Proceedings of the Language Technologies for All (LT4All) , pages 235–238
Paris, UNESCO Headquarters, 5-6 December, 2019.

Languages and Technology in Bhutan

Tshewang Norbu, Tenzin Namgyel
Secretary, Sr. ICTO
Dzongkha Development Commission, Thimphu, Bhutan
tshewangnorbu@yahoo.com, tenraj.1047@gmail.com
{tnorbu, tnamgyel}@dzongkha.gov.bt

Abstract

Bhutan, a small country with population less than a million, is linguistically rich with 19 different spoken languages. Dzongkha is the national language and also the official language of Bhutan. Dzongkha Development Commission is a government institute mandated to formulate language plans and policies, develop and promote the national language, and research, document, preserve and protect other indigenous languages of Bhutan. This paper attempts to provide an overview of the languages of Bhutan, language policies and plans, current status of technology development for Dzongkha, and the challenges being faced by Bhutan

Keywords: Dzongkha, Bhutan, Dzongkha Development Commission, Language Technology

Résumé

མི་རློབས་ཁྲི་གཅིག་ལས་ཉུང་བའི་འབྲུག་གི་རྒྱལ་ཁབ་ཆུང་ཀུ་ནང་ ཁ་སྐད་མ་འདྲཝ་ ༡༩ དེ་ཅིག་ཡོདཔ་ལས་ཁ་སྐད་ཀྱི་ཕྱུག་པའི་རྒྱལ་ཁབ་ཅིག་ཨིན། རྫོང་ཁ་འདི་འབྲུག་གི་རྒྱལ་ཡོངས་སྐད་ཡིག་ཨིན་པའི་ཁར་ གཞུང་འབྲེལ་གྱི་ཁ་སྐད་ཡང་ཨིན། རྫོང་ཁ་གོང་འཕེལ་ལྷན་ཚོགས་འདི་ སྐད་ཡིག་གི་སྲིད་བྱུས་དང་འཆར་གཞི་བརྩམ་ནི་དང་། རྒྱལ་ཡོངས་སྐད་ཡིག་གོང་འཕེལ་དང་དར་ཁྱབ་གཏང་ནི། དེ་ལས་ལུམ་པའི་ཁ་སྐད་ཚུ་ ཞིབ་འཚོལ་དང་ཐོ་བཀོད་འབད་དེ་ཉམས་སྲུང་དང་བདག་འཛིན་འཐབ་ནིའི་འགན་དབང་ཡོད་པའི་ གཞུང་གི་གཙུག་སྡེ་མཐོ་ཤོས་ཅིག་ཨིན། ཡིག་ཆ་འདི་ནང་ འབྲུག་གི་ཁ་སྐད་ཚུ་དང་། སྐད་ཡིག་གི་སྲིད་བྱུས་དང་འཆར་གཞི། རྫོང་ཁའི་དོན་ལུ་འཕྲུལ་རིག་གོང་འཕེལ། དེ་ལས་གདོང་ལེན་ཚུ་གི་སྐོར་ལས་ངོ་སྤྲོད་འབད་ནི་ཨིན།

1. Introduction

Guided by a unique development philosophy, Gross National Happiness, Bhutan, a tiny country sandwiched between China in north and India in south places so much importance to her rich cultural heritage and linguistic diversity. The land area of Bhutan is 38,394 square km and population is 734,374 (as of 2018).

1.1 Languages of Bhutan

According to the official survey carried out in 1991, there are 19 different spoken languages. The latest edition of Ethnologue have listed 23 languages. The list excludes Tibetan and includes two foreign languages, namely Kurux and Hindi, and Nupbikha, Lunanakha and Layakha as three additional languages of Bhutan. The Bhutanese languages are classified under Central Bodhish, East Bodhish, Bodic, and Indo-Aryan (van Driem 1998)

Dzongkha is the national language of Bhutan. It was declared as the national language in 1971. It is the native language of eight of the twenty districts of Bhutan, viz. Thimphu, Pünakha, Paro,Wangdi Phodrang, Gasa, Haa, Dagana and Chukha in western Bhutan, but Dzongkha is spoken as a lingua franca throughout the country.

According to Pema Wangdi (2015), all the languages of Bhutan with the exception of Dzongkha, Tshangla, and Lhotsham(Nepali), fall under the category of "endangered" languages. Three languages, namely Monkha, Lhokpu, and Gongduk are critically endangered. One dialect kown as Olekha, a variety of Monkha spoken in Rukha under Wangdue Dzongkhag, is a moribund.

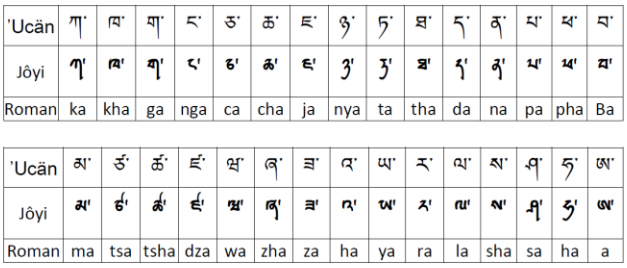
1.2 Dzongkha Development Commission

The Fourth King Jigme Singye Wangchuck established Dzongkha Development Commission (DDC) in 1986. It is now a premier government agency with the highest authority in the matters related to Languages. The broad mandates of the DDC are to formulate language plans and policies, to carry out the activities to develop and promote Dzongkha, as the national language and to carry out the activities to preserve and protect other indigenous languages of Bhutan as the rich linguistic and cultural heritage of Bhutan

The overall plans and policies of the DDC are guided by the commission which consists 10 members. The chairperson of the commission is the Honorable Prime Minister of Bhutan.

2. Writing Systems

The script used to write Dzongkha is the same script used for Tibetan. This writing system consists of 30 consonant symbols and 4 vowel symbols. It is called the Uchen Script and it is one of the two scripts first developed by Thonmi Sambhota in the 7th century, the other being Ume. Uchen is based on the Devanagari script. Another type of script known as Joyig, which is unique to Bhutan, was first developed by Demang Tsemang in Bhutan in the 8th century. Joyig has exactly the same number of consonant and vowel symbols and those Joyig symbols represent the same phonemes as the Uchen script.



'Ucän	ཀ་	ཁ་	ག་	ང་	ཅ་	ཆ་	ཇ་	ཉ་	ཏ་	ཐ་	ད་	ན་	པ་	ཕ་	བ་
Jôyi	ཀ་	ཁ་	ག་	ང་	ཅ་	ཆ་	ཇ་	ཉ་	ཏ་	ཐ་	ད་	ན་	པ་	ཕ་	བ་
Roman	ka	kha	ga	nga	ca	cha	ja	nya	ta	tha	da	na	pa	pha	Ba

'Ucän	མ་	ཙ་	ཚ་	ཛ་	ཝ་	ཞ་	ཟ་	འ་	ཡ་	ར་	ལ་	ཤ་	ས་	ཧ་	ཨ་
Jôyi	མ་	ཙ་	ཚ་	ཛ་	ཝ་	ཞ་	ཟ་	འ་	ཡ་	ར་	ལ་	ཤ་	ས་	ཧ་	ཨ་
Roman	ma	tsa	tsha	dza	wa	zha	za	ha	ya	ra	la	sha	sa	ha	a

Figure 1: 30 Consonants of Dzongkha

The four vowel symbols in Dzongkha are ི(གི་གུ་g'ikhu) representing the sound "i", ུ(ཞབས་ཀྱུ་zh'apju) representing the sound "u", ེ(འགྲེང་བོ་drengbo) representing the sound "e" and ོ(ན་རོ་naro) representing the sound "o". When no vowel is indicated, the vowel in a syllable is automatically "a". Dzongkha language is syllabic. A Dzongkha syllable can have 1 to 7 characters and most interestingly, up to four characters can stack on one another as shown in the figure below.

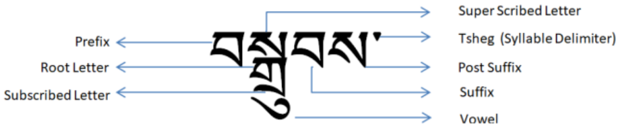


Figure 2: A Dzongkha Syllable

Since the number and the type of phonemic inventories of the Bhutanese languages are quite similar, albeit some with few extra ones, this script can be used to write Dzongkha and other languages of Bhutan with almost the same degree of consistency and economy. Dzongkha is the only officially written language of Bhutan. Other native languages are written sporadically for interpersonal communication in informal settings and in social media. Lhotshamkha is written using the Nepalese script (Nepal Bhasa), one of the Brahmic scripts of the Kathmandu valley of Nepal.

3. Language Policies, Plans and Actions

The language policy framework for Bhutan is enshrined in its Constitution, the mother of all laws. Section 8 of Article 1 of the Constitution states, "Dzongkha is the national language of Bhutan"; and, in Section 1 of Article 4, "language" and "literature" are enumerated along with the other cultural heritage of Bhutan to be preserved, protected and promoted. It provides a clear framework for developing and promoting Dzongkha as the national language of Bhutan and it also provides a clear framework for preserving the other languages of Bhutan as the rich linguistic and cultural heritage of Bhutan.

Pema Wangdi (2015), summarizes the language policy of Bhutan in the form of Quadrilingual Model:

Dzongkha, Chökê, English, and Mother Tongue

Mother Tongue in this model is the sum total of all the mother tongues in Bhutan.

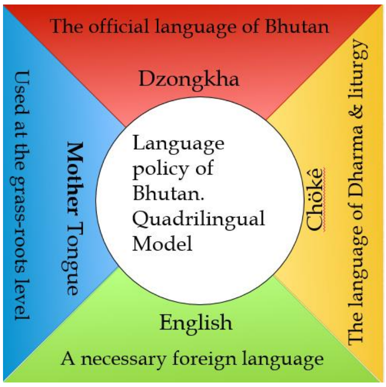

The official language of Bhutan
Dzongkha
Used at the grass-roots level
Mother Tongue
Language policy of Bhutan. Quadrilingual Model
Chökê
The language of Dharma & liturgy
English
A necessary foreign language

Figure 3: Quadrilingual Model

Dzongkha efficiently serves as the official language of Bhutan while Chökê serves as the language of Dharma and liturgy; and English is apparently used as the necessary foreign language while mother tongues are used at the grass-roots level.

Today, under the framework of 12th Five Year Plan, DDC is working to achieve following five objectives:

1. Develop and promote Dzongkha.
   To standardize the spelling and grammar, DDC writes grammars and dictionaries and other reference books. Various promotional activities such as competitions are also being carried out.

2. Research and document indigenous languages.
   Writing trilingual lexicons and phrase books between Dzongkha, English and indigenous languages and carrying out preliminary researches on the grammar of indigenous languages are some of the important activities carried out to preserve our indigenous languages.

3. Enhance Dzongkha usage in public service delivery.
   Despite Dzongkha being the national language, wide usage of English to disseminate information and deliver services to the general public gives unequal access to information and services and causes inequalities in the society. To tackle this and to protect our national identity, DDC provides support in making websites, applications and forms accessible in Dzongkha.

4. Enhance usage of Dzongkha in education system.
   Role of national language and other indigenous languages in mainstream education is limited as English is the only medium of instruction and Dzongkha was taught only as a language subject. DDC recommends and assists in including more Dzongkha subjects and also actively participates in curriculum development.

5. Develop language technology and promote Dzongkha through ICT.
   To keep Dzongkha digitally alive in the world of technology, DDC develops Unicode compatible Dzongkha keyboards and fonts. To bridge the digital

divide, solve other inequalities caused by the language barrier, and protect our national language, DDC has started to work on developing technology for Dzongkha which is covered in more detail in the following section.

4. Language Technology Development

In Bhutan, English is used very widely with technology and this can lead to digital extinction of national language and other language, and cause real extinction gradualy. Moreover, language barrier is the cause of unequal access to knowledge, information, services and digital divide which creates inequalities in the society. The best and the only solution would be to develop technology for Dzongkha.

4.1 Encoding, Input and Rendering Supports

Twenty years ago, there was no recognized or de-facto standard for encoding Dzongkha or Tibetan script characters. Word-processing applications and other programs adapted for Dzongkha used a variety of ad-hoc non-standardized encodings which gave codes in character sets actually meant for encoding Roman characters to Dzongkha letters. The biggest obstacle in using electronic Dzongkha data was the fact that files could not be easily shared by different Dzongkha word-processing programs and other applications without converting files from one encoding scheme to another.

In 2000, a 3-year project was carried out to develop a standardized system for Dzongkha based on the new Unicode / ISO 10646 character encoding standard. During the project, a standard keyboard layout was developed, a locale for Dzongkha and collation rules were developed and Unicode compatible fonts were also developed. Input and rendering support for Linux operating system was developed by Department of Information Technology and Telecom (DITT), Ministry of Information and Communication (MOIC) under the framework of PAN localization project phase I (2004-2007).

Today, major operating systems like Windows, MacOS iOS and Android have built-in Dzongkha rendering and input supports. However, as of now, except for older version of Linux, no language packs or localized versions of operating systems are available in Dzongkha.

4.2 Automation and Processing of Language

Currently, technology for Dzongkha is limited to input, storage and display. There are no working Dzongkha language processing tools: not even spelling or grammar checker. Except for few research works carried out by DITT, MOIC during the 2nd phase of PAN localization project (2007-2012), there is no history of much work done in the field of natural language processing in the past.

The DDC, in collaboration with College of Science and Technology, Phuntsholing, Bhutan started to develop part of speech (POS) tagged corpus since 2014. With support from Indian Institute of Technology (IIT), Guwahati, India, we have achieved the following.

4.2.1 Dzongkha Word Segmentation

Dzongkha does not have any word delimiter like space in English. Therefore, it is necessary for the computer to do word segmentation to be able to progress further in Dzongkha text processing. The Dzongkha word segmentation is done as a syllable tagging problem using various NLP toolkit and the best model is 95% accurate as of now.

4.2.2 Dzongkha Part of Speech (POS) Tagging

DDC has developed Dzongkha POS tagset and also annotated POS to around 2 lakhs Dzongkha words, manually. Using the manually POS-tagged corpus, we trained a model which is around 90% accurate as of now. We expect the performance to improve with increase in corpus size.

4.2.3 Dzongkha Automatic Speech Recognition

The first Dzongkha automatic speech recognition (ASR) prototype was developed during the ASR summer school conducted by IIT, Guwahati in the year 2017. With the knowldege and motivation gained, DDC has now increased the speech corpus to around 15 hours of recording and we are hoping to increase it.

5. Challanges

Bhutan as a least developed country, has many other priority areas to invest on for wellbeing of the citizen. Though Bhutan fully understands the importance of developing technology for language, not much could be done so far without enough fund. We also don’t have required expertise in this field. For instance, we do not have anybody with master’s degree or PhD. artificial intelligence; forget about in computational linguistic or natural language processing specialist. With the limited state funding, DDC is unable to provide long term training to staff and we are also unable to participate in international forums. Bhutan has not been able to secure any funding support from outside for development and promotion of language technology.

6. Conclusion

For a small landlocked country like Bhutan, cultural heritage and linguistic diversity is very important for national identity. Preservation of rich cultural heritage is one of the pillars of our development philosophy of Gross National Happiness. A sound language policy that ensures protection of all our languages is enshrined in our mother of law, the Constitution of Bhutan. To keep our national language and other languages alive, it is important that we make them usable in the digital world but due to the lack of expertise and fund, we could do very little so far to process by computer. Technology for our national language is limited to input, storage and display. We solicit support in terms of expertise, capacity building and funding from international organizations and donors.

7. Acknowledgements

We would like to thank Indian Institute of Technology, Guwahati, Assam, India for helping us in testing our corpus on various NLP toolkits and developing models for word segmentation, POS-tagging and ASR.

8. Bibliographical References

Pema, W., Language Policy and Planning in Bhutan, available at https://www.dzongkha.gov.bt/uploads/files/articles/A_Paper_on_Language_Policy_&_Planning_in_Bhutan_by_Pema_Wangdi_c8e8caeee831129a3be15aa6e99732c2.pdf accessed on 25th Dec, 2019

Chungku, C., Jurmey, R.,Gertrud, F. 2010. “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, pp. 103-110.

Department of Information Technology & Telecommunications (DITT), Bhutan, Research papers, Available online at “https://www.dit.gov.bt/research-paper,” Accessed on 25th Dec 2019.

Norbu, S., Choejey, P., Dendup, T., Hussain, S. and Muaz, A., 2010. "Dzongkha Word Segmentation", Proceedings of the 8th Workshop on Asian Language Resources, COLING2010, Beijing, China, April 3-8, pp. 200-209.