PAN Localization Project Phase II, Bhutan Component: Project Signing off Report

Pema Choejey, Tenzin Dendup

Department of Information Technology & Telecom

{pchoejey@dit.gov.bt, tdendup@dit.gov.bt}

Abstract

This document presents the project sign off report for the PAN Localization Project Phase II. Broadly the project aimed to build effective solutions to enable local language computing, to create, develop and deploy local content, to build human resources capacity for technology and content development and to drive policy changes to support local language development and enhance access to local language content and technologies. The project is funded by International Development Research Center (IDRC), Canada, administered by National University of Computer and Emerging Sciences (NUCES), Lahore, Pakistan and implemented by Department of Information Technology (DIT), Ministry of Information and Communications (MoIC).

1. Background

The tone for PAN Localization Project - Phase II has been set because of immense success of PAN Localization Project – Phase I. To the extent possible, Phase I has contributed significantly in raising regional awareness, capacity development, creating social networks of researchers and practitioners and building institutional support for local language development and access. While the focus of Phase I was on Access, PAN Localization Phase II not only aimed to consolidate and build on whatever has been achieved in Phase I but also focused on building complete info-structure for effective Access, Use and eventual Technological Appropriation.

Phase II Project for Bhutan Component started in 2007 and is to be completed by end of January 2010. Broadly, the project intended to achieve following objectives:

a) Examine effective means to develop digital literacy through use of local language computing and content.

b) Explore development of sustainable human resource capacity for research and development in local language computing as a mean to raise current levels of technological support for Asian languages.

c) Advance policy for development and use of local language computing and content.

d) Study and develop coherent instruments to gauge the effectiveness of multidisciplinary research concerning the adoption of local language technology by rural communities.

The major outcomes and deliverables expected at the end of second phase are:

- International Domain Names (IDNs) for Dzongkha
- Text to Speech system for Dzongkha
- Dzongkha Speech corpus
- Dzongkha Text Corpus
- Dzongkha Lexicon
- Training on usage of localized tools and software already developed during phase I of the project
- Application of Gender Evaluation Methodology within the framework on OM
- Local Language content development

2. Research Goals

2.1 Technology Development

Continue to develop standards to enable local language computing. Continue to advanced applications such as text-to-speech synthesis systems, optical character recognition, lexicon, text corpora and other resource. Conduct further research into localization of emerging mobile and open source platforms.

2.2 Development and Conduct of Training

Conduct survey on training need analysis, training planning and design, training conduction and assessment of training methods and strategies. Develop and trial reuse of training materials in local language for different end-user groups. Conduct end user training to determine effective strategies to generate culturally relevant content in local languages.

2.3 Content Development and Deployment

Investigate methods to determine end-user local language content requirements. Investigate challenges associated with development of local language content. Experiment with technological and social solutions for deployment of local language content.

2.4 Development of Sustainable Human Resource Capacity for R&D

Continue to develop human resources to conduct research and development in local language computing and its impact on ICT access. Continue to develop human resources to create, generate, translate and develop content in local language. Continue to develop human resources to use OM and GEM II methodologies to monitor and evaluate success of local language related projects.

2.5 Culture and Policy Promotion for Local Language Computing

Drive policy changes to support and promote local language technology, local language content and human resource capacities in local language. Establish regional network of researchers, practitioners and policy makers for collaborative learning in local language computing. Explore gender issues related to software development, software localization, training and content development. Evolve policy framework for local language computing through rigorous research publication program.

2.6 Measurement and Evaluation

Develop capacity to use OM and GEM II methodologies for project planning, monitoring and evaluation. Use adapted OM framework for localization related projects. Evaluate success of project against performance indicators and outcomes set at the beginning of the project implementation.

2.7 Marketing and Dissemination

Conduct marketing on local language technology and software to promote and create awareness among different end-user groups. Disseminate information on local language products through publications of pamphlets, research papers, advertisement, print media and training materials and end-user manuals.

3. Research Findings

3.1 Technology Development

As we went along project implementation, we have realized that lack of qualified professionals in open source software and technologies do often create hiccups. However, we were able to overcome such hindrances by providing appropriate and effective training to researchers engaged in research studies. Another problem is the lack of sufficient resources on Dzongkha language. We also had difficulties in finding very good and qualified local language and computational linguistics experts locally. In addition, we also found that collaboration and coordination among stakeholders and country partner institutions such as DDC and Sherubtse College is equally important smooth implementation. Furthermore, localization of applications is a continuous process which may not be sustainable in the long term if the qualified and experienced professionals leave the department.

3.2 Development and Conduct of Training

We had a major setback in development and conduction of training. This happened because no training need analysis was conducted. No baseline survey on Dzongkha Debian Linux end-user groups was conducted at start of the project implementation. There was also no proper planning and implementation of training programs. Assessment of training impacts after the training of training of trainers and end-user groups was also neglected. This could have happened in absence of dedicated project manager as the former project manager resigned and left the department. Delay in undertaking the training programs for OM and GEM II methodologies may be another factor where we were not able to properly assess the success of training except taking feedbacks from trainees which give very little clue on the impact of the training.

3.3 Content Development and Deployment

There was not much problems in design and development of content website. The only issue that we had was lack of local language content. Most of the Dzongkha contents are available in printed form and those available in soft copies are also written in legacy systems using legacy fonts which are incompatible and non portable to the Unicode. Creation of content by translation from English content to Dzongkha and typing manually into word application is a laborious task. Non willingness to provide content by the agencies that have content was another issue. In addition, we faced problems in rendering local content in browsers which do not have support for rendering Dzongkha.

3.4 Development of Sustainable Human Resource Capacity for R&D

Areas of research and development do require highly qualified people with specialization in specific research fields. Conducting research studies in basic language computing software and technology is not a big issue as it can be done by graduates who have bachelor degrees. However, when it comes to advanced applications such as machine translation, speech recognition and language assistive technologies, it does require qualified people at the level of PhDs. While provision of short term intensive training on particular areas does solve temporary difficulties, it is definitely not sustainable because this training is very specific and focused to one research area. Therefore, we feel that for long term sustainability, we need to have graduates with PhDs who can conduct independent research studies, mentor and supervise other researchers.

In order to sustain the human resources capacity, it is essential to establish networks and affiliation with other researchers and institutions that can provide technical support and guidance. Joining open communities and access to resources is also essential.

Another element of sustaining human resource capacity for research and development is the retention of qualified and experienced people. Retention of existing researcher in the department is going to be too difficult. At least in this department, long term sustainability of human resources is questionable. This is because at any point of time people who have been trained in local language development may leave and move to corporate and private sector.

3.5 Culture and Policy Promotion for Local Language Computing

Changing culture and policy promotion for local language computing is too daunting. People do show lots of reluctance to change their mindset and accept new things. The fundamental reasons for not readily accepting open source software or localized software and applications may be due to: a) In government administration 99 percent of the people do use proprietary software notably MS products, b) Open source software or localized software are seen not user friendly to use and c) small number of open source user base to influence the adoption of local language in mainstream governance.

3.6 Measurement and Evaluation

While we know that we achieved most of the expected project deliverables, we also know that we failed to achieve some of the outcomes. We were not able to measure and evaluate both tangible and intangible outcomes particularly in conduction of training, gender evaluation and policy matters. As stated earlier, we failed to conduct training need analysis, conduct baseline Dzongkha Linux user survey and survey on diffusion of local language software and applications. Our inability may be due to lack of knowledge and skills in planning and implementation of training and delay in getting appropriate training on OM and GEM II methodologies.

3.7 Marketing and Dissemination

If products have to sell like hot pizza marketing is the mantra. Within the project, we have done our best to raise awareness and promote Dzongkha Debian Linux and Open Office applications by having formal launching ceremony with wide media coverage. Information on local language software and applications has been disseminated not only using broadcast and print media, but also using pamphlets, brochures, locally designed T-shirts, training and users manuals.

4 Project Outputs

Optical Character Recognition

Dzongkha Optical character recognition based on Google's tesseract recognition engine has been designed and developed. The current system supports recognition of documents created using Jomolhari font. The accuracy of the system ranges from 80-90%. Below is the sample output of Dzongkha OCR.

Text-to-speech Synthesis

Text-to-speech synthesis has been designed and developed in close collaboration with NECTEC, Thailand. The accuracy of initial system based on subjective evaluation is about 3.5 out of 5. The synthesized speech also sounds quite robotic, flat and incoherent compared to human speech.

Text Corpora Database

A small text corpora database has been designed and developed. Text have been sourced from different sources such as online media – like BBS Dzongkha website – from books which are in electronic form, from news media, or from publish books typed manually. Text was classified or categorize into different domains, genres, styles and others. For example, a particular text on sports as domain may be categorise into indoor or outdoor sports and may fall into different genres like football, volleyball, lawn tennis, etc.

The current text corpora database contains about 4,00,000 words and have been further annotated with POS tagsets. The annotated corpora is being used for TTS development.

Word Segmentation

Dzongkha word segmentation algorithm has been designed and developed based on combined techniques of maximal matching and bi-grams methods. While the accuracy and performance of the system is good, it is fully dependent on lexical database and text corpora. Below is the sample out of word segmentation.

POS Tag Sets

Part of speech tagging, also called the grammatical tagging is defined as "The process of assigning a part of speech or other lexical class marker to each word in a corpus" [Jurafsky 2000] or "The process of marking up the words in text."

We have identified 45 POS tag sets to tag or mark the Dzongkha text. POS tag sets are essential for text annotation in Text-to-Speech Synthesizer, Word Segmentation, creating and building Corpora of Dzongkha text.

IDN

IDN stands for International Domain Names. It basically refers to domain names in different languages and scripts. While conducting research on IDN, Dzongkha character sets which are valid to be used for this application have been identified. Generic top level domains (gTLD) and country code top level domains (ccTLD) have been translated in Dzongkha. IDN test cases have been conducted to verify and test the IDN application to Dzongkha language.

Dzongkha Lexicon

Dzongkha lexicon containing 22,950 words have been collected from different Dzongkha dictionaries published by both private sector firms and Dzongkha Development Commission (DDC). Each word in the lexicon contains word meaning, word class, part-of-speech in Dzongkha, word pronunciation based on transcribed IPA. In future, lexicon will be annotated with POS tag sets.

Wordnet for Lexical Database

Preliminary research has been done on Wordnet lexical database. Classification of Dzongkha words into synsets equivalent to English word list and other relations have been partially done.

Content Development in Local Language

Bi-lingual content in local language had been designed and developed to commemorate 100 years of monarchy and 5th King's coronation. The website is called Bhutan 2008 (www.bhutan2008.bt). The website has been developed based on Drupal Content Management System (CMS) which has been fully translated and localized. Below is the sample output of local content development.

Dzongkha Debian Linux Version 3

Updated version of Dzongkha Debian Linux Version 3 based on Debian lenny is also released as part of the project deliverables.

3.1 Software

- Dzongkha Debian Linux Version 3
- Text-to-speech synthesis
- Dzongkha Optical Character Recognition
- Dzongkha Word Segmentation Algorithm
- Dzongkha POS Tagger

3.2 Research Reports

- Research Report on Character Set and Encoding Constrains for Dzongkha IDN

- Test Report on IDNs for Dzongkha
- Dzongkha Text-to-Speech Synthesis
- Dzongkha Part-of-Speech tag sets
- Dzongkha Phonetic Set Description
- Dzongkha Text Normalization
- Country Chapter – Language Processing
- Dzongkha Optical Character Recognition
- Dzongkha Word Segmentation

3.3 Research Publications

A research paper titled "Pioneering Dzongkha Text-to-Speech Synthesis" had been jointly published by researchers in DIT and NECTEC during the International Conference on Speech Database and Assessment, organized by National Institute of Information and Communications Technology (NIC) and Spoken Language Translation Research Laboratories (ATR), from November 25-27, 2009, in Kyoto, Japan.

3.4 Development of Training Material

Manuals on usage of Dzongkha Debian Linux 2.0 are published. The manual consists of two books namely: 1) Dzongkha Debian Linux 2.0 Manual – General and 2) Dzongkha Debian Linux 2.0 Manual – OpenOffice. These books are published in Dzongkha language. English versions of the manuals are also published as PDF in the Department of Information Technology's website.

5 Sustainability of work and Future Prospect

5.1 Human Resource Capacity Building

The project has provided wide range of opportunities from research trainings to conferences to meetings to gained required knowledge, skills and experiences to sustained human resources development within the department. It has also provided platforms to create social networks and affiliation with external institution like NECTEC and country partners involved in the PAN project. It is expected that future human resource capacity building requirements for translation, localization and training of users on basic level of computing in local language will be met from within the department. However, advanced and high level trainings in specific research areas may be required.

5.2 Adoption of Local Language Computing by End Users

Adoption of local language computing should not be left only to researchers, developers and programmers involved in the project. Within the project scope, department has tried its best to promote and infuse the use of local language tools and applications. However, much need to be done beyond. Hence, department is seriously thinking to take local language software and tools in particular and use of open source software in general to government administration bodies and to school education systems.

5.3 Influence on Language Computing Policy

At this moment, it is quite difficult to gauge the influence of project to local language computing policy. Neither department could initiate and nor draft policy for adoption and use of local language software and tools within the project period. However, given its merits compared to proprietary software, department is seriously thinking to draft local language computing policy required, in particular, to support and promote local language software and applications and in general use of open source software.

5.4 Beyond PAN Phase II

Beyond PAN Phase II, the intention of the department is to continue consolidation and system integration work. Support for Dzongkha TTS and OCR applications have to be integrated into system. Department also need to build usable online text corpora database, Dzongkha Wordnet and Dzongkha lexicon applications. In addition, department intends to conduct research studies in areas of speech recognition, machine translation, parallel text corpora, mobile interface localization and text retrieval system.

6 Acknowledgement

We would like to sincerely express our deepest gratitude to IDRC and NUCES for all your generous support and assistance that have been extended to us throughout the project period without which we would not have been in a position to sign off this project successfully. We look forward to IDRC and NUCES for similar landmark projects and research collaboration in future. We would also like to put in record to thank all our country partners and NECTEC for kind support and technical assistance.