PAN LOCALIZATION PROJECT OUTPUTS

Seminar on Dissemination of PAN Localization Project Outputs

March 26, 2012

Paro

Sithar Norbu
Department of Information Technology & Telecom

རྫོང་ཁ་ལི་ནགསི།

Aim of the Project

Broadly the project aimed to:

- *build effective solutions to enable local language computing*
- *create, develop and deploy local content*
- *build human resources capacity for technology and content development*
- *drive policy changes to support local language development and enhance access to local language content and technologies.*

Phase I (2004 - 2006)

Phase I has contributed significantly in:

- *raising regional awareness*
- *capacity development*
- *creating social networks of researchers and practitioners*
- *building institutional support for local language development and access.*

PHASE II (2007 - 2010)

Broadly, the project intended to achieve following objectives:

- *Examine effective means to develop digital literacy through use of local language computing and content.*
- *Explore development of sustainable HR capacity for R&D in local language computing as a mean to raise current levels of technological support for Asian languages.*
- *Advance policy for development and use of local language computing and content.*

Phase II Project Outputs are:

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%.

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.

POS Tag Sets

“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.

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.

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.

Software

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

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*

Research Publications

1. ***“Pioneering Dzongkha Text-to-Speech Synthesis”*** *at 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.*

2. ***“Building NLP resources for Dzongkha:A Tagset and A Tagged Corpus”*** *Proceedings of the 8th Workshop on Asian Language Resources, pages 103–110, Beijing, China, 21-22 August 2010.*

3. ***“Dzongkha Word Segementation”***

Proceedings of the 8th Workshop on Asian Language Resources, pages 95–102, Beijing, China, 21-22 August 2010.

4. ***“Dzongkha Speech Synthesis System – Phase II”***

Conference on Human Language Technology for Development, 2 – 5 May, 2010, Egypt.

Training Conduction & Development of Training Material

Manuals on usage of Dzongkha Debian Linux 3.0 are published. The manual consists of two books namely:

1) Dzongkha Debian Linux 3.0 Manual –General and

2) Dzongkha Debian Linux 3.0 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 and Telecom’s website (www.ditt.gov.bt).

Q & A

THANK YOU...

http://www.ditt.gov.bt

རྫོང་ཁ་ལི་ནགསི།