# Bringing Zero-resourced Languages of Myanmar to the Digital World

United Nations
Educational, Scientific and
Cultural Organization

2019 | International Year of
Indigenous Languages

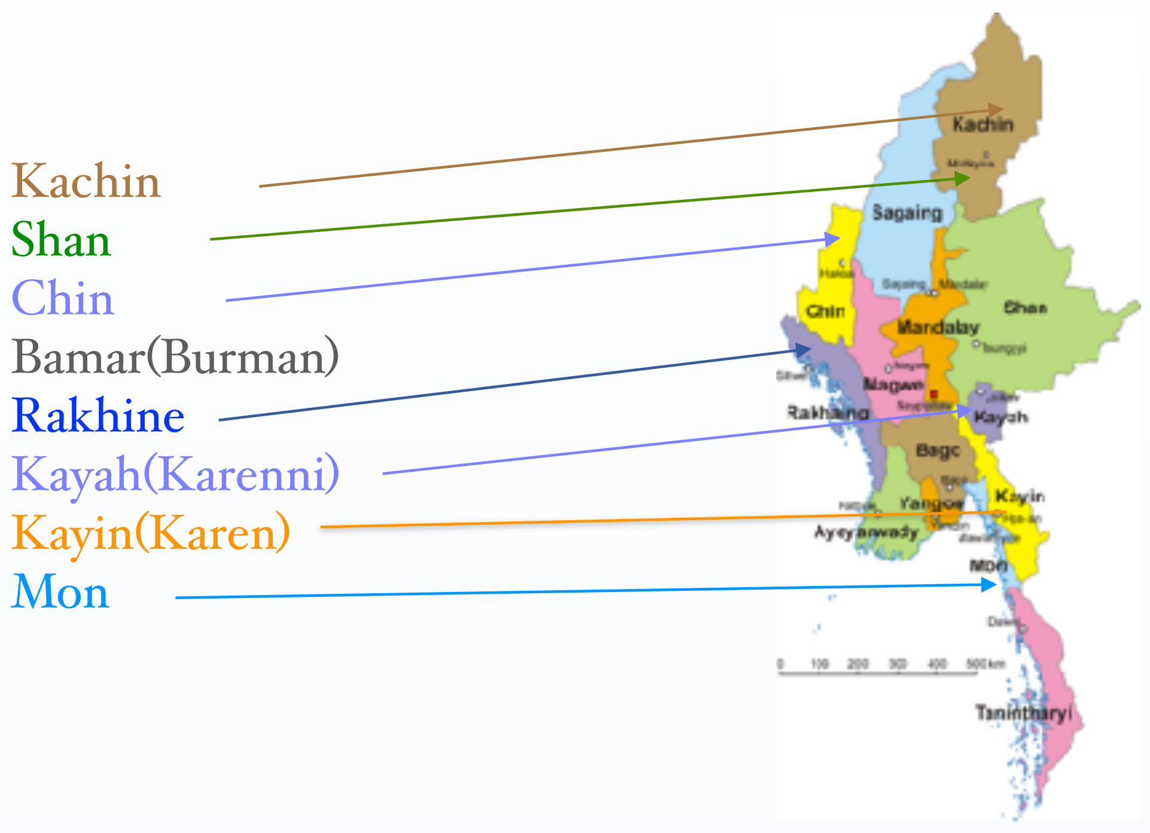
**Win Pa Pa**

**University Computer Studies, Yangon, Myanmar**

International Conference on Language Technologies for All (LT4All)
4-6 December 2019, UNESCO Headquarters, Paris, Franc

## Myanmar and its Languages

- a member of the Lolo-Burmese grouping of the Sino-Tibetan language family
- 135 distinct languages, generally grouped into 8 main ethnic language clusters, Kachin, Kayah, Kayin, Chin, Mon, Rakhine, Bamar and Shan.



## Dialects of Myanmar language

- The majority of Myanmar language speakers, who live throughout the Irrawaddy River Valley, use a number of largely similar dialects, while a minority speak non-standard dialects found in the peripheral areas of the country. These dialects include:
  - Tanintharyi Region: Merguese (Myeik), Tavoyan (Dawei), and Palaw
  - Magway Region: Yaw
  - Shan State: Intha, Taungyo and Danu

## Myanmar language is Low-resourced

- Until 2005, most Myanmar language websites used image-based text,
- No Digital Myanmar text available online as language resource for Natural Language Processing(NLP) before that time

## Language Technologies for all Myanmar People

## Machine Translation

- Rule-based English to Myanmar MT with the help of a dictionary till 2006.
- Statistical Machine Translation(SMT) from 2010 to 2016
- Joined ASEAN MT project 2011-2013
- Started Neural MT from 2018

## Resource Building for Machine Translation

- UCSYCorpus English-Myanmar bilingual 2017-2018
- manually collected from local news website, Wikipedia
- 220,000 parallel sentences
- Use for Myanmar - English Translation Task at Workshop on Asian Translation(WAT) 2018 & 2019 to promote Translation on Myanmar langauge

Ref: [Yimon Shwe Sin, Khin Mar Soe, *Syllable-based Myanmar-English Neural Machine Translation, ICCA2018]*

## Resources for Myanmar Dialects

- Rakhine-Myanmar Corpus (20k sentences)
- Tanintharyi Dialects (Dawei, Myeik - Myanmar) Corpus (20 k sentences)

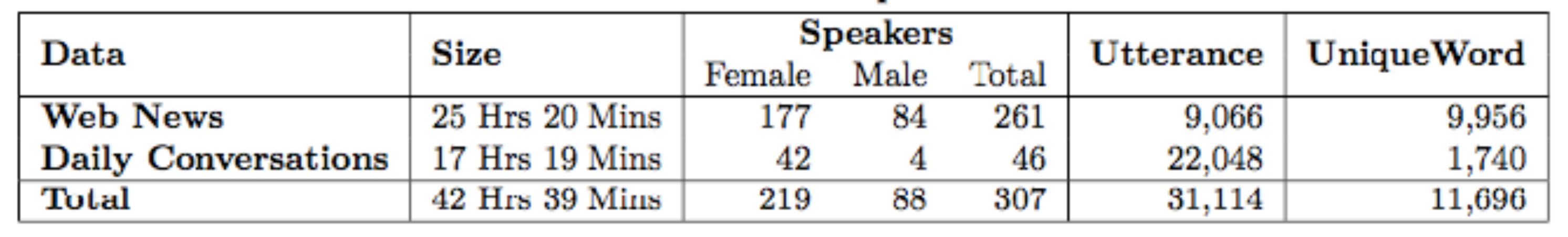
Table 1.Example parallel sentences of Myanmar Dialects

Ref: [Thazin Myint Oo, Ye Kyaw Thu and Khin Mar Soe, "*Neural Machine Translation between Myanmar (Burmese) and Rakhine (Arakanese)*", In Proceedings of the Sixth Workshop on NLP for Similar Languages, Varieties and Dialects, June 7th 2019]

## Automatic Speech Recognition

- collected speech data from broadcast news published online and manually transcribe them into text
- designing the text corpus first and recording the speech by reading the collected text
- Myanmar speech corpus (UCSY-SC1) for Automatic speech recognition

**Myanmar ASR Demo**
http://nlpresearch-ucsy.edu.mm/masr.html



| Data | Size | Speakers | | | Utterance | UniqueWord |
|---|---|---|---|---|---|---|
| | | Female | Male | Total | | |
| **Web News** | 25 Hrs 20 Mins | 177 | 84 | 261 | 9,066 | 9,956 |
| **Daily Conversations** | 17 Hrs 19 Mins | 42 | 4 | 46 | 22,048 | 1,740 |
| **Total** | 42 Hrs 39 Mins | 219 | 88 | 307 | 31,114 | 11,696 |

Table 2. Statistics of UCSY-SC1 corpus

| Model | Close domain TestSet | Open domain TestSet |
|---|---|---|
| Word-based ASR Model (SER%) | 9.7 | 16.8 |
| Syllable-based ASR Model (SER%) | 15 | 18.33 |

Table 3. Evaluation of ASR on UCSY-SC1 with CNN models

Ref: [Developing a Speech Corpus from Web News for Myanmar (Burmese) Language" *Aye Nyein Mon, Win Pa Pa, Ye Kyaw Thu and Yoshinori Sagisaka*, Oriental-COCOSDA 2017]

## Myanmar Text-to-Speech

- Tone is the integral part of the pronunciation of syllable and can affect the meaning of that syllable
- Tonal features in contextual information is important in the naturalness of synthesized speech
- Proposed a questionset for Myanmar language
- many pronunciations do not follow their orthography
- Graphic to phoneme(g2p) is one of the important task to Myanmar TTS
- LSTM-RNN and linguistics features are applied modeling Myanmar speech synthesis

**Applied in "Burmese Text-To-Speech Engine" desktop and mobile Myanmar TTS applications for visually impaired people**

Utterance:

နောက်ထပ် အားသာချက်ကတော့ မိသားစုနဲ့ ခရီးသွားဖို့ အဆင်ပြေတဲ့ ကျယ်ဝန်းမှုရှိတာပါပဲ ။

Word:

နောက်ထပ် | အားသာချက် | က | တော့ | မိသားစု | နဲ့ | ခရီးသွား | ဖို့ | အဆင်ပြေ | တဲ့ | ကျယ်ဝန်းမှု | ရှိတာပါပဲ | ။

Syllable:

နောက် | ထပ် | အား | သာ | ချက် | က | တော့ | မိ | သား | စု | နဲ့ | ခ | ရီး | သွား | ဖို့ | အ | ဆင် | ပြေ | တဲ့ | ကျယ် | ဝန်း | မှု | ရှိ | တာ | ပါ | ပဲ |
။

Phoneme in MLC symbol:(g2p)

n au' ht a' a: th a gy e' k a. d o. m i. th a: z u. n e. kh a- j i: th w a: b o. a- hs in pj ei d e. ky e w un: h m u. sh i. d a b a b e

**http://www.nlpresearch-ucsy.edu.mm/myantexttospeech.html**

Ref: {Aye Mya Hlaing, Win Pa Pa, Ye Kyaw Thu, "*Enhancing Myanmar Speech Synthesis with Linguistic Information and LSTM-RNN*", In Proceedings of 10th ISCA Speech Synthesis Workshop (SSW10), September 2019, Vienna, Austria}