lexicon

Corpora

Language processing tools

Visit of scientists, research scholars to other RCs and group centers

Prof. V.N. Jha demanded that the information of German Sanskrit activities be made available on site.

Session Recommendations:

It was unanimously felt that Intelligent Cognitive System (KUNDALINI) project proposal will result into new kind Knowledge Technologies and hence should be supported by the Government.

The panel discussion ended with the thanks to the chair.