

Pema Choejey, Department of Information Technology

{ mail2pemac@gmail.com }

Abstract

This paper presents an overview of localization of open source software undertaken by the Department of Information Technology in the past six years. It presents some of the technological tools and applications developed based on open source software to enable computing in Dzongkha.

Keywords: open source software, localization, operating system.

1. Introduction

Open source software¹ is: Software where the underlying “source code is freely distributed, allow modifications and derive works, license must not discriminate anyone, distribution of licenses must be equal for all users without further restriction, licenses for the same code may not differ if they are used in conjunction with other software and license must be technology neutral”.

Localization is the process of modifying products or services to account for differences in distinct markets.² Localisation thus means modification and customization of software products to make it culturally and linguistically appropriate to target locale – country, region and language – where it will be used and sold. Localisation goes beyond translation.

The Department of Information Technology (DIT) started research studies for localizing open source software (OSS) products since 2004. The program is supported by the International Development Research Centre (IRDC) and the National University of Computers and Emerging Sciences (NUCES), Pakistan. The program's ultimate goal is not only to have local language computing tools, technologies and applications in open source environment, but also to bring

¹ Open Source Initiative (OSI), “The Open Source Definition”, www.opensource.org. [Online]. Available: <http://www.opensource.org/docs/osd>. [Accessed: Dec. 3, 2009]

² The Localisation Industry Association, “The Globalisation Industry Primer”, www.lisa.org, [PDF]. Available: <http://www.lisa.org/Business-Decision-Da.512.0.html#c2077>. [Accessed: Dec. 9, 2009].

policy changes in use of open source software and advancement of open source technologies to enable content creation and access in local languages.³

As a result, DIT released two versions of Dzongkha Debian Linux⁴ in 2006 and 2007 respectively with the motto: “our language, our software”. The Dzongkha Debian Linux has the Tux/Penguin in monk’s red robes as mascot of the software, and interestingly this is “born in Bhutan” and “everyone can bring it home” for free. Moreover, this software is developed by Bhutanese developers and programmers for Bhutanese users.

Packaged applications which come together with Dzongkha Debian Linux are OpenOffice applications such as Writer, Impress and Calc, Firefox browser, Thunderbird email client and Gaim messaging application, to mention a few.

In the following sections, we describe some of the technologies and applications that we have designed, developed and localised so far.

2. Technology and Applications

2.1 Technology

Collation in Dzongkha⁵

Collation is the process and function of determining the sorting order of strings of characters. It provides a key function in a computer system; when a list of strings are presented to users, they would like to have it in a sorted order so that finding individual string would be easy and reliable. Therefore, collation is widely used in user interfaces. It is also ‘must-have’ for the operation of databases, not only for sorting records but also to select sets of records with fields within the bounds.

Rendering in Dzongkha⁶

Rendering is the process of converting the coded content to a required format for display or printing. When a character is pressed on the keyboard, it gets rendered on the display screen or printed on the paper.

X Input Method for Dzongkha⁷

³ PAN, “PAN Localisation Project”, www.panl10n.net, [Online]. Available: www.panl10n.net/. [Accessed: Dec. 5, 2009].

⁴ P. Geyleg, “Dzongkha Linux”. Dzongkha.sourceforge.net, Apr. 9, 2007. [Online]. Available: <http://dzongkha.sourceforge.net/html/web1024X768/home.html>. [Accessed: Dec. 5, 2009].

⁵ Pema Geyleg, “PAN Localisation Working Papers 2004-2007”, pp. 195

⁶ See footnote 4, pp. 200

This basically software tool to create Dzongkha keyboard. The keyboard enables users to input Unicode Dzongkha text using the normal English keyboard.

Fonts in Dzongkha⁸

Font is a collection of glyphs which represent a character or combination of character. Fonts are essential for the development of rendering software. Rescaling of fonts was carried out to be reused on Linux operating system while the actual fonts were developed by the Dzongkha Development Commission (DDC).

2.2 Applications

Open office applications

Open office applications such as Writer, Impress and Calc have been fully localized into Dzongkha. Users can now create, edit, modify Dzongkha document using those applications. The document format is compatible with Microsoft office applications and provides interfaces to save open office documents in MS format such as '.doc', '.ppt' and '.xls'.

Web browser

Browser application called Mozilla Firebox has been fully localized into Dzongkha. Dzongkha content can be now created, rendered and displayed properly on the browser client.

Email Reader

Email reader or client called Mozilla Thunderbird and Evolution has been localized into Dzongkha. Users can now create, send and receive electronic mail in Dzongkha.

Dzongkha Linux OS

Dzongkha Linux is a locally adapted version of the free Linux operating system which integrates full support for Dzongkha computing. It features fully translated user interfaces and provides support for computing in Dzongkha. That is, all fundamental activities like word processing, spreadsheets, presentations, emailing, web browsing, chatting etc. can all be carried out in Dzongkha.⁹

Dzongkha Computer Terms

⁷ See footnote 4, pp. 210

⁸ See footnote 4, pp. 206

⁹ P. Geyleg, "What is Dzongkha Linux", dzongkha.sourceforge.net, April 9, 2007. [Online]. Available: <http://dzongkha.sourceforge.net/html/web1024X768/home.html>. [Accessed: Dec 18, 2009]

As a result of localization, we also published a book on Dzongkha Computer Terms containing 35,000 computer terminologies and another 44,000 strings have been translated into Dzongkha.

Other Applications

Besides applications which we mentioned earlier, we have localized GAIM for messaging, Inkscape for creating vector graphics, GIMP for creating graphics which is similar to Windows Photoshop, X-Chat for Internet Relay Chat (IRC), Gnome Baker for CD burning.

2.3 Advanced Applications

Optical Character Recognition

Optical character recognition, usually abbreviated to OCR, is the mechanical or electronic translation of images of handwritten, typewritten or printed text (usually captured by a scanner) into machine-editable text. As most Dzongkha texts are in print or hand-written forms, there is minimal Dzongkha digital text which could be used for language processing and analysis. Development of OCR will help us to convert printed Dzongkha text into digital text making it editable and modifiable.

Text-to-speech Synthesis

Speech synthesis is the artificial production of human speech. A computer system used for this purpose is called a speech synthesizer, and can be implemented in software or hardware. A text-to-speech (TTS) system converts normal language text into speech; other systems render symbolic linguistic representations like phonetic transcriptions into speech.

Synthesized speech can be created by concatenating pieces of recorded speech that are stored in a database. Systems differ in the size of the stored speech units; a system that stores phones or diphones provides the largest output range, but may lack clarity. For specific usage domains, the storage of entire words or sentences allows for high-quality output. Alternatively, a synthesizer can incorporate a model of the vocal tract and other human voice characteristics to create a completely "synthetic" voice output.

The quality of a speech synthesizer is judged by its similarity to the human voice and by its ability to be understood. An intelligible text-to-speech program allows people with visual impairments or reading disabilities to listen to written works on a home computer.

Wordnet for Lexicon

Wordnet is a lexical database, an application for dictionary which will provide an interface where users can search dictionary word (keyword) and find out its class such as whether it is a noun or verb, etc. Similarly, this will also provide information on word senses – for example a word like father has senses which mean father as a head of family, father as a priest in church,

father as a God - synonyms, antonyms, etc. It will also provide word definitions and usage in sentences.

Currently we are analyzing the words in the Dzongkha dictionary to properly classify the word classes, categorize into synonyms and antonyms, identify different word senses, count word senses occurrences.

Once analysis is done, we will design the application, implement it and finally test the application for the quality, performance and accuracy.

Text Corpora Database

Text corpora database is a database – collection of text - of variety of text in electronic form. The text can be 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 can be 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.

Currently research studies is being conducted to analyse the text collected from different sources, to gather meta information on the text such as the kind of text, author of text, publisher of text, author of text, etc. Collection of text is being done currently from various sources on daily basis. We are also studying on aspects of text conversion format and encoding technologies.

Once proper analysis has been done, we will carry out text corpora database design and its implementation, and testing and quality control.

Text corpora database will have loads of applications in natural language processing. It will form the basic framework where researcher can conduct research on language to study about its structure and form, grammar analysis, count word frequencies, extract keywords to be used in dictionary, etc. It can be used in applications such as speech recognition systems, text-to-speech synthesis, to mention a few.

Spell Checker

Spell checker is an application where by incorrect or misspelled words in sentences are automatically flagged as an error. The flagged errors can be edited and replaced with possible suggestion of words provided by the Spell checker.

Currently we are conducting research study on word segmentation. Word segmentation refers to the process of segmenting known words that are predefined in a lexicon. The segmentation will be based on maximal matching algorithm.

Once word segmentation process is completed, spelling checking rules have to be defined. Then the Spell Checker will be integrated and configured in the Openoffice applications.

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.”¹⁰

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.

DIT has conducted research study on IDN and its applications. 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.

3. Conclusion

This paper presented an overview of localization in open source software. Tools and applications which DIT has developed to enable computing in Dzongkha have been briefly presented. As a result of localization initiatives, two versions of Dzongkha Debian Linux have been released. Now time has come for us to take benefits of localization beyond DIT and deploy it in government administration bodies.

Acknowledgement

I would like to sincerely thank DIT for giving me the opportunity to write this paper and present it during the first Annual ICT Conference 24-25, 2009. All relevant comments and suggestions from colleagues in DIT are heartedly acknowledged.

¹⁰ Wikipedia, “Part-of-Speech Tagging”, www.wikipedia.org, Oct. 21, 2009. [Online]. Available: http://en.wikipedia.org/wiki/Part-of-speech_tagging [Accessed: Dec. 18, 2009].