Research Capacity Building through PANL10n 2003-2010

Working Paper 03

Results of PAN L10n project evaluation research presented in preliminary form for discussion and critical comments

Sana Shams
Yasmeen Daud

Center for Language Engineering (CLE)
Al-Khawarizimi Institute of Computer Science (KICS)
University of Engineering & Technology (UET)

www.cle.org.pk

www.idrc.ca

Published by

Center for Language Engineering (CLE)
Al-Khawarizimi Institute of Computer Science (KICS)
University of Engineering & Technology (UET)
Lahore, Pakistan

This work was carried out with the aid of a grant from the International Development Research Center (IDRC), Ottawa, Canada, administered through the Centre for Language Engineering(CLE), Al-Khawarizimi institute of Computer Science(KICS), University of Engineering & Technology(UET) Lahore, Pakistan.

Acronyms

ACSA	Afghan Computer Science Association
BPPT	Agency for Assessment and Application of Technology
CRBLP	Center for Research on Bangla Language Processing
CRULP	Center for Research in Urdu Language Processing
DIT	Department of Information Technology
D.Net	Development of research center
ENRD	E-Network Research and Development
ICI	Information and Communication Technology
IMPP	Madan Puraskar Pustakalaya
MCIT	Ministry of Communications and Information Technology
MoEYS	Ministry of Education, Youth and Sports
MoIC	Ministry of Information and Communication
MUST	Mongolia University of Science and Technology
NiDA	National ICT Development Authority of Cambodia
NAST	National Authority for Science and Technology
NUCES	University of Computer and Emerging Sciences
NUM	National university of Mongolia
RCB	Research Capacity building
R&D	Research and Development
UCSC	University of Colombo School of Computing

Acknowledgments

PAN Localization Project

Enabling local language computing is essential for access and generation of information, and also urgently required for development of Asian countries. PAN Localization project is regional initiative to develop local language computing capacity in Asia. It is partnership, sampling eight countries from South and South-East Asia, to research into the challenges and solutions for local language computing development. One of the basic principles of the project is to develop and enhance capacity of local institutions and resources to develop their own language solutions.

The PAN Localization Project has three broad objectives:

I. To raise sustainable human resource capacity in the Asian region for R&D in local language computing
II. To develop local language computing support for Asian languages
III. To advance policy for local language content creation and access across Asia for development

Human resource development is being addressed through national and regional trainings and through a regional support network being established. The trainings are both short and long term to address the needs of relevant Asian community. In partner countries, resource and organizational development is also carried out by their involvement in development of local language computing solutions. This also caters to the second objective. The research being carried out by the partner countries is strategically located at different research entry points along the technology spectrum, with each country conducting research that is critical in terms of the applications that need to be delivered to the country's user market. Moreover, PAN Localizations project is playing an active role in raising awareness of the potential of local language computing for the development of Asian population. This will help focus the required attention and urgency to this important aspect of ICTs, and create the appropriate policy framework for its sustainable growth across Asia.

The scope of the PAN Localization project encompasses language computing in a broader sense, including linguistic standardization, computing applications, development platforms, content publishing and access, effective marketing and dissemination strategies and intellectual property right issues. As the Pan Localization project researches into problems and solutions for local language computing across Asia, it is designed to sample the cultural and linguistic diversity in the whole region. The project also builds an Asian network of researchers to share learning and knowledge and publishes research outputs, including a comprehensive review at the end of the project, documenting effective processes, results and recommendations.

Countries (and languages) directly involved in the project include Afghanistan (Pashto and Dari), Bangladesh (Bangla), Bhutan (Dzongkha), Cambodia (Khmer), Laos (Lao), Nepal (Nepali), Sir Lanka (Sinhala and Tamil) and Pakistan, which is the regional secretariat. The project started in January 2004 and will continue for three years, supporting a team of seventy five resources across these eight countries to research and develop local language computing solutions. Further details of the project, is partner organizations, activities and outputs are available from its website, www.PANL10n.net

Table of Contents

List of Tables

List of Figures

1 Introduction

At the beginning of the 21st century, a salient feature of the world is the socio-economic divide between different communities. Inequalities are quite pertinent among the two poles of the world. The most visible of these are related to the world development problem of inequalities among nations (Avgerou, 2003). ICT is identified as the most promising and fundamental driver for economic growth and the improvement of social conditions consistently in contemporary discourses on development (Avgerou, 2003, Lutz, 2003). It is due to its potential that enabled the under-privileged to leapfrog the barriers to information access, improve livelihood opportunities and communicate with people across the globe (World Bank, 2002). It is expected that information and communication technologies (ICTs) would play a key developmental role in poor and developing countries. The potential of these technologies could turn around uncompetitive industries and dysfunctional public administration, and provide unprecedented opportunities for the information intensive social services, such as health and education (Sahay & Avgerou, 2002).

As reported in Human Development Report (UNDP, 2001), the developing countries are being rapidly adopted information and communication technologies (ICTs) to make improvement in social and economic conditions and enhancing the standards of people's life. However, studies on utilization of ICTs in national development have been shown that investment in ICT are not contributed significantly in social and economic development (Wellenius et al. 2000, Yang 2001). Such studies have been raised the question whether ICTs have any real effect on national development (Heeks, 1999).

This question becomes more significant when (ITU, 2011) reveals that only 21% of developing economies can access information on the internet. This ratio is quite low as compared to the developed countries. The most striking reason of this digital divide is language barrier as English is the lingua franca for ICTs (Pimienta, 2005). Language barrier is hampered this access as local population is barely literate to adequately read and write in foreign (English) language.

While in the case of Asia, this language based digital divide is very pronounced. About 2,322 languages are spoken (Lewis, 2009) only in Asia. Among them, just 2% people know how to read and write in English, a pre-requisite for ICT usage. In this scenario, the transformation of technology or Local language computing in local languages called localization is much needed for the efficient exploitation of ICTs in Asia (Gul, 2004).

As defined by Hussain and Mohan (2007), localization is "*The process of developing, tailoring and/or enhancing the capability of hardware and software to input process and output information in the language, norms and metaphors used by the community*." It is a three step process. First, the linguistic analysis is required to document (and standardize) language conventions that are to be modeled. Second, localized applications (both basic and intermediate level) e.g. fonts, keyboard, locale, spell checkers, etc. need to be developed to enable basic input and output of text in a local language. Thirdly to provide comprehensive access and assist content development, advanced applications like translation systems, speech dialogue applications, etc., need to be developed.

Localization therefore requires significant knowledge of linguistics (phonetics, phonology, morphology, syntax, semantics and pragmatics), signal and speech processing, image processing, statistics, computational linguistics and advanced computing (Hussain et al, 2007). This research being language

dependent, entails nurturing indigenous *research capacity* (Breen et al, 2004, DFID, 2007) at the levels of individuals, organizations, and systems to sustain.

2 PAN Localization Project

Strengthening indigenous research capacity versus technology transfer (Harris, 2004; Nokolov and Illieva, 2008) is the most effective process for advancing research in localization. In this context, PAN Localization project was conceived. It was a regional initiative to develop local language computing capacity in Asia. It was the joint venture of eight countries from South and South-East Asia, to research into the challenges and solutions for local language computing development. One of the basic principles of the project was to develop and enhance capacity of local institutions and resources to develop their own language solutions.

The scope of the PAN Localization project was encompassed language computing in a broader sense, including linguistic standardization, computing applications, development platforms, content publishing and access, effective marketing and dissemination strategies and intellectual property rights issues. As the PAN Localization project researches into problems and solutions for local language computing across Asia, it was designed to sample the cultural and linguistic diversity in the whole region. The project will also build an Asian network of researchers to share learning and knowledge and publishes research outputs, including a comprehensive review at the end of the project, documenting effective processes, results and recommendations.

Countries (and languages) directly involved in the project include Afghanistan (Pashto and Dari), Bangladesh (Bangla), Bhutan (Dzongkha), Cambodia (Khmer), Laos (Lao), Nepal (Nepali), Sri Lanka (Sinhala and Tamil) and Pakistan, which is the regional secretariat.

3 Research Capacity Building (RCB)

Capacity Building within the context of Research is enhancing the abilities of individuals, organizations and systems to undertake and disseminate high quality research efficiently and effectively (Department for International Development, 2010). "*Capacity building is a process whereby people are enabled to better perform defined functions either as individuals, through improved technical skills and or professional understanding, or as groups aligning their activities to achieve common purpose"* (Breen. et.al., 2004). Capacity building needs to be evaluated and many have significant research interest in this area within the field of evaluation (Powell & Boyd, 2008).

In PAN Localization project, a significant effort has been done in research to investigate the challenges associated with building capacity for localization in the partner countries. Thus, to meet the challenges regarding research capacity building in the country components of PAN Localization projects specifically objectives were developed to assess the project contribution in building Research capacity on local language computing in the country partner institutions (CPI's) discussed below.

I. The extent to which the required deliverables have been delivered by each CPI
II. How many research publications on localization have been published by each CPI
III. The level of technical learning through training in local language computing
IV. The extent of collaboration and online participation by each CPI
V. The level of participation to disseminate research work by each CPI
VI. The Infrastructure that developed during PAN localization project for research capacity building

4 Research Capacity Building Model for PAN Localization Project

In order to study the effectiveness of the PAN Localization project in building research capacity in its partner institutes, a comprehensive survey of literature was conducted to adopt framework in assessing evaluation objectives made to investigate the project contribution in building capacity in local language computing in its partner institutions. Based on the review it was noted that Research capacity building (RCB) frameworks define levels and set of practices that address project objectives set to investigate capacity building.

RCB frameworks available in literature (Cooke 2005, Neilson and Lusthaus 2007, Wignaraja 2009) largely recommend three structural levels and six basic principles upon which capacity building must be designed as discussed in the sub-sections below.

Structural level of RCB defines the point of view upon which capacity development initiatives must be targeted. They include individual, organizational and system levels (Neilson and Lusthans 2007, Breen et al, 2004). Some frameworks follow a hierarchal categorization of these levels (Potter and Brough, 2004) and others form a phase-wise development plan (Wibberley, Dack, & Smith, 2002, Breen et al, 2004) where capacity building at certain prior level necessitates capacity development at the next level. Interventions however cannot be carried out at a certain level in isolation. Every activity accomplished at a certain level has impact on the other levels.

Cooke (2005) recommends six principles of capacity building that include focusing interventions on: skill development; focus on close to practice research; establishment of linkages, partnerships and collaborations; development of capacity for dissemination and impact; sustainability and continuity of research and development of infrastructure. Each principle is briefly described below.

4.1 Skill Development

RCB requires a multi-faceted skill development process through training and supervision to primarily develop technical, managerial, and publishing skills (Harris 2004, Raina 2007). Skill development can also be viewed in the context of career development and generating opportunities to apply research skills in practice (Rhee and Riggins. 2007).

In PAN Localization project, skills development was addressed by building technical skills to conduct and publish localized research outputs. Indicators used to measure skill development were as follows.

I. Completion of localized software
II. Publications of research papers

Each country team had developed a project plan and agreed to complete a set of deliverables as set out in the project contract. These deliverables involved expertise in Computer Sciences, Linguistics or computational linguistics. Requiring each project team to deliver specific research outputs served as a persistent capacity building process. Completion of project deliverables enabled researchers to work on real localization problems and find research solutions through involvement in problem identification, project designing, implementation, quantitative and qualitative analysis.

Ability to publish research in the form of research papers is a salient indicator for measuring the researcher's research capacity. Thus the number of research publications produced by PAN Localization

project teams at various national as well as international research conferences is used as the second indicator for analyzing research capacity enhancement.

4.2 Training on Close to Practice Research

A foremost principle of RCB is in directing researchers' ability to produce research that is useful for informing policy and practice (Cooke, 2005). Thus capacity building interventions ensure that research is "close to practice" such that new knowledge generated can directly impact development.

Through PAN Localization project's research it was envisioned that localized technology being developed must be deployable and of direct use to the communities. Data was s collected from the project beneficiaries in communities of Pakistan, Nepal, Cambodia and Bangladesh where this research work was being conducted to show the relevance and closeness to practical needs of the communities for which the research work was being undertaken. Indicators used to measure the progress on this principle in the context of the project were as follows

I. Proficiency of end users in the language spoken at home
II. Proficiency of end users in the language spoken at work
III. Reading skill of end users in English vs. local Language
IV. Writing skill of end users in English vs. local Language

Specific question was asked from the 228 rural community members in Bangladesh, Nepal, Pakistan, and Cambodia regarding the language that they speak at home and at their work. The respondents were also asked to rate their reading and writing skill in English vs. local language on a scale ranging from excellent to poor.

4.3 Development of Linkages

Developing linkages, partnerships and collaborations is a reciprocating process of involving organizations in the knowledge information chain for fostering development and diffusion of quality research (Wignaraja 2009, Breen et al 2004). It also harnesses an increased knowledge base for research development and enhancement.

Research groups often operate in isolation, limiting the scope and success of their work. Thus in order to enhance the capacity, resources must be appropriately linked up and connected with active groups working on similar initiatives for robust and collaborative learning. Experiences of researchers who are working successfully under similarly resource-constrained conditions engender trust and motivation.

The following indicators have been used to see the extent to which country project teams have also been focusing on building their capacity by developing appropriate linkages, partnerships and collaborations. They were:

I. No. of formal organizational collaborations
II. Extent of Participation of teams on research groups

Partner teams were encouraged to establish partnership and collaboration with institution that had more expertise in a specific field. These collaborations enabled the partners to collectively plan the technical and financial details, exchange data and technology and formalize shared intellectual property regimes, building institutional capacities in the context. The project teams were also encouraged in participating in online research networks, discussion groups, communities and forums for collaboration, knowledge sharing and sharing.

4.4 Dissemination and Impact

Dissemination of research, through peer reviewed publications and presentations at academic conferences, is essential for sharing knowledge (Harris 2004, Breen et al 2004). Capacity building for wider research dissemination incorporates instruments of publicity through factsheets, the media and the Internet (Cooke 2005) for a variety of stakeholders, including public, policy makers and the relevant research community.

Dissemination is an essential part of undertaking research. Research is as credible as much as it is referenced, cited in other publications, brought to people knowledge and properly disseminated. Indicators used to measure the dissemination capacity enhancement for the country components include:

I. Development of a local project website
II. Organization of awareness seminars
III. Creation of promotional materials
IV. Participation of teams in events and competitions

Project required each country team to provide local content for centrally maintained multilingual website www.pan110.net. In addition the teams also hosted their separate website providing detailed information about their respective groups. This has given global access to project outputs.

The project has organized awareness seminars to disseminate and publicize research results to local community. Through these seminars partner institutions have been regularly presenting their work to the key stakeholders from government, IT industry, academia, media and end user communities. Development of promotional material has been an integral strategy for research dissemination. In addition to publicity flyers, teams have distributed CDs containing the project outputs. Teams participated in events and competitions and many of the project outputs have been presented at national and international forums and have been awarded.

4.5 Infrastructure Development

Rhee and Riggins (2007) defines infrastructure as a set of structures and processes that are set up to enable the smooth and effective running of research projects. These include availability of technical resources including equipment, books, connectivity, etc. as well as sound academic and managerial leadership and support for developing and sustaining research capacity.

Following Indicators were used to measure the Infrastructure development included:

I. Acquisition of academic resources
II. Procurement of equipments
III. Provision for Operating expenses

The teams have been capacity build to develop the appropriate localization research infrastructure by providing funds for acquiring academic resources, e.g. books and journals, and specialized software and to some countries with supporting the recurring expenses of the organizations for the initial research group set-up.

4.6 Sustainability and Continuity

RCB must ensure strategies for maintenance and continuity of the acquired skills and structures to undertake research. Wignaraja (2009) defines capacity development as a process of transformation that

emerges from within the individuals, organizations and systems. Long term sustainable capacity development requires consolidation of local systems and processes through practice.
Building research capacity for skill development enhances the competency to enable sustainability and continuity of the research being undertaken. Indicators to measure sustainability and continuity two indicators have been used:

I. Degree of Organizational skill development
II. No of trained recourses in the different domains of localization

Organizational capacity enhancement as a result of team skill building is another salient factor in measuring the capacity enhancement of teams for sustainability of research. Thus organizations required to enhance their knowledge to gain advancement in different domains of local language computing. Through the PAN Localization project, teams required to train the significant number of technical developers, linguistics and social scientist.

5 Methodology

In order to collect data for the evaluation of capacity building, Mix method approach was adopted. Structure questionnaires (attached in appendix B) were sent through e-mail to each of the relevant country project coordinators for getting information on capacity building. In addition, the content analysis of the annual progress reports and research reports respectively developed by the regional secretariat of the project and the partner institutions has been also conducted too. After collecting data, it was evaluated on the indicators developed through RCB model to see each partner institution contribution in building capacity on local language computing.

6 Findings

This section presents the case study of each country component that participated in PAN Localization project. Through PAN Localization, an effort was made to build and enhance the capacity of partner institution in local language computing. These efforts were analyzed through RCB model adopt to evaluate the performance of each country component. Results are discussed below in detail.

6.1 Afghanistan

In Afghanistan PAN Localization project has been working with collaboration of Afghan Computer Science Association (ACSA, http://www.acsa.org.af/) and Ministry of Communications and Information Technology (MCIT, http://mcit.gov.af/en)

In Afghanistan, Computer literacy and ownership rates were estimated at less than 10 percent of the population in 2010 (Bureau of Democracy, Human Rights, and Labor, 2010). This underdevelopment is mainly attributed to the political environment of the country prevailing during last three decades. When ICT was introduced to the world, Afghan society was in the process of reshaping itself after the war. The resultant digital divide is now one of the major problems of the social sector of Afghanistan. To overcome this problem, development is essentially needed in information technology sector and for quick development, it was required that technology should be understandable; this means that the content should be available in most commonly spoken local language of the country, so that Afghanis can conveniently acquire latest technology and use it.

Through PAN Localization project, the situation gradually improved in ICT sector of Afghanistan. Marjan (2009, Pg. 132) explained that *ACSA (Afghan Computer Science Association), in collaboration with the MCIT (ministry of Communications and Information Technology) and Microsoft, completed the Pashto version of Microsoft Windows XP and Office 2003 in December 2007. Work on Font, lexicon, and spell check development is still conducting. The ACSA team has likewise prepared the initial feasibility report and produced the localization version of International Domain Names in the Pashto language. All of these are expected to boost the capacity of the Afghan people to develop digital content*.

The scrutiny of the work done by Afghanistan country component was highlighted that PAN Localization project had enormously contributed to enhance the capacity of Afghanistan country team in developing digital content and also boosted them in building research capacity on localization. The following sections presented information showing capacity development of each project team assessed though Research Capacity Building model.

Skill Development

The prime focus of the project country component was the development of specific localized software. These localized software involved expertise in linguistics, computer science and computational linguistics. In Phase I, Pashto Font and Pashto Keyboard whereas in phase II, Pashto Sea monkey and Pashto Character Set for IDNs were requisite localized software to deliver and Afghanistan country component successfully accomplished these research outputs. Computer science and Linguistics competence have been involved in the localized software namely Pashto Font, Pashto Sea monkey and Pashto Character Set for IDNs while "Pashto Keyboard" involved expertise only in Computer science.
Skill set pertaining to competence in linguistics, computer science or computation linguistics has also been highlighted for each of localized software in the table below:

Afghanistan				
Localized software	Ling.	CL	CS	Status
Pashto Font	*		*	Completed
Pashto Keyboard			*	Completed
Pashto Sea Monkey	*		*	Completed
Pashto Character Set for IDNs	*		*	Completed

Table 1: Afghanistan team's status regarding Localized Software

The status mentioned in above table showed that the project team has been able to deliver all localized software as per the contract and Afghanistan country team's skill has enormously enhanced over the project implementation. It is also worth mentioning that the country project team has also advanced in their over-all local language computing skill set through continual research and development of local language software and its components.

A comparison of the accomplished localized software in PAN Localization project's phase 1 and phase 2 reveal the fact that the project country component was researching on development of intermediate complexity local language computing application as compared to Phase 1 in which team was only focusing on the development of basic complexity software. During the first phase of the project, Afghanistan Country Component had worked in the areas of character set finalization and keyboard layout development. In the second phase work on terminology translation and character set definitions for Internationalized Domain Names had been planned. During phase II of PAN project Afghanistan Country Component had updated Pashto keyboard for Windows XP and Vista along with Pashto fonts. In

addition to that, Pashto SeaMonkey comprising of web browser and email client has been released. The team had the honor of winning the Manthan award (2008) in the E-Localization category for Pashto SeaMonkey. Thus it is clearly evident that as the team has gained more technical skills over the project implementation.

Afghanistan country component developed a glossary of around 3000 terms in Local language to ease and standardize the translation of Microsoft LIP Localization and this performance manifested skill capacity enhancement of their team.

Training on Localization Essentials, conducted at the National University of Computer and Emerging Sciences, Lahore, Pakistan, from 4th - 8th May 2005. This training was held in CRULP, Pakistan to introduce and familiarize the Afghanistan team formally with the essentials of local language computing. Discussion was focused on the topics related Localization Requirements, Font Development, Keyboard, Collation, Lexicon, XML Framework, Spell Checker and Advanced Localization Applications. This training was very helpful in providing knowledge of basic localization topics to the Afghanistan team as most of the team did not have any prior experience in local language technology development. The training was organized in NUCES, Lahore from December 3rd, 2007 till December 7th, 2007 to enhance the skill of Afghanistan country component in developing project plans or designing project evaluation by using social evaluation tools. The training was mainly focused on different steps involved in Font development. The discussion during the training covered subjects broadly classified as Font Development, Open Source Software (OSS) Localization, and Outcome Mapping. The discussion had also been focusing on localization of open source Internet suite, SeaMonkey, by using the open source Computer Aided Translation tool, OmegaT. The training was enormously helpful in providing Afghanistan country component evaluative thinking and an orientation to planning, monitoring and evaluation through outcome mapping methodology. The training was very accommodating for the trainees to enhance their skill capacity on localization. Hameed Sherani (Project Leader), Mr. Sharifullah Mahboob (Linguist), Mr. Habibur Rahman(Senior Software Developer), Mr. Munir Paykan (Calligrapher) were the trainees from Afghanistan. Afghanistan country component's performance regarding localized software was significant in spite of the fact that their project team could not manage publication of the relevant research papers.

Figure 1: Trainees of Afghanistan, 2007

Development of linkage

Development of appropriate linkages, partnerships and collaborations is being considered another significant matter in capacity development. Afghan computer science association (ACSA) collaborated with Ministry of communication and information technology (MCIT) at national level. The Ministry of Communication (MoC) addresses the ICT issues in the country and it helped to start the localization project in the country. At international level, Afghanistan country component collaborated with Microsoft localization program (MLP) and this collaboration was advantageous to conduct localization project in Afghanistan and it has also been contributing to the ICT sector of country. In addition, Microsoft assisted to Ministry of communication and information technology regarding the development of localization project. These collaborations enabled the partners to collectively plan the

technical and financial details, exchange data and technology and discuss and formalize shared intellectual property regimes, building institutional capacities in the context.

Dissemination

The credibility of research has been assessed though it's referencing and citation in other publications. In other words, dissemination is an essential part of undertaking research.

The main and sustained source of information and outputs of the project has been the project website. The core site has been maintained by the project's regional secretariat and one person from country team of Afghanistan act as a website coordinator and provides local content for the centrally maintained multilingual website www.panl10n.net.

In addition the team also hosted their separate websites www.acsa.org.af/ providing detailed information about their respective research groups, hosted by their organizations, which are linked from the main website as well. This has given global access to project outputs.

Figure 2: Home page of Local website of ACSA

The project has organized awareness seminars to disseminate and publicize research results to local community. The first ever National Computational Linguistics Seminar in Afghanistan was conducted on August 12th-13th, 2006. It was the first step to promote awareness of local language computing standards on such a large scale. The objective of this seminar was to provide a platform for networking between the linguists and ICT professionals.

The work done by Afghanistan country component of PAN Localization project was also highlighted in the seminar by the country leader Mr. Omar Mansoor Ansari. The expected outcome of this seminar is contribution to the awareness of public about the latest development of this project through media outreach. Moreover it was a quick and stable move towards advancement of the localization program by meeting with high authorities of the Government of Afghanistan. The gap between linguist scholars and ICT professional is also to be bridged by creating a joint committee from the members of these two sectors.

Figure 3: Awareness seminar 2006

In phase II Afghanistan Country Component has updated Pashto keyboard for Windows XP and Vista along with Pashto fonts. In addition to that, they have successfully released Pashto SeaMonkey

comprising of web browser and email client. The Afghanistan component participated in Manthan Award and ICT Mela South Asia 2008.

The team had the privilege and honor of winning the award in the E-Localization category for Pashto SeaMonkey. This work has successfully completed and ended in November 2008. The project manager of Afghanistan Country Component of PAN Localization, Mr. Rafiqullah Kakar had the honour and privilege of obtaining the Manthan Award South Asia 2008 in e-localization category. This work has successfully completed and ended in November 2008**.**

Figure 4: Mr. Rafiqullah Kakar is receiving Manthan Award South Asia 2008

Infrastructure development

In building capacity, infrastructural development plays a vital role. To develop the appropriate localization research infrastructure, funds were needed for acquiring academic resources, e.g. books and journals, and specialized software.

In phase I, the team of Afghanistan utilized funds for different components of the project; Operational field (trained resources), acquisition of the equipments and books related to different disciplines like linguistics, language processing and computer science. Equipments included PCs, scanners and printers. In the operational field, participants regarding different domains of the PAN Localization project were trained. In phase II, the Afghanistan country partner institution also utilized funds for purchase of the equipments like scanners, PCs and for arranging trainings. In both phases of PAN Localization project, country component focused on procurement of computer hardware, development of networking and conducting trainings and available funds were mostly used for these activities. The accessibility of these funds helped developing appropriate localization research infrastructure and enhanced research capacity in Afghanistan.

Sustainability and Continuity

Organizational capacity enhancement as a result of team skill building is another salient factor in measuring the capacity enhancement of teams for sustainability of research. Thus organization has focused on enhancing their knowledge base to gain advancement in other domains of local language computing as well. This has been a contributing factor for organization to acquire more projects on localization technology development. Afghanistan Country component focused on the development of standardization and basic localization during PAN Localization project.

Through PAN Localization project a significant number of technical developers, linguists and social scientists have been trained to enable sustainability and continuity of the research being undertaken. Afghanistan country component trained 7 participants from different domains like management, technology and linguistics.

The findings showed that PAN Localization project has been successful to build research capacity on localization in Afghanistan country component and research capacity has been enhanced successfully over the project implementations.

6.2 Bangladesh

In Bangladesh, project collaborated with Center for Research on Bangla Language Processing (CRBLP, http://crblp.bracu.ac.bd/) and Development of research center (D.Net, http://www.dnet.org.bd/bid.htm).

ICT is considered to be an important tool in the development of country. The usage of Personal computer per 100 populations in Bangladesh was 0.02 in 1997, which rose to 0.34 in 2002(Government of Bangladesh & United Nation, 2005).These figures regarding lower usage of computer show that Bangladesh was suffering from the digital divide. The lack of local content is also a barrier to increased use of ICT (Raihan, 2009); therefore, for rapid development in ICT sector of Bangladesh, localized software should be made available. This means that the digital content should be available in local language of the Bangladesh. With the collaboration of Center for Research on Bangla Language Processing (CRBLP), PAN Localization project worked on the development of Localized software. Development of Research Center (D.Net) arranged training for end users to learn the use of the localized software. This project enabled people of Bangladesh to understand the technology, thus the situation gradually improved in the ICT sector of Bangladesh.

The perusal of the work done by Bangladesh country component would show in findings section that PAN Localization project has been enormously helpful to enhance the capacity of Bangladesh country team to develop localized digital content and also boosted them to build research capacity on localization. The following section highlighted the contribution in enhancing capacity by Bangladesh component assessed through RCB model.

Skill development

Bangladesh country team was required to deliver specific localized software. These software involved expertise in linguistics, computer science and computational linguistics. In phase I, Bangla Lexicon, Bangla OCR system and Bangla Spell Checker while In phase II, TTS for Bangla, Bengali SMS to Speech Application, Bengali Wordnet, Bengali Diphone Database, Bengali Speech Database, English-Bengali Parallel and Aligned Tagged Corpus, 5M Word tagged corpus, Bangla gTLD and ccTLD were the requisite localized software to submit and Bangladesh country component successfully submitted requisite localized software. Linguistics, computational linguistics and computer science competence have been involved in TTS for Bangla, Bengali SMS to Speech Application and Bengali Diphone Database. Detailed information regarding eleven required localized software pertained to competence in linguistics, computer science or computation linguistics has also been highlighted in the table below:

Bangladesh				
Localized software	Ling.	CL	CS	Status
Bangla Lexicon				Completed
Bangla OCR system				Completed
Bangla Spell Checker				Completed
TTS for Bangla	*	*	*	Completed
Bengali SMS to Speech Application	*	*	*	Completed

Bengali Wordnet	*	*		Completed
Bengali Diphone Database	*	*	*	Completed
Bengali Speech Database		*	*	Completed
English-Bengali Parallel and Aligned Tagged Corpus	*		*	Completed
5M Word tagged corpus	*			Completed
Bangla gTLD and ccTLD	*		*	Completed

Table 2: Bangladesh team's status regarding Localized software

The team of Bangladesh has been able to deliver all localized software as per the contract. The accomplished localized software in Pan Localization project revealed the fact that that Bangladesh country team's capacity has developed skill enhancement over the project implementation. It is evident that the country project team has also advanced in their over-all local language computing skill set through continual research and development of local language software and its components.

In phase I, Bangladesh country component is focusing on basic, intermediate and advanced complexity local language computing applications and in phase 2, project team is researching on intermediate and advanced complexity software. During the first phase of the project Bangladesh Country Component worked on the development of collation and locale. In addition to that Bangla Optical Character Recognition, Bangla lexicon, spell checking and sorting software were developed during first phase of the project. With the help of research carried out during PAN Localization project, Bangladesh team has also released Language Interface Packs for Microsoft platform. Bangladesh team has successfully released Bangla-English version of PENN Treebank parallel corpus for first 100,000 words. In addition to this work a Bangla corpus of 5 million words has also been gathered covering various domains such as scientific, medical, humanitarian, newspaper articles and samples from novels, stories, textbooks as well as transcribed speech. The work on Part of speech tagset has also been carried out and a tagset of 55 tags has been developed. Using this tagset, 25,000 words have been manually tagged. The Brill tagger has been trained for automatic tagging of the Bangla text. The reported accuracy of this tagger is around 70.6 %. Bangla Parallel corpus has been tagged with this tagset. The Bangla Country Component has worked on language table and translations of gTLDs and ccTLDs in Bamgla for IDNs. The terminology translations of gTLDs and ccTLDs can be accessed online (http://www.panl10n.net/english/OutputsBangla2.htm). Major issues faced during the design of character set, gTLD and ccTLDs have also been reported. In the second phase, the team has also been working on the development of the Bangla WordNet (BWN) based on English WordNet(distribution of Princeton University). The primary focus of BWN has been on design and implementation of a framework which could be used to build and use Bengali WordNet. For the development of the BWN, the bottom up approach has been used which translate the words in the target language. The high frequency 6,000 words from the Prothom Alo corpus have been selected for the BWN development. The synsets have been compiled in lexical source files, which have been then included into the WordNet database using a "grinder", and the resulting system could be used through a set of interfaces. The Bangla WordNet is online available at (http://www.bracuniversity.net/research/crblp/demo/bangla_wordnet/bwnV1.50/). The research paper

titled "BWN - A Software Platform for Developing Bengali WordNet" has been published in International Joint Conferences on Computer, Information, and Systems Sciences, and Engineering (CISSE 08).

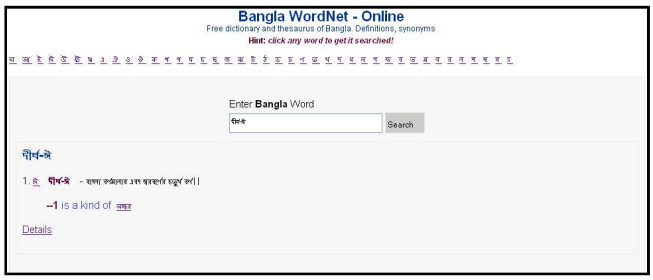


Figure 5: Bangla Word Net

Bangla team has developed two speech corpora; one is read speech corpus and other is diaphone corpus. To develop this speech corpus 106,860 words text corpus has been collected from different domains such as magazine, novels, blog, legal text, a small part of constitution of Bangladesh, history, and different types of news. A professional speaker had been hired for recording. After recording, the sentence level labeling is performed on the cleaned corpus. This corpus has almost 10,000 sentences and 18,000 unique tokens. This speech corpus has been used to develop acoustic models for speech recognition, to analyze the intonation pattern, and to develop a TTS by unit selection technique. In addition to this speech corpus, the diphone database has also developed. It contains 4,355 sentences, which are typically nonsense sentences. These sentences have been formed by combining the nonsense words with 4,355 diphones. Bangladesh component has been working on Text to Speech systems during first phase of project. In phase II a working Text to Speech application has been released as an output of this research. The system is based on Festival; an open source TTS engine. During the development of this system, various language resources have been gathered including speech corpora, pronunciation lexicon and text to sound rules. The system has been thoroughly tested for two aspects (Intelligibility and Naturalness) and three levels of results have been reported. Synthesized speech is 85% accurate for sentence level, 84% for phrases and 57% for words in terms of intelligibility. The degree of naturalness is 90% for sentences, 85 % for phrases and 57% for words. Future directions and improvements have been suggested by researchers in published paper "Text To Speech for Bangla Language using Festival", in Proc. of 1st International Conference on Digital Communications and Computer Applications (DCCA2007), Irbid, Jordan, 2007. System is available online for testing purpose (http://crblp.bracu.ac.bd/demo/tts/). The Bangla TTS is winner of the 2010 innovation award of BASIS.

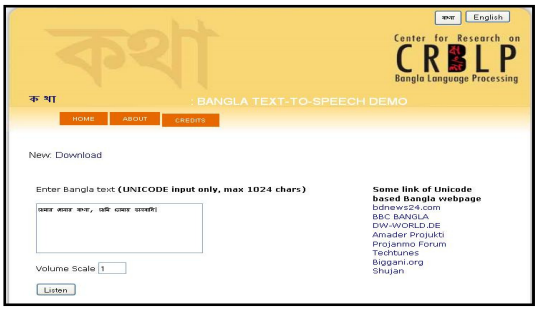


Figure 6: Bangla Text to Speech System

The work has also done on development of the Optical character recognition (OCR) for Bangla. At the start the HMM model has been used for the training and recognition. This system has been matured by adding the some preprocessing and post processing modules. The preprocessing modules have been

used to correct the input image and in the post processing module dictionary and some rules have been used to correct the recognized word. After attaching these modules, the reported accuracy is about 98%. The Bangla OCR has been integrated with open source OCR framework OCRopus. The research work for the Bangla OCR has been in the form of three research papers titled "Integrating Bangla script recognition support in Tesseract OCR", "Rule based segmentation of lower modifiers in complex Bangla scripts" and "Elimination of splitting errors in printed Bangla scripts" and published in Conference on Language and Technology 2009 (CLT09).In the second phase, Bangladesh has officially released Bangla language processing software package including Text-to-Speech and Bangla OCR for Linux, Windows and Mac OS. Bangladesh component is now working on better integration with screen readers in collaboration with the vision impaired community. All of the above mentioned outputs are available at http://www.panl10n.net/english/OutputsBangla2.htm) and (http://crblp.bracu.ac.bd/demos-downloads.php).

To enhance skill of country project team, BRAC University and D.Net received first training on Outcome Mapping framework in November 2007 in Bangladesh. Regional monitoring and evaluation team at Regional Secretariat was in continuous contact with country team in Bangladesh to build their evaluation capacity. The different communication channels such as Skype meetings, email, and telephonic conversation were used for that purpose.

D.Net organized a training titled “Local Language Content Development” from 25-30 March, 2008. The objective of this workshop was to train content developer on Content Management System, Wiki, and blog.

Figure 7: Training on Local Language Content Development

D.Net organized a training titled “Accessing and Intermediation of Local Language Content” from January 23-31, 2008 for Infomediary and trained them on word processing, spreadsheet Fundamentals, and basics of Word Wide Web. In Bangladesh, D.Net (Development Research Network) has been focusing on empowerment of the rural poor through its rural livelihood Information network. In this network, content developers were trained to develop locally relevant content and information workers/ Infomediaries were trained to enable the rural community to access the required information.

D.Net conducted refreshers training for content developer and Infomediary. Training for Infomediary was conducted from 22-23 August, 2008. D.Net also conducted refresher training for content developers from 24-25 August, 2008. 9 participants from different telecentres and government institutes attended this training and shared their field experiences. At the closing session, the project coordinator with all participants decided to make an email group to share their activities in the field.

Figure 8: Infomediary Training on Local Language Content, January 23-31, 2008

These trainings and workshops helped Bangladesh team in enhancing their skill enormously as indicated by their momentous progresses on Localized Content Development.

Ability to publish research in the form of research papers is a salient indicator for measuring the researcher's research capacity. Thus the number of research publications produced by PAN Localization project teams at various national as well as international research conferences is used as the second indicator for analyzing research capacity enhancement. The project team of Bangladesh published 8 research papers covering MT, Script and Speech processing during the project's phase 2 which is the second highest among the all participating countries.

Detailed list of research report publication by project team of Bangladesh is presented in Appendix A.

Training to Conduct Close to Practice Research

Through PAN Localization project's research it was envisioned that localized technology being developed must be deployable and of direct use to the communities.

In order to establish the need for localized application, specific question was asked from the communities regarding the language that they speak at home and at their work. Answers from this question would ascertain their preference of language to undertake everyday communication, both written and verbal. When end-users were asked regarding the language spoken at home and work, 100% respondent indicated that they only use local language for communicate at home as well as at their workplace. This response clearly indicated that the language most convenient for communication for the specific communities was their respective local language. Thus researching for development of local language ICT applications becomes directly useful and relevant to the subject communities, because in order to communicate electronically, and for work, the communities would require applications developed in local languages of the communities.

The respondents were also asked to rate their reading skill and writing skill in English on a scale ranging from Excellent to poor. 11 respondents in total answered this question and 4 of them rated their reading skill in English as excellent and no one of respondents rated their reading skill in English as poor. 3 respondents rated their writing skill in English as excellent and only 1 respondents rated their writing skill in English as poor. Similarly the respondent were also asked to rate their reading skill and writing skill in Local Language on a scale ranging from Excellent to Poor. 11 respondents in total answered this question and majority of them 7 rated their reading skill in local language as excellent and no one of

respondents rated reading skill in local language as poor. A large majority of them 7 rated their writing skill in local language as excellent and no one of respondents rated writing skill in local language as poor.

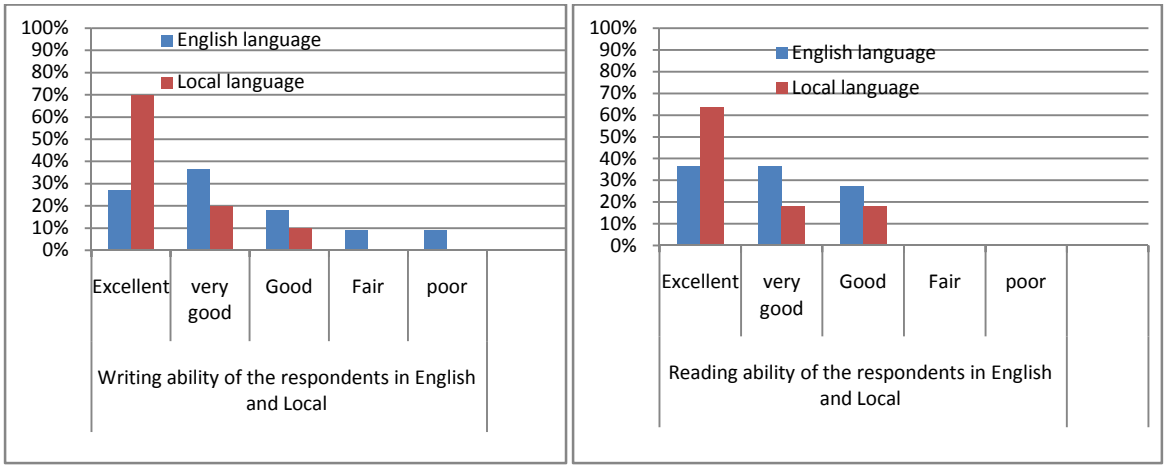


Figure 9: Graphs showing reading and writing ability of the respondents of Bangladesh, in English vs. Local Language

Development of linkage

Bangladesh was encouraged to establish partnerships and collaboration with institutions that had more expertise in a specific field. Through PAN Localization project, at national level, Center for Research on Bangla Language Processing (CRBLP) collaborated with Development of research center (D.Net). This collaboration enabled D.Net to train rural communities of Bangladesh by means of trainings and workshops on localized software, developed by Center for Research on Bangla Language Processing CRBLP, as discussed earlier in Skill Development section.

The project team also has been participating in online research networks, discussion groups, communities and forums for collaboration, knowledge sharing and learning. The work they have performed has given them confidence not only to learn but also contribute on these online forums. The project created an online support network to encourage project partners to be a part of an online learning culture. Bangladesh has been participating on this forum, sharing their project experiences with each other. Nepal and Bangladesh team discussed their challenges in developing spell checker for open source software for Brahmic scripts. The solution based on HunSpell by Nepalese helped the team develop Bangal spell checker in Bangladesh.

Dissemination

Dissemination is an essential part of undertaking research. Research is as credible as much as it is referenced, cited in other publications, brought to people knowledge and properly disseminated.

The main and sustained source of information and outputs of the project has been the project website. The core site has been maintained by the project's regional secretariat and one person from country team of Bangladesh act as a website coordinator and provides local content for the centrally maintained multilingual website www.panl10n.net. In addition CRBLP team at BRAC University Bangladesh also created awareness about its work through its local website http://crblp.bracu.ac.bd/. This has given global access to project outputs.

Figure 10: Homepage of local website of Bangladesh

The project team of Bangladesh followed different strategies to disseminate its work and created awareness of the advantages of local language technology. In training seminar on Content Development Methodologies organized by D.Net on August 26, 2007 key stakeholders from NGO sector were invited and the country team provided an overview of the various activities undertaken as part of the project. Another awareness Seminar on "Computational Linguistics in the Bangla Language: Current Situation and Future Prospects", Organized by The Linguistics Association of Bangladesh, 14 September, 2008 at Linguistic Department, Dhaka University.

The Center for Research on Bangla Language Processing (CRBLP) at BRAC University launched the first official release of the Bangla language processing software packages for Text to Speech (TTS) and Optical Character Recognition (OCR) named “Katha” and BanglaOCR respectively, on February 19, 2009 at 3 at BRAC University. The guests at the launching ceremony included Dr. Salehuddin Ahmed, Pro Vice Chancellor, BRAC University, Monsur Ahmed Choudhury, MD, Jatyo Protibondhi Unnayan Foundation (JPUF), Saiful Islam Khan, Project Officer, Joint project ‘Strengthening Government Integrated Education program for the Blind’ of the Shomaj Sheba Odibhoptor and Sight Savers International. The activities of CRBLP have been widely covered by the newspapers, IT magazines and electronic media such as The Daily Star, The New Age, the Prothom-Alo, the Monthly Computer Jagat and Bangladesh Television (BTV).

Figure 11: TTS Launching Seminar at BRAC University

These seminars have been helpful to create awareness and Bangladesh project team won awards due to their significant work on localization. Center for Research on Bangla Language Processing (CRBLP) research received e-Content and ICT for Development contest 2010 award as a finalist.CRBLP developed software Katha to translate Bangla text to speech and it was appeared as a finalist in the category of e-localization. For the first time e-Content and ICT for Development contest was held in Bangladesh. It was

held in 15 categories both for on-line and off-line content products and ICT interventions and applications.

Ministry of Science and ICT, Government of People Republic of Bangladesh was the host of the contest and D.Net was the organizer of the contest. The Katha research team consists of Firoj Alam, S. M. Murtoza Habib, Kamrul Hayder. Professor Mumit Khan, Head of the Centre for Research on Bangla Language Processing (CRBLP), received the award on behalf the CRBLP research team at the ceremony.

Figure 12: Professor Mumit Khan is receiving award

In addition, Bangladesh Software Industries Association (BASIS) selected Bangla TTS, developed under PAN Phased II, for BASIS IT Innovation Search Award in SOFTEXPO 2010.

BASIS selected three finalists for this award, and showcased the projects in their 2010 expo. Interestingly, two out of three were PAN projects - Bangla OCR and Bangla TTS. Professor Mumit Khan, Head of the Centre for Research on Bangla Language Processing (CRBLP), received the award on behalf the CRBLP research team at the ceremony.

Figure 13; Professor Mumit Khan is receiving award

Participants of Bangladesh project team also attended three conferences. Abdul Hasnat and Firoj Alam from Bangladesh attended Conference on Language and Technology 2009 held in Pakistan from 22nd-24th January, 2009. In addition, 42 participants (Bangladesh, Bhutan, Cambodia, China, Indonesia, Japan, Laos, Mongolia, Nepal, Pakistan, Philippines, South Africa, Sri Lanka and Thailand) attended Regional Conference on Localized ICT Development & Dissemination across Asia in Laos from 11th-16th January, 2009.

Infra structure development

To develop the appropriate localization research infrastructure, funds were needed for acquiring academic resources, e.g. books and journals, and specialized software. In phase I, the team of Bangladesh utilize funds for different components of the project; Operational field (trained resources), acquisition of the equipments, making of software Bangla and books related to different disciplines like linguistics, language processing and computer science. Equipments included PCs, scanners and printers, switch and networking. In the operational field, participants regarding different domains of the PAN Localization project were trained. In phase II, the Bangladesh country partner institution utilized funds for purchase of the equipment like scanners, PCs, Laser Printer and cartridges, DVD writer, USB, Amplifier preamp, Large Speaker, Small Speakers, Headphones, Switch and networking. Funds also

utilized for arranging trainings and acquisition of books and journals. In both phases of PAN Localization project, Bangladesh country component focused on procurement of computer hardware, development of networking and conducting trainings and available funds were mostly used for these activities. The accessibility of these funds helped developing appropriate localization research infrastructure and enhanced research capacity in Bangladesh.

Sustainability and continuity

Organizational capacity enhancement as a result of team skill building is another salient factor in measuring the capacity enhancement of teams for sustainability of research. Thus organization has focused on enhancing their knowledge base to gain advancement in other domains of local language computing as well. This has been a contributing factor for organization to acquire more projects on localization technology development. Bangladesh Country component focused on the development of standardization, basic localization, language processing, script processing and speech processing during PAN Localization project.

Through PAN Localization project a significant number of technical developers, linguists and social scientists have been trained to enable sustainability and continuity of the research being undertaken. Bangladesh country component trained 22 participants from different domains like management, technology, linguistics and social science.

6.3 Bhutan

In case of Bhutan, the project involved partners within Department of Information Technology (DIT, http://www.dit.gov.bt/). Department of Information Technology (DIT) is the lead department working under Ministry of Information and Communication (MoIC) for the development and coordination of all ICT-related activities in the country.

Information Technology (IT) is relatively new and embryonic industry in Bhutan (Jurmi & Wangchuk, 2010). (In Bhutan) the Computer usage per 100 inhabitants was 1.45 in 2002 (Government of Bangladesh & United Nation, 2005). The lower usage of computers shows that Bhutan was suffering from digital divide. To reduce the digital divide, development was needed in ICT sector of Bhutan. To boost ICT development, it was mandatory that digital content should be available in the local language (Dzongkha) so that technology might be understandable for Bhutanese because most spoken language in Bhutan is Dzongkha. Through PAN Localization project, localized software has been developed e.g. Dzongkha Desktop, Dzongkha keyboard and Dzongkha Linux. Jurmi & Wangchuk (2009, p. 157) highlighted that *The Dzongkha Desktop is good alternative for users who cannot read and write in English because it has an interface in Dzongkha as well as in English. Dzongkha Linux, which was funded under the PAN Localization project, was a success story in the research and development (R&D) front*. Bhutan is making considerable progress in implementing numerous ICT activities.

The PAN Localization project has been helpful for Bhutan country team in developing localized software and also helped them in building research capacity in local language computing. The following sections presented information showing capacity development of each project team assessed though Research Capacity Building model.

Skill Development

Project country component was required to deliver specific localized software. These localized software involved expertise in linguistics, computer science and computational linguistics. In phase I, Keyboard in Linux and Linux Distribution with open Office whereas In phase II, Dzongkha gTLDs and ccTLDs, Dzongkha Speech Corpus, Dzongkha TTS, Dzongkha Diphone Database, Dzongkha Corpus, Dzongkha Linux 3.0 and Dzongkha Lexicon were requisite localized software to deliver .Linguistics, Computational Linguistics and computer science competence were involved in Dzongkha TTS where as Computational Linguistics and computer science competence was required in Dzongkha Speech Corpus and Dzongkha Diphone Database . Bhutan submitted successfully all localized software except the "Dzongkha Speech Corpus".

Skill set pertaining to competence in linguistics, computer science or computation linguistics has been highlighted in detail for each of localized software of both phases in the table below:

Bhutan				
Localized software	Ling.	CL	CS	Status
Keyboard in Linux			*	Completed
Linux Distribution with open Office	*		*	Completed
Dzongkha gTLDs and ccTLDs	*			Completed
Dzongkha Speech Corpus		*	*	Not Completed
Dzongkha TTS	*	*	*	Completed
Dzongkha Diphone Database		*	*	Completed
Dzongkha Corpus	*		*	Completed
Dzongkha Linux 3.0			*	Completed
Dzongkha Lexicon		*		Completed

Table 3: Status of Bhutan's team regarding Localized Software

Tables showed that Bhutan has accomplished the larger objective of the required localized software and has been submitted 90% localized software by project country component as per the contract.

A comparison of the accomplished localized software in PAN Localization project's phase 1 and phase 2 reveal the fact that Bhutan country component was researching on development of basic, intermediate and advanced complexity local language computing applications as compared to Phase 1 in which team was only focusing on the development of either basic or intermediate complexity software. Thus it is clearly evident that as the Project team has gained more technical skills over the project implementation, therefore they have advanced development from intermediate to more sophisticated software development. Bhutan country component developed many standardized applications during Phase I of project including keyboard, fonts, collation and locale. All of these utilities had been incorporated in Dzongkha Linux which was released on 2nd June, 2006. It is a Debian based Linux operating system with localization of Gnome and OpenOffice.org suite. As a part of this project DIT released a book on Dzongkha Computer Terms containing 5,000 phrases.

Bhutan team has been working on corpus collection from different domains such as arts, religion, official documents and sports. The collected corpus contains 400,000 words (600,000 syllables). The collected texts have been sourced mainly from dictionaries, printed books, the print and broadcast media, and from relevant websites. A lexicon of approximately 23,000 unique words has been extracted out of this corpus, containing meaning, pronunciation and part of speech tag of each word. The corpus is available online (http://panl10n.net/english/OutputsBhutan2.htm).

DIT has also been working on Part of Speech tagset development. Proposed Dzongkha tagset contains 41 tags which are thoroughly defined with examples. Selection of tags, for this tagset, is carried out by complying PENN Treebank guidelines. The major issue faced during development of automatic tagging system was absence of word segmentation utility. In order to solve this problem 20,000 words have been segmented and tagged manually. After that Tree Tagger is trained with the help of this manually annotated sub corpus. The test results show that automatic tagging can be done with an accuracy of nearly 85 %. Future directions of this work are incorporation of word segmentation module and increase in the size of trained data.

In the second phase, work has been done on Dzongkha IDNs. Language tables and lists of gTLDs and ccTLDs have been released as final output. In addition to development of these lists, a comprehensive testing exercise of IDNs has also been carried out in DIT. A lab has been set up with domain name server, web server and client machines. The test exercise has been done on Local Area Network. Client machines have been provided with Mozilla Firefox plugin for local language domain name translation to punycode. The prototype successfully worked for all modules of IDNs process.

Research on Text to Speech System (TTS) has also been started by DIT in Phase I. Dzongkha phonetic set description and diphone inventory have been developed. The completed milestones after Phase I is text to diphone conversion module. In Phase II Initial work has been carried out on process of speech synthesis. The work is done in collaboration with Human Language Technology (HLT) team at NECTEC. A complete model of TTS has been proposed, using Hidden Morkov Models, as an output of this activity. Unfortunately due to unavailability of skilled technical resources the work faced progress issues. In order to resolve this hurdle and enhance ICT capacity, two team members have been sent to NECTEC, Thailand. Both of them stayed there for two months and received training on complete developmental cycle of TTS. A working prototype of TTS system is available now in which input string is converted into diphones and then speech is synthesized. Improvement of this system is continued in DIT after Phase II. All reports and research material is available online (http://panl10n.net/english/OutputsBhutan2.htm).
Bhutan team initiated work on OCR system development during first year of PhaseII. A team member from Bhutan went to attend regional training of OCR at NECTEC in July, 2007 with support from PAN Localization project. It helped in analyzing Tibetan script and a technique of ligature segmentation was proposed. Research on segmentation process has been continued after that training. Due to lack of image processing and machine learning knowledge, productive outcome for segmentation and recognition modules was not enough to cope with script challenges. Eventually, another training of Bhutan team had been arranged in 2009, in Lahore to provide the appropriate mentoring. The training was helpful in development of prototypical OCR system for Dzongkha Jomolhari font. The work on this project is being carried forward at DIT after completion of Phase II.

In the second phase most of the work done on Dzongkha Linux has been enhancement of the localization of existing Open Source Softwares. The work has been done in terminology translation of Gnome Interface, OpenOffice.org, FireFox, ThunderBird, Debian Installer and CD burning applications, etc. (http://dzongkha.sourceforge.net/) These improvements have been carried out according to the

feedback of users of Dzongkha Linux. Also, the reported bugs have been removed for new release e.g. in previous versions broadband modem was not supported, which has been incorporated in DzongkhaLinux 3.0. The release of live CD is expected towards the mid of 2010.

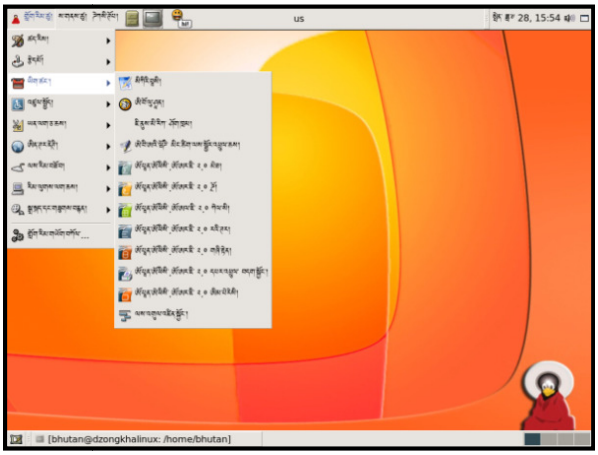


Figure 14: DzongkhaLinux Desktop

Based on the project experience, the country project leaders were asked to prorate the ability and skill development/enhancement of the organization's researchers during project phase 2 on a scale of 1-5; where 1 represent *Challenged*; 2 represent *Fair*; 3 represent *Average*; while 4 represents *Good* and 5 represents *Excellent* enhancement in the team's performance.

The following table presents those comparative figures for assessment of project team's capacity by the project leader from Bhutan country component collected for the team's performance at the beginning of project Phase 2 in 2007 and towards the end of this phase in 2009.

DIT

Skill Development	Start of project, early 2007	Towards project end, Mid 2009
LLC Project development	2	4
LLC Project design	2	4
Problem identification	2	4
Project implementation	2	4
Ability to do analysis	2	4
Ability to communicate results	3	5
Multi disciplinary research	1	3
Quantitative analytical skills	2	5
Qualitative analytical skills	2	4

Table 4: Performance of Bhutan's team regarding Skill Development

The table shown above presents that the country project leader has confirmed the enhancement of team's skills starting from project development and design to its implementation and analysis within the 3 year span of the project specifically in over-all project execution.

The publications of research paper produced by PAN Localization project teams at various national as well as international research conferences is used as the second indicator for analyzing research capacity enhancement. Research team in Bhutan produced 1 research paper covering TTS. Detailed list of research reports publication by each country team is given in Appendix A.

Development of Linkages

Partner teams were encouraged to establish partnerships and collaboration with institutions that had more expertise in a specific field. These collaborations enabled the partners to collectively plan the technical and financial details, exchange data and technology and discuss and formalize shared intellectual property regimes, building institutional capacities in the context. For example, in Bhutan, Department of IT (DIT), the primary partner institute of the PAN L10n project collaborated with NECTEC, Thailand at international level whereas at national level DIT collaborated with Dzongkha Development Authority, their national language development and language standardization authority to develop the technical terminology translations for the software. The advantage of that collaboration was that once the terminology was developed by DDA, it would become a national standard for such terminology translation.

The Pan Localization work paved the way for a Memorandum of Understanding (MoU) between Bhutan's Department of Information Technology (DIT) and Thailand's National Electronics and Computer Technology Center (NECTEC) to promote R&D in the area of ICTs. The MoU included a plan to strengthen Open Source Natural Language Processing, Image Processing Technology and Speech Processing Technology in Bhutan. The project work also resulted in development of a permanent research and development division at DIT. The Bhutan project team also has been participating in online research networks, discussion groups, communities and forums to access to technical support.

Dissemination

Dissemination is an essential part of undertaking research. Research is as credible as much as it is referenced, cited in other publications, brought to people's knowledge and properly disseminated.

The main and sustained source of information and outputs of the project has been the project website. The core site has been maintained by the project's regional secretariat and one person from country team of Bhutan act as a website coordinator and provides local content for the centrally maintained multilingual website www.panl10n.net. In addition the team also hosted their separate websites http://www.dit.gov.bt/ providing detailed information about their respective research groups, hosted by their organizations, which are linked from the main website as well. This has given global access to project outputs.

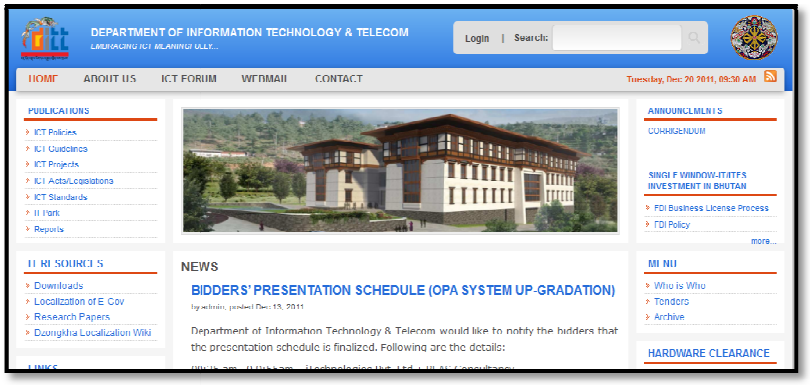


Figure 15: Homepage of Bhutan local website

The team also supported the development of Bhutan Digital Library Project (http://www.e-bhutan.net.bt/ndlb). The mission of the Digital Library project is to preserve and promote the cultural heritage. The website contains large volume of content in Dzongkha. The team has supported the development of a Centenary website (http://www.bhutan2008.bt/**)** which contains the content about 100 years of Monarchy, centenary events and general information about culture, political system and constitution of Bhutan. The Bhutan component has also created a CIC web portal, (http://dzongkha.sourceforge.net) comprising information in various areas including health, education, government services and agriculture etc.

Figure 16: Homepage of **www.bhutan2008.bt**

The project has organized awareness seminars to disseminate and publicize research results to local community. These seminars have been attended by a large number of participants from academia, public and private sectors. Through these seminars partner institutions have been regularly presenting their work to the key stakeholders from government, IT industry, academia, media, and end user communities. Awareness Seminars and end user trainings have been organized by the country component to disseminate their work on localized software.

The launch ceremony of DzongkhaLinux was organized on June 2nd, 2006 and the launch ceremony of this Linux distribution received vast media coverage. In addition, the training ON DzongkhaLinux, was organized in Department of Information Technology, Ministry of Information and Communications, Thimphu, Bhutan from November 19th to November 27th, 2007. The objective of the training was to train the end user trainers and it was expected from trainers that they go back to their own organizations (Private IT institutes and Government offices) and train the other users.

Another, Dzongkha Linux End User training was conducted from December 3-16, 2008 at Dzongkha Language Institute (DLI), Thimphu and Computer Management Institute (CMI), Phuentsholing. Both institutes were authorized by the Dzongkha Development Commission (DDC) and Department of Information Technology (DIT) for conducting trainings. About 40 participants from various agencies including Royal Bhutan Army and Royal Bhutan Police. The objective of this training was to provide basic knowledge about installation and operation of Dzongkha Linux.

Figure 17: End user training (December 3-16, 2008)

Participants were also trained on open office applications, Firefox and Thunderbird and other applications. Participants were expected to further transfer their knowledge and skills to their colleagues and create awareness about Dzongkha Linux and open source software in general. Training manuals on Dzongkha Linux Installation, Open Office and other applications were provided to the participants. Trainees were also provided lessons materials in the form of PowerPoint Presentations. More detail can be viewed at http://www.panl10n.net/english/Outputs Phase 2/CCs/Bhutan/Papers/2008/0802/Training Conduction Report.pdf

A two day's training session on Dzongkha Linux was conducted in Lhuntse DYT hall on 28th May, 2009 for the teachers, monks and staffs Dzongkhag Administration. The training was funded by Dzongkha Development Commission (DDC). DIT planned to organize an advance level Dzongkha Linux training from 11th to 25 January 2010. The training detail can be viewed at http://www.moic.gov.bt/pdf/progress20Report for December 2009.pdf. 500 CDs of Dzongkha Linux 3.0 and key boards were also distributed by the country team. These trainings and distribution of CDs have been helpful to disseminate Bhutan country team's work on localized software.

Infrastructure Development

The team of Bhutan has been capacity to develop the appropriate localization research infrastructure by providing funds for acquiring academic resources, e.g. books and journals, and specialized software. In building capacity, infrastructural development plays a vital role. In phase I, the team of Bhutan utilize funds for different components of the project; Operational field (trained resources), acquisition of the equipments and books related to different disciplines like linguistics, language processing, computer science and TALIP. Equipments included PCs, scanners, printers and servers. In the operational field, participants regarding different domains of the PAN Localization project were trained. In phase II, the Bhutan country partner institution also utilized funds for purchase of the equipment like scanners, PCs, UPS, Amplifier, Mics, speakers, CDs and for arranging trainings. In both phases of PAN Localization project, country component focused on procurement of computer hardware, software and conducting trainings and available funds were mostly used for these activities. The accessibility of these funds helped

developing appropriate localization research infrastructure and enhanced research capacity in Afghanistan.

Sustainability and Continuity

Organizational capacity enhancement as a result of team skill building is another salient factor in measuring the capacity enhancement of teams for sustainability of research. Thus organization has focused on enhancing their knowledge base to gain advancement in other domains of local language computing as well. This has been a contributing factor for organization to acquire more projects on localization technology development. Bhutan Country component focused on the development of standardization, basic localization and script processing during PAN Localization project.

Through the PAN Localization project a significant number of technical developers, linguists and social scientists have been trained. This has further strengthened the organizational capability in maturing their research capacity technically, and able to commit and acquire more projects. The Bhutan country team trained 13 trainers which is the second lowest among the all participant countries.

6.4 Cambodia

In Cambodia, the project involved partners within National ICT Development Authority of Cambodia (NiDA, http://www.nida.gov.kh/) and Ministry of Education, Youth and Sports (MoEYS, http://www.moeys.gov.kh/) during phase I. further, in Cambodia, the PAN Localization project also collaborated with Institute of Technology in phase II.

According to view of Cambodia website, "Khmer is the official language of Cambodia where 90 percent of the populations, about 6 million people speak it as the first or second language". It depicts that the most spoken language in the country is Khmer rather than English and Sorasak & Konsona (2009, p. 173) highlighted that "Cambodia does not have English language skill". English language competence issue leads to underdevelopment of ICT sector in the country because all available software was in English language. That's why for the development in ICT sector, it was necessary to develop localized software so technology should be understandable for the local people of Cambodia and they feel convenient to use technology. Through PAN Localization project, MoEYS and NiDA effectively contributed towards local language computing policy in the country. "In 2005, the Ministry of Education, Youth, and Sport (MoEYS) started to implement the policy and strategies on Information and Communication Technology in Cambodia" (Sorasak & Konsona ,2009, P.167). Ministry of Interior (MoI) and National Assembly (NA) adopted the Khmer Unicode and the applications developed by the project teams. NiDA also regularly invited government officials in its end users trainings to create awareness among policy makers of the benefits of localized technology. All these activities show that Government of Bhutan highly interested regarding development in ICT sector of Bhutan. "Over the last 10 years, the government has been proactive in the development of ICT" (Sorasak & Konsona, 2009, 167).

The PAN Localization project has been contributing to develop digital content in local language of the country and also helped the Cambodia country team to build research capacity in local language computing. The following sections presented information showing capacity development of each project team assessed though Research Capacity Building model.

Skill Development

Project country component was required to deliver specific localized software. These localized software involved expertise in linguistics, computer science and computational linguistics. In phase I, Encoding conversion utility, Sorting utility for Khmer, Terminology Translation for Khmer Computer Interface, Khmer Lexicon and Khmer spell check whereas in Phase II, Translation of gTLDs and ccTLDs in Khmer, English-Khmer Parallel and Aligned Tagged Corpus 100k words, SMS in Khmer Application , Khmer TTS, Khmer Diphone Database and Khmer Speech Corpus are requisite localized software to deliver. Cambodia country component successfully submitted all the localized sofware. Linguistics, Computational Linguistics and computer science competence were involved in English-Khmer Parallel and Aligned Tagged Corpus 100k words whereas Translation of gTLDs and ccTLDs in Khmer, SMS in Khmer Application, Khmer TTS, Khmer Diphone Database and Khmer Speech Corpus involved linguistics and computer science competence. Skill set pertaining to competence in linguistics, computer science or computation linguistics has also been highlighted for each of localized software of both phases in the table below:

Cambodia				
Localized software	Ling.	CL	CS	Status
Encoding conversion utility			*	Completed
Sorting utility for Khmer	*		*	Completed
Terminology Translation for Khmer Computer Interface	*		*	Completed
Khmer Lexicon		*		Completed
Khmer spell check		*	*	Completed
Translation of gTLDs and ccTLDs in Khmer	*		*	Completed
English-Khmer Parallel and Aligned Tagged Corpus 100k words	*		*	Completed
SMS in Khmer Application	*		*	Completed
Khmer TTS	*	*	*	Completed
Khmer Diphone Database	*		*	Completed
Khmer Speech Corpus	*		*	Completed

Table 5: Status of Cambodia's team regarding Localized Software

The above figure showed that Cambodia country component has been able to deliver all localized software as per the contract.

In the following table, a comparison of the accomplished localized software in PAN Localization project's phase 1 and phase 2 reveal the fact that the project country component was researching on development of intermediate complexity and advanced complexity local language computing

applications as compared to Phase 1 in which they were only focusing on the development of either basic or intermediate complexity software. In first phase the MoEYS component of Cambodia developed various Unicode based language processing applications including fonts and standardized keyboard. Some text processing applications like encoding conversion utility, word segmentation, sorting utility, find & replace and spell checking have also been released. These utilities have been developed for windows platform only. In phase II MoEYS focused on conducting research of advance NLP areas like Optical Character Recognition (OCR), Text to Speech, Internationalized Domain Names (IDN) and Part of Speech Tagging. They have also reviewed HTML standards for English and Khmer.

The initial work on the IDNs started in Phase 2. This work contributes the research reports for defining the character set and encoding constraints for IDNs in Khmer. Comprehensive lists of gTLDs and ccTLDs have been released as output of IDN project. Part of speech tagset design and documentation has been carried out by PLC team. It has also gathered a corpus of 150,000 words tagged with their parts of speech. This corpus is annotated with the help of probabilistic part of speech tagger (TnT). A trained model of TnT tagger is released along with corpus.

The project of Khmer-English parallel corpus was planned but unavailability of appropriate translators was a hurdle for its completion. A number of interviews were carried out, by country project leader, to hire translators but the quality of translations was not good enough to kick off this project. Another alternative was to hire professional companies to do the work but there were two major issues in this approach. First, there was not any capacity building incentive involved and secondly the cost of that outsourcing was going way beyond the budget. Eventually this project was closed and budget was reallocated for training activities with consent of regional project leader. During second year of phase II PLC has been working on Khmer OCR. Due to resource retention issues the progress of project was quite slow. In order to resolve this issue a mentor placement program for Cambodia was arranged. Based on the guidance of the mentor, functional OCR of Limon R1 & S1 fonts have been released (http://panl10n.net/english/OutputsCambodia2.htm). The reported accuracy of recognition process is nearly 98% for ligature level. The incorporated modules in this system are Noise removal, Preprocessing, Recognition and Unicode mapping. Following screen shot elaborated the output of Khmer OCR system.

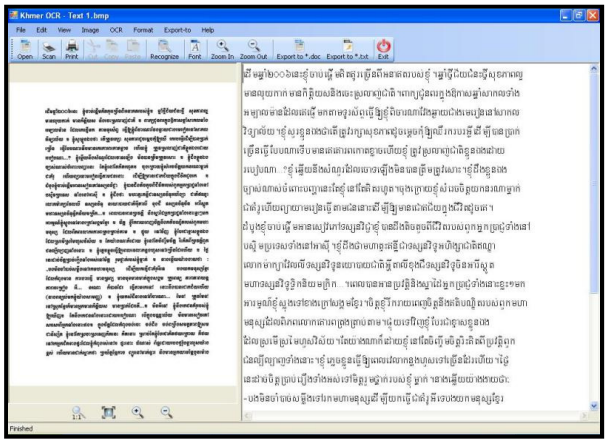


Figure 18: Khmer OCR

Text processing utilities, developed in phase I, were proved to be very useful for end users. It was decided to make them available across the platforms i.e. Windows and Open source OS. Hence, very extensive work has been done in order to release those utilities with multiple interfaces. All of these utilities were ported to Java platform and their Linux and Windows versions were release. In addition to

that, these applications have been incorporated in OpenOffice.org writer application. After covering the diversity of end users, windows based APIs are also developed and released to facilitate programmers working in Khmer Language Processing area. In addition to that Khmer Collation support for MySQL has also been provided.

PLC team also worked on development of Khmer mobile SMS application. Research on mobile font development, keyboard design and J2ME framework has been carried out. This activity was planned for one year duration and a complete SMS application in Khmer language was released along with mobile font. This activity was helpful in exploring the domains of font rendering on mobile platform. Another bright aspect was availability of Khmer font for other mobile applications. Some small but effective sub projects have also been carried out at PLC including Khmer Typing Game, One Click Installation Package and active PAN Cambodia website (http://www.pancambodia.info/index.php).

The MoEYS component has a second team of ITC as well which carried out very extensive research in speech processing area. A complete Text to Speech system has been released as an outcome of phase II. The successful development of this system has been possible because of research on Phonetic and Phonological Analysis of Khmer, NLP engine and Letter to Sound Conversion. The recordings for this system were carried out by professional Khmer speaker in isolated studio environment. This activity was good initiative for development of speech processing area in Khmer. It helped not only in gathering language resources in the area of speech but also provided an opportunity to establish a speech lab in ITC. A long term output is formation of research environment and origination of ongoing NLP research activities. Highlighted issue faced in this project was unavailability of language resources e.g. Diphone database and programming skills for language processing.

In Cambodia, National ICT Development Authority of Cambodia (NiDA) conducted training on localized technology. The training program focused on a diverse group of trainees including students, teachers, farmers, journalists, ministry officials and other staff from government organizations. At first stage, NiDA trained its partners. At second stage, NiDA conducted the training and its trained partner also assisted the NiDA team. At third stage, the trained partner conducted the training with the support from NiDA team.

In all those trainings NiDA also provided training material developed in Khmer language. NiDA also conducted pre and post training surveys to determine the competency level of the participants. The NiDA team also observed growing demand of ICT training. The Help Desk facility is expected to play an important role in sustainability of the training program.

Figure 19: Khmer FOSS Training in Kampong Speu (60 Trainees) December 2008

Training on Localization and Khmer Language Processing has been conducted in Cambodia from 20th June 2004 till 27th December 2004.

The team received training on basic to advanced programming and basic to advanced language processing techniques. The discussion during training covered the topics regarding Open Type font development, Unicode language Processing ,Lexicon Development, Project life cycle from design to execution and testing, Advance programming in C++ and Visual Basic.NET.

Figure 20: Training on localization and Khmer Language Processing 2004

In addition, the training was arranged in Institute of Technology of Cambodia from May 16th, 2007 to June 16th, 2007, to introduce the fundamental concepts of Khmer Language Processing and the Usage of Khmer Unicode in the Web Site Development.

Training for Optical Character Recognition and Khmer Language Processing was held in PAN Cambodia Office from 1st July 2008 till 10th January, 2009. The objective of the training was to introduce Cambodia team with basic local language computing technologies. The training covered topics on Khmer OCR, Open office plug-ins of Khmer applications, Khmer collation support for MySQL, Automatic POS tagger for Khmer, Khmer Lexicon development, Khmer Tagged Corpus, Khmer SMS software for Java based mobile phones. These trainings have been helpful for skill enhancement of Cambodia project team. Based on the project experience, the Cambodia country component was asked to prorate the ability and skill development/enhancement of the organization's researchers during project phase 2 on a scale of 1-5; where 1 represent *Challenged*; 2 represent *Fair*; 3 represent *Average*; while 4 represents *Good* and 5 represents *Excellent* enhancement in the team's performance.

The following table presents those figures for assessment of its team's capacity by the project leader from country component collected for the team's performance at the beginning of project Phase 2 in 2007 and towards the end of this phase in 2009.

MoEYS

Skill Development	Start of Project, Early 2007	Towards project End, Mid 2009
LLC Project development	1	4
LLC Project design	1	4
Problem identification	1	4
Project implementation	1	4
Ability to do analysis	1	4
Ability to communicate results	1	4
Multi disciplinary research	1	4
Quantitative analytical skills	1	4
Qualitative analytical skills	1	4

Table 6: Performance of Cambodia's team regarding Skill Development

The table shown above presents that the project leader has confirmed the enhancement of team's skills starting from project development and design to its implementation and analysis within the 3 year span of the project specifically in over-all project execution.

Training to Conduct Close to Practice Research

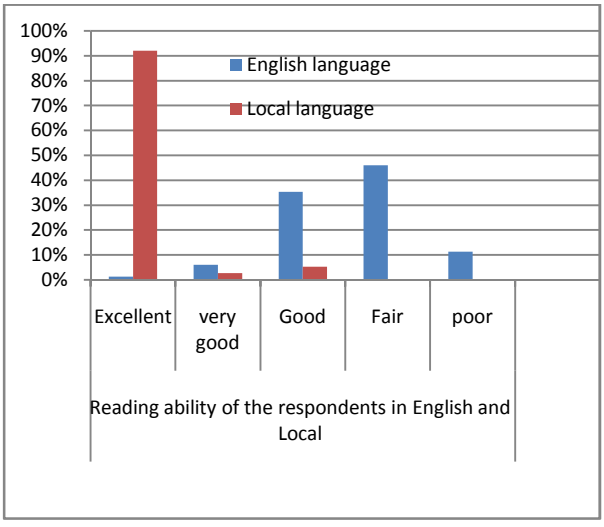
Through PAN Localization project's research it was envisioned that localized technology being developed must be deployable and of direct use to the communities.

In order to establish the need for localized application, specific question was asked from the communities regarding the language that they speak at home and at their work. Answers from this question would ascertain their preference of language to undertake everyday communication, both written and verbal. When end-users were asked regarding the language spoken at home and work, 100% respondent indicated that they only use local language for communicate at home as well as at their workplace.

This response clearly indicated that the language most convenient for communication for the specific communities was their respective local language. Thus researching for development of local language ICT applications becomes directly useful and relevant to the subject communities, because in order to communicate electronically, and for work, the communities would require applications developed in local languages of the communities.

The respondents were also asked to rate their reading skill and writing skill in English on a scale ranging from Excellent to poor. 150 respondents in total answered this question and only 2 rated their reading skill in English as excellent and 17 respondents rated their reading skill in English as poor. Only 3 of them rated their writing skill in English as excellent and 27 of them respondents rated their writing skill in English as poor.

Similarly the respondent were also asked to rate their reading skill and writing skill in Local Language on a scale ranging from Excellent to Poor. 150 respondents in total answered this question and majority of them (138) rated their reading skill in local language as excellent and no one of respondents rated reading skill in local language as poor. A large majority of them (137) rated their writing skill in local language as excellent and only 1 respondent rated writing skill in local language as poor.



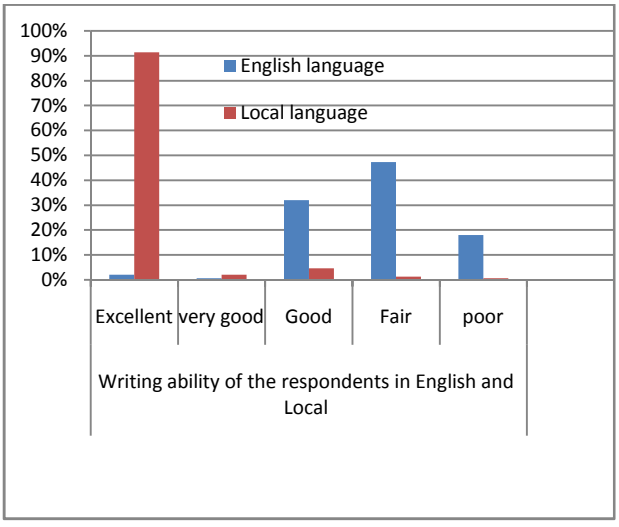


Figure 21: Graphs showing reading and writing ability of the respondents of Cambodia, in English vs. Local language

Training on localized technology was conducted by National ICT Development Authority of Cambodia (NiDA) conducted. The training program focused on a diverse group of trainees including students, teachers, farmers, journalists, ministry officials and other staff from government organizations. Rural communities were trained on SuSE Linux, Open Office, use of Internet and email using NiDA Khmer Standard Unicode Keyboard. The training was conducted in eight different provinces of the country.

Development of Linkages

Development of appropriate linkages, partnerships and collaborations is being considered another significant matter in capacity development. Partner teams were encouraged to establish partnerships and collaboration with institutions that had more expertise in a specific field. These collaborations enabled the partners to collectively plan the technical and financial details, exchange data and technology and discuss and formalize shared intellectual property regimes, building institutional capacities in the context. Through PAN Localization project, Cambodia country component collaborated with National Institute of Language and Institute of Technology, Cambodia (ITC) at national level. Institute of Technology, Cambodia (ITC) that had professors working on localization research and students are taking up localization research projects in their BS final year projects.

The Cambodia project team also has been participating in online research networks, discussion groups, communities and forums to access the technical support and to enhance the quality of their work regarding localization.

Dissemination

Dissemination is an essential part of undertaking research. Research is as credible as much as it is referenced, cited in other publications, brought to people knowledge and properly disseminated.

The main and sustained source of information and outputs of the project has been the project website. The core site has been maintained by the project's regional secretariat and one person from country team of Cambodia act as a website coordinator and provides local content for the centrally maintained multilingual website www.panl10n.net. Ministry of Education, Youth and Sports (www.pancambodia.info), Cambodia team also disseminated its work through its website. This has given global access to project outputs.

Figure 22: home page of local website of Cambodia

In Cambodia, The project followed different strategies to market and disseminate the research. Project partners created awareness of the advantages of Khmer Unicode and Khmer based applications among the computers users and developers. Cambodia country component of PAN Localization project conducted a seminar on 23rd December, 2005 in Cambodia, to introduce the work done under the project. The seminar was attended by a large number of participants from teacher and student communities. CD containing the research outputs (Khmer Smart Typing, Encoding Conversion Utilities, Collation and Sorting Utilities, Word Wrapping Utilities, Spell Checker Utilities) was also freely distributed. The seminar was also followed by technical session.

NiDA also invited government officials in its end users trainings to promote awareness of localized technology among stakeholders and general public. In his visits to Cambodia, Project Leader also created awareness among stakeholders and sensitize policy makers about the value of research being done under the project.Mr. Chea Sok Huor, Project Lead PAN Localization Cambodia Component, was interviewed by Natural Magazine. The magazine has nominated him as a "Legendary Man" who has made constructive efforts for Cambodian people.

Infrastructure Development

The team of Cambodia has been capacity to develop the appropriate localization research infrastructure by providing funds for acquiring academic resources, e.g. books and journals, and specialized software. In building capacity, infrastructural development plays a vital role. In phase I, the team of Cambodia utilize funds for different components of the project; Operational field (trained resources), acquisition of the equipments and books related to different disciplines like linguistics, language processing and computer science. Equipments included PCs, printers, servers, notebook and networking. In the operational field, participants regarding different domains of the PAN Localization project were trained. In phase II, the Cambodia country partner institution utilized funds for purchase of the equipment like PCs, Printer, Amplifier, Speaker, Mica, Speakers, Headphones and networking. Funds also utilized for arranging trainings and acquisition of books and journals. In both phases of PAN Localization project, Cambodia country component focused on procurement of computer hardware, development of networking and conducting trainings and available funds were mostly used for these activities. The accessibility of these funds helped developing appropriate localization research infrastructure and enhanced research capacity in Cambodia.

Sustainability and continuity

Organizational capacity enhancement as a result of team skill building is another salient factor in measuring the capacity enhancement of teams for sustainability of research. Thus organization has focused on enhancing their knowledge base to gain advancement in other domains of local language computing as well. This has been a contributing factor for organization to acquire more projects on localization technology development. Cambodia Country component focused on the development of standardization, basic localization, language processing, script processing and speech processing during PAN Localization project.

Through the PAN Localization project a significant number of technical developers, linguists and social scientists have been trained. This has further strengthened the organizational capability in maturing their research capacity technically, and able to commit and acquire more projects. The project country team trained 62 trainers which is the highest among the all participant countries.

6.5 China

In China, PAN Localization project collaborated with Institute of Science and Technology, Tibet Academy of Agricultural and Animal Husbandry Sciences and Tibet University.

The progress of China component of PAN Localization project was not very satisfactory. The English language competency issue caused communication problems from the beginning of the project, and the project staff did not respond to regular mails and reminders. The China component could not be able to send the deliverables from start of the project and situation did not improve. Mr. QunNuo from China attended the regional conference of project and he assured Project leader to follow-up, however there was not much progress. Eventually the country component is closed.

6.6 Indonesia

In Indonesia, PAN Localization project collaborated with Agency for Assessment and Application of Technology (BPPT, http://www.bppt.go.id/) and University of Indonesia (http://www.ui.ac.id/)**.** The PAN Localization project in Indonesia was initiated in April 2008 during phase II.

In Indonesia, the usage of Computers per 100 inhabitants is 2.8 in 2007 (UN-APCICT, 2007) which is quite low. This is understandable because"Economy of Indonesia was badly hit by the Asian crisis in late 1990s" (Chaunka Wattegama, 2011, P. 09). This was time when ICT was being introduced to the world, Indonesia was suffering from economic crisis and it might be reason for this underdevelopment in ICT sector. Development in ICT sector is a must for improvement in all sectors of the country. "In a presidential lecture held on 9th May, 2008, with Microsoft Chairman Bill Gates in the audience, Indonesian President Susilo Bambang Yudhoyono stated that ICT can help resolve many of the country's problems, such as poverty, natural disaster and mismanagement" (Donny & Mudiardjo, 2009, P.208). For some notable development in ICT sector of the country, it was necessary to make software in local language, so that Indonesian could conveniently use the technology and fast development could be possible.

Through PAN Localization project, Indonesia country component made localized software. The results given below show that significant progress was ongoing in ICT sector of the country during currency of the project and in future more progress is expected. PAN Localization project also boosted Indonesia project team to build research capacity in local language computing. The following sections presented information showing capacity development of each project team assessed though Research Capacity Building model.

Skill development

Project country component was required to deliver specific localized software. These localized software involved expertise in linguistics, computer science and computational linguistics. POS Tagger for Indonesia, 1 Million POS Tagged Corpus, and Statistical Machine Translation System were the requisite localized software to deliver and Indonesia successfully submitted all localized software. Linguistics, computer science and computation linguistics competence has been involved in Statistical Machine Translation System. Skill set pertaining to competence in linguistics, computer science or computation linguistics has also been highlighted for each of localized software in the table below:

Indonesia				
Localized software	Ling.	CL	CS	Status

POS Tagger for Indonesia	*		*	Completed
1 Million POS Tagged Corpus	*		*	Completed
Statistical Machine Translation System	*	*	*	Completed

Table 7: Status of Indonesia team's regarding Localized Software

The above status showed that the project country component has been able to deliver all localized software as per the contract.

The Indonesia country component was researching on development of advanced complexity local language computing application during phase II. The project aimed to develop linguistic resources and English to Bahasa Indonesia Machine Translation System. There were two corpora developed during the project. Initially BPPT collected 500,000 words of Bahasa Indonesia in order to build a corpus. This collection has been translated sentence by sentence into English. Along with this activity, UI team translated first 500,000 words of PENN Treebank into Bahasa Indonesia in collaboration with the Faculty of Arts, University of Indonesia. Finally, by combining these two corpora a parallel corpus of one million words has been made which is annotated with part of speech.

The UI team also worked on development of Part of Speech Tagset for Bahasa Indonesia. They have developed a tagset containing 37 Part of Speech tags. A statistical POS tagger has been developed using Maximum Entropy. The tagger is trained on approximately 150,000 manually tagged words and used for automatic tagging of one million words corpus. This Part of Speech tagger is ready to use and is available online (http://panl10n.net/english/OutputsIndonesia2.htm). The research paper on Indonesian POS tagger was presented by Dr. Mirna Adriani of Indonesia component, at the Third International MALINDO Workshop, ACL IJCNLP in Singapore in 2009.

Statistical Machine Translation (SMT) is the technique based on sentence-level aligned parallel corpora. The team has used open source SMT tools including Giza and Moses to achieve translations. The system has been trained on one million words corpus. Bleu Scores are computed for evaluation of MT system and achieved accuracy of 0.938 for ENG-IND translation and 0.926 for IND-ENG. Details of training and testing are given in the report released with outputs (http://panl10n.net/english/OutputsIndonesia2.htm). The online system is available for general public use (http://translator.iptek.net.id/PANL/). The screen in Figure 2 below shows the input sentence from headlines on BBC Indonesia on 30th Jan. 2010 and its corresponding English translation through this online translation tool (which translates in both directions).

Figure 23: Online SMT for Bahasa Indonesia

The 100 % accomplishment of the project deliverables showed that Indonesia country team's skill has enormously enhanced over the project implementation.

Thus the publications of research papers produced by PAN Localization project teams at various national as well as international research conferences is used as the second indicator for analyzing research capacity enhancement. The project team of Indonesia published 2 research papers covering MT, SLP, POS during the project's phase 2 which is the second highest among the all participating countries. Detailed list of research report publication by project team of Indonesia is presented in Appendix A.

Dissemination

Dissemination is an essential part of undertaking research. Research is as credible as much as it is referenced, cited in other publications, brought to people knowledge and properly disseminated.

The main and sustained source of information and outputs of the project has been the project website. The core site has been maintained by the project's regional secretariat and one person from country team of Indonesia act as a website coordinator and provides local content for the centrally maintained multilingual website www.panl10n.net. In addition the team also hosted their separate websites http://www.bppt.go.id/ providing detailed information about their respective research groups, hosted by their organizations, which are linked from the main website as well. This has given global access to project outputs.

Figure 24: home page of local website of BPPT

To disseminate research outputs, Dr. Hammam Riza, CPI leader of BPPT has also presented the paper Building Parallel Text Corpora for Multi-Domain Translation System at 7th Workshop on Asian Language Resource, ACL-IJCNLP Singapore in August 2009. The local website and such workshops helped Indonesia to disseminate their work on Localization.

Infrastructure Development

The team of Indonesia has been capacity to develop the appropriate localization research infrastructure by providing funds for acquiring academic resources, e.g. books and journals, and specialized software. In building capacity, infrastructural development plays a vital role.

The team of Indonesia utilize funds for different components of the project; acquisition of the equipments and books related to different disciplines like linguistics, language processing and computer science. Equipments included computer hardware like PCs, scanners, printers and server. Indonesia country component also focused on development of networking and available funds were mostly used for these activities. The accessibility of these funds helped developing appropriate localization research infrastructure and enhanced research capacity in Indonesia.

Sustainability and Continuity

Organizational capacity enhancement as a result of team skill building is another salient factor in measuring the capacity enhancement of teams for sustainability of research. Thus organization has focused on enhancing their knowledge base to gain advancement in other domains of local language computing as well. This has been a contributing factor for organization to acquire more projects on localization technology development. Indonesia Country component focused on the development of basic localization and language processing during PAN Localization project.

Through PAN Localization project a significant number of technical developers, linguists and social scientists have been trained to enable sustainability and continuity of the research being undertaken. Indonesia country component trained 18 participants from different domains like management, technology and linguistics.

6.7 Laos

In Laos, the project involved partners within National Authority for Science and Technology (NAST, (http://www.laol10n.info.la/), policy making body in the country. So the work done under the project contributed had direct influence on local language computing policy in the country.

In Laos, the majority (73 percent) of the country's 5.65 million people lives in the rural areas and the usage of computer per 100 habitants is just 1.7 in 2007(Phissamay, 2009). The low ICT diffusion in the respective country might be due to the English language competence issue because only a few people in Laos understand English and all available software was in English language. Although, the Government of Laos has been encouraging the officials and students to learn English, yet, the official and dominant language in Laos is Lao (Bureau of East Asian and Pacific Affairs, 2011). "Government of Laos aims to bring the country into information age by increasing general access to ICT" (Phissamay, 2010, P.243. Firstly, in order to facilitate access of common Lao people to ICT, it was needed that software should be developed in local language so that majority of the people in Laos could easily use the latest technology.

Through PAN Localization project, Laos country team made localized software and paved the way for significant and fast development in Information and communication technology sector of the country.

The PAN Localized project encouraged the country team to build research capacity in local language computing as well. The following sections presented information showing capacity development of each project team assessed though Research Capacity Building model.

Skill development

Project country component was required to deliver specific localized software. These localized software involved expertise in linguistics, computer science and computational linguistics. In Phase I, Sorting Utility for Lao, Spell Checker, Lao Lexicon and Grammar Checker whereas in Phase II, Lao Keypad Standard, Lao Natural Language Processor for TTS, List of gTLDs and ccTLDs in Lao, English-Lao Parallel and Aligned Tagged Corpus 100k words, Lao Open Office with Lao Line Breaking and Collation, Phonetic module for Lao TTS, Lao speech corpus, Lao TTS, Lao diphone database and 100 mb Lao content development were the requisite localized software to deliver. Linguistics, computer science and computational linguistics competence have been involved in Lao Natural Language Processor for TTS, Phonetic module for Lao TTS and Lao TTS. Skill set pertaining to competence in linguistics, computer science or computation linguistics has also been highlighted for each of localized software of both phases in the table below:

Laos				
Localized software	Ling.	CL	CS	Status
Sorting Utility for Lao	*		*	Completed
Spell Checker		*	*	Completed
Lao Lexicon		*	*	Completed
Grammar Checker		*	*	Not completed
Lao Keypad Standard	*		*	Not Completed
Lao Natural Language Processor for TTS	*	*	*	Not Completed
List of gTLDs and ccTLDs in Lao	*		*	Completed
English-Lao Parallel and Aligned Tagged Corpus 100k words	*		*	Completed
Lao Open Office with Lao Line Breaking and Collation	*		*	Completed
Phonetic module for Lao TTS	*	*	*	Not Completed
Lao speech corpus	*		*	Not Completed
Lao TTS	*	*	*	Not Completed
Lao diphone database	*		*	Not Completed

100 mb Lao content development			*	Not Completed

Table 8: Status of Lao's team regarding Localized Software

Laos however has not been able to submit the required localized software due to the unavailability of trainers and the country project team has been able to submit only 40 % software as per the contract. Laos is still continuing the development and training is being organized at the regional secretariat to train the team members from Laos and capacity builds them to submit their incomplete localized software.

A comparison of the submitted localized software in PAN Localization project's phase 1 and phase 2 reveal the fact that project country component was researching on development of intermediate complexity and advanced complexity local language computing applications as compared to Phase 1 in which they were only focusing on the development of either basic or intermediate complexity software. Thus it is clearly evident that as the team has gained more technical skills over the project implementation, therefore team has advanced development from intermediate to more sophisticated software development.

Laos Country Component worked on proposals for standards for Lao fonts and Keyboard during the first phase of the project. Various utilities such as open type fonts, encoding conversion, collation sequence and word segmentation were developed with the guidance of regional mentor (http://panl10n.net/english/activity_11.htm). In Phase II they have been working on advanced applications of Language Processing.

Laos Country Component has worked on the development of the Lao-English Parallel corpus of PENN Treebank. The team has successfully released approximately 50,000 words corpus. The research has also been done for the Part of Speech of Lao language. The tagset of 41 tags has been developed. The tagging of corpus could not be carried out because of lack of linguistic knowledge. In addition, the lexicon of around 67,000 words has been developed from different dictionary books and non electronic sources.

The Laos team has worked on language table and terminology translation of gTLDs and ccTLDs in Lao for IDNs. The Lao translated gTLDs and ccTLDs can be accessed online (http://www.panl10n.net/english/OutputsLaos2.htm).

Lao OCR has been initially developed in the first phase with the support of Sri Lanka team. Later on, it has been enhanced by adapting the algorithm developed by the NECTEC for Thai OCR. The neural network has been used for the pattern matching. This newer version of Lao OCR has provided the facility of input image of multiple formats and it generates the respective Unicode of the characters. This OCR could recognize printed Lao characters of ten mostly used fonts namely Alice0Lao, Alice1_2000, Alice2_2000, Chanthabuli Lao, Chanthabuli 95, aysettha Lao, Saysettha 2000, Saysettha 95 and FontLao1. The reported accuracy of the system is around 98.7%.

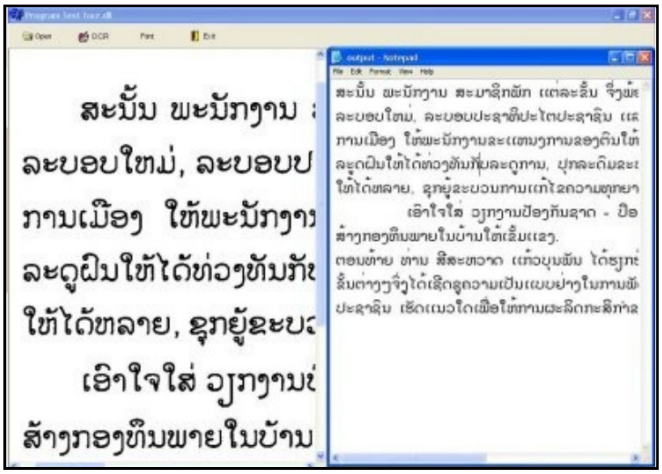


Figure 25: Lao OCR

Lao team has also started the work for the TTS. Work on phoneme analysis of Lao language, which is a precursor to the development of TTS application has been completed. Though much progress has been made in terms of application development, it has been largely achieved through the mentor placement program for porting applications and with the support of NECTEC for OCR and TTS. Team capacity still needs to be built significantly for the sustainability of this work. However, the country level agreement between Lao team and the Thai team at NECTEC will certainly contribute to this end after completion of PAN L10n project.

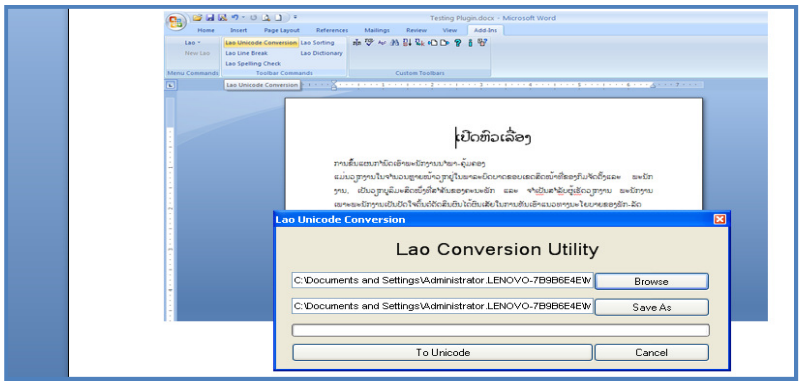


Figure 26: Lao OpenOffice.org Plug-in

The Lao team has also worked on the Plug-in of the Microsoft Office and OpenOffice.org writer. This process has two modules. In the first module, the algorithms for Encoding Standardization and Lao Syllabification has been used which are implemented with the help of APIs of all these utilities developed in phase 1. The second module is embedding of this work which has been developed to extract the data from the office application. After applying the languages dependant algorithm, data is also saved back into the respective places. This module has been developed with the help of Cambodian component by customizing the Cambodian code for Lao language. In addition to this work OpenOffice.org interface Localization has also been started and currently 7,000 words have been translated. The work will be continued at STEA (NAST) after closure of PAN Localization process. All of the above mentioned research outputs are online available at (http://www.panl10n.net/english/OutputsLaos2.htm).

The Lao team of PAN Localization project planned to conduct two types of end user trainings i.e. content providers training and local community training. The country team decided to train 20 to 30 content

providers on Open office, Email and Mozilla Firefox and content provider would be selected from different organizations such as newspapers, ministries such as agriculture, health, culture. Local community would be trained on email and how to copy files to CD. The team also developed a user guide on how to access, upload and download content from web, which is freely available at http://www.dokuwiki.org/manual/.

Training on Localization in Lao, conducted in Laos by Mr. Aamir Wali, from 4th – 11th September 2004. The main topics covered during training were Keyboard Layout creation, Collation and Line/word breaking, basic concepts in C++, Open Type Font Development. The training had been organized to enables Laos country team to develop algorithm for line/word breaking and collation.

In 2005, training on Computing for Localization, conducted at the Science Technology and Environment Agency (STEA), Laos by Mr. Nadir Durrani, from 27th January 2005 till 27th June. The main objective of this training was to capacity build the Laos team in local language computing and enables them to localize software components in Lao.

The discussion during the training covered subjects broadly classified Basic Programming and algorithm development, Advanced Concepts in C++, Visual Basic DOT NET, Microsoft Visual Studio, Using Microsoft Access with VB.NET, Development of Line Breaking Algorithm, Development of Collation Sequence and Sorting Algorithm, Encoding conversion of Non-Unicode fonts to Unicode fonts, Development of Lao-English-Lao Lexicon, Development of Find & Replace Utility and Development of Spell Checker.

Figure 27: Training on Computing for Localization in 2005

Dissemination

Dissemination is an essential part of undertaking research. Research is as credible as much as it is referenced, cited in other publications, brought to people knowledge and properly disseminated.

The main and sustained source of information and outputs of the project has been the project website. The core site has been maintained by the project's regional secretariat and one person from country team of Laos act as a website coordinator and provides local content for the centrally maintained multilingual website www.panl10n.net. In addition the team also hosted their separate websites (http://www.laol10n.info.la/) providing detailed information about their respective research groups, hosted by their organizations, which are linked from the main website as well. This has given global access to project outputs.

Figure 28: Website of Lao L10n

The second website was developed for local community. This website was produced by using Dokuwiki web content development tool and contains 100 MB of content including text and images (http://laocontent.info.la). The website contains various types of information: Public health, Information Technology, Agriculture, Law, Environment and PAN Lao localization. The content focused on the community with low literacy level or basic computer/internet skills.

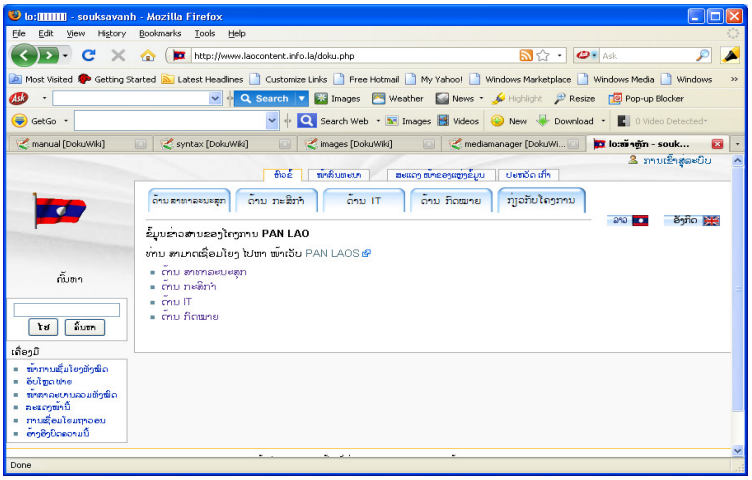


Figure 29: Homepage of http://laocontent.info.la

National Authority for Science and Technology (NAST) organized a meeting of policy makers and IT people on January 10, 2008. In this meeting Pan Localization team in Laos highlighted the research work done under the project and raised the awareness of decision makers and other relevant stakeholders about the potential and impact of local language computing. The president of NAST appreciated the efforts being done under the project and showed his commitment to adopt a vigorous localization policy. NAST also decided to work for E application applying the research of Pan Localization team and to seek government approval for national standard on localization.

In addition to disseminate work, the Laos country team also organized an ICT fair and prime minister of Laos was invited for inauguration. The team conducted a technical workshop and 70-80 government officials, students and journalists also participated in this workshop. CD containing fonts, keyboard utilities and open office application was also distributed to disseminate work on localization.

Infrastructure Development

In building capacity, infrastructural development plays a vital role. To develop the appropriate localization research infrastructure, funds were needed for acquiring academic resources, e.g. books and journals, and specialized software. In phase I, the team of Laos utilize funds for different components of the project; Operational field (trained resources), acquisition of the hardware, development of localized software, and books related to different disciplines like linguistics, language processing and computer science. Equipments included PCs whereas software included Visual Studio.Net and Win Snoori. In the operational field, participants regarding different domains of the PAN Localization project were trained. In phase II, the Laos country partner institution also utilized funds for purchase of the equipment like PCs, laptop, switch, PC Upgrades, amplifier, headphones, mics, speaker, LCD, digital camera and for arranging trainings. In both phases of PAN Localization project, country component focused on procurement of computer hardware, books and conducting trainings and available funds were mostly used for these activities. The accessibility of these funds helped developing appropriate localization research infrastructure and enhanced research capacity in Afghanistan.

Sustainability and Continuity

Organizational capacity enhancement as a result of team skill building is another salient factor in measuring the capacity enhancement of teams for sustainability of research. Thus organization has focused on enhancing their knowledge base to gain advancement in other domains of local language computing as well. This has been a contributing factor for organization to acquire more projects on localization technology development. Laos Country component focused on the development of basic localization and script processing during PAN Localization project.

Through PAN Localization project a significant number of technical developers, linguists and social scientists have been trained to enable sustainability and continuity of the research being undertaken. Laos country component trained 23 participants from different domains like management, technology and linguistics.

6.8 Mongolia

In Mongolia, project collaborated with Mongolia University of Science and Technology (MUST, http://www.must.edu.mn/beta3/), National university of Mongolia (NUM, http://old.num.edu.mn/) and InfoCon Co. Ltd (www.infocon.mn). The PAN Localization project in Mongolia was initiated during phase II.

In Mongolia, the usage of computer per 100 habitants is 2.5 in 2.8 in 2008 (Ariunaa & Uyanga, 2009). The figures regarding lower usage of computer showed that Mongolia is suffering from digital divide. One reason might be English language competence issue and this is understandable because most spoken language in the country is Mongolia whereas all software was available in English language. “Mongolian is national language of the Mongolia and over 2.7 million people speak Mongolian throughout the Mongolia” (Bayanduuren, D, 2007). To reduce the digital divide, one solution could be the development of localized software so that technology is accessible for all Mongolians.

Through PAN Localization project, Mongolia country team has successfully developed localized software. PAN Localization project has contributed for the ICT sector of the country by developing localized software and this project also boosted the country team in building research capacity regarding local language computing. The following sections presented information showing capacity development of each project team assessed though Research Capacity Building model.

Skill development

Project country component was required to deliver specific localized software. These localized software involved expertise in linguistics, computer science and computational linguistics. During Phase II, List of gTLDs and ccTLDs in Mongolian, Localized FireFox (seamonkey) Web Browser, Localized Thunderbird or other Email Client, 100K Word Manually Tagged Corpus, 5 Million Word Mongolian Corpus, Tagged 10k Word Lexicon, Mongolian Spell Checker, Localized Open Office Word Processor, Localized Open Office Spread Sheet, Mongolian POS Tagger, 3K Word ASR for Mongolian, Localized Open Office Presentation Software, Localized GAIM Chatting Program (Pidgin), Integrated OO in Mongolian, Mongolian Sentence Parser, Mongolian Prototype Computational Grammar, Large Vocabulary ASR for Mongolian and Mongolian Speech Corpus were required localized software to deliver.

Skill set pertaining to competence in linguistics, computer science or computation linguistics has also been highlighted for each of localized software in the table below:

Mongolia				
Localized software	Ling.	CL	CS	Status
List of gTLDs and ccTLDs in Mongolian	*		*	Completed
Localized FireFox (seamonkey) Web Browser	*		*	Completed
Localized Thunderbird or other Email Client	*		*	Completed
100K Word Manually Tagged Corpus	*		*	Completed
5 Million Word Mongolian Corpus	*		*	Completed
Tagged 10k Word Lexicon	*		*	Completed
Mongolian Spell Checker	*	*	*	Completed
Localized Open Office Word Processor	*		*	Not Completed
Localized Open Office Spread Sheet	*		*	Not Completed
Mongolian POS Tagger	*		*	Completed
3K Word ASR for Mongolian	*		*	Completed
Localized Open Office Presentation Software	*		*	Not Completed
Localized GAIM Chatting Program (Pidgin)	*		*	Completed
Integrated OO in Mongolian	*		*	Not Completed

Mongolian Sentence Parser	*	*	*	Not Completed
Mongolian Prototype Computational Grammar	*	*	*	Not Completed
Large Vocabulary ASR for Mongolian	*			Completed
Mongolian Speech Corpus	*		*	Completed

Table 9: Status of Mongolia's team regarding Localized Software

The above status shows that the project team of Mongolia has been able to submit 67% localized software as per contract.

National University of Mongolia (NUM) has been working on Mongolian corpus collection from different sources such as literature, law and news paper. Total of 5 million words corpus has been collected and cleaned using the corpus cleaning tools. In addition, some spelling mistakes are corrected with the help of spell checker. The corpus is available at http://panl10n.net/english/OutputsMongolia2.htm.
The NUM team has also developed 10,000 word lexicon from this corpus and is online available for downloading. Mongolian team NUM has also worked for the development of Mongolian spell checker. The main focus of developing this tool is to correct the spelling mistakes in the corpus of 5 million words. For this, a dictionary based spell checker is developed. The Figure given below shows screen shot of Mongolian spell checker.

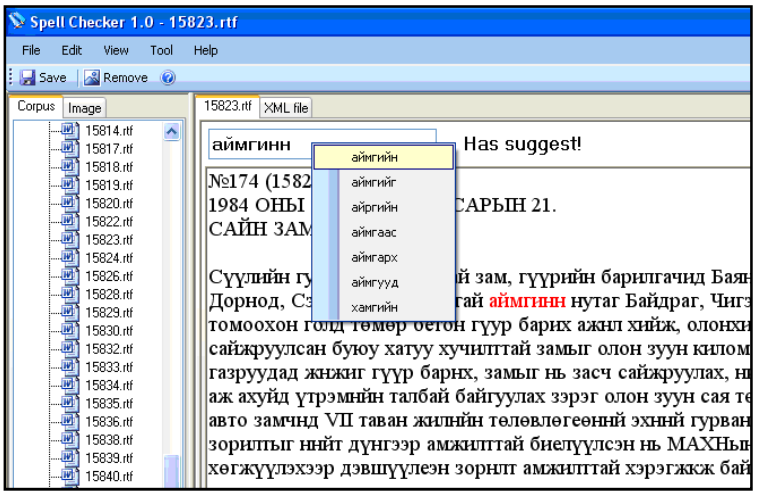


Figure 30: Working View of Spell-Checker

NUM has also been working on Part of Speech tagset. The tagset of 81 tags has been developed. The 100,000 words corpus is tagged manually and is online available. To manually tag the corpus, the tool for manual tagger is also developed. The statistical tagger is developed for automatic tagging of Mongolian text which is based on the HMM model. The reported accuracy is this system is around 81 % for 260,000 words. The paper title “Part of speech Tagging for Mongolian Corpus” is published in 7th Workshop on Asian Language Resources, (ACL-IJNLP2009).

The Mongolia Country Component has already worked on terminology translation through involvement of a committee defined through ICT Agency of Mongolia. During the first year of second phase of the project, the Infocon team has worked on translations of gTLDs and ccTLDs in Mongolian.

The Mongolian country component of PAN has also worked on the localization of tools. In the second phase, The Infocon team has carried out the localization of the web browser, email client and chat tools. A complete version of SeaMonkey internet suite which consists of email client, composer, html editor, chat, web browser has been localized with 43,000 strings. The terminology translation of chat program (PIDGIN) has also been done and released. These all localized software is downloadable from project website.

Mongolian University of Science and Technology (MUST) started research for the development of an Automatic Speech Recognition system (ASR). At start of the research, team conducted surveys for comparative evaluation of ASR toolkits. They had developed a prototype recognition system for Mongolian on HTK and Sphinx toolkits. The HMM based approach has been used to recognize the speech file. The team has developed a 6,000 word ASR based on HMM by using HTK toolkit.

The system has been enhanced by increasing the size of speech corpus. Total of 6,000 high frequency words have been selected from the corpus. One wave file has 10 words and each word has been repeated 10 times. Therefore the speech corpus has 60,000 wave files. The wave file has sampling rate of 16 kHz and sampling size of 16 bits. For the recording of these words 80 speakers have been selected. The test data consists of 100 words spoken by 20 native speakers who had not participated in the training data. The reported accuracy of the system is above 90% for this test data. The paper titled “A Large Vocabulary Speech recognition System for Mongolian language” is published in proceedings of Oriental-COCOSDA 2008.
All of the outputs discussed above are available online at http://panl10n.net/english/OutputsMongolia2.htm
In Mongolia, the PAN Localization team has been interacting regularly with Information and Communications Technology Authority (ICTA). The meetings deliberated on open source software localization, standardization of terminology, IDNs and other related issues.

The country team worked on IT terminology translation and the work was approved by ICTA committee. The country team worked on IDNs standards and reviewed and released generic TLDs (gTLDs), country-code TLDs (ccTLDs) for Mongolian. The Pan Localization project also supported the development of Speech Lab at Mongolian University of Science and Technology (MUST) and Center for Language Processing at National University of Mongolia (NUM) with a view to ensuring the sustainability of localization work. The project also contributed to realization of an MOU signed between NUM and NECTEC.

The accomplished deliverables in Pan Localization project reveal the fact that the project team of Mongolia was focusing on the basic, intermediate and advanced complexity local language computing applications. Based on the project experience, the country project leaders were asked to prorate the ability and skill development/enhancement of the organization’s researchers during project phase 2 on a scale of 1-5; where 1 represent *Challenged*; 2 represent *Fair*; 3 represent *Average*; while 4 represents *Good* and 5 represents *Excellent* enhancement in the team’s performance.

The following table presents those comparative figures for assessment of team's capacity by the project leader from partner country collected for the team's performance at the beginning of project Phase 2 in 2007 and towards the end of this phase in 2009.

MUST,Mongolia

Skill Development	Start of Project, Early 2007	Towards project End, Mid 2009
LLC Project development	3	5
LLC Project design	3	5
Problem identification	3	5
Project implementation	3	5
Ability to do analysis	3	4
Ability to communicate results	2	4
Multi disciplinary research	3	4
Quantitative analytical skills	3	4
Qualitative analytical skills	3	4

Table 10: Performance of Mongolia's team (MUST) regarding Skill Development

Num,Mongolia

Skill Development	Start of Project, Early 2007	Towards project End, Mid 2009
LLC Project development	1	4
LLC Project design	3	5
Problem identification	2	5
Project implementation	3	5
Ability to do analysis	3	5
Ability to communicate results	2	4
Multi disciplinary research	3	5
Quantitative analytical skills	1	3
Qualitative analytical skills	2	4

Table 11: performance of Mongolia's team (NUM) regarding Skill Development

InfoCon,Mongolia

Skill Development	Start of Project, Early 2007	Towards project End, Mid 2009
LLC Project development	2	5
LLC Project design	3	5
Problem identification	3	5
Project implementation	3	5
Ability to do analysis	2	4
Ability to communicate results	2	4
Multi disciplinary research	3	5
Quantitative analytical skills	3	4
Qualitative analytical skills	3	4

Table 12: Performance of Mongolia's team (InfoCon) regarding Skill Development

The table shown above presents that the country project leaders have confirmed the enhancement of team's skills starting from project development and design to its implementation and analysis within the 3 year span of the project specifically in over-all project execution.

The publications of research paper produced by PAN Localization project teams at various national as well as international research conferences is used as the second indicator for analyzing research capacity enhancement. The project team of Mongolia published 6 papers covering POS, Corpus, and Speech during the project's phase 2 which is the third highest among the all participating countries. Detailed list of research report publication by Mongolia project team is presented in Appendix A.

Development of Linkages

Project country component has been focusing on building capacity by developing appropriate linkages, partnerships and collaborations and for this purpose. Partner teams were encouraged to establish partnerships and collaboration with institutions that had more expertise in a specific field. These collaborations enabled the partners to collectively plan the technical and financial details, exchange data and technology and discuss and formalize shared intellectual property regimes, building institutional capacities in the context. MUST, NUM and InfoCon collaborated with each other at national level. In addition, InfoCon collaborated with Mongolian information, Communication Technology and Post Authority and MUST collaborated with Institute of language and literature of Mongolian Academy of Science. Communication Technology and Post Authority had been helpful in making IT related policy and implementing agency whereas Institute of language and literature of Mongolian Academy of Science helped in research regarding Linguistics.

The project team also has been participating in online research networks to enhance the quality of their work regarding localization and also to acquire latest information on Local Language Computing development.

Dissemination

Dissemination is an essential part of undertaking research. Research is as credible as much as it is referenced, cited in other publications, brought to people knowledge and properly disseminated.

The main and sustained source of information and outputs of the project has been the project website. The core site has been maintained by the project's regional secretariat and one person from country team of Mongolia act as a website coordinator and provides local content for the centrally maintained multilingual website www.panl10n.net. In addition the team also hosted their separate websites www.infocon.mn providing detailed information about their respective research groups, hosted by their organizations, which are linked from the main website as well. This has given global access to project outputs.

Figure 31: Website of INFOCON

The project has organized awareness seminars to disseminate and publicize research results to local community. These seminars have been attended by a large number of participants from academia, public and private sectors. One seminar has been organized by the Mongolia country component to disseminate their work on localized software.

The research paper titled "Corpus Building for Mongolian Language" is published in 6th Workshop on Asian Language Resources from January 7-12, 2008, the 3rd International Joint Conference on Natural Language Processing (IJCNLP2008). The research work of the project was presented by the country team at different workshops and conferences such as ALRN 2007 (Asian Language Research Network) from March 1-2, 2007, Tokyo, Japan, 7th Workshop on Asian Language Resources from August 2-7, 2009, Suntec, Singapore, Asian Language Resource Summit from March 20-21, 2009, Phuket, Thailand and Workshop on Applied NLP and Language Resource development from February 23 – 27, 2009, Bangkok, Thailand. Project partners have been involved in designing, developing and disseminating the material developed, which has contributed to mutual capacity to disseminate research.

Infrastructure Development

The team of Mongolia has been capacity to develop the appropriate localization research infrastructure by providing funds for acquiring academic resources, e.g. books and journals, and specialized software. In building capacity, infrastructural development plays a vital role. The team of Mongolia utilize funds for different components of the project; acquisition of the equipments and books related to different disciplines like linguistics, language processing and computer science. Equipments included computer hardware like PCs, scanners, printers and speech equipments (amplifier, mics, etc). Mongolia country

component also focused on development of networking and available funds were mostly used for these activities. The accessibility of these funds helped developing appropriate localization research infrastructure and enhanced research capacity in Mongolia.

Sustainability and Continuity

Organizational capacity enhancement as a result of team skill building is another salient factor in measuring the capacity enhancement of teams for sustainability of research. Thus organization has focused on enhancing their knowledge base to gain advancement in other domains of local language computing as well. This has been a contributing factor for organization to acquire more projects on localization technology development. Mongolia University of Science and Technology (MUST) focused on the development of language processing, speech processing and InfoCon focused on Standardization and Basic Localization during PAN Localization project. National university of Mongolia (NUM) focused on all Local Language Computing (LLC) Domains namely Standardization, Basic Localization, Lang. Processing, Script processes and Speech Processing.

Through PAN Localization project a significant number of technical developers, linguists and social scientists have been trained to enable sustainability and continuity of the research being undertaken. Mongolia country component trained 27 participants from different domains like management, technology and linguistics.

6.9 NEPAL

In Nepal, project collaborated with Madan Puraskar Pustakalaya (MPP, http://madanpuraskar.org/) and E-Network Research and Development (ENRD, http://www.enrd.org/).

In Nepal, the usage of computer per 100 populations was 0.35 in 2001(Government of Bangladesh & United Nation, 2005). The deep rooted nature of social exclusion in Nepal has hampered the development of ICT sector. Besides, English language competence issue might also be one of the reasons for less progress in ICT sector because most spoken language in Nepal is Nepali but all available software was in English. For fast development regarding Information and Communication Technology, it was required that digital content should be transferred in local language of the country. So that Nepali could easily use software and acquire latest technology.

Through PAN Localization project, the Nepal country team has successfully developed localized software e.g. NepaLinux. The impact of the PAN Localization Project in Nepal has been very positive in the local language computing. The work in itself was technically challenging because for the first time, everything was being done in the Nepali Language. Within a time span of around 11 months, the first result of this project came out in December 2005 with the release of the NepaLinux 1.0. This was a major breakthrough in the history of ICT sector of Nepal. In Nepal, PAN Localization project has also been helpful to build research capacity in local language computing.

The following sections presented information showing capacity development of each project team assessed though Research Capacity Building model.

Skill Development

Project country component was required to deliver specific localized software. These localized software involved expertise in linguistics, computer science and computational linguistics. In Phase I, Nepal country component was required to deliver specific research outputs such as NepaLinux, Nepali Spell Checker and Nepali Grammar Checker whereas in phase II, NepaLinux Training Kit, NepaLinux 3.0, Nepali

Grammar Checker, Nepali Spell Checker, Reviewed List of gTLDs and ccTLDs in Nepali, Nepali Computational Grammar and Nepali OCR were requisite localized software to deliver. Linguistics, computer science and computational linguistics competence have been involved in Nepali Grammar Checker, Nepali Spell Checker and Nepali Computational Grammar. Skill set pertaining to competence in linguistics, computer science or computation linguistics has also been highlighted for each of localized software of both phases in the table below:

Nepal				
Localized software	Ling.	CL	CS	Status
NepaLinux			*	Completed
Nepali Spell Checker		*	*	Completed
Nepali Grammar Checker		*	*	Not completed
NepaLinux Training Kit			*	Completed
NepaLinux 3.0			*	Completed
Nepali Grammar Checker	*	*	*	Completed
Nepali Spell Checker	*	*	*	Completed
Reviewed List of gTLDs and ccTLDs in Nepali	*		*	Completed
Nepali Computational Grammar	*	*	*	Completed
Nepali OCR		*	*	Completed

Table 13: Status of Nepal team's regarding Localized Software

The above status showed that the Nepal country component has been able to submit all required localized software as per the contract except Nepali Grammar Checker.

During the first phase of the project, Nepal made excellent progress by working on translation of terminology, locale and collation. A complete Nepali Linux Distribution was developed and released on 22nd December, 2005 as well (http://www.nepalinux.org/). The launch was a huge success in terms of public response. More than 5 national newspapers did the coverage of the launching ceremony.

Parallel corpus generation for languages of PAN partner countries is one of the prime objectives of Phase II. Nepal country component has carried out this work in collaboration with the Pakistan team. The end goal was to develop a tri-lingual tagged corpus for Urdu-English-Nepali languages. This project was supported by PAN Localization project through funding of Language Resource Association (GSK) of Japan. It was very extensive activity including translation of text, research and development of part of speech tagset for Nepali and Urdu languages. Parallel teams in Nepal and Pakistan worked together to accomplish this task in time. Eventually the target of translating 100, 000 words is achieved and tagged parallel corpus has been released (http://www.crulp.org/software/ling_resources.htm).

MPP has been working on development of computational part of speech (POS) tagset for Nepali language. Initially a tagset was designed comprising of 112 tags. Results of semi automatic tagging showed that the designed tagset is error prone because of its depth. The tagset is carefully pruned and essential 43 tags have been selected. A corpus of 80,000 words has been manually tagged and TnT tagger is trained on it. The reported accuracy of this tagger is 97% for known and 56% for unknown words. This tagger is used to annotate developed corpus of one million words.

In the second phase, work has been done on Nepali IDNs. Language tables and lists of gTLDs and ccTLDs have been released as final output. A meeting with policy makers was organized on 24th March, 2009 to discuss the challenges faced during development of Nepali Internationalized Domain Names in Applications (IDNA). Some of these challenges are usage of IDNA for Nepali and alternative of symbols such as period "." and www. Efforts made by international and national bodies for enabling IDNA are also brought into the knowledge of policy makers. The meeting concluded on recommendations for future directions.

MPP has worked on Nepali Spell Checker which is based on Hunspell open source framework. A lexicon of 37,000 words has been incorporated in Hunspell along with 1800 affix rules. The spell checker is incorporated in OpenOfffice.org and tested for approximately 2000 words. Reported accuracy for this test data is around 90%. The system has coverage of 6.2 million Nepali words. Nepali spell checker stand alone application has been developed and released (http://panl10n.net/english/OutputsNepal2.htm).
Nepali team has successfully released Nepali Computational Grammar Analyzer (NCGA) during Phase II. A rule based chunker has been developed along with 30 manually extracted chunking rules. Currently there are 11 chunk tags in chunkset. Nepali Grammar Analyzer is an integrated application that comprises POS tagger, chunker and a parser. Currently the NCGA parses and analyzes declarative sentences with only one verb. Around 700 verbs have been handled in this grammar checker. The accuracy of grammar analyzer depends on the accuracy of chunker. Reported results show that it parses around 90% sentences accurately.

The work on Nepalinux has continued in second phase and its version 3.0 has been developed and formally released on 25th May, 2008. Recent research in advanced areas of Natural Language Processing has been incorporated in this version. One of such utilities is Nepali Text-to-Speech application (not developed directly through PAN Localization projected). Nepali Sabdakos (Dictionary) has also been added which contains meaning of 8,000 words with examples. Other useful applications which are included in this version are offline English dictionary, Gcompris, TuxType, Nepali Spell Checker and KTouch typing tutor. MPP has also separately released an educational version of Nepalinux 3.0 (Educational) on 6th July, 2008. It contains basic Nepalinux and essential educational applications such as Spell Checker, Gcompris and KTouch typing tutor.

Madan Puraskar Pustakalaya (MPP) has also released the Nepali Office CD. It comprises localized OpenOffice.org suite in Nepali along with a few other useful free and Open Source Software like Gimp, FireFox and Thunderbird etc. The Nepali Spell Checker is also integrated in the OpenOffice.org suite.
Work is being done in the area of Nepali OCR and a beta version has already been developed by the team. This has been done with the guidance and direct training from the Bangladesh team. Although complete OCR system was planned to release but the person working on this project had left the team. The work could not be transferred to other resources. Eventually the OCR project was closed with release of beta version. Under the initiatives of MPP and Kathmandu University (KU), efforts are continuing for developing a Tesseract based Nepali OCR. More information on the activity is available at (http://nepalinux.org/index.php?option=com_content&task=view&id=46&Itemid=53).

The accomplished localized software in Pan Localization project revealed that in phase I, Mongolia country component had been focusing on inter-mediate complexity software while in phase 2, the country project team researched on advanced complexity local language computing applications.

Nepal Country Component (Madan Puraskar Pustakalya(MPP) in collaboration with its Country Partner Institution , E-Network Research and Development(ENRD) successfully conducted a Training (November 9-17, 2008, Nangi, Myagdi, Nepal) for skill enhancement of project team. During training discussion was focused on Nepalinux Feedback, shortcomings, keyboard layouts, new spellchecker, machine translations, nepali sabdakos, tuxtype, gcompris, blog, CMS, domain registration, domain hosting, harddisk partitioning etc.

Based on the project experience, the country project leaders were asked to prorate the ability and skill development/enhancement of the organization's researchers during project phase 2 on a scale of 1-5; where 1 represent *Challenged*; 2 represent *Fair*; 3 represent *Average*; while 4 represents *Good* and 5 represents *Excellent* enhancement in the team's performance.

The following table presents those figures for assessment of its team's capacity by the project leader from country component collected for the team's performance at the beginning of project Phase 2 in 2007 and towards the end of this phase in 2009.

MPP

Skill Development	Start of Project, Early 2007	Towards project End, Mid 2009
LLC Project development	3	5
LLC Project design	3	4
Problem identification	3	5
Project implementation	3	5
Ability to do analysis	4	5
Ability to communicate results	3	5
Multi disciplinary research	3	4
Quantitative analytical skills	3	4
Qualitative analytical skills	3	5

Table 14: Performance of Nepal's team regarding Skill Development

The table shown above presents that the project leader has confirmed the enhancement of team's skills starting from project development and design to its implementation and analysis within the 3 year span of the project specifically in over-all project execution.

Ability to publish research in the form of research papers is a salient indicator for measuring the researcher's research capacity. Thus publications of research paper produced by PAN Localization project teams at various national as well as international research conferences is used as the second indicator for analyzing research capacity enhancement. Nepal country component published one

research paper covering NLP during second phase of PAN Localization project. Detailed list of research report publications by team of Nepal is given in Appendix A.

Training to Conduct Close to Practice Research

Through PAN Localization project's research it was envisioned that localized technology being developed must be deployable and of direct use to the communities.

In order to establish the need for localized application, specific question was asked from the communities regarding the language that they speak at home and at their work. Answers from this question would ascertain their preference of language to undertake everyday communication, both written and verbal. When end-users were asked regarding the language spoken at home and work, 100% respondent indicated that they only use local language for communicate at home as well as at their workplace.

This response clearly indicated that the language most convenient for communication for the specific communities was their respective local language. Thus researching for development of local language ICT applications becomes directly useful and relevant to the subject communities, because in order to communicate electronically, and for work, the communities would require applications developed in local languages of the communities.

The respondents were also asked to rate their reading skill and writing skill in English on a scale ranging from Excellent to poor. 102 respondents in total answered this question and no of respondents rated reading skill in English as excellent and 10 respondents rated their reading skill in English as poor. 8 respondents rated their writing skill in English as excellent and only 1 respondent rated their writing skill in English as poor.

Similarly the respondent were also asked to rate their reading skill and writing skill in Local Language on a scale ranging from Excellent to Poor. 102 respondents in total answered this question and no one of respondents rated reading skill in local language as excellent and 14 respondents rated their reading skill in local language as poor. Only 5 rated their writing skill in local language as excellent and 10 respondent rated writing skill in local language as poor.

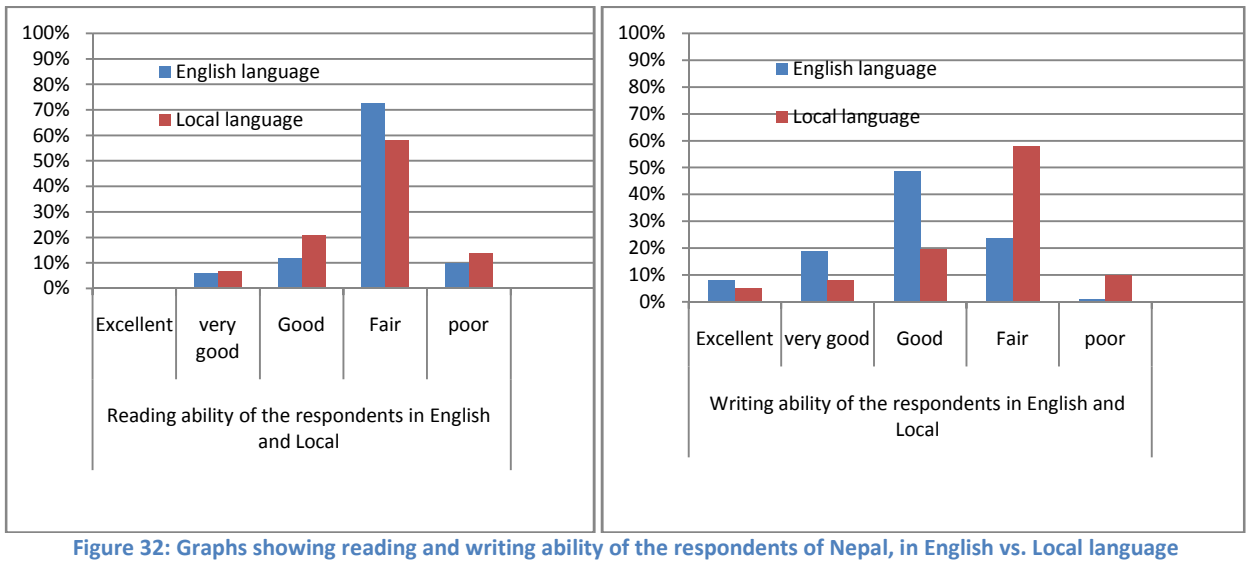


Figure 32: Graphs showing reading and writing ability of the respondents of Nepal, in English vs. Local language

In Nepal, the project involved partners within Madan Puraskar Pustakalaya (MPP), E-Network Research and Development (ENRD). Training on local technology was conducted by E-Network Research and Development (ENRD).The training program focused on farmers, women, students, and youth groups. The training was conducted in five different locations of the country. End users were trained on NepaLinux and other localized software, such as Content Management System.

Development of Linkages

Partner teams were encouraged to establish partnerships and collaboration with institutions that had more expertise in a specific field. These collaborations enabled the partners to collectively plan the technical and financial details, exchange data and technology and discuss and formalize shared intellectual property regimes, building institutional capacities in the context.

At national level, Nepal project team collaborated with its Country Partner Institution, E-Network Research and Development (ENRD) whereas at international level, Nepal country component collaborated with Center for Research in Urdu Language Processing (CRULP), and Center for Research in Bangla Language Processing (CRBLP), BRAC University. CRULP helped Nepal country team in administration of the PAN Localization Project and provided them technical help on Language and Script Processing. CRBLP, BRAC University has been helpful for Nepal project team in developing Optical Character Recognition (OCR).

The project teams have been participating in online research networks, discussion groups, communities and forums for collaboration, knowledge sharing and learning. The work they have performed has given them confidence not only to learn but also contribute on these online forums. The project created an online support network to encourage project partners to be a part of an online learning culture. The project partners have been participating on this forum, sharing their project experiences with each other. Nepal and Bangladesh team discussed their challenges in developing spell checker for open source software for Brahmic scripts. The solution based on HunSpell by Nepalese helped the team develop Bangal spell checker in Bangladesh.

Dissemination

Dissemination is an essential part of undertaking research. Research is as credible as much as it is referenced, cited in other publications, brought to people knowledge and properly disseminated.

The main and sustained source of information and outputs of the project has been the project website. The core site has been maintained by the project's regional secretariat and one person from country team of Nepal act as a website coordinator and provides local content for the centrally maintained multilingual website www.panl10n.net.MPP (http://www.mpp.org.np/pannepal/) and ENRD (http://www.enrd.org/panproject/) also disseminate their work through their websites.

Figure 33: Website of ENRD

Figure 34: Website of MPP

The project has organized awareness seminars to disseminate and publicize research results to local community. These seminars have been attended by a large number of participants from academia, public and private sectors. Through these seminars partner institutions have been regularly presenting their work to the key stakeholders from government, IT industry, academia, media, and end user communities. The project has organized awareness seminars to disseminate and publicize research results to local community.

Nepalinux had been launched by Nepal country component of PAN Localization project, at Kathmandu, Nepal on 23rd December, 2005. To announce the release Mr. Kamal Mani Dixit, president of Madan Puraskar Pustakalaya (MPP) presented a complete CD of NepaLinux to Mr. Sanjeev Rajbhandri, the chief of ISP and IT company Mercantile Communications. Speaking at this launch Mr. Sanjeev said that the Nepali operating system would gain popularity once the government and private institutions endorsed and public embraced it.

Figure 35: launching ceremony of Nepalinux

NepaLinux, had been awarded the prestigious international APC Chris Nicol FOSS Prize 2007, jointly with Free Geek (an organization based in the United States of America and working for the promotion of FOSS), The APC Chris Nicol FOSS Prize is a biennial prize established to honor a long time FOSS advocate and activist as well as APC member. NepaLinux has been put in the annual exhibitions like CAN Info Tech organized by the Computer Association of Nepal (CAN) for the last four years continuously. An estimated 3000 copies of CDs/DVDs of NepaLinux both the downloadable and CD/DVD burnt versions have been distributed to the end users. NepaLinux is currently deployed in around 10 telecenters (Fulchowki, Dhading, Sindhupalchowk, Myagdi, Kaski, Rasuwa, Dailekh and other places - established partly under the PAN Localization Project(http://panl10n.net), Bhasha Sanchar Project(http://bhashasanchar.org) and other collaborations)and in the process of deployment in another 16 telecenters ;Bhaktapur and Butwal and 14 others under the Rato Bangla Public Private Partnership Network and facilitated by Madan Puraskar Pustakalaya in direct partnership with Nepal Telecommunication Authority. More detailed information is available at http://nepalinux.org.

In addition to disseminate research outputs, rural community in Nepal was trained on localized technology in five different locations of the country. The locations included Danda Gaun in Rasuwa district, Jhuwani in Chitwan district, Tolka in Kaski district, Nangi and Shika in Myagdi district. The training program focused on farmers, women, students, youth and other groups.

Training program adopted the train-the-trainer format and training was conducted in three stages. At first stage, in November 2007, Madan Puraskar Pustakalaya (MPP), in collaboration with E-Network Research and Development (ENRD) trained telecenter operators and teachers in 10 day trainer's training program. The main objective of the training was to make the participants familiar with NepaLinux and other localized software such as Content Management System so that they can successfully provide technical support and further training to the local community. These participants were further trained in 9 day trainer's training program in home village of Magsaysay Award winner - Mr. Mahabir Pun. At second stage, every trained teacher/ tele-center operator developed his/her own training outline and course and nominated a group of five participants. Each group has representation from target population of women, farmers, students, youth and teachers. Telecentre operator/teacher trained his/her group. The trainees at this level were defined as local level leadership. At third stage, each local level leader trained his/her own community members. Trainees at third layer were identified as end-user community. 25 end-users were trained at each location and total 125 end-users were trained in that process.ENRD also evaluated the effectiveness of these trainings.

A large majority of the end users were not familiar with computer before the training and at the completion of training they were to create content. ENRD also observed that teachers and students had relatively more learning capacity and their content requirements were easier to address. ENRD also conducted further trainings on basic computer skills, open office, Net Meeting, Instant messenger and web browser (Mozilla Firefox).

Figure 36 : Second TOT Training at Nangi, Nepal

Telecenter operators and teachers trained at first stage of training also developed five websites (www.shikha.com.np, www.nangi.com.np, www.jhuwani.com.np, www.tolka.com.np, www.dandagaun.com.np) on which rural community uploaded the content. This content included educational material, poem, stories, and advertisement of the local products, local news and tourism. The partners in Nepal used two approaches for content development. One was top-bottom approach and the other was bottom–up approach. In the first approach, the content was produced by ENRD, MPP and other organizations. In the second approach content was produced by the community including teachers, students, villagers and local government. These trainings had been helpful to disseminate research outputs.

Infrastructure Development

In building capacity, infrastructural development plays a vital role. To develop the appropriate localization research infrastructure, funds were needed for acquiring academic resources, e.g. books and journals, and specialized software. In phase I, the team of Nepal utilize funds for different components

of the project; Operational field (trained resources), acquisition of the equipments and books related to different disciplines like linguistics, language processing and computer science. Equipments included PCs, PDA and printers. In the operational field, participants regarding different domains of the PAN Localization project were trained. In phase II, the Nepal country partner institution also utilized funds for purchase of the equipment like PCs and printers. In both phases of PAN Localization project, country component focused on procurement of computer hardware and conducting trainings and available funds were mostly used for these activities. The accessibility of these funds helped developing appropriate localization research infrastructure and enhanced research capacity in Nepal.

Sustainability and Continuity

Organizational capacity enhancement as a result of team skill building is another salient factor in measuring the capacity enhancement of teams for sustainability of research. Thus organization has focused on enhancing their knowledge base to gain advancement in other domains of local language computing as well. This has been a contributing factor for organization to acquire more projects on localization technology development. Nepal Country component focused on the development of basic localization, Language processing, script processing and speech processing during PAN Localization project.

Through PAN Localization project a significant number of technical developers, linguists and social scientists have been trained to enable sustainability and continuity of the research being undertaken. Nepal country component trained 31 participants from different domains like management, technology and linguistics.

6.10 Pakistan

In Pakistan, the project collaborated with Center for Research in Urdu Language Processing (CRULP, http://www.crulp.org/) and University of Computer and Emerging Sciences (NUCES, http://www.nu.edu.pk/) to build research capacity. The PAN Localization project in Pakistan was initiated during phase II.

In Pakistan, the usage of Personal computer per 100 populations in Pakistan was 0.4 in 2001 (Government of Bangladesh & United Nation, 2005). It shows that ICT sector was in developing stage. English Language competence issue might be reason behind the slow development in ICT sector as Ansari, S and Saleem, S (2009, p.297) highlighted that “In Pakistan, Urdu is the dominant language of instruction”. Through PAN Localization project, training was conducted regarding language spoken by Pakistanis and result showed that they only use local language for communication. To cope with language competence issue, it was needed to develop localized software so that majority of Pakistanis could conveniently understand and use technology.

The development of local language content was a major challenge because for the first time content was being developed in Urdu language and through PAN Localization project, Pakistan country team successfully developed localized software. This project has been helpful to build research capacity in local language computing. The following sections presented information showing capacity development of each project team assessed though Research Capacity Building model.

Skill development

Project country component was required to deliver specific localized software. These localized software involved expertise in linguistics, computer science and computational linguistics. Pakistan country component were required to deliver six localized software namely Localized Email Client, Localized Internet Browser, Localized OpenOffice.Org Writer, Localized OpenOffice.org Writer, Localized OpenOffice.org Draw, Localized Web Composer and Localized Psi. Skill set pertaining to competence in linguistics, computer science or computation linguistics has also been highlighted for each of localized software in the table below:

Pakistan				
Localized software	Ling.	CL	CS	Status
Localized Email Client	*		*	Completed
Localized Internet Browser	*		*	Completed
Localized OpenOffice.org Writer	*		*	Completed
Localized OpenOffice.org Draw	*		*	Completed
Localized Web Composer	*		*	Completed
Localized Psi	*		*	Completed

Table 15: Status of Pakistan's team regarding Localized Software

The above status shows that the Pakistan country component has been able to submit all localized software as per the contract.

The accomplished localized software in Pan Localization project reveal that CRULP has been focusing on basic, intermediate and advanced complexity local language computing applications.

Pakistan team has worked on development of English-Urdu parallel corpus by translating first 100,000 words of PENN Treebank. This work has been supported by PAN Localization project through funding of Language Resource Association (GSK) of Japan. As a part of this project, a Part of Speech tagset is also designed following the PENN Treebank guidelines. Developed tagset contains 46 tags which are properly defined with examples. Finally, translated corpus of PENN Treebank is manually tagged with this tagset.

An additional work has been carried out to computationally enhance Urdu POS tagset for automatic tagging. Tagset is reduced to 32 tags by selection of only those tags which give contextual information instead of lexical information. Urdu corpus translated from PENN Treebank is tagged with this tagset. An additional corpus of 100,000 words from news data is also tagged. The tagging process has been semi automatically carried out with the help of TnT tagger. Trained tagger showed accuracy around 91%. This work is published in 7th Workshop on Asian Language Resources, ACL- IJCNLP in Singapore in 2009 (http://crulp.org/research/papers.htm).

In 2009, an Urdu rule based stemmer is also developed. In order to cater 140,000 words, 174 prefixes and 712 postfixes are identified. Rules for affix removal are developed with the help of these lists. These rules are helpful for stemming process of Urdu words. The reported accuracy of the system is 91.2 %. Urdu stemmer is also available online (http://www.crulp.org/software/langproc/UrduStemmer.htm). The research paper on Urdu stemmer has been presented by Dr. Sarmad Hussain at 7th Workshop on Asian Language Resources, ACL- IJCNLP in Singapore in 2009 (http://www.crulp.org/research/papers.htm). The Figure 1 shows the stemming of one Urdu word.

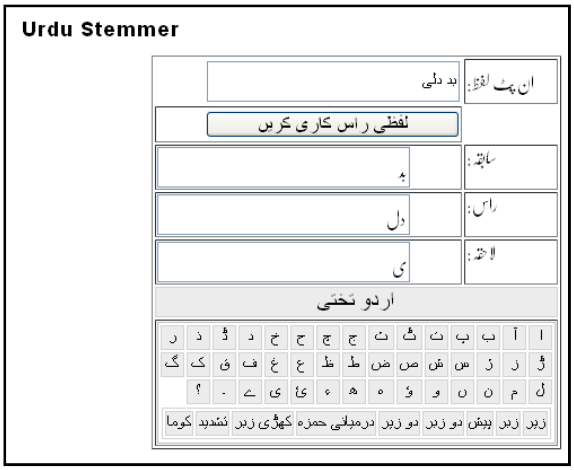


Figure 37: Online Stemming of the Urdu Word "بد دلی"

In addition to POS tagger and Stemmer, CRULP has been working on development of Collation sequence, Spell Checking and Text Normalization applications. Some of the work in these areas has already been done and improvements have been partially supported by PAN Localization project. All of these utilities are available online (http://www.crulp.org/software/langproc.htm).

Pakistan team has also initiated work on development of Urdu OCR system during Phase II. The work started with research and development of multilingual framework for OCR systems. Basic design has been made in this activity. In October 2007, Google released first version of their multilingual OCR framework OCROpus (http://code.google.com/p/ocropus/). It was decided to participate in research and development of OCROpus and build working Urdu modules for that system.

Work on preprocessing engine has been carried out and a technique for font size independent OCR is proposed. A working solution for isolated Urdu characters has been made and tested as a part of this

project. The process starts by taking the boundary of the image. After outline approximation using splines of the boundary the scaling is applied to have the ligature size according to the 36 font size of the respective ligature. The boundary of outline is filled and then converted into the image form. The reported accuracy of the system is 96 % for single character ligature.

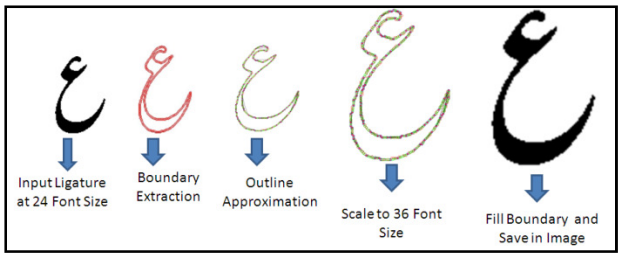


Figure 38: Prototype of Urdu OCR

The paper titled “Font size independent OCR for Noori Nasstaleeq" has been presented in Proceedings of Graduate Colloquium on Computer Sciences (GCCS), Department of Computer Science, FAST-NU Lahore, Volume 1, 2010.

In addition to that, research on OCR development of Noori Nastaleeq font has been carried out. Prototype of this research is developed using HMM based recognizer. The developed system is an advancement of already working system which had partial alphabet coverage. The system contains different modules of OCR system as segmentation, recognition and post processing units in order to cater all combinations of Urdu characters. The system identifies ligatures of Noori Nastaleeq of 36 point size. The application has been tested for 5,000 frequent words of Urdu and showed accuracy around 80%. This work is partially supported by PAN Localization project. Following figure depicts the interface of application.

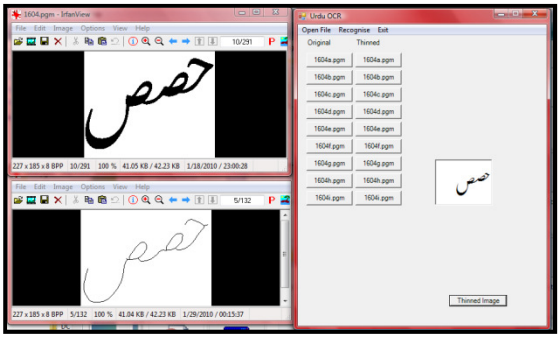


Figure 39: Prototype of Urdu OCR

Work has also been done in development of Machine Translation (MT) system for English-Urdu pair. A rule based MT system has already been developed by CRULP for Urdu. In order to enhance the efficiency and accuracy of that system, research on statistical MT system has been carried out. It is very difficult to cover diversity of Urdu language with rule based system. The proposed work is an enhancement of the system to cater longer and more complex sentences. Research on Semi-Automatic Lexical Functional Grammar Development has been published in the Proceedings of the Conference on Language and Technology 2009 (http://www.crulp.org/research/papers.htm).

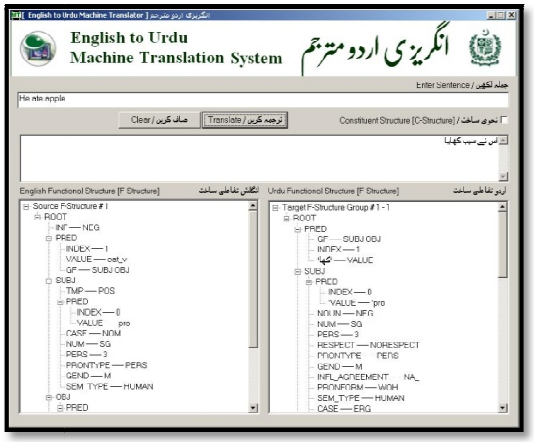


Figure 40: Rule based Machine Translation system

The Urdu Localization Terminology Glossary is developed for the localization of open source software. It is based on the Electronic Dictionary of Localization of Computer Applications (English-Urdu), 2005 by the Center of Excellence for Urdu Informatics, National Language Authority, Islamabad (Pakistan). In addition to this Mozilla Urdu Language Pack, OpenOffice, FireFox and Thunderbird (for Urdu-India) have also been consulted. This glossary is online available (http://www.crulp.org/software/localization/OSS/ossGlossary.html). The Figure shows the Urdu translation of the word "Abbreviation" along with the source information from which word translation is taken.

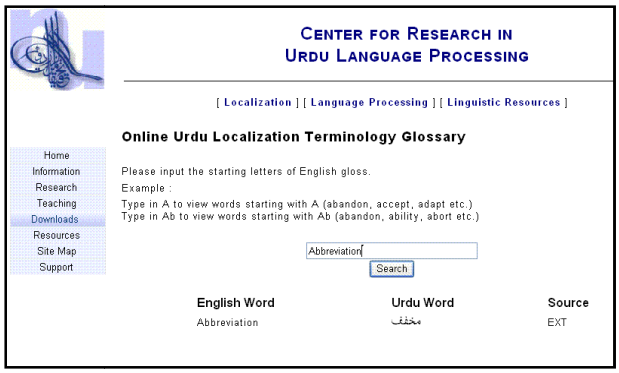


Figure 41: Online Urdu Translation of the word "Abbreviation"

Workshops and trainings have been organized by CRULP for their team's skill enhancement. The first workshop organized on IDNs for Pakistani Languages at FAST-NUCES, Lahore on April 20th, 2008. The objectives of the workshop were to identify the confusable characters within a language, characters required for each language represented and to identify the characters in these languages which are not represented in Unicode.

Figure 42: workshop on IDNs for Pakistan Languages, 2008

Second Workshop on Internationalized Domain Names for local content development in Pakistani languages organized by National IT Development and Promotional Unit (NIDU) under the Ministry of Information Technology, held at FAST-NUCES from 15th - 16st May, 2009.

The purpose of this meeting was to provide a platform for the discussion of policy and standardization issues relating to internationalizing domain names in Pakistani languages. Participants from various language communities were invited to provide suggestions and share ideas on conclusions drawn in the previously held workshop.

Figure 43: Workshops 2009

Based on the project experience, the Pakistan country component was asked to prorate the ability and skill development/enhancement of the organization's researchers during project phase 2 on a scale of 1-5; where 1 represent *Challenged*; 2 represent *Fair*; 3 represent *Average*; while 4 represents *Good* and 5 represents *Excellent* enhancement in the team's performance.

The following table presents those figures for assessment of its team's capacity by the project leader from Pakistan country component collected for the team's performance at the beginning of project Phase 2 in 2007 and towards the end of this phase in 2009.

CRULP

Skill Development in	Start of Project, Early 2007	Towards project End, Mid 2009
LLC Project development	5	5
LLC Project design	5	5
Problem identification	4	4
Project implementation	4	4
Ability to do analysis	4	5

Ability to communicate results	2	5
Multi disciplinary research	3	5
Quantitative analytical skills	3	4
Qualitative analytical skills	2	4

Table 16: Performance of Pakistan's team regarding Skill Development

The table shown above presents that the project leader has confirmed the enhancement of team's skills starting from project development and design to its implementation and analysis within the 3 year span of the project specifically in over-all project execution. According to the project leader, the maximum skill enhancement has been in the project development and in the project design which includes the researcher's ability to conceptualize new projects. Teams have also significantly improved in other project areas including the ability to conduct multi-disciplinary research, problem identification and quantitative as well as qualitative analysis of the research.

Ability to publish research in the form of research papers is a salient indicator for measuring the researcher's research capacity. Thus publications of research papers produced by PAN Localization project teams at various national as well as international research conferences is used as the second indicator for analyzing research capacity enhancement. The project team of Pakistan published 5 research papers covering M&T, POS, and IDN during the project.

Pakistan is one of those countries who have published maximum number of research papers. A closer observation of team structure reveal that the project is housed within universities and the educational qualification of the country team leader is doctorate. The educational qualification of the country project team leader is a significant factor that has influenced more academic research within the teams. Detailed list of research report publication by each country component is presented in Appendix A.

Training to Conduct to Practice Research

Through PAN Localization project's research it was envisioned that localized technology being developed must be deployable and of direct use to the communities.

In order to establish the need for localized application, specific question was asked from the communities regarding the language that they speak at home and at their work. Answers from this question would ascertain their preference of language to undertake everyday communication, both written and verbal. When end-users were asked regarding the language spoken at home and work, 100% respondent indicated that they only use local language for communicate at home as well as at their workplace.

This response clearly indicated that the language most convenient for communication for the specific communities was their respective local language. Thus researching for development of local language ICT applications becomes directly useful and relevant to the subject communities, because in order to communicate electronically, and for work, the communities would require applications developed in local languages of the communities.

The respondents were also asked to rate their reading skill and writing skill in English on a scale ranging from Excellent to poor. 61 respondents answered this question and majority of them (20) rated their reading skill in English as excellent and only 2 respondents rated their reading skill in English as poor. A

large majority of them (35) rated their writing skill in English as excellent and no one of respondents rated their writing skill in English as poor.

Similarly the respondent were also asked to rate their reading skill and writing skill in Local Language on a scale ranging from Excellent to Poor. 61 respondents in total answered this question and majority of them (24) rated their reading skill in local language as excellent and no one of respondents rated reading skill in local language as poor. 10 respondents rated their writing skill in local language as excellent and nobody rated writing skill in local language as poor.

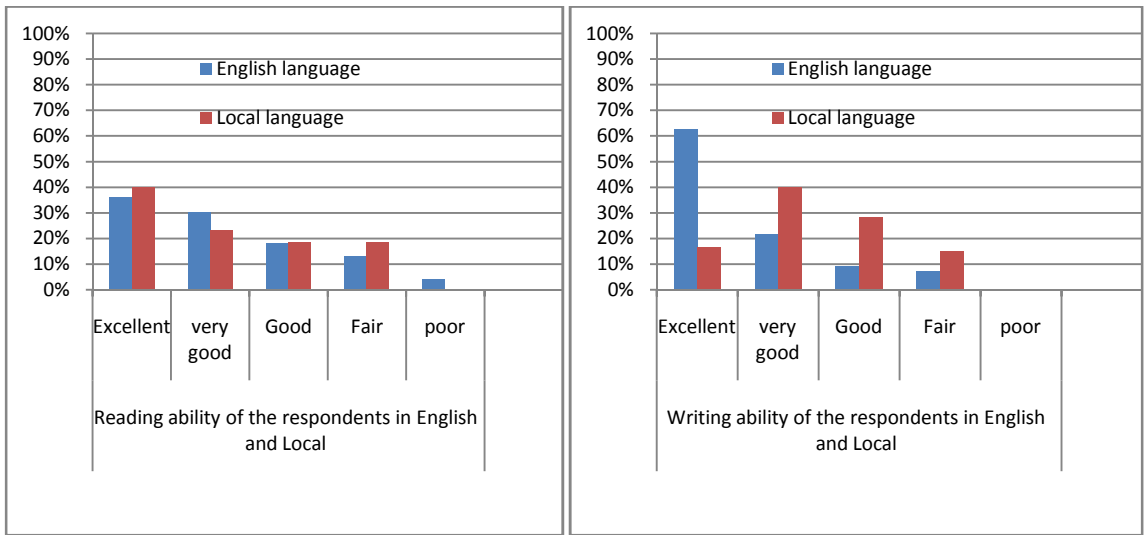


Figure 44: Graphs showing reading and writing ability of the respondents in English vs. Local language from Pakistan.

In Pakistan, CRULP conducted training on localized technology. The training program focused on students from rural areas. The participants were trained on basic computer skills and localized applications, including OpenOffice (Word Processor, Graphics Editor), SeaMonkey (Browser, Email Client, and Web Editor) and Psi (Instant Messaging Client).

Development of linkages

Project country component has been focusing on building capacity by developing appropriate linkages, partnerships and collaborations. Partner teams were encouraged to establish partnerships and collaboration with institutions that had more expertise in a specific field. In Pakistan, center for research in Urdu Language Processing (CRULP) collaborated with 20 international organizations and they all were Pan Partners. These collaborations enabled the partners to collectively plan the technical and financial details, exchange data and technology and discuss and formalize shared intellectual property regimes, building institutional capacities in the context.

The project team has been participating in online research networks, discussion groups, communities and forums for collaboration, knowledge sharing and learning. The work they have performed has given them confidence not only to learn but also contribute on these online forums. The project created an online support network to encourage project partners to be a part of an online learning culture. The project partner has been participating on this forum, sharing their project experiences with each other. Pakistan country component, through CRULP, is also an active member of the Arabic Script IDNs Working Group (ASIWG), a self-organizing group of individuals representing language communities that

use Arabic script, created in March 2008. Ever since, CRULP has been a keen contributor towards the standards being developed for Arabic script in general and Pakistani languages in particular. Four ASWIG meetings have been held so far with major outlines related to the development of language and normalization tables for Arabic. Pakistan team has provided major feedback on IDNA standard as it was being developed. Pakistan team has also provided considerable inputs on various IDN issues such as mixing of digit sets, confusable characters, and exclusion of unwanted character from the standard and inclusion of characters that are mandatory for some Pakistani languages. CRULP has also been involved with ICANN in their IDN implementation and new gTLD registration process. Comprehensive feedback reports on ICANN similarity testing algorithm for gTLD program have been submitted in this regard.

Dissemination

Dissemination is an essential part of undertaking research. Research is as credible as much as it is referenced, cited in other publications, brought to people knowledge and properly disseminated.

The main and sustained source of information and outputs of the project has been the project website. The core site has been maintained by the project's regional secretariat and one person from country team of Pakistan act as a website coordinator and provides local content for the centrally maintained multilingual website www.panl10n.net. In addition, Pakistan country team has been highlighting its work through its local website (http://www.crulp.org/dareecha/).

Figure 45: Website of Dareecha Project

The project has organized awareness seminars to disseminate and publicize research results to local community. These seminars have been attended by a large number of participants from academia, public and private sectors. Through these seminars CRULP has been regularly presenting its work to the key stakeholders from government, IT industry, academia, media, and end user communities. In Pakistan, the country team has been active in interaction with various stakeholders such as Ministry of Information Technology and Telecommunications (MoITT), Ministry of Education, Pakistan Telecommunication Authority (PTA), Universal Service Fund Guarantee Ltd (USF), Pakistan Software Export Board (PSEB), All Software Houses Association (PASHA), Internet Service Providers Association of Pakistan (ISPAK), and Computer Society of Pakistan, National Language Authority (Muqtadra Qaumi Zaban, Urdu Science Board. In these meetings, the country team created awareness about the benefits of localized technology.

CRULP conducted a pilot training of a group consisting of seven 8 grade students in May 2008. The participants were trained on basics computer skills and web browsing. This exercise was beneficial in many ways; it helped the team improve the localized training material, and it also gave the experience to the trainers.

A very large project, named Dareecha, involving a team of 14 human resources has been carried out during Phase II of PAN Localization project. Dareecha project aimed to determine and evaluate sustainable strategies for Information and Communication Technology (ICT) access and local language content generation in rural schools. The project covered a diverse range of activities including local language resource development, local language technology deployment and project monitoring and evaluation. Dareecha work can be divided into three phases.

In the first phase, localization of web browsing and emailing (Sea Monkey), word processing (Open Office), chatting (Psi) and web page development (Web Composer) was done in Urdu during the first year using OSSs. It aimed to provide the support for those people who do not understand English but Localization (Urdu translation) can easily make these tools available for them. The training material comprising of seven books was developed in Urdu language and it is freely available online at http://www.crulp.org/dareecha/Training-Books.htm.The internet related applications, SeaMonkey was selected for the localization of the GUI. The reasons behind this tool selection were 1) it is a complete Internet suite including a web browser and an e-mail client (because a browser and e-mail client is integrated into a single application, the overall localization effort would be less as compared to localizing a browser and e-mail client separately), 2) it has Unicode (UTF-8) and bidirectional language support which was required for Urdu, 3) it is localizable, and 4) it is supported on Windows, Linux and Mac OS X. Openoffice.org was selected because of its wide range of usage on multiple platforms including Windows, Linux and Mac. In addition to its universal usage, defined localization process and updated support make the localization process very smooth. Still there were many challenges in this process. Some of the hurdles faced during this project were presence of very large translation strings, selection of appropriate (understandable for layman) meaning and building of office suite. Translation process was carried out with the help of OmegaT translation management tool. In order to extract strings from SeaMonkey suite, Mozilla translator was also used. Different translation resources were consulted to accomplish localization process. These resources included (higher to lower priority) glossary of National Language Authority Pakistan, Firefox 1.0.6 ur-IN glossary and OpenOffice 2.0.3 ur-IN glossary etc.

In the second phase of Dareecha, the material for training of these tools was developed. For the assessment of the training material, competency levels were defined. The competency levels followed a step-by-step approach towards learning new concepts. Starting from beginner's level, difficulty increased with each level. Competency levels enabled us to devise lecture plans for students and later, assessments were also designed on the basis of these competency levels. Each series of competency levels consisted of a set of entry criteria, which must be met before users could proceed with learning, and a set of exit criteria, which must be achieved at the end of the learning associated with that learning area. Within the entry and exit criteria a series of skills was needed to be acquired corresponding to the learning area, and as the users transition through the steps, they were going through increasing levels of difficulty. Each skill level was then further broken down into a more specific skill set. The Competency levels for training of basic computing, web browsing, email, instant messaging, word processing, graphics editing and web page development were defined and made available online at (http://www.crulp.org/dareecha/Training.htm).

In the third phase, training for these tools was provided in ten (five boys, five girls) schools of rural areas from 3rd November 2008 and ending on 12 June 2009. The Training sessions for students (grades eight and nine) was conducted across six locations. At least 14 students were nominated by each school to attend the training sessions. Nominated teachers from these schools received trainings along with helping material before the training sessions of students. The purpose was to prepare them for the assistance of Project Dareecha trainers during the student training sessions. There were three five-day training sessions for students at each school, conducted by Project Dareecha team with the assistance of the trained school teachers. Between the training sessions, the students were left under the supervision of the school teachers. It gave them opportunity to continue practicing the skills they have learnt during training sessions. School teachers were provided with additional training material (work plans, exercises etc.) that can be used between training sessions. Teachers from each school were trained together at a single location whereas students were trained at their respective schools. After the teacher training session, the country team replicated the same training session at each school for the students. After each student training session, the students worked under the teachers' supervision teachers at their school. Training material including books, presentation slides were distributed freely among training participant and books were also provided to each school to use in further trainings.

Figure 46: Students during Dareecha Training Sessions

Recognizing the cultural norm, a team composed of exclusively female members conducted trainings in the girls' schools, and a team composed of exclusively male members conducted trainings in the boys' schools. The schools were encouraged to develop business models to use the labs in the evenings to generate income. The program was initially expected to train 140 students but considering the growing demand for training, 228 students were trained. Among them 140 were girls and 88 were boys. 20 teachers were trained. Among them 10 were male teachers and 10 were female teachers. The detail of training program is available at http://www.crulp.org/dareecha/Training.htm.

Figure 47: Training session for teachers

CRULP team also trained Master trainers/teachers in Punjab IT Labs started by the Punjab Government in 2009 on Urdu usage on computer. In the Punjab IT labs project, IT labs were deployed in 4,286 secondary and higher secondary schools of the province. CRULP contributed to improve the usability of Urdu on computers. A phonetic keyboard layout and fonts (Nafees Nastaliq and Nafees Web Naskh) package was provided for easy use of Urdu on the computer on the MS Server 2008 platform deployed by the Punjab IT Labs Project. The detail of this activity is available at http://www.crulp.org/Downloads/PunjabITLPS/UrduUsageGuide.pdf.

The Center for Research in Urdu Language Processing (CRULP) actively participates in developments taking place in IDNs worldwide. A workshop was organized by CRULP in year 2008, gathering participants representing the various languages spoken in Pakistan (http://www.panl10n.net/english/Pakistani IDNs Workshop.htm). An initial attempt was made to draft character sets for different languages including Balochi, Pashto, Punjabi, Saraiki, Sindhi and Torwali. A follow-up workshop on IDNs was arranged at NUCES on behalf of the Ministry of IT Pakistan in May 2009, to build on the earlier for Pakistani languages (http://www.panl10n.net/english/PakistaniIDNsWorkshop2nd.htm). This workshop was organized in two sessions: the first session was an open discussion where general public was invited through advertisements placed in Urdu and English newspapers. The second session was a closed group meeting where experts finalized decisions on major issues regarding implementation of IDNs for Pakistani languages. CRULP is also a member of the technical committee formed by the Ministry of IT (MoIT), Pakistan to implement IDNs. Plan of a third meeting in collaboration with the MoIT is under way.

In Pakistan, it was strategically planned to develop locally relevant content using bottom - up approach. For this purpose, teachers and students were trained to develop simple web pages. They were encouraged to develop websites. At the end of training, an inter-school web development competition was organized. Entries were invited in three categories: Community website (by a group of students), School website (by a group of teachers) and Individual website (locally relevant website by individual students). 57 entries in total were received, 10 websites were developed by teachers and 47 were developed by students. Women also participated in significant number in the competition. Among 47 websites developed by students, 36 were developed by female students whereas 11 websites were developed by male students. All websites are available at http://www.crulp.org/dareecha/competitions/competitions.htm.

These websites were evaluated by a panel of judges from the IT industry, academia, government, media and other organizations. After evaluation of websites, a prize distribution ceremony was held on January 23, 2010 at the NUCES Lahore campus and 250 project participants attended the ceremony. The websites awarded with prizes are available at http://www.crulp.org/dareecha/winners.htm.

Figure 48: Homepage of Website winning First Prize

The Pan Localization Pakistan component Dareecha has been presented at the following events at national and international levels to disseminate their research outputs:

Gendered OM methodology selected for demonstration at International Conference on Information and Communication Technologies and development ICTD2009, in CMU Doha, Qatar from April 17-19, 2009. Sustainable Development Policy Institute (SDPI) Study Group Meeting on Women and ICTS: Exclusion or Empowerment on 13 August, 2009.Internet Governance Forum (IGF) Workshop on Equality in access to knowledge society, through language and cultural diversity held at Sharm El Sheikh, Egypt on 18 November, 2009. 2nd GEM Global Exchange, will be held in Bali, Indonesia from November 24-30, 2009. Seminar on Integrating IT in Education: Language, Curriculum and Training Challenges in Government Schools of Punjab held on 17 December, 2009 at NUCES, FAST Lahore.

David M. Malone, President of Canada's International Development Research Centre, IDRC president visited CRULP on 5 February, 2010 and highly appreciated the achievements of country team. The project activities in Pakistan have been widely covered by print and electronic media. Spider, popular IT Magazine in Pakistan published articles highlighting progress made by the project in the country.

Infrastructure Development

The team of Pakistan has been capacity to develop the appropriate localization research infrastructure by providing funds for acquiring academic resources, e.g. books and journals, and specialized software. In building capacity, infrastructural development plays a vital role. The team of Pakistan utilize funds for different components of the project; acquisition of the equipments and books related to different disciplines like linguistics, language processing and computer science. Equipments included computer hardware like PCs, laptop, scanners, printers, CD-RW Drive, USB Drive and mobile phone for end user. Pakistan country component also focused on development of networking and available funds were mostly used for these activities. The accessibility of these funds helped developing appropriate localization research infrastructure and enhanced research capacity in Pakistan.

Sustainability and Continuity

Organizational capacity enhancement as a result of team skill building is another salient factor in measuring the capacity enhancement of teams for sustainability of research. Thus organization has focused on enhancing their knowledge base to gain advancement in other domains of local language computing as well. This has been a contributing factor for organization to acquire more projects on localization technology development. Pakistan Country component focused on the development of

standardization, basic localization, Language processing, script processing and speech processing during PAN Localization project.

Through PAN Localization project a significant number of technical developers, linguists and social scientists have been trained to enable sustainability and continuity of the research being undertaken. Pakistan country component trained 44 participants from different domains like management, technology and linguistics.

6.11 Sri Lanka

In Sri Lanka, the project collaborated with University of Colombo School of Computing (UCSC, http://www.ucsc.cmb.ac.lk/) to build research capacity.

The usage of computer per 100 populations in Sri Lanka was 8.2 in 2008 (Weerasinghe, R & Desilva, C, 2009). The figure depicts that the usage of computer in Sri Lanka was comparatively more than other countries discussed above. However, development of localized software could be helpful to maximize the use of ICT because most of the people speak in their local language "Sinhala". According to Statistics by Scenic Sri Lanka (2007), "Sinhalese make up about 74 percent of the Sri Lankan people and the language they speak is Sinhala, which is the official language". When technology is available in their local language then much more people would easily use latest technology.

Through PAN Localization project, Sri Lanka country component worked on the development of localized software and this project boosted them to build research capacity in local language computing. The following sections presented information showing capacity development of each project team assessed though Research Capacity Building model.

Skill development

Project country component was required to deliver specific localized software. These localized software involved expertise in linguistics, computer science and computational linguistics. In Phase I, Encoding Conversion Utility, Sinhala Lexicon, Multilingual Lexicon (Sihala, English, Tamil), Sinhala TTS System and Sinhala OCR System whereas List of gTLDs and ccTLDs in Sinhala, English- Sinhala Parallel and Aligned Tagged Corpus 100k words, Sinhala Wordnet 5000 words, TM Application, Sinhala H/W Recognition System for PDAs and Tamil Language Learning Tool were required localized software to deliver. Skill set pertaining to competence in linguistics, computer science or computation linguistics has also been highlighted for each of localized software of both phases in the table below:

Sri Lanka				
Localized software	Ling.	CL	CS	Status
Encoding Conversion Utility			*	Completed
Sinhala Lexicon		*	*	Completed
Multilingual Lexicon		*	*	Completed
Sinhala TTS System		*	*	Completed

Sinhala OCR System		*	*	Completed
List of gTLDs and ccTLDs in Sinhala	*		*	Completed
English- Sinhala Parallel and Aligned Tagged Corpus 100k words	*		*	Completed
Sinhala Wordnet 5000 words	*	*	*	Completed
TM Application	*	*	*	Completed
Sinhala H/W Recognition System for PDAs	*		*	Completed
Tamil Language Learning Tool	*	*	*	Completed

Table 17: Status of Sri Lanka's team regarding Localized Software

The status mentioned in above table shows that the project team has been able to submit all localized software as per the contract and the 100 % accomplishment of the localized software shows that Sri Lanka country team's skill has enormously enhanced over the project implementation. It is also worth mentioning that the country project team has also advanced in their over-all local language computing skill set through continual research and development of local language software and its components.
A comparison of the accomplished localized software in PAN Localization project's phase 1 and phase 2 reveal the fact that the project country component was researching on development of intermediate and advanced complexity local language computing application as compared to Phase 1 in which team was only focusing on the development of basic complexity software.

Sri Lanka Country Component worked on some basic text processing utilities in Phase I such as Unicode conversion, Collation and Syllabification. In addition to these areas, work on advanced NLP applications such as Text to Speech and Optical Character Recognition was also carried out during first phase. Sinhala Text to Speech system was awarded as the "Most Innovative Product" at the Biennial Infotel Trade Exhibition held in Colombo, Sri Lanka on 1st November 2008 (http://www.itpro.lk/node/1554).

In second phase of project Sri Lanka team has worked on translation of PENN Treebank to build a Parallel Corpus of Sinhala and English. First 100,000 words has been translated and released as an output of Parallel Corpus. Issues faced during the translation process have also been documented. The work on Part of Speech Tagset and Tagger has already been carried out by Language Technology Research Laboratory (LTRL) and a paper titled "A Stochastic Part of Speech Tagger for Sinhala" has been published in Proceedings of 6th International Information Technology Conference. Colombo, Sri Lanka (2004). A complete document containing tags, meanings and examples has been released as a part of parallel corpus activity. Sinhala version of parallel corpus has also been tagged using developed stochastic part of speech tagger. In addition to that, 100,000 words from 10 Million words corpus of Sinhala have been annotated with POS, resulting a tagged corpus of 200,000 words in total.

The team has also worked for the development of the Sinhala WordNet. For this, different Sinhala sources such as Maha Sinhala Sabdakoshaya by Dr Harishchandra Wijetunge (Main Source), Sinhala-English Dictionary by Charles Carter, Sinhala-English Dictionary by Benjamin Clough, Sinhala-English

Dictionary by Dr A. P. De Soysa, Sanskrit- English Dictionary by Monier Williams, Technical Glossaries published by the Official Language Department, Sinhala Namawaliya (An ancient Sinhala thesaurus) and Ruwanmala (An ancient Sinhala thesaurus) have been used. Previously, very little effort was made for the semantic aspect of Sinhala language. Therefore some English sources were used to extract the semantic aspect of words and their relations and classifications. These sources are Princeton English WordNet (Fellbaum, 1998), Roget's Thesaurus (Roget, 1962) and Suggested Upper Merged Ontology (SUMO) Ontology (www.ontologyportal.org).
1100 high frequency synsets (Approximately 5,000 words) have been selected from the UCSC Sinhala corpus. All those words have been considered as the separate entries for the WordNet which have different semantic information. In the next step, sense identification process has been performed in which a linguist has determined the list of senses from the dictionary and English WordNet. Maha Sinahla Sabdakoshaya and the Princeton English WordNet are the main sources for this purpose. All the sense relations have been extracted from the database instead of defining them from scratch. This procedure helps to develop the Sinhala WordNet using relations in less effort. After this, these relations have been stored in the human readable form. The Sinhala word has been translated into the English, and then the synset identifier of the translated word is obtained from English WordNet. The Sinhala word along with the synset identifier is then inserted into the Sinhala WordNet. In addition to wordnet, a trilingual dictionary (Akaradi) has also been made for English, Tamil and Sinhala users. These resources are available at (http://www.ucsc.cmb.ac.lk/ltrl/?page=services&lang=en&style=default).

Sri Lanka team has worked on language table and terminology translation of gTLDs and ccTLDs in Sinhala for IDNs. The translated Sinhala gTLDs and ccTLDs can be accessed online (http://www.panl10n.net/english/OutputsSri-Lanka2.htm). These efforts have contributed in posting a formal application to ICANN for registration of Sinhala and Tamil country codes. It has been drawn from work carried out by Information and Communication Technology Agency (ICTA) with technical support from University of Colombo. ICTA of Sri Lanka and the LK Domain Registrar have been part of the global discussion on the fast-track Internationalized Domain Names (IDNs) process. An ICTA commissioned working group, consisting of experts from within itself, the UCSC, the UOM and the LK Domain Registry (a subset of its Local Language Working Group) has been looking into issues relating to this process. Major issues have also been discussed such as homoglyps, mixing of scripts, multiple forms of the same word, spelling variants and browsers issues. In addition, a systematic process, considering both human and technical factors, for providing a sustainable IDNs solution for languages belonging to the Indic family has been proposed.

Sri Lanka team has also been working on developing web applications in order to disseminate the research and get feedback from common users. A useful application of this category is EnSiTip which is a Firefox Plug-in (http://www.ucsc.cmb.ac.lk/ltrl/projects/EnSiTip/). It facilitates users by providing Sinhala translations for English words in a popup menu. The plug-in has been downloaded 15,942 times until 3rd March, 2010. In addition to this plug-in Online Encoding Converters for Sinhala have also been made available for public usage (http://www.ucsc.cmb.ac.lk/ltrl/services/feconverter/). Sinhala typing has been made easy by providing an online version of Unicode keyboard. Text can be written online using this keyboard and copied to any desired text box. This keyboard can be accessed at (http://www.ucsc.cmb.ac.lk/ltrl/services/keyboard). Sinhala Spell Checker (Subasa) will be made hosted online in near future. Development of these applications has been partially supported by PAN Localization project.

Sri Lanka team has also been working on development of Translation Memory tool. The purpose of this application is to facilitate translation process from English to Sinhala with word suggestions and central

memory management. OmegaT has been chosen for this purpose because of source code availability. But OmegaT does not support Sinhala Unicode Rendering (as it has been developed in Java Swing). Therefore, OmegaT's core modules were used to create a web application which supports most of the scripts.

A central system OpenTM, based on OmegaT, has been made in Language Technology Research Laboratory to facilitate the translation process. The system is able to maintain the centralized memory backup and translations are updated after feedback from translators. Following figure shows the working of this system.

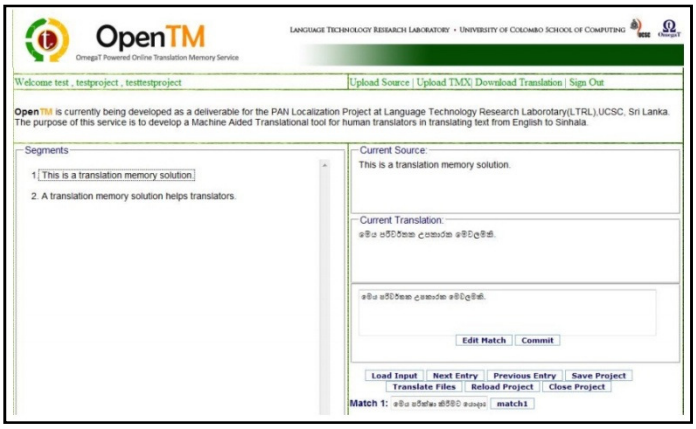


Figure 49: Open TM (An OmegaT powered system)

Another project on development of Tamil Language Learning Tool has been carried out during Phase II. It has been very extensive exercise which included development of training content for Tamil in Sinhala and English (Official languages of Sri Lanka). The content was generated with the help of already existing books like “An Introduction to Spoken Tamil by James W. Gair, S. Suseendirarajah and W. S. Karunatilaka” which has been translated in Tamil by Rev. Kadurugamuwe Nagitha Thero, a lecturer at the University of Kelaniya, Sri Lanka. Interactive question answers are selected to cover daily life scenarios in form of 25 lessons. There are three different sections of each lesson; Dialogs, Grammar and Exercise (For details please see Sri Lanka section of Content chapter).

A parallel activity of application development has also been carried out along with learning content generation. Complete training data is stored in XML format which is used by Language Learning Tool. For development purpose, Macromedia Flash has been selected because of its support of Unicode and portability on desktop and websites. The tool facilitates learner with Sinhala and English interfaces for all lessons. Following figure displays the application layout.

Figure 50: Tamil Language Learning Tool Application

In addition to this tool an online trilingual glossary (English-Sinhala-Tamil) has also been developed to facilitate localization of GUIs for Sinhala. The online framework also provides the facility to vote for the terms.

Sri Lanka team has also initiated research on Online Handwriting Recognition system. The objective of the work is to develop an isolated recognizer for real time input. The proposed model is based on Template Matching with Dynamic Time Warping (DTW) distances. The project has reached its first milestone and data capturing, data analysis and classification modules have been developed. Train data has also been collected from 12 different users on PDA. The training set contains three samples of 1,500 strokes. An error analysis has been carried out on collected samples to prune train data. The research on data cleaning and system training is being continued after completion of PAN localization project.

Sinhala Text to Speech system was developed during Phase I. In second phase a screen reader facility has been incorporated in this TTS application. All of the outputs discussed and relevant research material is available on PAN Laicization website (http://panl10n.net/english/OutputsSri-Lanka2.htm) and regular updates can be found on (http://www.ucsc.cmb.ac.lk/ltrl/).

To enhance skill development of Sri Lankan team, training on Phonetics and Phonology for TTS, conducted at the University of Colombo, School of Computing, Colombo, Sri Lanka, by Dr. Sarmad Hussain, from 21st - 25th June 2004. Objectives of the training were to give requisite background to the Sri Lankan PAN Localization team and others interested on Phonetics, Phonology, Acoustic Phonetics and how all this fits in with TTS process.

Figure 51 :Training on Phonetics, Sri Lanka ,2004

The publications of research papers produced by PAN Localization project teams at various national as well as international research conferences is used as the second indicator for analyzing research capacity enhancement. The project team of Sri Lanka published 10 papers covering MT, Lexicon, Speech and IDN during the project's phase 2 which is the highest among the all participating countries.

Detailed list of research report publication by Sri Lanka project team is presented in Appendix A.

Dissemination

Dissemination is an essential part of undertaking research. Research is as credible as much as it is referenced, cited in other publications, brought to people knowledge and properly disseminated.

The main and sustained source of information and outputs of the project has been the project website. The core site has been maintained by the project's regional secretariat and one person from country team of Sri Lanka act as a website coordinator and provides local content for the centrally maintained multilingual website www.panl10n.net. The county team also disseminated its work through its website (http://ucsc.cmb.ac.lk/ltrl/).

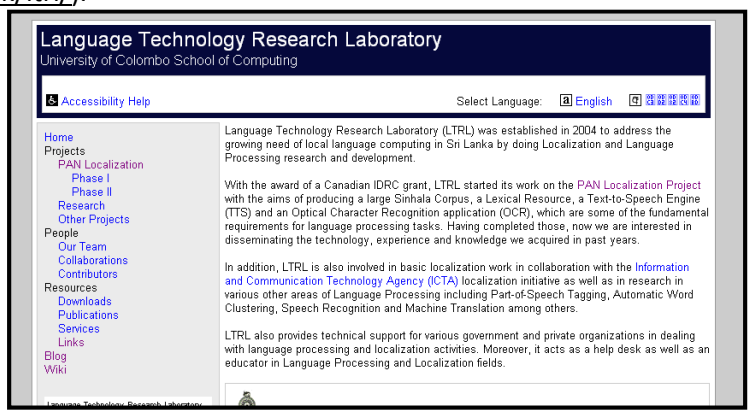


Figure 52: Website of LTRL

The country team has been at the centre of efforts to develop local content. Sri Lanka component conducted several trainings for University staff and fresh graduates to disseminate their research outputs. The training participants were trained to write the articles to Sinhala Wikipedia and blogs. The main objective was to the increase the local language contents of the web. University staff was from the University of Sabaragamuwa while students were following computer awareness program under IRQUE project (http://www.irque.lk). Trainings for university staff was held in two stages. One workshop was conducted for academic staff and at the other was conducted for non-academic staff. These trainings covered the topics including Sinhala Unicode, Structure of internet and Web 2.0, Wikis and blogs. The country component developed a Language Learning Tool (LLT) in Sinhala and English to facilitate Tamil language learning. The main objective of this project was to make language learning process less arduous and the framework followed is flexible enough to extend it to be used by other project partners to teach their languages. The tool contains twenty five lessons. Each lesson has dialog, grammar section and exercises. First five chapters describe the basics in the Tamil language. Exercises can be found at the end of each chapter and reviews can be found after every five chapters. The content for this tool was taken from Sinhala translation of "An Introduction to Spoken Tamil" by James W. Gair, S. Suseendirarajah and W. S. Karunatilaka. The translation work was done by Rev. Kadurugamuwe Nagitha Thero, a lecturer at the University of Kelaniya, Sri Lanka. According to the author, the book increases awareness about Tamil language and also gives an idea about its culture. The book is helpful to learn spoken as well as written. Tamil LLT material is available at http://www.panl10n.net/english/OutputsSri-Lanka2.htm.

Figure 53: Main Page of Language Learning Tool

The country team conducted first training on Introduction to Unicode and Sinhala Wikipedia for 55 graduates of University of Colombo on 26 March, 2008. Second training on Introduction to Unicode and Sinhala Wikipedia was conducted for 51 graduates on 10 June, 2008. Third training on Introduction to Web 2 and its application in local languages was conducted for Academic staff, University of Sabaragamuwa, Sri Lanka on 12 June 2008. 30 participants were trained in this training. Fourth training was conducted on Introduction to Web 2 and its application in local languages was conducted for Non-Academic staff, University of Sabaragamuwa, Sri Lanka on 13 June 2008. 48 participants were trained in this training. Fifth training on Introduction to Unicode and Sinhala Wikipedia was conducted for 24 graduates on 22 July, 2008. Sixth training on Introduction to Unicode and Sinhala Wikipedia was conducted for 25 graduates on 6 March, 2009. Some of the details are available at http://ucsc.cmb.ac.lk/ltrl/blog/.

Figure 54: Non Academic Staff Training at University of Sabaragamuwa on 13 June, 2008

The country team was also part of DAISY related trainings from the beginning, and focused on DAISY compatible digital book production in local languages. Training conducted at University of Peradeniya, at the request of Professor Weerakkody.

The Sri Lankan team also won "Most Innovative Product" award for Sinhala Text-to-Speech System at the Biennial Infotel Trade Exhibition held in Colombo, Sri Lanka on 1 November, 2008 (http://ucsc.cmb.ac.lk/ltrl/blog/). Project partners have been involved in designing, developing and

disseminating the material developed, which has contributed to mutual capacity to disseminate research.

Infrastructure development

In building capacity, infrastructural development plays a vital role. To develop the appropriate localization research infrastructure, funds were needed for acquiring academic resources, e.g. books and journals, and specialized software. In phase I, the team of Sri Lanka utilize funds for different components of the project; Operational field (trained resources), acquisition of the equipments, and books related to different disciplines like linguistics, language processing and computer science. Equipments included PCs. In the operational field, participants regarding different domains of the PAN Localization project were trained. Funds also utilize on software namely Visual Studio.Net and Win Snoori. In phase II, the Sri Lanka country partner institution also utilized funds for purchase of the equipment like PCs, Laptops, PC Upgrades and for arranging trainings. In both phases of PAN Localization project, country component focused on procurement of computer hardware and conducting trainings and available funds were mostly used for these activities. The accessibility of these funds helped developing appropriate localization research infrastructure and enhanced research capacity in Sri Lanka.

Sustainability and continuity

Organizational capacity enhancement as a result of team skill building is another salient factor in measuring the capacity enhancement of teams for sustainability of research. Thus organization has focused on enhancing their knowledge base to gain advancement in other domains of local language computing as well. This has been a contributing factor for organization to acquire more projects on localization technology development. Sri Lanka Country component focused on the development of standardization, Language processing and Script processing during PAN Localization project.

Through PAN Localization project a significant number of technical developers, linguists and social scientists have been trained to enable sustainability and continuity of the research being undertaken. Sri Lanka country component trained 14 participants from different domains like management, technology and linguistics.

7 Discussion:

Prior to initiating the capacity development program, baseline study of the existing research capacity in partner countries was conducted to help in devising the strategies (Hussain, 2004). The study showed that while teams had very limited experience in basic localization (except in a couple of countries) and many countries had no work done in localization policy development and development of intermediate and advanced localization. Also there was hardly any experience in inter-disciplinary research and development across computing, engineering and linguistics. Regarding team management, only two country team leaders had experience in running long-term multi-disciplinary projects.

Faced with the above capacity challenges, appropriate measures had to be undertaken, focusing across the six principles of research capacity building to target holistic improvement. PAN Localization project's regional secretariat had a focused approach on capacity building the teams in ensuring sustainability of the research being conducted. In this context, the project had instituted specific strategies that are detailed below:

7.1 Skill Development Strategies

The skill enhancement of project teams based on the completion of project plans as set out in the country contracts and the number of research publications that have been produced. While the country teams tremendously worked hard to realize the above goals, however the regional secretariat has also steered the overall effort by instituting relevant measures when they were challenged and stopped short of skills to meet the planned research outputs. In this context, four specific strategies were used to facilitate teams in filling the knowledge gap to achieve their deliverables. These are explained in the section below.

Specific strategies employed by the project to meet these ends were:
An outreach component for the project research work was specifically implemented with most of its project partners during the second phase of the project, while the first phase had focused on development of the technology. For this purpose, in the second phase the project also developed partnerships with civil society organizations to specifically focus on dissemination of technology to end users in the partner countries, with explicit funding allocations to support the partnerships. For example, Nepalinux developed by Madan Puraskar Pustakalaya (MPP) was used by E-Network Research and Development (ENRD) to train five rural communities around Nepal, which included farmers, mothers' group, and retired army men. They used the Nepali language applications to communicate with their relatives abroad and to develop online community portals

The partnerships have enabled partners focusing on outreach to appreciate technical challenges and helped the technical partners to appreciate the end user dissemination and adoption challenges. Both lessons significant for planning research they would undertake in the future.

7.1.1 Training through Summer School in Local Language Computing

RS had envisioned that the project plans and contracted deliverables in PAN Localization project phase 1 were focusing more on basic or intermediate complexity localization software applications. However in project's phase 2, the teams were developing more advanced localized applications for which teams would require advanced skills in local language computing and sound theoretical background.
Thus to facilitate the teams in such development, an innovative form of their technical training called Summer School in local language computing was organized during the time when the country teams were transiting from completion of the phase 1 project outputs and the inception of the project phase 2. This was a semester equivalent (three month long) extensive academic program with credit for five graduate courses in linguistics and computational linguistics that were not offered in the partner countries. This program was only offered to those team members who agreed to work on the project for at least one year after the completion of the training. In addition, as an incentive to go through the training, it was arranged in collaboration with other universities in South Asia that the credit hours earned through the semester were transferrable in any other graduate program in those universities. The course instructors selected to teach these courses were experts in their fields chosen from around the world. This helped quickly boost the capacity of the partner teams, enabling the transition from undertaking research in localization in Phase 1 to more advanced research in language computing in Phase 2.

7.1.2 Short Term Training

Short term training was another capacity building strategy designed by RS. Short term training were organized as a week-long activity targeting training on a specific research area in which the team was

lacking in technical competence. **Six** short term training were conducted during the project, covering a varied set of topics, for example, FOSS localization, OCR development, linguistics and monitoring and evaluation using outcome mapping framework.

As a specific example a five day long training of the Afghanistan country team was held in Pakistan. http://panl10n.net/english/Afghanistantraining.htm During this training, the team was trained on outcome mapping framework for planning, monitoring and evaluating their projects. In addition specific sessions were conducted on font development and open source software localization that helped the team to initiate their country project.

In addition to building individual's capacity, this mode of training also helped build institutional capacity. Trainees receiving the short term training were not limited to project staff only but would also include additional relevant staff where this training was organized.

7.1.3 Mentor Placement Program

Where the country team required longer training to address capacity challenges, mentor placement programs were initiated through the RS. Through this program a mentor would be place with the partner country that would provide technical and management support to the recipient team. Two different models were adopted in this context. In first model (referred to as mentor Placement I in Table 2), a mentor from within the partner countries was sent to partner needing support. Three such mentor placements were conducted from 2004-2007, and 2 were held during the second phase of the project. With the mentor placement held in Bhutan during the project phase 1, country team was guided by Mr. Guntupalli Karunakar, on localization of Linux and Open Office in Dzongkha. Following this training, the country team was able to develop a live debian based Dzongkha Linux distribution.

In second model (referred to as mentor Placement II in Table 2) respective country component nominated one or two persons from team to stay with mentoring organization for the training duration. One such placement was initiated in the project's first phase of the project, while 5 such placements were done in the Phase 2. While both models worked out equally well, an extension of first model has also been tried by providing the remote mentoring facility after completion of training, which has also proved effective in achieving the research outcomes.

7.1.4 Support to Present at Workshops and Conferences

As presented above a number of research publications have been produced by the partner countries. This was a testimony of the maturity of their research skills developed during the project. As a strategy for motivating teams to produce more publishable research, RS provided teams with an incentive for covering the conference registration expenses as well as their travel and stay for presenting the research paper at the conference. This strategy significantly motivated teams to produced publishable research. The following table summarizes the number of times each intervention was conducted during the project's phase 1 and phase 2.

Training Strategy	Phase 1 (2004-07)	Phase 2 (2007-10)
Short Term Training	6	-
Mentor Placement (I)	3	2
Mentor Placement (II)	1	5

Summer School	1	-
Conference Participation	12	40

Table 18: Capacity Building Interventions during Project Phases 1 and 2

As presented in the table above, it is evident that the project's Phase 1 focused more on short term training and mentor placements I. However as the teams gained more in their technical skills the project's Phase 2 strategies targeted collaborations, e.g., Mentor Placement II, summer school and conference participation, and longer term impact, e.g. through support for higher studies.

7.2 Strategies regarding training to conduct to practice research

An outreach component for the project research work was specifically implemented with most of its project partners during the second phase of the project, while the first phase had focused on development of the technology. For this purpose, in the second phase the project also developed partnerships with civil society organizations to specifically focus on dissemination of technology to end users in the partner countries, with explicit funding allocations to support the partnerships. For example, Nepalinux developed by Madan Puraskar Pustakalaya (MPP) was used by E-Network Research and Development (ENRD) to train five rural communities around Nepal, which included farmers, mothers' group, and retired army men. They used the Nepali language applications to communicate with their relatives abroad and to develop online community portals. Similarly, Pakistan component collaborated with District Governments of Sargodha, Chakwal and Attock to deploy localized open source applications in ten rural schools, training more than 200 school students and teachers on information access, communication and content generation.

The partnerships have enabled partners focusing on outreach to appreciate technical challenges and helped the technical partners to appreciate the end user dissemination and adoption challenges. Both lessons significant for planning research they would undertake in the future.

7.3 Strategies for the development of linkages

As a strategy to develop international collaborations, RS has been organizing regional training, conferences and workshops, in which experts from the region are invited. These have provided opportunities to meet and discuss opportunities for collaboration. As a salient example, project partners have been interacting with NECTEC, Thailand, which have eventually resulted in formal bi-lateral and multi-lateral partnerships. The project has also worked with researchers from Korea, India, Japan and regional organization like Asian Federation of Natural Language Processing. Such interactions have also resulted in direct partnerships between Microsoft and country partners resulting in the development of Language Interface Packs (LIP) for MS Office and Windows in Urdu, Pashto, Bangla, Sinhala, Khmer, and Lao, by the project partner countries.

7.4 Strategies to disseminate research work

As a continued effort to enhance country team's capacity to disseminate research work, Regional secretariat has been channeling the regional project funds for research publication and dissemination to country teams for organizing their national events as well. In this regard, the initial Nepalinux and Dzongkha Linux release events were organized through additional funding provided to Nepal country component by diverting centrally available un-spent funds to the country components. Similarly regional marketing budget has been mobilized to bear expenses for organizing country awareness seminars and printing and publication of project promotional materials.

7.5 Strategies for the sustainability of the research work

PAN Localization project's regional secretariat had a focused approach on capacity building the teams in ensuring sustainability of the research being conducted. In this context, the project had instituted specific strategies that are detailed below:

7.5.1 Support for Higher Studies

As a strategy for continued and advanced training in local language computing, the project provided completed or partial scholarships for many team members for pursuing higher studies in disciplines related to localization research. Specific researchers were funded in Pakistan, to accomplish their academic research through working on Project. In addition, project facilitated these team members by providing time for studies and examinations and in certain instances by supporting the tuition fee for their degree program. This approach improved the organizational research capacity by having more trained resources in the local language computing domain. While it was also used for retention of researchers, as these team members would remain with the Project until degree completion

7.5.2 Developing research centers: to advance Research Capacity

For sustainability and advancement of localization research, fully equipped research labs were developed, with equipment, specialized resources, including software, books, etc. such that country projects housed within universities and government institutions that would continue the research beyond the project duration. Dedicated research centers were established through the project, for example. In Pakistan, Center for Language Engineering at University of Engineering and Technology was established. This has instigated further localization collaboration and research.

7.5.3 PAN L10n Multilingual Chair in Local Language Computing

To sustain and consolidate the regional momentum of localization research capacity building initiated through the project, a permanent research chair for multilingual computing has been established at the project regional secretariat in Pakistan funded by International Development Research Center (IDRC), Canada. Establishment of this research chair would provide a sound foundation to sustain, nurture and grow the network of localization researchers, and to provide direct support in language computing community, including researchers and policy makers.

8 Recommendations on RCB Model for localization

Though the six principles of RCB holistically address the challenge of RCB in localization, however appropriate sequencing within the six principles must be done in order to foster maximal impact (Potter & Brough, 2004). Based on the project experience RCB for localization must follow a developmental cycle within the defined capacity building principles.

Initially *Skill Building* and *Infrastructure development* must form the focus of RCB interventions for localization. Different countries require different research skills and infrastructure needs owing to the existing competencies. Thus a participatory need analysis should be performed to ensure skill development is based on national priorities and capacity. Based on the identified RCB needs, appropriate mentorship structure and organizational resources may be planned to ensure development of the research base. Secondly, localization RCB can be built to carry out *close to practice research*, of direct benefit in practice, after development of technology the basic and intermediate levels of localization research as a pre-requisite, only then can the research be conducted that harness solutions that can be readily used by the benefitting populations. At the same time, capacity building initiatives must target to form *linkages and partnership* with relevant academic, policy making, regional

standardization bodies, and public and private sector bodies. This would follow skill development at level 1 as synergetic and mutually benefitting collaborations can only be developed if the local teams are able to contribute back to the knowledge network. Finally, *research dissemination and sustainability* must be targeted as these RCB dimensions provide research maturity to publish in technical forums, and compete for funding.

9 Conclusion

The existing ICT human resource capacity indicators signify a steep demand for localization skills in ICT professionals with the increasing ICT diffusion in the Asia Pacific (Rhee and Riggins, 2007, Raina, 2007). UN-APCICT/ESCAP (2010) speculates that existing institutions for ICT education and training in the region cannot fulfill this sharply increasing demand. Localization is the most effective tool for providing ample opportunity to the communities to avail the benefits of ICT revolution. It is, therefore, utmost necessary that Localization RCB is taken up as a national and regional priority in order to bridge the demand- supply gap of the required localization in ICT in developing Asia.

Pan Localization project aimed at building the requisite capacity in localization of Information and Communication Technology. Under this project various strategies based on the six principles laid down in the Research capacity building framework (RCB) were devised to build the capacity of the local professionals of the partner countries in local language computing. The performance indicators to assess the accomplishment in various tasks to be carried out by the country components were determined by the regional secretariat of the PAN Localization project.

The need of capacity building in different partner countries was different, depending upon their respective social and political environment that has been prevailing in the immediate past as well as their literacy rate. Each country component was therefore required to transform the available software in to local languages, train the end users on the use of localized software and create a team of trainers for dissemination of knowledge on sustained basis. The achievement of various country components when measured according to specified performance indicators shows that the implementation of the project definitely resulted in enhancement of the partner countries' capacity in accordance with each principle of RCB model.

References

Ansari,S., & Saleem,S.(2009). Pakistan. In Shahid Akhtar & Patricia Arinto (Eds.), *Digital Review of Asia Pacific*, 2009-10 (pp. 294-301).New Delhi: Sage Publications India Pvt. Ltd..

Ariunaa, L., & Uyanga,S .(2009). Mongolia. In Shahid Akhtar & Patricia Arinto (Eds.), *Digital Review of Asia Pacific ,*2009-2010(pp. 268-273) New Delhi: Sage Publications India Pvt. Ltd.

Baliamoune-Lutz, Mina. (2003). An Analysis of the Determinants and Diffusion of ICTs in Developing Countries. *Information Technology for Development*, 10:151–169

Bayanduuren, D. (2007), PAN Localization Project Phase II, Mongolian Speech Recognition. Retrieved 2010, December 21from www.**pan**l10n.net/Presentations/Bhutan/**Phase2**/MUST.pdf

Boyle,H.P.S. (2008). A Multidisciplinary Model of Evaluation Capacity Building. America Journal of Evaluation(2008)1-17

Breen, C. M., Jaganyi, J. J., Van Wilgen, B. W., and Van Wyk, E. (2004). Research Projects and Capacity Building. *Water SA*, 30(4):429-434.

Bureau of Democracy, Human Rights, and Labor. (2010). Human Rights Report" in U.S Department of state. Retrieved 2011, December 20 from http://www.state.gov/g/drl/rls/hrrpt/2010/sca/154477.htm]

Bureau of East Asian and Pacific Affairs .(2011). Background Note: Laos in U.S department of state.[2010, December 20] Retrieved 2011, December 19 from http://www.state.gov/r/pa/ei/bgn/2770.htm

Connolly,P., & York, P. (2002). Capacity-Building efforts for Nonprofit Organization. 34: 33-39

Cooke, Jo. (2005). A Framework to Evaluate Research Capacity Building in Health Care. *BMC Family Practice*, 6(44)

Department for International Development (DFID). (2008), Research Strategy 2008-2013. Working Paper Series: Capacity Building. Retrieved 2011, December 19 from http://www.dfid.gov.uk/r4d/PDF/Outputs/Consultation/ResearchStrategyWorkingPaperfinal_capacity_P1.pdf

Department for International Development (DFID). (2010), *How to note Capacity Building in research. Retrieved 2011, December 20 from growthandemployment.org/.../...*

Donny B.U., & Mudiardjo,R.(2009). Indonesia. In Shahid Akhtar & Patricia Arinto (Eds.), *Digital Review of Asia Pacific*,2009-2010 (pp. 201-209).New Delhi: Sage Publication India Pvt. Ltd.

Earl, Sara., Carden, Fred., & Smutylo, Terry. (2001), Outcome Mapping: Building Learning and Reflection into Development Programs. Ottawa, Canada: International Development Research Centre.

Government of Bangladesh and United Nation. (2005), Millennium Development Goal. Retrieved 2011, December 21from *www.searo.who.int/LinkFiles/MDG_**Reports_Bangladesh**MDG.pdf*

Gul, Sana. (2004). Dilemmas of Localisation in Asia. *I4D Online*, 2(12)

Harris, Eva. (2004). Building Scientific Research Capacity in Developing Countries. *European Molecular Biology Organization Reports 2004*, 5(1):7-11

Hussain, Sarmad. (2004). Developing Local Language Computing. *I4D Online*, 2(6)

Hussain, Sarmad., & Gul, Sana. (2004). Localization in Pakistan. Localisation Focus: *An International Journal for Localization*, 3(4)

Hussain, Sarmad, Gul, Sana., & Waseem, Afifah. (2007). Developing Lexicographic Sorting: An Example for Urdu. *ACM Transactions on Asian Language Information Processing (TALIP)*, 6(3)

Hussain, Sarmad., & Mohan, Ram. (2007). Localization in Asia Pacific. in Librero, Felix (Eds.), *Digital Review of Asia Pacific*, 2007-08 () New Delhi: Sage Publications India Pvt. Ltd.

International Telecommunication Union (ITU), (2011). The World in 2010, ICT Facts and Figures. Reterived from http://www.itu.int/ITU-D/ict/material/FactsFigures2010.pdf

Jurmi,K., & Wangchuk, S. (2009). Bhutan. In Shahid Akhtar & Patricia Arinto (Eds.), *Digital Review of Asia Pacific*, 2009-2010(pp152-159) New Delhi: Sage Publications India Pvt. Ltd.

Lennie. (2005). *An evaluation capacity-building process for sustainable community IT initiatives. Evaluation,* 11(4),390-414.

Lewis, M. Paul. (2009). *Ethnologue: Languages of the World*. SIL International.

Marjan,M.A. (2010). Afghanistan. In Shahid Akhtar & Patricia Arinto (Eds.), *Digital Review of Asia Pacific*, 2009-2010(pp. 129-134) New Delhi: Sage Publications India Pvt. Ltd.

Naccarella, L., Pirkis, J., Kohn,F., Morley,B., Burgess,P., & Blashki,G. (2007). Building evaluation capacity: Definitional and practical implication from an Australian case study. *Evaluation and Program Planning* 30 (3),231-236.

Neilson, Stephanie., & Lusthaus, Charles.(2007). *IDRC Supported Capacity Building: Developing a Framework for Capturing Capacity Changes.* Universalia.

Nikolov, Roumen., & Illieva, Sylvia.(2008). *A Model for Strengthening the Software Engineering Research Capacity*, SEESE'08, Germany.

Phissamay,P.(2009), Laos. In Shahid Akhtar & Patricia Arinto (Eds.), *Digital Review of Asia Pacific*, 2009-2010(pp. 241-248) New Delhi: Sage Publications India Pvt. Ltd.

Pimienta, Daniel. (2005). Linguistic Diversity in Cyberspace – Models for Development and Measurement, in UNESCO Institute for Statistics (eds), *Measuring the Linguistic Diversity on the Internet*, UNESCO

Potter, C., & Brough, R. (2004). Systemic Capacity Building: A hierarchy of Needs. *Health Policy & Planning,* 19(5),336-345.

Powell, E T., & Boyd, H. H. (2008). Evaluation capacity building in complex organizations. *Program evaluation in a complex organizational system: Lessons from Cooperative Extension*. New Directions for Evaluation, 120, 55–69.

Raihan, A. (2009), Bangladesh. In Shahid Akhtar & Patricia Arinto (Eds.), *Digital Review of Asia Pacific*, 2009-2010(pp. 144-151) New Delhi: Sage Publications India Pvt. Ltd.

Raina, Ravi. (2007), ICT Human Resource Development in Asia and the Pacific: Current Status, Emerging Trends, Policies and Strategies. UN-APCICT. Retrieved from http://www.unapcict.org/ecohub/resources/ict-human-resource-development-in-asia-and-the

Rhee, Hyeun-Suk., & Riggins, Frederick J. (2007). Development of a Multi-Factor Set of Country-Level ICT Human Resource Capacity Indicators. UN-APCICT. Retrived from http://www.unapcict.org/ecohub/resources/development-of-a-multi-factor-set-of-country-level

Shams, Sana., Hussain, Sarmad., & Mirza, Atif. (2010). Gender and Outcome Mapping, In Belawati, T. and Baggaley, J. (Eds) *Policy and Practice in Asian Distance Education*, New Delhi: Sage Publication India Pvt. Ltd.

Sorasak, P., & Konsona, C.(2010), Cambodia. In Shahid Akhtar & Patricia Arinto (Eds.), *Digital Review of Asia Pacific*,2009-2010(pp.167-174)New Delhi: Sage Publications India Pvt. Ltd.

Scenic Sri Lanka (2007). Retrieved 2011, December 21 from http://www.scenicsrilanka.com/people-of-sri-lanka.html

UN –APCICT, 2007. (United Nations Asian and Pacific Training Centre for Information and Communication Technology for Development). Retrieved 2011, December 20 from http://www.unapcict.org/member-countries/indonesia

View of Cambodia. 2011, Retrieved 2011 December 20 from http://www.cambodianview.com/khmer-language.html

Wattegama,C. (2011). ICT Sector Performance review for Indonesia.[2011, December 20] *lirneasia.net/wp.../ID_SPR_**Indonesia**_Wattegama_revised-ver-1.pdf*

Weerasinghe, R and de Silva,C.(2010), Sri Lanka. In Shahid Akhtar & Patricia Arinto (Eds.), *Digital Review of Asia Pacific*,2009-2010(324-334)New Delhi: Sage Publications India Pvt. Ltd.

Wibberley, C., Dack, L. M. F., & Smith, M. (2002). Research-minded Practice in Substance (mis) Use Services. *Journal of Substance* Use, 7(1):19-23.

Wignaraja, Kanni. (2009). *Capacity Development: A UNDP Primer.* Retrieved from http://content.undp.org/go/cms-service/download/asset/?asset_id=2222277

Wing,T.K. (2004). Assessing the effectiveness of Capacity-Building Initiatives: Seven Issue for the Field. vol.33, 2004 153-160

World Bank. (2002). Information and Communication Technologies: A World Bank Group Strategy. World Bank Group.

Appendix A

Country-Wise list of Research Publications through PAN Localization project

Bangladesh

I. Khan Md. Anwarus Salam, Mumit Khan and Tetsuro Nishino, "Example Based English-Bengali **Machine Translation** Using WordNet", TriSA 2008, Japan.

II. Md. Abul Hasnat, Muttakinur Rahman Chowdhury and Mumit Khan, "Integrating Bangla Script Recognition Support in Tesseract **OCR**", Proc. of Conference on Language and Technology 2009 (CLT09), Lahore, Pakistan, January 22-24, 2009.

III. Md. Abul Hasnat and Mumit Khan, "Rule Based **Segmentation** of Lower Modifiers in Complex Bangla Scripts", Proc. of Conference on Language and Technology 2009 (CLT09), Lahore, Pakistan, January 22-24, 2009.

IV. Md. Abul Hasnat and Mumit Khan, "**Elimination of Splitting Errors** in Printed Bangla Scripts", Proc. of Conference on Language and Technology 2009 (CLT09), Lahore, Pakistan, January 22-24, 2009.

V. Firoj Alam, S. M. Murtoza Habib and Mumit Khan. **Text Normalization System** for Bangla. Conference on Language and Technology 2009 (CLT09), NUCES, Lahore, Pakistan, January 22-24, 2009. [poster]

VI. Altaf Mahmud, Kazi Zubair Ahmed and Mumit Khan, "**Detecting Flames and Insults in Text**", Proc. of 6th International Conference on Natural Language Processing (ICON-2008), CDAC Pune, India, December 20 - 22, 2008.

VII. Farhana Faruqe and Mumit Khan. "BWN - A Software Platform for Developing Bengali **WordNet**", International Joint Conferences on Computer, Information, and Systems Sciences, and Engineering (CISSE 08), December 5 - 13, 2008.

VIII. Firoj Alam , S. M. Murtoza Habib and Mumit Khan. "**Acoustic Analysis of** Bangla Consonants", Spoken Language Technologies for Under-resourced language (SLTU'08), Hanoi, Vietnam, May 5 - 7, 2008.

Indonesia

I. Adriani, Mirna, Ruli Manurung, and Femphy Pisceldo. **Statistical Based Part Of Speech Tagger for Bahasa Indonesia**. Workshop, Co-located Event ACL-IJCNLP 2009. Singapore, August 1, 2009. Third MALINDO International MALINDO.

II. Budiono, Hammam Riza, Chairil Hakim. Resource Report: Building **Parallel Text Corpora for Multi-Domain Translation** System. 7th Workshop on Asian Language Resource, ACL-IJCNLP 2009, Singapore, August 2009.

Mongolia

I. A.Altangerel, Journal Scientific Transactions. A Design and Implementation Mongolian **Speech** Recognition System. MUST № 5/102, 2008.

II. J. Purev and Odbayar. **Corpus B**uilding for Mongolian Language. The 6th Workshop on Asian Language Resources, 2008.

III. Purev Jaimai and Odbayar Chimeddorj. **Corpus B**uilding for Mongolian Language. Proceedings of the 6th International Workshop on Asian Language Resources (ALR)- Jan 11-12, 2008, Hyderabad, India.

IV. Purev Jaimai, Tsolmon Zundui, Altangerel Chagnaa, and Cheol-Young Ock. PC-KIMMO-based Description of Mongolian Morphology. International Journal of Information Processing Systems, Vol. 1 (1), pp. 41-48. (2007).

V. Purev Jaimai and Odbayar Chimeddorj. **Part of Speech** Tagging for Mongolian Corpus. 4th International Joint Conference on Natural Language Processing. (IJCNLP). The 7th Workshop on Asian Language Resources. August 2-7, 2009, Singapore.

VI. Purev Jaimai and Odbayar Chimeddorj. **Resources for Mongolian Language**. Proceedings of the 3rd International Universal Communication Symposium. December 3-4, 2009, Tokyo, Japan.

Nepal

I. Bal Krishna Bal.Towards Building Advanced Natural Language Applications – An Overview of the Existing Primary Resources and Applications in Nepali. Proceedings of the 7'th Workshop on Asian Language Resources, Association for Computational Linguistics, Suntec, Singapore, August, 2009, pp.165-170.

Pakistan

I. Hussain, S., Gul, S., Waseem,A. Developing lexicographic sorting: An Example for Urdu. In ACM Transactions on Asian Language Information Processing (TALIP), Volume 6 Issue 3, 2007.

II. Hussain, S. Resources for Urdu Language Processing. In the Proceedings of the 6th Workshop on Asian Language Resources, IJCNLP'08, IIIT Hyderabad, India, 2008.

III. Hussain, S., Karamat N., Mansoor, A. Arabic Script Internationalized Domain Names. In the Proceedings of the CIIT Workshop on Research in Computing, CWRC'08, CIIT Lahore, Pakistan, 2008.

IV. Sana Shams. Gendered Outcome Mapping. Outcome Mapping Learning Community Newsletter 2009. No.2.

V. Sana Shams, Sarmad Hussain and Atif Mirza, Gender and Outcome Mapping. Published in PANdora Distance Education Guidebook (1st edition), 25th October 2008.

Research Publications

I. Ruvan Weerasinghe, Asanka Wasala, Viraj Welgama and Kumudu Gamage. Festival-si: A Sinhala Text-to-Speech System. Proceedings of Text, Speech and Dialogue, 10th International Conference, TSD 2007, Pilsen, Czech Republic, September 3-7, 2007.

II. Ruvan Weerasinghe, Asanka Wasala and Samantha Mathara Arachchi. Facilitating Information Accessibility for the Print Disabled. Diriya 2007 - a Conference on "Mainstreaming Disability into Development". Colombo, Sri Lanka.

III. Asanka Wasala, Ruvan Weerasinghe. EnSiTip: A Tool to Unlock the English Web. 11th International Conference on Humans and Computers, Nagaoka University of Technology, Nagaoka, Japan, 20-23 November 2008.

IV. Ruvan Weerasinghe, Asanka Wasala, Dulip Herath and Viraj Welgama. NLP Applications of Sinhala: TTS & OCR. 3rd International Joint Conference on Natural Language Processing. (IJCNLP). Exhibitions & Demonstration Session. January 7-12, 2008, Hyderabad, India.

V. Harsha Wijayawardana, Asanka Wasala, Ruvan Weerasinghe and Chamila Liyanage. Implementation of Internet Domain Names in Sinhala. International Symposium on Country Domain Governance. Nov, 20-22, 2008, Nagaoka, Japan.

VI. Silva, A. M. and Weerasinghe, A. R. Example Based Machine Translation for English-Sinhala Translations. In Proceedings of the 09th International IT Conference (IITC 2008), Colombo, Sri Lanka, 27-28 October 2008.

VII. Asanka Wasala, Ruvan Weerasinghe, Randil Pushpananda, Chamila Liyanage and Eranga Jayalatharachchi . An Open-Source Data Driven **Spell Checker** for Sinhala. e-Asia 2009. Colombo, Sri Lanka, 2-4 December 2009.

VIII. Weerasinghe, A. R., Liyanapathirana, J. U., Asanka Wasala, Dulip Herath, Viraj Welgama . **OpenTM:** A Translation Memory System for Complex Script Languages. International Conference on Machine Translation Twenty-Five Years On, Bedfordshire, UK, 21-22 November 2009.

IX. Ruvan Weerasinghe, Dulip Herath and Viraj Welgama**. Corpus-based Sinhala Lexicon**. 4th International Joint Conference on Natural Language Processing. (IJCNLP). The 7th Workshop on Asian Language Resources. August 2-7, 2009, Singapore.

X. Ruvan Weerasinghe, Asanka Wasala and Kumudu Gamage. A Rule Based **Syllabification Algorithm** for Sinhala. 2nd International Joint Conference on Natural Language Processing (IJCNLP-05), Jeju Island, Korea, 2005.

XI. Asanka Wasala, Ruvan Weerasinghe and Kumudu Gamage . Sinhala **Grapheme to Phoneme** Conversion and Rules for Schwa Epenthesis. In Proceedings of the COLING/ACL on Main Conference Poster Sessions (Sydney, Australia, July 17 - 18, 2006). Annual Meeting of the ACL. Association for Computational Linguistics, Morristown, NJ, 890-897.