Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 81 (2016) 237 – 242

Procedia Computer Science

5th Workshop on Spoken Language Technology for Under-resourced Languages, SLTU 2016, 9-12 May 2016, Yogyakarta, Indonesia

The Development of an Audible Pattani Malay-Thai Electronic Phrasebook for Military Purposes

Prachya Boonkwan[a,*], Thepchai Supnithi[a], Wandee Tosuwan[b], Chai Wutiwiwatchai[a]

[a] *National Electronics and Computer Technology Center, 112 Thailand Science Park, Phahonyothin Road, Khlong Nueng, Khlong Luang, Pathum Thani 12120 Thailand*

[b] *Foreign Language Center, Department of Defense Science and Technology, Ministry of Defense, Nonthaburi, Thailand*

Abstract

Pattani Malay is a minority language spoken by ethnic Malays in southernmost provinces of Thailand, and it plays a strategically and politically crucial role in municipal governance in the area. This paper presents the development of a multimodal Pattani Malay-Thai corpus and its application to an audible electronic phrasebook for military purposes. The bilingual corpus contains 10,000 parallel sentences and sound recordings in both languages. A mobile application based on Android platform is developed on top of the corpus and offers the search engine for on-field usage and reconnoiters, as well as language lessons for military personnels.

Peer-review under responsibility of the Organizing Committee of SLTU 2016

Keywords: electronic phrasebook, Pattani Malay, Thai, parallel corpus, multimodal data

1. Introduction

Language resource construction for ethnic minority languages is a nontrivial task, linguistically and politically. Under-resourced and not widely spoken, these languages pose quite a challenge to natural language processing, in which a sufficient amount of data is required to statistically represent the language of interest. Despite relatively small numbers of speakers and the lack of language resources, some of them have tremendous impact on regional goverance and the World politics, where language barrier is an obstacle. Moreover, some of these languages are also endangered due to its gradual decrease of native speakers.

To surmount these issues, the need of bilingual parallel corpora is imminent for the development of statistical MT systems[1]. Parallel corpora can be constructed in three approaches: (a) extraction from existing parallel texts[2,3,4], (b) automatic collection from comparable texts[5,6,7,8,9,10], and (c) manual construction from scratch[11,12,13,14,15]. Because there are very little linguistic resources for Pattani Malay-Thai parallel text, we opt for the last approach in constructing the parallel text for this language pair.

* Corresponding author. Tel.: +66(0)-2-564-6900 ext 2213
E-mail address: Prachya.Boonkwan@nectec.or.th

Peer-review under responsibility of the Organizing Committee of SLTU 2016
doi:10.1016/j.procs.2016.04.055

Jawi script	بهاس ملايو ڤطاني
Rumi script	Bahasa Melayu Patani
Thai script	บา'ซอ 'นายู 'ตานิง
IPA transcription	/baˈsɔ ˈnaːju ˈtaːniŋ/

Fig. 1. Writing systems of Pattani Malay

In this paper, we present the construction of a Pattani Malay-Thai bilingual corpus and its application as an electronic phrasebook for military purposes. The constructed corpus is multimodal, containing 10,000 sentence pairs along with corresponding sound clips for both languages. It offers a search engine for ease of information access in the phrasebook. It also incorporates language lessons built on top of the keyword-oriented practical sentences.

The rest of the paper is organized as follows. We elaborate some background knowledge of Pattani Malay in §2. The design and construction of our Pattani Malay-Thai are explained in §3. We describe an implementation of a military-purposed mobile application based on this parallel corpus in §4. Finally, we conclude our paper in §5.

2. Background

Pattani Malay, as known as *Yawi* and *Bahasa Jawi*, is a subregional dialect of Malay spoken in the southernmost provinces of Thailand and Kelantan State of Malaysia. Geopolitically influenced by the majority of Thai speakers, it contains a considerable number of loan words from Thai that seamlessly intermingle with Malay lexicons. Phonetic differences between Pattani Malay and Standard Malay are prominent, particularly in vowel nasalization, stark vowel changes, and coda drops. These make cross-linguistic comprehension between both languages difficult.

Unlike Standard Malay that adopts the Latin alphabet (called Rumi script), Pattani Malay strictly preserved its Arabic-based writing script called *Jawi*. The use of Thai script, however, has gradually gained some interest among its young-generation speakers as a means of advanced education in Thailand. All of these scripts, as well as an IPA transcription of local pronunciation, are provided in figure 1.

Pattani Malay, similar to Standard Malay, is *mildly* agglutinative and non-total. Despite the rigid word order of SVO, head-initialness, and an extensive use of adverbs and modals for grammatical functions, words can also be transformed with morphological affixation and reduplication. In Pattani Malay the semantics of utterances are distinguished not by tones but by pitch-accent intonation. The intonation is not represented in Jawi script, but it is sometimes approximated by the tone marks when written in Thai script.

Pattani Malay plays a strategically and politically crucial role in municipal governance in the southernmost provinces of Thailand. While its urban native speakers are Thai-Malay bilingual, the rural ones are Malay monolingual and speak very little of Thai. This issue enlarges the wall of communication, rooting tremendous misunderstanding between Thai soldiers and the native people. As a result, national separatism and terrorism have caused regional turmoil and many losses on both sides.

In spite of its importance, teaching materials for Pattani Malay are considerably scarce and hard to prepare owing to the lack of human resources with teaching experience. Language training for military personnels is time-consuming and requires some dedication to acquire practical linguistic proficiency and cultural insights. To solve these problems, practical language lessons and teaching materials for Pattani Malay have to be portable and instantaneous for on-field communication and reconnoiters.

3. Design and Construction of Multimodal Bilingual Corpus

We designed the Pattani Malay-Thai corpus to comply with the needs of on-field communication and language mastering. The schema of the corpus is conceptually represented as an entity-relation diagram in figure 2.

There are three entity tables in the schema: Thai for Thai texts, PtMalay for Pattani Malay texts, and Lesson for lesson categorization. All texts are encoded in the UTF-8 format. Each table has an integer surrogate key suffixed