Developing Information Technology Solutions in Indian Languages: Pros and Cons

Madhuresh Singhal
Aurigene Discovery Technologies Ltd.
Electronic city, Bangalore- 562158
madhuresh_s@aurigene.com

T S Prasanna
Intel Technologies India Ltd.
Bangalore
t.s.x.prasanna@intel.com

Sharad Kumar Sonker
Rajiv Gandhi University of Health Science
IV T Block, Jayanagar, Bangalore
sksonker@rguhs.ac.in

S Shashinath
National Centre for Science Information
IISc, Bangalore - 560012
shashi@ncsi.iisc.ernet.in

Abstract

A very large part of the content in our country is in various local languages. Language is a barrier to get the full advantage of this knowledge. In order to remove the language barrier, computer and IT solutions can play a major role. It is needed to create a system of multilingual content knowledge base so that it can serve all-regional community requirements. However this is not an easy task as there are various technological hurdles and lack of commitment. Commercial companies are not interested in developing such solutions, as there is no such big market in doing that. This paper gives an insight about developing the local language solutions and it's pros and cons. It also discusses about the initiatives taken by the government and its supported organisatons.

1.1 Introduction

The Indian Constitution mentions 18 languages as languages of India. Each language has its own literature, comprising great novels, drama and poetry. It is indeed these differences that make India an interesting country. India has over 1683 languages and dialects and an estimated 850 languages in daily use. All communities have their own culture, rooted from their language and all languages have their literature. Indian education system has been given preference to its regional languages from beginning itself. Ancient history of Indian scholarly achievement, like Takshshila, Nalanda, Ancient Gurukul Sytems etc can be taken as proof, which was based on our languages. India was a leader of knowledge, wisdom and cultural development in the past when we were using our languages as a way for research, learning and education. Saying this is not that we should not learn or should not use other language for expression or for learning, Should be, but if regional languages will get preferences in varsity than it will be useful for general people.

According to an UNESCO study involving world's 140 most published authors; 90 out of 140 were English writers in 1994 compared to 64 out of 140 in 1980. More than 80 percent of the information on the Internet is in English – even though only 8 percent of world populations speak English as first language. [UNESCO, 1998]

In a large multilingual society like ours where there are vast diversity of culture and languages, human communication is a major issue. As the trade and business are widening, the people had to migrate to expand their business activities. In such a scenario, every human being is forced to learn more than one language to be able to communicate with others.

It will be useful for knowledge development rather than language learning & development. A very large part of the databases and content in our country is in these languages. It is important to be able to communicate knowledge seamlessly without language being a barrier. A major effort would therefore have to be directed at ensuring that IT can deliver its potential in local languages. By providing linguistically cooperative environment, which facilitates smooth communication across different linguistic groups, Information Technology (IT) emerges as a catalytic agent in this process.

1.2 LOCALISATION: Indian Scenario

Impact of Information Technology was felt as early in 1970s. Solutions towards adaptation of rapidly growing Information Technology for Indian languages were developed. Input-output problems and coding schemes were analysed. In 1990-91, Government launched the program on TDIL (Technology Development of Indian Languages) under which projects were supported for development of corpora, OCR, Text-to-Speech, machine translation and generic software for Information processing. Standards for keyboard layout and internal Code for Information Interchange were also evolved. This resulted into confidence in having solutions for Information processing in Indian languages.

In order to remove the language barrier in the use of computers in various Indian Languages, it was felt desirable to have the complete environment of computer to be localized in Indian Languages. This would help to increase the penetration of computers in the Society in large proportion. This requires localization of Operating Systems, and also the Application Softwares in various Indian Languages. With this in view, programmes are now underway involving Indian and Multi national Organisations

To find various solutions to India's "Digital Divide" by making Information Technology, Information Communication Technology, and the Internet, work in Local Indian Languages in order to offer - to the people at the edge of the digital revolution - social awakening and uplift, economic and commercial viability as well as opportunity, and the possibility to be integrated, and be a part of the global village.

1.3 Why Information Technology hasn’t reached masses?

Linguistic Diversity in India

Bio-diversity is the characteristic of nature in balance. Similarly the linguistic diversity is the characteristic of the evolving mankind that is geographically dispersed. Linguistic-based division into states ensures use of the official language of that state in governance and education.

India's average literacy level is about 52 percent. Less than 5 percent of people can either read or write English. Over 95 percent population is normally deprived of the benefits of English-based Information Technology. Interestingly, all Indian languages owe their origin to Sanskrit; hence they have in common rich cultural heritage and treasure of knowledge. Following may be the reason for Information Technology hasn't reached masses -

- ✓ IT and the Internet belong mostly to English knowing and speaking people, and it has further diversified into those languages, the economics of which is better, such as European languages, Japanese and some other.
- ✓ The content that is available on the Internet for masses are mostly in the elite languages, the highest of which belongs to English.
- ✓ There is no tool/Interface that has been developed to allow the have-nots to publish and create content in their local language.
- ✓ There has been no standard software/application interface to allow the have-nots to use in their local language, such as, Operating Systems, and Office Suits are not available in local languages.
- ✓ Lack of industry involvement due to constrained demand; There has been no basic standard for developing or enabling the IT and the Internet for local languages.
- ✓ Most of the organisations that are working towards developing language solutions, or doing researches on the requirements of IT/ICT/Internet, do not even discuss with the representative of the people at the other side of the digital divide.
- ✓ No strategy for language technology marketing.
- ✓ Unable to check import of IT products and services, which don't support Indian language(s).
- ✓ No Consensus on standardization Standards in use; ISCII-88, ISCII-91, UNICODE, many propriety code; Content is largely glyph-coded, not (ISCII) character-coded.
- ✓ Slow pace of transfer of language technology from academia to industry.

1.4 Initiatives for Overcoming Language Barrier and Providing IT Solutions in Local Languages:

Overcoming Language Barrier: In a country like India, communication overcoming language barrier is crucial to the growth of society and in preventing the Digital Divide. The first step in this direction was the launch of TDIL (Technology Development for Indian Languages) Programme in 1991 by Ministry of Information Technology to develop information processing tools to facilitate human machine interaction in Indian Languages and to create and access multilingual knowledge resources and integrating them to develop innovative user products and services. The next milestone has been the setting up of Resource Centres for Indian Language Technology Solutions. These centres will develop technologies for providing solutions with citizen interface in Indian languages selectively and thus covering all Indian languages.

Standardisation: Standardization of 8 bit ISCII (Indian Script Standard Code for Information Interchange) was developed by erstwhile Department of Electronics, Government of India, in 1988 and later on the revised version was published by the Bureau of Indian Standards in 1991.

Knowledge Tools: Multi-lingual e-mail Client has also been developed at CMC. Font-based multilingual packages, multilingual word processor, transcription facility, Font based Indian script enabling DTP packages, Database packages, Indian script enabling packages, Data entry packages, e-mailing system, Machine Translation Systems, application software packages in Indian languages such as Address management system, Indian language learning system, Management Information Systems in Government, business management system, etc have been developed at various organizations like Modular, Sonata, Softek, Summit, NCST, CCE, ER&DC/N, NIC, TCS, IITK, IBM, Oracle, etc. Indian language support is also becoming available on operating systems, Windows 2000 and Linux.

Evaluation: Systematic and objective evaluation of Natural Language technologies and products, though not easy, is a necessary and effective mechanism to establish and extend the state of the art.

Translation Support Systems: Mantra is a Machine-aided Translation System (English to Hindi) for Government notifications at C-DAC. Angalabharati, at IIT Kanpur & ER&DCI/N, a Machine-aided Translation System (English to Hindi) for public health domain is being developed for the Anti-Malaria Campaign.

Human Machine Interface Systems: An alpha version of "Hindi Vani" software that is PC based Unlimited Vocabulary Text-to-Speech Conversion Software for Hindi for DOS platform has been developed which is being ported to Windows platform. Line printers were enabled for printing Devanagari [at Lipi Data Systems & Transmetic Systems].

Localisation: Localisation of the existing generic software was carried out by designing Indic script enabling interface software. Indian script support is now being provided at Operating System level also in DOS, Windows and Linux. Localisation of e-Content involves use of local language enriched with locale specific cultural values. Simple Inexpensive Multi-lingual ComPUTER (Simputer) has been designed that enables use of Smartcard, Text-to-Speech, and Information Markup Language for Internet applications.

1.5 Implementation Strategy

The implementation strategy can be given as *Consolidation, Integration, Embedding and Innovation*, which could be further elaborated as-

- Technology Integration and Localization of solutions through Resource Centres.
- Focus on user products, services and total solutions
- Public Domain/General Public License (GPL) approach for faster development.

- IT localization clinics for wider dissemination and internship training.
- Bilateral/International cooperation in Language Technology and Applications.

1.6 Pros of having IT Solutions in Local Languages

Because the 'e' in e-Era does not stand for 'English'. So in order to felicitate people of all regions in India, irrespective of their languages and distance barriers, a system is very much required to facilitate the students and researchers for enhancing their skills, and acquire a degree whenever or wherever they want. This can be achieved by creating a system of multilingual content knowledge base so that it can serve all-regional community requirements. Therefore by creating system only in English, we should not restrict the facilities to any communities, as today technology and tools are already available, where it needs to put together and integrate them in a single system. So advantages of having localized IT solutions are:

- Encouragement to do more research in local languages
- Spurred by information technology spending by the state governments, Indian language software and hardware market was slated to touch Rs 100 billion by 2001, according to late Devang Mehta, president of NASSCOM.
- Market level of IT products and solutions will increase if continuous uses of local language sources are sought.
- IT spending for e-governance by state governments in local languages is growing rapidly and the language software market would grow automatically.
- It might increase the IT literacy factor at all levels of masses.
- Since Indian languages are based on similar phonetics preparation of interface so template for one language might reduce the work for other local languages
- We can have solutions/services based on IT like Knowledge tools, Knowledge resources, Translation Support System, Human machine interface system.
- To avoid language barriers to acquire knowledge.
- To promote Indian languages as communication channels for wisdom & scholar communication.
- To serve Indian community who knows English and Indian languages as well.
- To provide access to all sources of information available in Indian languages.
- To facilitate learning in national and regional language as a matter of self-respect of our country and our languages.
- Speech recognition technology can be a major breakthrough in serving the rural and illiterate community, as now they also will have the access of information.

It is true that English is spoken by a very small fraction of the people (about 10%), and yet widely perceived as the language of aspiration but Automatic Translation Systems are available today on the Web, which enable automatic translation of messages and content from English to several other languages. Development can be made by which people can access the Net in their own language and hence the dependencies on an English translator will be reduced.

Speech based system can be helpful in this regard particularly for the rural community. Illiterate people can interact with the computer system in their own language and computer system, after processing, can respond in the same language and can accomplish the task being told.

1.7 Downside of having IT Solutions in Local Languages

The major hurdle in developing these technologies is that it is difficult to standardize these efforts. However there are many language portal and other IT solutions are there but there is lack of standardization.

- Developing an interface in local languages needs more knowledge and thorough understanding about local languages which is time consuming
- To develop IT solution on a particular language we need an expert of local language.
- In the absence of compatibility among various font packages users had problems, limiting the use of IT in Indian languages. So language software companies should tie up with hardware manufacturers, especially printers, to capture large share of market in the country.
- Information Technology Market will be limited to particular local area
- Local languages are at the level of machine, i.e., machines are only ASCII text understandable
- Initial effort is needed more in the development stage. IT Solutions/Services to existing language like English is more easily compared to local languages.
- Process itself is a slow as more time will be taken for translating local language to ASCII machine language.
- There is a risk of services being stagnant when it doesn't reach the mass properly.
- People want to employ easily implemented quick-n-dirty retrofit solutions. They want the results now and for them standards are damned.
- They are difficult to implement. Needs technology development. There is a Lack of initiative on part of implementers.

The major problems for such efforts can be –

- One can display web pages in Indian languages using just fonts. But can one search for information in Indian languages? It is an extremely difficult task, if not impossible, to design a search engine that will understand all the proprietary formats used in today's Indian language web pages. Information organization and retrieval on the web will remain a pipe dream, and advanced web based methodologies such as cross language information retrieval, machine assisted translation, text to speech etc. will be impossible to even think of.
- From the perspective of a web site and content developer, it is a nightmare to maintain non-standard source documents. The web is the medium for collaborative effort. As tools become available to retrieve information from multiple sites, it would be a tremendous waste of resources to build into them the ability to read different proprietary formats.
- Again there will be lack of compatibility in different languages. Suppose for example there is a balance sheet or any technical report available in 'Telugu' language, which is for publicly accessible. Now the people who don't know that language cannot use it. The argument can be that they can convert it into their

own language using some tools but again the question is how they will come to know that there is such a balance sheet or technical report exists in such language, which can be useful for them.

- Also there will be a problem in upgrading of such tools. If any new version with some added features or any patch in previous version of the software tools has to be released, it has to be made available for all the languages, which requires more effort and time.
- Still the URL of the website has to be typed in English. Even though NASSCOM and iDNS.Net are pursuing to register domain names in Indian languages, it will be difficult to communicate these addresses to other users.

1.8 Future Directions

With the recent advancement in computer science, the evolution of Information Technology (IT) has sowed its applicability into the soil of the society. With more and more real life applications of computer systems for public utilities, the need for local language applications is growing exponential. There is a great demand for localization of software and regional language interfaces. With this background the future applications of this frontier technology relies on major R & D efforts from the national laboratories, academia and research organisations with proper support from the Government. The future thrust areas and applications are summarised below:

Standardisation

Nowadays the Indian market is flooded with various products and application systems in Indian languages. There is a need to standardise these systems and products in order to have portability and compatibility. In order to synergies these efforts, standardisation is considered essential in the following areas.

1. Standardisation of IT terminologies in various Indian Languages: This requires a consensus view of various experts (linguists and IT) to arrive a nomenclature for various glossaries in IT.

2. Standardisation of transliteration rules for names and proper nouns: This is very essential for data processing and data base creation. This allows creation of database in only one language (say English) but report generation and query evaluation in various Indian languages. e.g. The spellings ‘Moorthy’, ‘Murty’, ‘Moorty’ should transliterate into only one word in any of the Indian languages.

3. Standardisation of code for various applications like Graphics, Natural Language Processing, String Processing etc. in Indian Languages.

4. Standardisation of Cultural clip arts: This is essential when multi-lingual documents are prepared using Harward Graphics, MS Office etc. for representing cultural symbols etc. e.g. Symbols of standard shapes and colours for representing various cultural arts like ‘Ohm’, ‘Swastik’, ‘Namasthe’, ‘Kalash’, ‘Temple’ etc.

Internet-Based Applications

The Internet users are growing day by day. Almost all information ranging from travel, education, entertainment, hotels, appointments, announcements, news bulletins etc. is now available on Internet. There is great demand to develop tools and applications to support Indian languages on Internet which include:

Development of HTML Plug-ins for Internet browsers and editors to enable them for Indian Language fonts.

Development of Multi-lingual Email servers which facilitate sending and receiving of emails in various Indian languages.

Identification of various news groups and Creation of Bulletin boards for various user groups.

Content creation and Web-page creations for various applications such as Tourism, Industries, medical etc. It is expected that the initiative would come from the respective Government departments.

Machine Aided Translation Systems

In a large multi-lingual society like ours, there is a great demand for translation of documents from one language to another. Most of the state governments work in the respective regional languages where as the union Government's official documents and reports are in bilingual (Hindi/English). In order to have a proper communication there is a need to translate these reports and documents in the respective regional languages. With the limitations of human translators, most of this information (reports/documents) is missing and not percolating down. A machine assisted translation system or a translators' workstation would increase the efficiency of the human translators. In order to realise a MT system, development of domain specific translation systems could be identified as:

Government Administrative procedures and formats.

Parliamentary Questions and Answers

Pharmaceutical information

Legal terminology and important judgements etc.

Human-Machine Interface Systems

In order to create a lot of content for Internet and database applications, large chunks of texts from the printed documents need to be keyed-in. In such circumstances, Human-Machine Interface Systems (HUMIS) would be of great help. The specific areas to be explored and deployable systems need to be developed are:

Optical character recognition systems for various fonts for different scripts of Indian languages.

Speech Syntheses systems for Indian languages: These are very useful for public address systems, broadcasting stations etc. To achieve this there is a need to standardize the speech database (acoustic and phonetic features) already developed for Hindi words and design co-articulation rules, prosodic rules and Knowledge Base for synthesis of phonemes for various Indian languages. It may be noted that the phoneme 'Bha' is same in all Indian languages. However, its acoustic-phonetic features like its duration, intonation and pitch varies with the context (previous phonemes/syllables and the language).

Multimedia-Multilingual Based Systems

Multimedia is one of the emerging technologies which especially catching the education and entertainment sectors to a large extent. In order to motivate people to learn various Indian languages, and assist them in communicating in the local/regional languages these systems would be of great help. Some of the systems that could be developed in a short period of 2-3 years are:

Multi-media Based system for adult education: This could be developed for broadcasting through CCTV and TV networks.
Electronic gadgets for terminology and translation of small sentences/phrases for Tourists.
Language Learning systems.

1.9 Conclusion

In our opinion the following adage would sum up the dilemma of this issue.

Thirty spokes share the wheel's hub;
It is the center hole that makes it useful.
Shape clay into a vessel;
It is the space within that makes it useful.
Cut doors and windows for a room;
It is the holes that make it useful.
Therefore profit comes from what is there;
Usefulness from what is not there.

- Lao-Tse

India being a multilingual country had already recognized the potential of multilingual computing and some of the programs to build competency were initiated more than a decade ago. Since then slowly and steadily many research, development and application oriented activities have been built up at government public and private organizations. This has resulted in creating awareness about the use of computers in the areas of language analysis, understanding and processing. Preparatory work for building corpora of contemporary text have led to take up the development of potential applications like morphological analyzers, spell checkers etc. Defining and refining standards, development of operating systems, human machine interfaces, Internet tools and technologies, machine-aided translations and speech related efforts are some of the major thrust areas identified for attention in the near future. Standardization of terminology for use in regional languages is also receiving considerable attention. The challenges ahead require cooperative efforts in the many upcoming areas such as automatic translations of web-based information, search engines, multimedia content generation and refinement of human machine interfaces. It is well recognised that these efforts need to be accelerated particularly to meet the objective of deeper and wider penetration of IT in the country.

The challenge before us now is to see how best we can harness this emerging technological capability to make it more effective and efficient. In this, the development of language-oriented programmes plays a very important part.

Finally, we can say that by developing the solutions keeping in mind the market size certainly will increase the awareness among the users and will create another market for hardware and software, where the IT sector can generate the profit. Also it will serve the less privileged person and will provide a means to access and develop knowledge base, so it is a win-win situation for both.

References:

1. Frederick Noronha and Partha Pratim Sarker. " Showcasing People Oriented IT Practices: An Alternative Model in South Asia." http://www.bytesforall.org/Egovernance/html/bytesforall_intro.pdf
2. Om Vikas, “ Language Technology Development in India”. Ministry of Information Technology, New Delhi, India. http://www.emille.lancs.ac.uk/lesal/omvikas.pdf
3. Osama Manzar. “Digital divide” could be the biggest business opportunity”. http://www.itcd.net/itcd-2001/papers/doc_pdf/doc_31.PDF
4. http://tdil.mit.gov.in/newsletter1.htm
5. http://www.ethnologue.com/show_country.asp?name=India
6. Going Native on The Net. K. Sunil Thomas, The WEEK, September 24, 2000.
7. http://www.mithi.com/
8. http://www.apnic.net/mailing-lists/s-asia-it/archive/2001/12/msg00030.html
9. http://www.cdacindia.com/