PAN
Localization
www.PANL10n.net

PAN Localization Project
(www.PANL10n.net)

A regional initiative to develop local language computing capacity in Asia

Information has now become such an integral part of our society, that its access is considered as a basic human right. This is because development of rural and urban developing populations is getting increasingly dependent upon access to information. This is specifically applicable to Asia which houses the largest developing population. ICTs, including the Internet, is the largest repository of this information. And, though Asians have become the largest group of Internet users since 2001, these users still form only about 4.5 % of the total Asian population. This shows that there is enormous potential for Internet usage in Asia.

However, in addition to being most populous, Asia is also the most culturally and linguistically diverse region of the world. There are 2197 languages spoken in Asia, which is the largest number of languages spoken in any one region. Only about 20% of these people can communicate in English. This makes English language content available on ICTs inaccessible to a large majority of Asians. This particularly affects those living in the rural areas of developing countries in Asia.

Investments have been put into developing ICT infrastructures in Asia. Nevertheless, the persisting digital divide attests that the current path towards providing connectivity and technology infrastructure alone would not enable the majority of Asian populations to benefit from the present information availability. There are multiple problems perpetuating this divide. One obvious reason is that these populations cannot circumvent the obstacle of English language content. Unless these large non-English speaking populations have the ability to generate and access content in their native languages, they will not be able to use ICTs for their development effectively.

Enabling ICTs in the local language of the user is known as "localization". Specifically, it is enabling computing experience in linguistic culture of the user. Linguistic culture is not just limited to the language but how the language is used by the environment of the user. Thus, for Punjabi speakers in India, the computer should display the language in Gurmukhi script and for Punjabi speakers in Pakistan, the same language should be displayed in Arabic script.

Localization of ICTs requires definition and implementation of standards. These standards include character set encoding, keyboard (and keypad) layout, collation/sorting sequence, locale and ICT terminology. In addition to definition of standards, applications also need to be developed for local language computing to support access and generation of local language content. There is a large variety of applications required, some being more fundamental in nature, while others are more advanced and complex but equally vital for end users. These applications include fonts, lexicon, thesaurus, spell checker, grammar checker, text-to-speech system, speech recognition system, machine translation system and optical character recognition system.

Survey of State of Localization in Asia[1]

Localization and development of applications is only starting for many Asian languages. The reasons have been lack of commercial incentives (as these markets do not promise of large financial returns for software vendors) and the complexity of the local Asian languages. A survey was conducted during a recent training on “Fundamentals of Local Language Computing” held as part of the PAN Localization project (details presented later) at Lahore, Pakistan, in January 2004. Localization experts and developers from 13 different countries participated in this training and provided the data collated in tables below[2] (‘√’ indicates “ work completed”, ‘*’ indicates “work in progress/partially done”, ‘?’ indicates “do not know” and blank indicates “no work done”).

Before any content can be generated or any application is developed, some basic standards for encoding the language must be developed. These include character set encoding (e.g. Unicode), keyboard layout, key pad layout (e.g. for mobile telephones), collation sequence (to enable applications like databases), terminology translation and locale definition (to enable computer interface in local language). The survey responses are tabulated in Table 1.

Table 1: Localization Standards

Country	Language	Char. Set	KB	Key pad	Coll. Seq.	Interface Term.	Locale
Myanmar	Burmese	√				*	*
Cambodia	Khmer	√	√				
Mongolia	Mongolian	√	√	?	*	*	√
Laos	Lao	√	√				
Nepal	Nepali	*	*		*	*	*
Sri Lanka	Sinhalese	√	√	*	*		*
Thailand	Thai	√	√	*	*	*	√
Bhutan	Dzongkha	√	√		*		*
China	Tibetan	√	√		?		
Japan	Japanese	√	√	√	√	√	√
Bangladesh	Bangla	√		?	*	*	√
Afghanistan	Pashto	√	√		?		*
Iran	Farsi	√	√		*	*	*
Pakistan	Urdu	√	*		*	*	*

The responses indicate that the encoding and keyboard layouts are standardized for most languages. This would allow developing basic desktop publishing capability, and has been achieved through national and international efforts, e.g. organization like Unicode Consortium (www.unicode.org). However, much work needs to be done to define other standards needed to further process the data. For example, collation sequences for the languages have to be defined to enable applications which sort linguistic data, like voter lists, etc.

[1] This survey is limited to only the countries from which representatives attended PAN Localization Project training (see the “Training” link at www.PANL10n.net).

[2] The data has been provided by the training participants and has not been independently verified, therefore some variation may exist from the responses received. The data is still representative of the bigger picture for Asian region.

Based on the standards, the applications may be developed on Microsoft or Linux platforms, two most popular end-user desktop operating systems. The survey also tried to determine the level of application support on these two platforms. The questions were divided into two categories of applications: basic applications which realize the standards and allow basic desktop publishing for the end-user, and advanced applications used to assist user to generate and access content in local languages.

The basic applications include utilities which enable to realize the encoding standard (Keyboard and Fonts), sort and search data (Collation and Find/Replace utilities) and allow basic word processing facilities, like spelling checker, thesaurus and Natural Language Processing component (e.g. Word/Line Break Determiner for languages like Lao and Khmer, Bidirectional Algorithms for Arabic script based languages like Urdu and Farsi, etc.). The responses for Microsoft platform are tabulated in Table 2 and for Linux platform are given in Table 3 below.

Table 2: Basic Localized Applications on Microsoft Platform

Country	Language	KB Driver	Font	Collation	Find/ Replace	NLP	Spell Check	Thesaurus
Myanmar	Burmese		*	*			*	
Cambodia	Khmer	√			√			
Mongolia	Mongolian	√	√		√		*	?
Laos	Lao	√	√					
Nepal	Nepali	√	√	*	*	*	*	*
Sri Lanka	Sinhalese	*	√			*	*	
Thailand	Thai	√	√	√	√	*	√	√
Bhutan	Dzongkha	*	*					
China	Tibetan	*	*		*			
Japan	Japanese	√	√	√	√	*	√	√
Bangladesh	Bangla	√	*	*	*		*	*
Afghanistan	Pashto	√	√		?	√		
Iran	Farsi	*	*	*	*	√	*	*
Pakistan	Urdu	√		*	*	√		*

Table 3: Basic Localized Applications on Linux Platform

Country	Language	KB Driver	Font	Collation	Find/ Replace	NLP	Spell Check	Thesaurus
Myanmar	Burmese		*					
Cambodia	Khmer							
Mongolia	Mongolian	√	√	*	*		*	?
Laos	Lao							
Nepal	Nepali	√						*
Sri Lanka	Sinhalese	*	√				*	
Thailand	Thai	√	*	√	√	*		*
Bhutan	Dzongkha							
China	Tibetan	*			*		?	
Japan	Japanese	√	√	√	√	?	?	?
Bangladesh	Bangla	√	*	*				?
Afghanistan	Pashto		*		?	√		
Iran	Farsi	√	*	√	*	√	*	*
Pakistan	Urdu	*	*			√		

As the responses indicate, there is currently more support on Microsoft platform for keyboard and fonts to do basic local language data processing. Linux is catching up as more solutions are being developed but does not provide the same level of support at this time. It should be noted that all the support indicated for Microsoft platform is not developed by Microsoft and is sometimes developed by third parties.

Additional utilities, out of which collation is perhaps most necessary for data processing (NLP is also significant for some languages), are still missing for most of the languages and much work needs to be done to fill this gap. Japanese and Thai are ahead of all other languages surveyed.

For wider access to content to literate and illiterate non-English speaking population of Asia and for quicker content generation, some advance applications must be developed. Automatic Machine Translation systems can provide instant access to existing English data on the Internet. Text-to-speech systems can provide access to illiterate populations. Automatic speech recognition can help create local language and culturally meaningful content quickly and similarly optical character recognition system can help convert published material into electronic content for exchange. These applications can be instrumental in bridging the digital divide. The status of these applications for Asian languages is given in Table 4 (for Microsoft platform) and Table 5 (for Linux platform).

Table 4: Advanced Local Language Applications on Microsoft Platform

Country	Language	Grammar Check	Lang. Corpus	TTS	ASR	MT	OCR
Myanmar	Burmese				*		*
Cambodia	Khmer						
Mongolia	Mongolian	*	?	?	?		
Laos	Lao						
Nepal	Nepali						
Sri Lanka	Sinhalese	*	*	*		?	*
Thailand	Thai		*	√	?	*	*
Bhutan	Dzongkha						
China	Tibetan					*	
Japan	Japanese	√	√	√	√	√	√
Bangladesh	Bangla		*				*
Afghanistan	Pashto						
Iran	Farsi	?	*	*	?	*	*
Pakistan	Urdu	*	*	*	*	*	*

Table 5: Advanced Local Language Applications on Linux Platform

Country	Language	Grammar Check	Lang. Corpus	TTS	ASR	MT	OCR	Linux Distr.
Myanmar	Burmese							
Cambodia	Khmer							
Mongolia	Mongolian	*	?	?	?	*	*	√
Laos	Lao							
Nepal	Nepali							*
Sri Lanka	Sinhalese		*	*		?	*	*
Thailand	Thai		*	*	?	*	*	√
Bhutan	Dzongkha							
China	Tibetan	?	?					
Japan	Japanese	?	?	?	?	?	?	√

Bangladesh	Bangla		*					*
Afghanistan	Pashto							
Iran	Farsi	?	*	*		?	?	*
Pakistan	Urdu	*	*	*	*	*		*

As can be seen from the responses of the survey, though many initiatives are underway for various languages, there is hardly any significant development completed in this area. Only Japanese language applications are currently available. For many of the languages, there is not even efforts which have started looking in these areas.

Most of the times research and development in these areas are guided by policies. Relevant policies which would address local language computing consist of linguistic policies of the country, their ICT policies and specifically their localization policies. The survey also included questions regarding existence of such policies. The responses are tabulated in Table 6 below.

Table 6: National Policy Related to Local Language Computing

Country	Language	Ling. Policy	IT Policy	Localization Policy
Myanmar	Burmese	*	*	*
Cambodia	Khmer	√		
Mongolia	Mongolian	*	*	*
Laos	Lao	?	*	?
Nepal	Nepali	*	*	*
Sri Lanka	Sinhalese	*	*	*
Thailand	Thai	*	*	*
Bhutan	Dzongkha	?	*	
China	Tibetan	?	?	?
Japan	Japanese	?	√	√
Bangladesh	Bangla	?	√	
Afghanistan	Pashto	√		*
Iran	Farsi	√	√	
Pakistan	Urdu	√	√	√

Interestingly though few countries have localization policy, many are working toward developing one. Most countries also have Linguistic and IT policies, but they may or may not drive localization policy. This needs to be further investigated.

As the survey indicates, localization initiative have not been either rigorously taken up or pursued consistently in many Asian countries. To address these issues a regional effort to develop local language computing capacity for Asia has been taken up by the International Development Research Centre (IDRC) through its Pan Asia Networking PAN programme, in collaboration with National University of Computer and Emerging Sciences (NUCES) through its Centre for Research in Urdu Language Processing (CRULP), and is called **PAN Localization** project.

PAN Localization Project (www.PANL10n.net)

PAN Localization project focuses on documenting the problems and researching the solutions to enable localization of ICTs. This project is unique as it will be the first study of its kind which looks at the common problems faced by Asian region and research into a comprehensive solution. It is thus a timely and an urgently needed initiative for Asian underdeveloped populations and will be instrumental towards providing an equitable access to information is this digitally divided information society.

The core objectives of PAN Localization project are to research into the following three fundamental dimensions of localization for Asian languages.

- to develop sustainable human resource capacity in the Asian region for R&D in local language technology
- to raise current levels of technological support for Asian languages
- to advance policy for local language content creation and access across Asia for development

In this project, CRULP is coordinating efforts across Asia within ICT researchers, practitioners, linguists and policymakers from the governmental agencies, universities and private sector of six countries of Asia including:

- **BANGLADESH:** BRAC University, working for Bangla
- **BHUTAN:** Department of IT, Ministry of Information & Communications, working for Dzongkha,
- **CAMBODIA:** National ICT Development Authority, working for Khmer
- **LAOS:** Science, Technology and Environment Agency, working for Lao
- **NEPAL:** Madan Puraskar Pustakalya, working for Nepali
- **SRI LANKA:** University of Colombo School of Computing, working for Sinhala & Tamil

Specific details of standards and local language applications being developed by these countries through the project are given in Table 7 below.

Table 7: Project Output Matrix for Local Language Software Being Developed Through PAN Localization Project

x – PAN Localization project deliverable on Microsoft Platform
xx – PAN Localization project deliverable on Linux Platform
x or xx – only basic algorithms will be implemented

	Nepal	Laos	Cambodia	Sri Lanka	Bangladesh	Bhutan
Standards						
Character Set						
Keyboard						
Keypad (Telephone)	x		x			
Collation Sequence	x	x	x		x	
Interface Terminology			x			xx
Handheld Device Interface Terminology	x		x			
Locale						
Software Applications						
Keyboard driver						
Normalization and Sorting Utility	xx	x	x		xx	xx
Find/Replace Utility		x			xx	
Fonts						xx
Mobile/PDA Fonts	x		x			
Lexicon	xx	x	x	x	xx	
Thesaurus	xx					

Natural Language Processor			**x**		**xx**	
Spell Checker	**xx**	**x**	**x**		**xx**	
Grammar Checker	**xx**	**x**	**x**			
Text Corpus Development for NLP Applications				**x**		
Text to Speech System				**x**		
Speech Recognition System						
Machine Translation System						
Optical Character Recognition System				**x**	**xx**	
Linux Distribution	**xx**					**xx**
GNU Cash (Open source accounting System)	**xx**					

Other than the above mentioned technical outputs, PAN Localization project also aims for the following:

- to develop baseline of the status of local language computing in Asia
- to conduct research into linguistics, computing and language processing for selected local languages
- to develop training material and provide training in local language computing
- to evolve a process framework for developing local language computing
- to experiment with marketing strategies to promote the use of local language tools for content development
- to develop a regional network of researchers, practitioners and policy-makers for collaborative learning in local language computing
- to contribute to the state-of-practice in local language computing through a rigorous research publication program
- to evolve a policy framework for local language computing and generation and usage of local language content for development

Conclusion

Local language computing development is a precursor to bridging the persisting digital divide. Without local language solutions the rural and underdeveloped populations of Asia, which do not understand English or other popular languages on the internet, will remain isolated from the information available and will be unable to use ICTs effectively for development. Though work is being done in local language computing, most of the vendors are focusing on rich markets and underdeveloped and poor populations (which need this most urgently) are suffering from continuous delays, as these markets do not provide financially viable investment opportunities. The few efforts which have been made in these directions have been individual and amateurish and, at best, organizational. Very limited large national or regional initiatives have been made to look at the general and specific problems across Asia and to come up with effective strategies to address them.

PAN Localization project is thus a timely and an urgently needed initiative for Asian underdeveloped populations and would be instrumental towards providing an equitable access to information is this digitally divided information society. This project will work towards developing sustainable local human resource capacity for localization, assess and develop technology for some Asian languages and, looking at the problems at regional level, advance policy recommendations to alleviate the issue.

Research findings and developments through this project will be put in use to directly benefit rural developing populations in Asia. They will also be used for addressing the areas and populations not included in this project. The peer network developed through this project will create an active force which can sustain this regional effort and generate a collective regional voice to guide vendors, national and international policies and local human resource for the promotion of local language computing. This three year project will sample Laos, Cambodia, Bangladesh, Nepal, Bhutan and Sri Lanka as test cases for this purpose.

The findings from this project will be published in the form of a comprehensive PAN Localization Handbook, in addition to the numerous research reports, all of which will be available for all relevant people and organizations regionally and globally. The findings will be disseminated through the multilingual website of this project: www.PANL10n.net.