

- endangered languages, the politics of language
- > and minorities; and other basic technologies like writing and
- > printing; will also cover the critical technologies

LANGUAGE TECHNOLOGY: A ROAD-MAP FOR SOUTH ASIAN LANGUAGES

**UDAYA NARAYANA SINGH
CENTRAL INSTITUTE OF INDIAN LANGUAGES
MYSORE**

**THEME PAPER TO BE PRESENTED AT THE
SCALLA (Sharing Capability in Localization and Human Language Technologies)
Working Conferences 2004**

Jan 6-9, Kathmandu

(Funded in part by the European Union in the Asia IT&C programme)



0. The Scope

This paper will try to look into the contexts within which development of human language technologies and software localization takes place. It considers both space and time factors and tries to outline both big and small pictures to trace the legends associated with human language technology. First, I shall identify a few situations where language technology can have some decisive role as agents of change or development, and then will talk about its role in South Asia of our times in particular in exploratory terms.

1. The legend: Role of Language Technology

1.1. Linguistics as a discipline and Language Technology: What a linguist does when confronted with a language under investigation is that he feigns himself to be like any other new learner of this complex system. Like a newborn human being with his share of wonder and awe at being exposed to what we call 'language', the linguist now unpacks his tools and techniques to measure this exposure. As anybody familiar with field techniques would know, different areas of this system require different degrees of sophistication on the part of the linguist and this precision deployment of techniques comes from three elements – (i) the particular structure of a given language which may show greater complexities in sound systems rather than its word formation rules or in none of these, but in syntactic layering and order of words, phrases and clauses, (ii) wide experience of the field worker or researcher who is further helped by a plethora of background studies in both language universals and typologies and in linguistic model-building, and (iii) from the tools he employs to record, transcribe, segment, represent, and analyze data at hand. It is in the third area that current advances in language technology can be a great help.

1.2. Specialized Language Learning and Language Technology: Mastery of any language structure depends on our realization that language learning is a 'lifelong learning.' And yet, a linguist does not necessarily have the time like a native speaker to expend a lifetime on a single language. (Some linguists do, but most think it profitable to face newer and newer challenges.) In comparison, an interpreter, critic, translator, researcher (using this given language as a resource) or a prospective writer find it both necessary

and important that one makes a commitment to spend a major part of one's life in learning, chiseling, changing, and redirecting the language he is dealing with. Such specialized learners can take recourse to language technology to quicken and sharpen their linguistic instincts, or in finding the right expression or the texts or sequences he is looking for. Therefore, this is another area where language technologists need to concentrate on. There are, of course, questions that often puzzle those who are newly exposed to text-books in linguistics or grammar-books of a particular language (because they find it so very different from what they were exposed to in schools): Are the methods of linguists and those employed by lifelong learners as outlined above going to remain the same? And, what about their motivations at different crossroads? If the answer is in the negative (or even if it is positive), will language technology respond to these two needs differently? There may arise many other subsidiary questions at this point, but the important point is that the instruments used for this kind of learner (and that includes text-books and courses) must be very different from a linguist's highly specialized grammatical descriptions.

1.3. Language Teaching and Language Technology: As we are aware, with introduction of each new theory, technique and technology (if we may align them together), there had been a great revolution that followed or accompanied it in the processes or methods of teaching. I am sure each one of us remembers with excitement the introduction of chalk and a blackboard, or chalk and slate in the initial days of our schooling. The instrument of writing or the abstract concept of a written symbol were both novelties that engage all who pass through the literacy lane – whether as a child, or as an adult in non-formal education. We could recall the 1922 prediction of Edison, namely, that motion picture would revolutionize education, and eventually supplant books, as narrated by Stoll (1995) in a book, titled '*Silicon snake oil: Second thoughts on the information highway*' (Anchor Books, NY). With each new technology breakthrough – television, computers & multimedia, similar prognostications were heard. However, in the context of a larger part of South Asia, when we look at our classrooms, nothing much seems to have changed. This is difficult to accept considering the burgeoning economy as well as tremendous strides made by this region in IT and AI-related work. In some ways, distance education has tried to bring in a change of some sort – by changing the learning environment – by taking education away from the confines of four walls of classrooms. But that, to my mind, is not enough. Greater changes can and need be introduced, and these can arrive from the lessons we have learnt in Language Technology.

2. Life in an Information Age

As we all know, a time came in the course of all-around development for countries when for good or bad, fertilizers of synthetic variety, blue jeans and alternative medicines spread all over the world. Technology has spread in the same manner, and has now come to stay. However, a great deal of value is still attached to retaining, reviving, and furthering our conventional knowledge bases – many of which are either oral, or confined to unread and difficult to decipher manuscripts. Many of us have rather become more conscious of these knowledge bases and in their documentation than we used to be. Documentation of minor and minority languages and cultures have thus assumed great significance. Teaching and learning (and that would include learning the mastery of languages), are no more a matter of elite formation or passports to “reach” somewhere – where one could not go otherwise. Learning of any kind is closely associated with a number of skills now. I would say that we now stand at a crossroad called the 'Knowledge Square' – with *Knowledge, Learning, Information, and Skilled Application of Intelligence* being the four pillars of all-round development and economic rejuvenation. But as one could easily see, technology drove us to each of these roads, impacting all of them in some ways. Structuring and exploring into knowledge would require technology sophistication that would take us beyond a mere database design, as they would demand critical technologies of various kinds such as training software and meta-crawlers which are already in place. Using various tools and programs for learning of any kind underscore the need of technology once again. Information today is stored, gathered and presented through a combination of different kinds of technologies, which are fast integrating. Finally, if we as a knowledge community want to achieve the competitive edge, we will have to plan and implement reforms in all spheres where with advancements in technology, numerous applications and skills would take the front-seat.

However, social engineering needs our planners to ensure that such reforms and changes are not limited to only those who can afford – because of affluence, age, gender, sheer number, or location in

space. This also implies that languages – whether they are spoken by lesser or greater number of speakers, and languages that are spread in far flung places – must also be covered under any planned activities that would affect Language Technology in South Asia. It is a different matter that one has to see whether one would like to create technologies for such language communities that are easy to use and reasonable to afford. Even if they are provided by the state, one must be able to use devices like ‘simputer’ (ref:) or ideas like ‘shellbooks’ (ref:) for this purpose. It goes without saying that publishing books and materials in such smaller languages will not be cost-effective, and creating virtual systems or Radio/TV lessons would be more beneficial in the ultimate analysis. However, everything needs money, and we must make biggest ever investment – both in terms of money as well as in manpower and material (or, teaching tools) development plans at all levels of education. This will turn out to be our best investment, too.

I would like to mention here that the advanced cultures have already been taking initiatives as the internet announcement of the US Government shows:

Advanced Education Technology Initiative

“The U.S. Department of Commerce and the U.S. Department of Education announced the launch of the Advanced Education Technology Initiative on October 23, 2003. As part of the initiative, an Interagency Working Group on Advanced Technologies for Education and Training will foster the development, application, and deployment of advanced technologies in education and training in the U.S. The working group is co-chaired by Under Secretary of Commerce for Technology Phillip J. Bond and the Department of Education's Director of Educational Technology John Bailey.”

This would give rise to the most obvious question in South Asia: Should the developing world lag behind in taking a major technology initiative in both school and higher education sectors, e-governance, e-commerce, e-publications and e-zines? The obvious answer is an overwhelming “No.” Of course, that means that a lot of software would need localization, and at times, one might need to write specialized programs for certain kind of learning situation. But never mind this extra bit of labor.

3. The Challenge of Technology Integration

Granting that relating our country's education system with technology is critical to our social and economic well being, how do we effectively integrate technology into our class-rooms? As technology will shape the way we will all do our work in future, we will need to overcome whatever may be hurdles in this integration. Today's children are more adaptable to new technology than many of us. We know that for a teacher, it will not be easy to master the tools, its grammar, learning to harness its enormous potential and yet be someone who would provide the students with a road-map to have technology access. Even in USA, until late '80s –a study "Teachers and Technology: Making the Connection" shows that relatively few of its 2.8 million teachers used technology in their classrooms regularly. Even there, mostly, it is a skill that one learns from one's peers – out of sheer interest that initiates new learners into such technologies.

One reason why this kind of integration is difficult to achieve is because of People's, or let's say, participants' negative attitude toward technology. Such attitudes can derive from the experiences they have had during encounters as active innovators, researchers, and facilitators or as consumers. Alternatively, negative views are often related to different kinds of failures. To recount a few - unrealistic expectations, immature products, lack of support, non-robust design, and a myriad of stresses that inhibit success. However, any one who wishes build a technology, and this applied to educational technology as well, must remember that failure is integral to technology planning. It is obvious that one must design for low failure rates. However, we could set aside certain types of problems arising out of situations such as personal failures, catastrophes, and others hindrances. These are typical irritants. But we must be wary of and arrest partial failures, or failures to implement decisions, and try and plug these loopholes.

The biggest challenge comes from the anti-technologists, and there are many. There are thinkers like John Zerzan who would argue that <http://www.spunk.org/texts/writers/zerzan/>, it is technology that causes us to fail as humans and as caretakers of the natural world. Then there are others like Jerry Mander, who would object because of the negative effects of certain mass-based technology such as television - and to a great extent, he is right. But here we are to work on the effects on a social plane. Some may argue that

technology at its best only provides second hand experiences. Clifford Stoll (1995: 116) laments, "All of us want children to experience warmth, human interaction, the thrill of discovery, and solid grounding in essentials: reading, getting along with others, training in civic values. Only a teacher, live in the classroom, can bring about this inspiration." Once again, the assumption being made is that technology will replace human beings, which is wholly incorrect, as it would take out the drudgery and mundane elements from teaching and training but would never ever try and substitute teachers. Many argue against it saying we are not ready yet – not civilized enough to implement great changes.

However, whenever such objections arise, one must ask oneself: In answer to the question as to '*Who survives*', Charles Darwin had stated – quite rightly that "It is not the strongest of the species that survive, nor the most intelligent, but the one most responsive to change." I would also add here the old Chinese proverb: "Is it not too late to begin digging a well until one feels thirsty?" For ages, a teacher's world was defined by the printed word. The printed word determined the manner in which information was organized. But times are changing now and we need to walk with the changing times.

4. Non-Linearity of Knowledge

We know that the modern times has seen knowledge being presented or represented mainly in the print mode, and it is this shift from oral preservation and transmission culture to the written mode that defines the change over from pre-modern to modern times. Those who produce knowledge (or wish to impart it – without saying new things themselves) now write books, articles, reports, and papers, the students come to class, read materials, engage themselves in discussions, write term papers and assignments, and take tests. All these activities were centered around print, and this is what faces a great challenge with the advent of Computers, AI, digital libraries and the World Wide web. We must understand that this is as important a shift in paradigm as change over from orality to literacy. New technologies are changing how information can be presented and ordered, or how learners must interact with technologies, and how they need to communicate among the peers or with the teacher, as texts, audio, video, and graphics are all rolled into one now. Learners have to be quick now, because each person has his or her own time limits and constraints, and consequently the learners' views about how they should learn are also changing. All concerned realize now that information is neither linear, nor is there any clear beginning and end. It is only a starting point from which to make additional connections to other materials. However, in essence, nothing substantial is changed in critical reading and referencing as used to happen in the class-rooms with great teachers who would often move back and forth from the text to its CON-text earlier, with profuse references and allusions. Now all this action is available in the hypertexts as we browse through a text because the net can bring about much more information in a cyclical and evolving manner rather than in a linear pattern. Readings are no more one after the other, as one could jump on to chapter 7 much before one has been through chapters 2 through 6, as cross references and hyper-linking would make it easy to know about what preceded easily. Skimming through a text has become so much easier.

5. Technology – from nominative to instrumental

Technology changes are still not appearing to be obvious to many. To such people, I would point out how technology has become- from a mere subject of learning in the '70s to an invaluable tool of learning in end-'90s. What was 'Computer Literacy' earlier in '80s for a linguist or for any social scientist was a working knowledge of the computer – how it works and how it is structured. One who mastered some bit of programming was considered a "giant among pygmies" – as it were. It was not surprising, therefore, that road-side teaching shops offering programming and other computational courses blossomed and schools also tried to mould themselves accordingly. Taylor's (1980) *Computer in the School: Tutor, Tool, and Tutee* gives a good account of this trend. But now, technology has moved from being in the nominative pedestal to act as an instrument of change, as it seems

- (i) to have already been streaming across the curriculum and syllabi,
- (ii) has become an integral part of a student learning any discipline, with
- (iii) a seamless integration

It goes without saying that technology is never neutral, howsoever liberating it may be. Let's admit that schools and colleges are themselves perceived as an important move in technology by many communities who think that these are liberating institutions. Any kind of technology (starting from EVMs employed in the recent elections in India even among illiterate voters to digital media for teaching how to write, or fill up a form) could be employed to train future citizens to be organized, and transmit information to one another, learn to accept an authority or inculcate certain values. Even if we agree that some of these functions are performed better by technology, we must not be under this false impression that technology is better because it is value-free. That I think is a great falsity as technology is never neutral. Its values and practices must always either support or subvert those managing the organizations where they are placed. Therefore, technology must be brought into action – whether in language analysis and teaching or elsewhere - with a clear picture of the social need and social action in mind, and not because it is a mere fashion. Let's face the reality head on.

6. CIIL and the Technology Angle

At CIIL, we noticed quite early that the institution must make an impact on what I would call three 'T's in language education.

- It must train and orient teachers.
- It must produce the state of the art text-books, and
- It must create appropriate technology for language teaching.

Accordingly, we have been doing several things to fulfill these needs, e.g. 10 months of intensive L2 teaching (through which 8000 school teachers have been trained so far – for whom we hold National Integration Camps, Orientation and Refresher courses for trained teachers every year), offering Distance Courses:Tamil/Telugu/Bengali/Urdu, Specialized Courses in Communication, Orientation Courses for Mother-tongue teachers, Refresher Courses (under the UGC Academic Staff College), creating and beaming Radio & cassette Courses, and publishing books and electronic materials. To achieve this end, we have set up a 150-node LAN at CIIL plus an Itanium Web server and database server at CIIL for hosting various sites, established a High speed V-SAT connection through STPI, set up both state of the art Analog and Digital audiotek language laboratories at SRLC, ERLC, and NRLC, and brought in 4200 Electronic Journals for browsing in the library.

However, we do realize that what we need to do more is stupendous because in India, we teach about 58 to 69 languages either as subject or as media of instruction, and any language technology plan must take them into its fold. Further, there are 3954 newspapers and periodicals in 35 languages with varying degrees of regularity and readership. If they are to be modernized in terms of production and sheer data generation as well as in terms of becoming archives and information bases, we need to account for them, too. This is not to forget that there are 14 major writing systems in use. In addition, in a recent survey conducted by Padmanabha, Mahapatra, Verma and McConnel (1989)), we are told that out of the 96 languages surveyed out of the 114 populous languages (with 10k plus speakers) listed in Census 1981, about 50 were found to have written modes of expression – even though they may be using only known scripts which may or may not be adequate compared to their phonological geniuses. The visual and audio media employ 146 speech varieties & 24 “major languages” of India as in our radio network. India awards highest literary prizes in 22 languages, which indicates the level of sophistication needed in respect of these languages. Surely, we need to create rich lexical resources for them – at least electronic dictionaries, thesauri, word-finders, collection of technical terms, and idioms and collocation glossaries. All these make our task all the more difficult. Let us not forget that Census 1991 had recorded 1,576 rationalized mother-tongues and 1,796 other mother-tongues.

What we have done or are doing at CIIL so far are as follows:

- Archiving of data of 118 languages which are now being digitized with the help of Computerized Speech Lab (Model 4100) with Main Programme Version 2.5.2 with Real-Time Spectrogram (Model 5129), Video Phonetics Program and Databases, Multi-Dimensional Voice Program

(including the Advanced Model), Real-Time Pitch, and Analysis Synthesis Laboratory plus Web-enabled IPA version 1.7.ink and Sensimetrics CD-ROM: Speech Production and Perception.

- Creating Spoken Language Corpora – the biggest of its kind – in collaboration with various agencies for specific applications in mind, samples of which are to be posted on the net. An Indo-Swedish collaborative project has been already begun (www.ciil-spokencorpus.net)
- Studied 80 Tribal/Border languages, grammars and lexicon of which are to be brought out.
- Made 35 grammars available on-line (in both .pdf & pmk at www.ciilgrammars.org)
- Cassette Courses in : Assamese, Urdu, Bengali & Marathi – being digitized now. Kashmiri course already on the net.
- Radio courses in Hindi through Kannada beamed through many stations
- National translation initiatives available on-line (www.anukriti.net)
- Cassette Courses in : Assamese, Urdu, Bengali & Marathi – being digitized now. Kashmiri course already on the net (www.koshur.net)
- VTLS software and Sun server for digital library project.
- 4200 on-line journals
- 44 million word Indian languages corpus (also available with Emille Corpora) and various tools as well as Parallel corpora being developed with Sahitya Akademi (National Academy of Letters) and National Book Trust.
- E-zines (Translation Today and IJLM) and E-books on the net (www.ciilebooks.net)
- Radio courses in Hindi through Kannada beamed through many stations
- Digital Documentation of minority Indian languages and cultures.

Finally, if we are to introduce technology angle to developing the minor languages here, we must list out what the smaller languages require. In all these cases, language technology can play a major role. These endangered and lesser-known languages need the following:

- Decision on Scripts
- Print standards & fonts
- Basic Grammatical Sketch
- Word-book with pronunciation keys
- Pictorial glossaries
- Primers
- Language Games
- Literacy materials
- Post-literacy Texts
- Rhymes & Riddles
- Recall vocabulary
- Cultural glossary
- Style Manuals

How do we do it?

- Through creation and tailor-making of E-Tools
- Literary Promotion and large-scale translation (for which, again, we need tools)
- Shell-book model (<http://www.fil.org/shellbooks/SBMaker.htm>) as proposed by the Foundation of Endangered Languages and as in use for numerous Papua New Guinea languages. These **Shellbooks** are a sequence of resources that provide an interactive *shell* (framework) of resource information on a particular topic, and are designed to enable indigenous communities to culturally adapt the shell content and publish it as a **Shellbook Version** in their own language. Shellbooks are formatted so that one can browse and interact with their content using *Shellbook* software.