

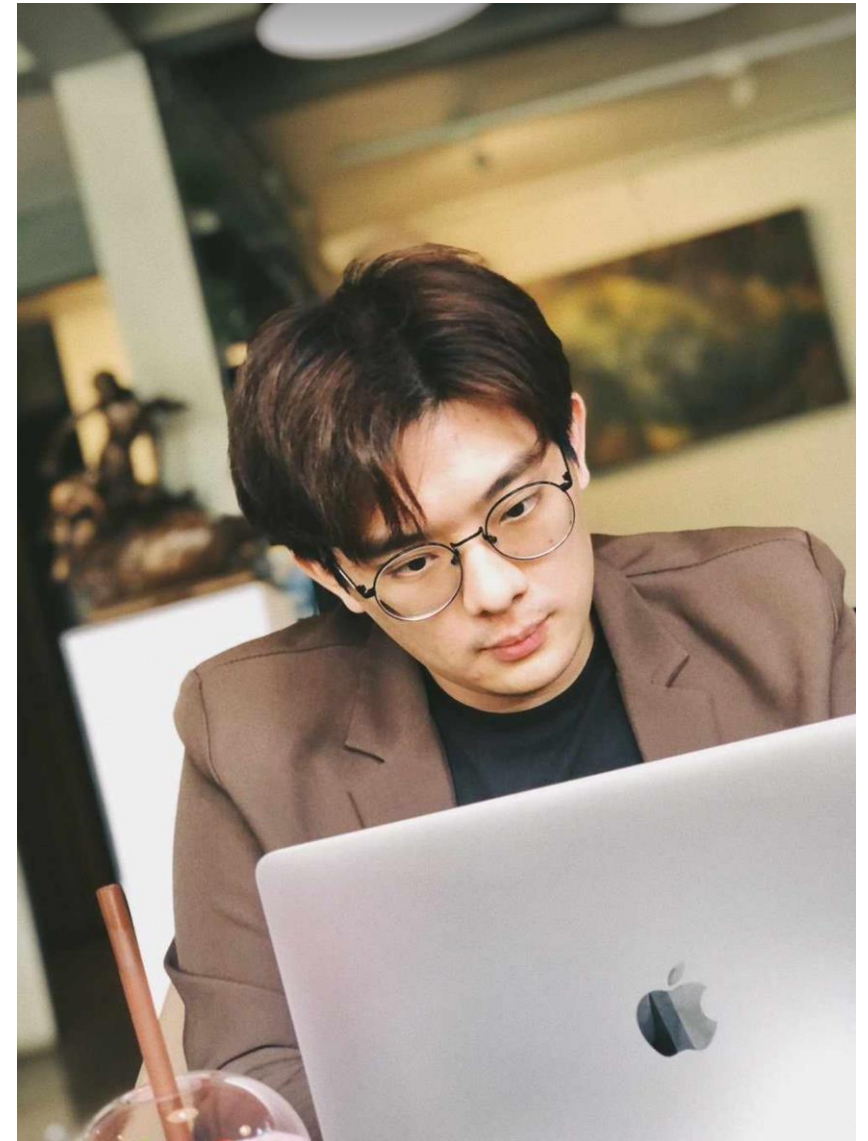
Tanintornlimp@gmail.com



# Me

---

- Nickname : Zen
- Born: Sept 1992
- Work: Freelance Researcher (such SAC, MU) (2014-Now)
- Education
  - Bachelor in Archaeology at Silpakorn University, Thailand (2011-2014)
  - Did a PhD in Linguistics (theory) at Mahidol University, Thailand from (2016-2024)



# Outline (talking topics)

---

- Background of work
- Linguistic work – phonetic and phonology
- Corpus Creation concerning NLP
- Phonology to NLP

# Background of work

# Background of work

---

Linguistic analysis

Linguistic feature analysis

- **AN AUTOSEGMENTAL – METRICAL ANALYSIS AND PROSODIC ANNOTATION IN MYANMAR: IMPLICATIONS FOR MYANMAR – THAI MACHINE TRANSLATION IN MEDICAL COMMUNICATION**

Subjective Language

Corpus creation & MT experiment

# Background of work

---

- Topic of Thesis :
- **AN AUTOSEGMENTAL – METRICAL ANALYSIS AND PROSODIC ANNOTATION IN MYANMAR**: IMPLICATIONS FOR **MYANMAR – THAI MACHINE TRANSLATION IN MEDICAL COMMUNICATION**
- Machine translation (MT) nowadays plays role for the language barrier situation for cross-language communication. >>> addition of the linguistic feature on the MT can improve MT such POS-tag (Koehn & Hoang, 2007; Nguyen et al., 2016).
- By the same idea of adding linguistic feature on MT, the phonological feature can take into account for additional feature.
- >>> **Tone and Break Indices (ToBI)** is an approach of annotation system that annotate prosodic feature such Intonation.

# Background of work

---

- **Prosodic annotation** such as Tone marker can be marked on text such

ฉัน|4| กิน|1| ข้าว|3|

*where specific number represents specific tone*

## I chose Intonation .... Why not tone?

- Here I found that tone can be varied into many type of tone shape in actual speech
  - Intonation is selected in this case to reflect the actual speech.
  - Autosegmental-Metrical analysis suits for intonation analysis.
- 
- **Intonation** refers to the variation in spoken pitch (melody) when uttering a sentence or phrase.

# Background of work

---

- My Hypotheses then “the approach of the **Tone and Break Indices (ToBI)** (in **Autosegmental-Metrical theory**) can be worked as **POS-tag**”

## POS-tag (Part-Of-Speech tagging)

The|DT| quick|JJ| brown|JJ| fox|NN| jumps|VBZ| over|IN| the|DT| lazy|JJ| dog|NN| .|\.|

## Myanmar ToBI-tag

အနာ|?| ပတ်|??| လည်ကို|???| သန့်|???| ရှင်းရေးလုပ်ပါ|???|

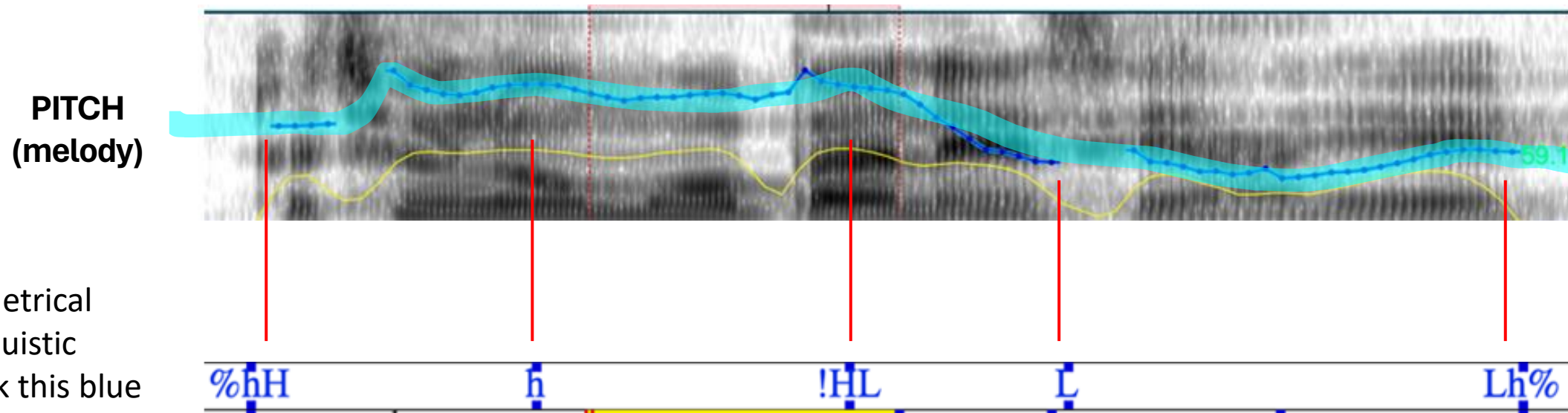
- Question: How prosodic work in MT ? Is it the same as POS-tag ?



# Linguistic work:

## Phonetic and phonology analysis

# Linguistic work: the analysis



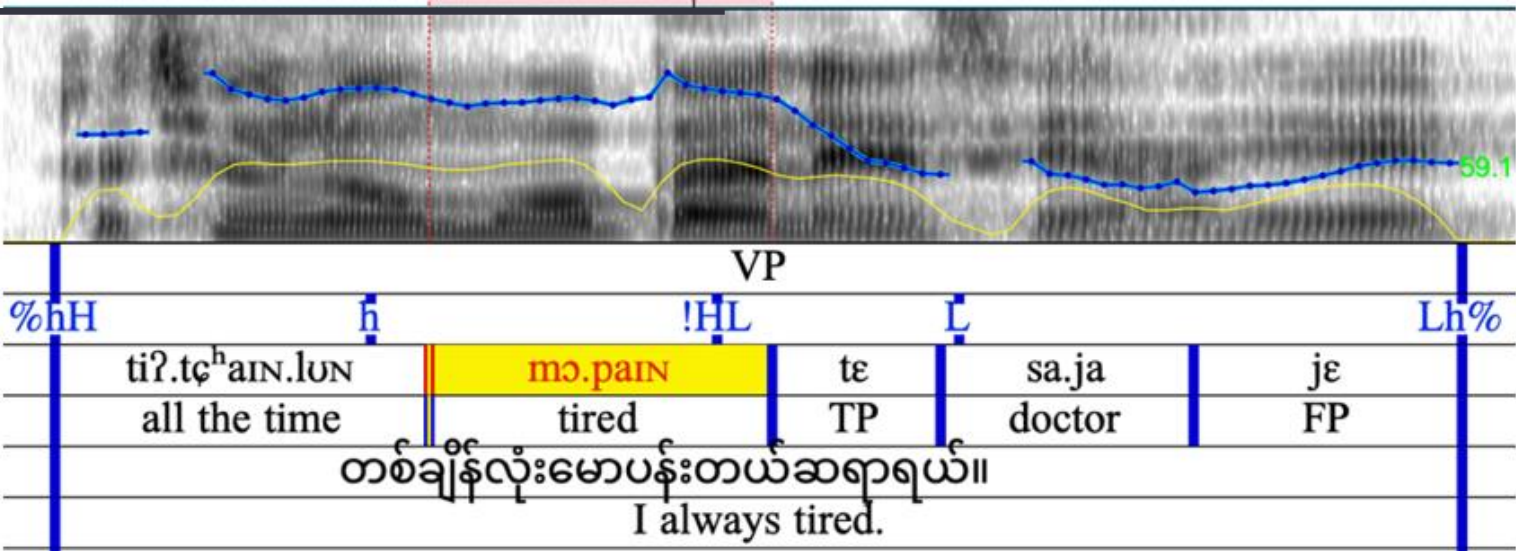
## What they told us ?

- It's tell us the movement of melody (**Intonation**) of speech in an utterance.
- “If these intonations occur repeatedly in many different sentences, it indicates that Which intonation occurs most often in which sentence?”

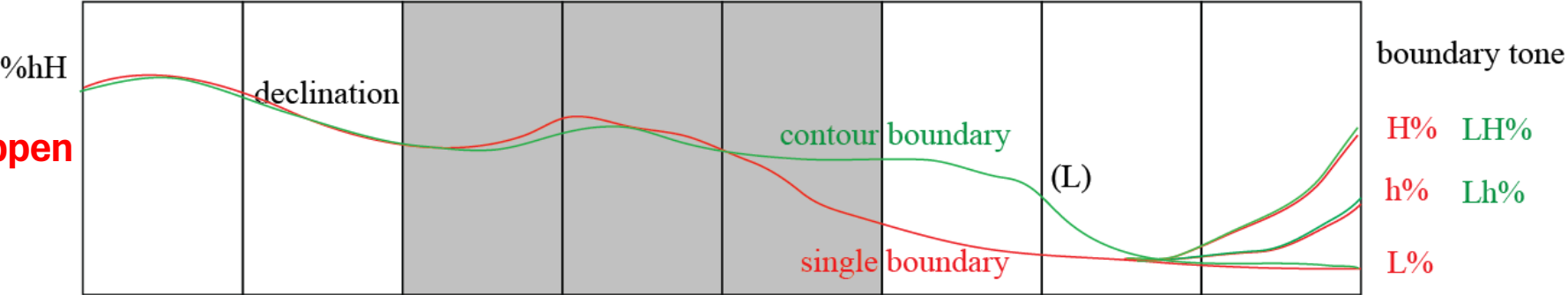
Intonation of Myanmar (the example)

BROAD FOCUS STATEMENT

- %hH -initial boundary tone with mid-high to High
- h -Mid-high tone
- !HL -Falling tone with downstepped
- L -Low tone
- L% -Low offset final boundary tone



Most intonation patterns happen

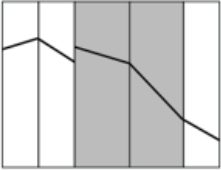
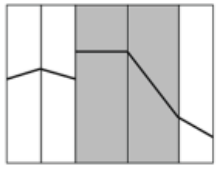
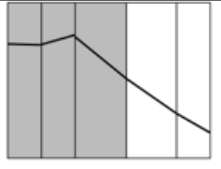
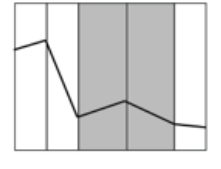
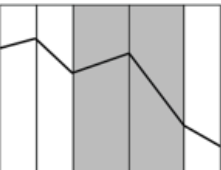
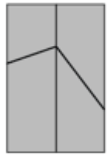


<div>-Broad-focus statement</div> <div>-Yes/no question</div>	<div>%h !H L L%</div>	
---	-----------------------	--

# The intonation patterns of Myanmar

There are 7 patterns of intonation in the Myanmar language. There are differences according to the phenomenon of intonation in each sentence type.

The presented intonation patterns (yellow highlight) are the most pattern that occurred in the given type of sentence.

<div><div>-Broad-focus statement</div><div>-Yes/no question</div></div>	<div>%h !H L L%</div>	<div></div>
<div><div>-Verb Intensification-focus statement</div></div>	<div><div>%h h H* L%</div><div>%h L H* L%</div></div>	<div></div>
<div><div>-Object intensification-focus</div><div>-Imperative</div></div>	<div>%h H*L L%</div>	<div></div>
<div><div>-Wh-question with /ba ... le/ (what, why)</div><div>-Wh-question with /be ... le/ (How, When, Where, Which)</div></div>	<div>%hH L !h L L%</div>	<div></div>
<div><div>-Negation-focus statement</div><div>-Negation answering (with NP)</div><div>-Wh-question with /ba ... le/ (what, why)</div><div>-Wh-question with /be ... le/ (How, When, Where, Which)</div><div>-Yes-No question</div></div>	<div>%hH h !H* L L%</div>	<div></div>
<div>-Negation answering (without NP)</div>	<div>%h H* L%</div>	<div></div>

## From Phonological analysis to ToBI-tag

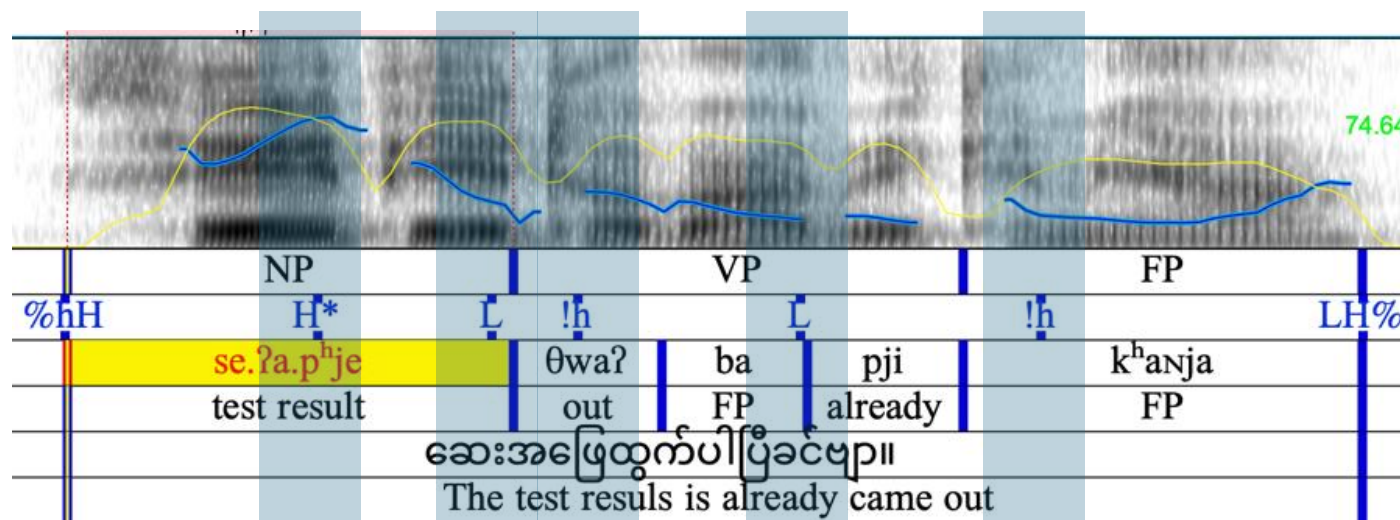
**Myanmar Tone and Break Indices (MY-ToBI) model and prosodic annotation system.****Tone tier****AP tone**

L

H : H, !H, H\*

h: !h, h\*, !h\*

HL : !HL, !H\*L, H\*L

**Boundary tone**

%H

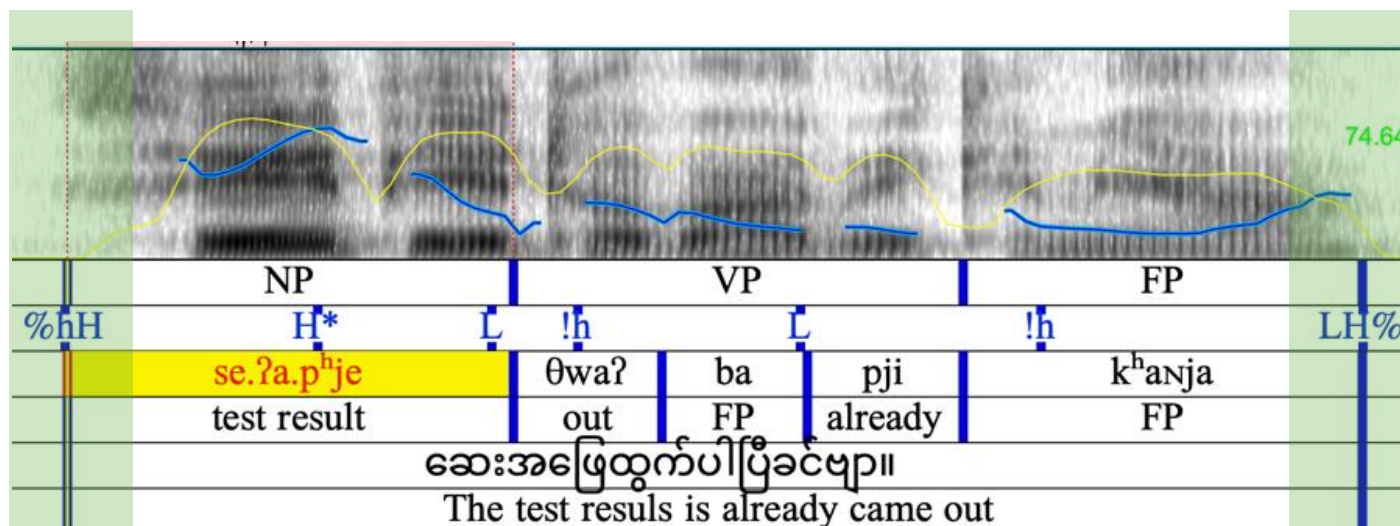
%h

%hH

L%

H%, h%

Lh% , LH%



# Corpus Creation concerning NLP

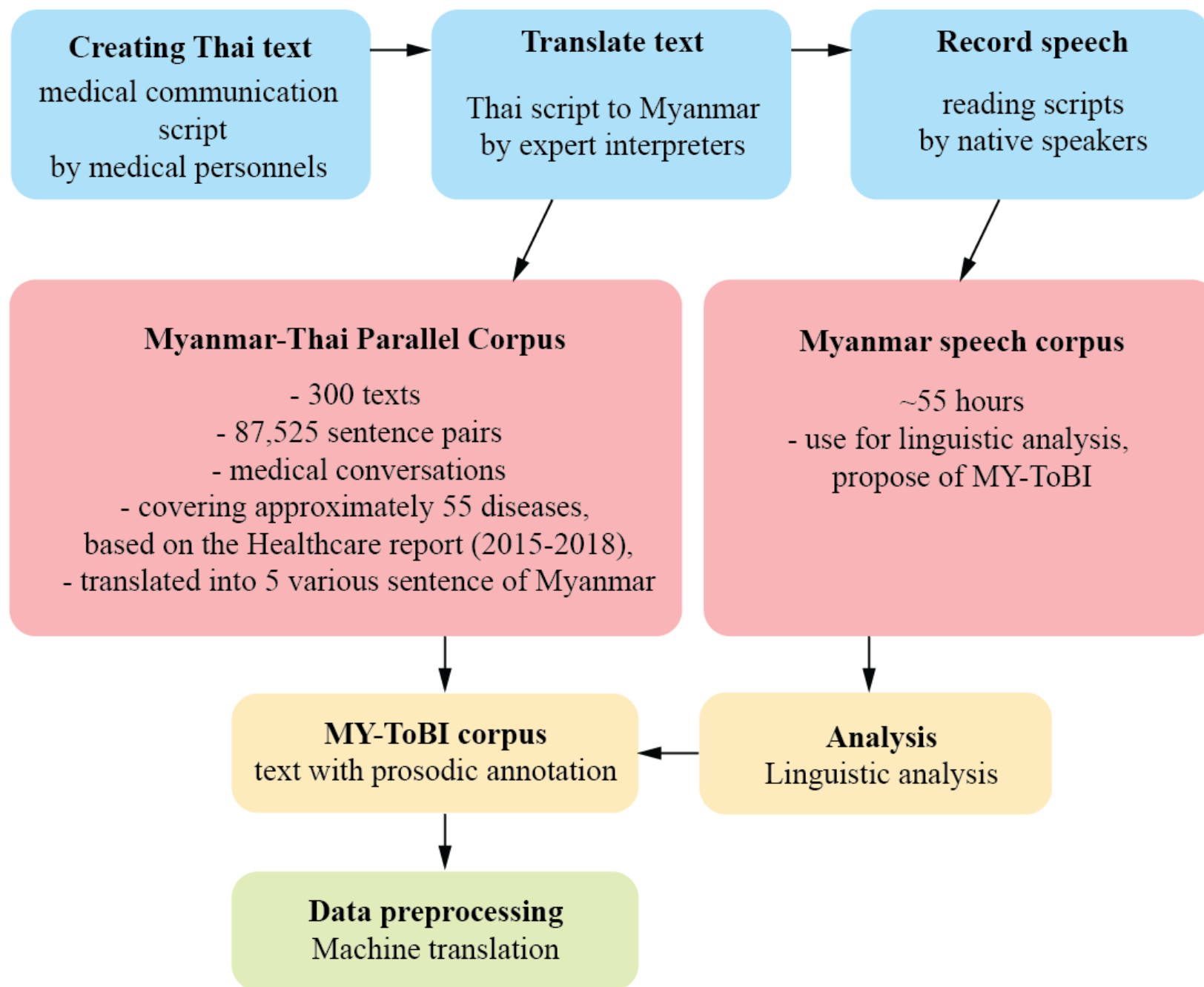
# Corpus Creation concerning NLP

---

## *Corpus Design*

- What Corpus ? : Parallel Corpus , Specific Corpus , annotation corpus, phrase/sentence corpus
- Corpus size ? : 87k
- Parallel corpus text : MY-TH
- What corpus about? What are text about? : Medical communication / healthcare service conversation
- Speech corpus ? : recording speech for phonological analysis
- How to record ? : according to linguistic fieldwork methodology
- Who will tag / annotating? manual ? or Auto-? : Manual tag







# Corpus Creation Methodology

---

## METHOD

## RESULT

1

**Text and speech data preparation: script and simulation**

Script preparation: script and simulation

Simulation preparation: Participant and Place



**Myanmar-Thai parallel text corpus**  
*[MYTHmed parallel corpus]*

2

**Recording procedure**



**Myanmar Speech Corpus**

3

**Analysis**

Autosegmental-Metrical phonological analysis



**Prosodic structure and intonation structure**

Myanmar Tone and Break Indices (MY-ToBI) Convention



**MY-ToBI proposal**

Transformer-based Neural Machine-Translation experiment



**The model with MY-ToBI feature augmented**

# Corpus Creation Methodology

---

## 1 Text and speech data preparation: script and simulation

Script preparation: script and simulation

Simulation preparation: Participant and Place

Myanmar-Thai parallel text corpus

**87,525 sentences**

**Cover 55 diseases**

### The TH and MY script

The conversation scripts are created from Medical professionals such as nurses, doctors, or physician in Thai language.

The conversation Thai scripts are translated by bilingual native Burmese interpreters

### The scenarios

according to Report of Diseases Surveillance in Foreigners in 2015-2018 from Ministry of Public Health, 2015-2018)

- General disease examination
- Dental treatment,
- Respiratory disease group,
- Infectious disease,
- Psychiatric disease,
- Sexually transmitted disease group
- Emergency room service group

# Corpus Creation Methodology: Speech corpus

## 2 Recording procedure

### 55 hours of Myanmar Speech Corpus

the speeches are recorded from reading script

#### Participant

**Four main speakers were selected.**

The qualification of participants must be:

- Be native and fluent in Myanmar language, official language in Myanmar (Yangon dialect, standard dialect of Burmese)
- Be fair bilingual Myanmar-Thai skill of language
- Have normal speech organs
- Understand the simulated situation set in script

#### Recording

Per recording, two participants will speak following the role in the Q&A conversation scripts. Then switch the role to make variant in sentence style and intonation pattern. The other two participants will do the same for additional variables.

#### Recording tool:

A Zoom H5 Portable recorder together with X/Y mic capsule placed about 25 cm from speaker's mouth 30 degrees off axis. The sampling rate was set to **44,100 Hz with 16 bit per sample. The audio are save as .wav file.**



# Analysis Methodology

---

## 3 Analysis

### Autosegmental-Metrical phonological analysis

Stress, rhythm, tone, intonation

By using acoustic phonetic as tool for analysis [Intensity, duration, pitch]



**Prosodic structure and intonation structure**

### Myanmar Tone and Break Indices (MY-ToBI) Convention

Prosodic labelling system



**MY-ToBI proposal**

### Transformer-based Neural Machine-Translation experiment



**The model with MY-ToBI feature augmented**

Comparative study between parallel plain text corpus and MY-ToBI augmented parallel text corpus.

## DEVELOPING MYANMAR-THAI PARALLEL CORPUS (MYTHmed parallel corpus)

### Parallel text creation example

#### Thai script translate into Myanmar

Example of Myanmar-Thjai parallel corpus in healthcare communication

#### 1 sentence of Thai script to 5 sentences of Myanmar

One-to-five variation of sentences

No	Role	Myanmar Script	Thai script	Thai	Myanmar
1.	แพทย์ ဆရာဝန်	မင်္ဂလာပါခင်ဗျာ။ ဒီနေ့ ဘယ်လိုလက္ခဏာတွေ ခံစားရပါလဲခင်ဗျာ။	สวัสดีครับ วันนี้มีอาการอะไรไม่สบายครับ	ยาตัวนี้จะเป็นยาที่ใช้ทำความสะอาดลำไส้	ဒီဆေးကအူလမ်းကြောင်းကိုရှင်းပေးတဲ့ဆေးဖြစ်ပါတယ်။
2.	ผู้รับบริการ ဧည့်သည်(လူနာ)	ကျွန်မပင်ပန်းတယ်ခံစားနေရတယ်ရှင်။	ฉันรู้สึกเหนื่อยค่ะ	ยาตัวนี้จะเป็นยาที่ใช้ทำความสะอาดลำไส้	ဒီဆေးကအူကိုရှင်းပေးတဲ့ဆေးဖြစ်ပါတယ်။
3.	ဆရာဝန်	ဒီလိုဖြစ်နေတာဘယ်နှရက်ရှိပြီလဲခင်ဗျာ။	เป็นมากี่วันแล้วครับ	ยาตัวนี้จะเป็นยาที่ใช้ทำความสะอาดลำไส้	ဒီဆေးကအူကိုသန့်ရှင်းအောင်လုပ်ပေးတဲ့ဆေးဖြစ်ပါတယ်။
4.	ဧည့်သည်(လူနာ)	ဒီအတောအတွင်းပါဘဲရှင်။	ก็ช่วงนี้แหละค่ะ	ยาตัวนี้จะเป็นยาที่ใช้ทำความสะอาดลำไส้	ဒီဆေးကအူကိုသန့်ရှင်းအောင်လုပ်ပေးတဲ့ဆေးဖြစ်ပါတယ်။
5.	ဆရာဝန်	လက္ခဏာတွေဘယ်လိုပြလဲခင်ဗျာ။	อาการยังงีครับ	ยาตัวนี้จะเป็นยาที่ใช้ทำความสะอาดลำไส้	ဒီဆေးကအူလမ်းကြောင်းကိုရှင်းလင်းအောင်လုပ်ပေးတဲ့ဆေးဖြစ်ပါတယ်။
6.	ဧည့်သည်(လူနာ)	ဘာလုပ်လုပ် ပင်ပန်းတယ်လို့ ခံစားရတယ်။	ทำอะไรก็ดูจะเหนื่อยไปหมดเลยค่ะ		
7.	ဆရာဝန်	ပင်ပန်းတဲ့အလုပ်များလုပ်နေလို့များလားခင်ဗျာ။	ทำงานหนักอะไรหรือเปล่าครับ		
8.	ဧည့်သည်(လူနာ)	မဟုတ်ပါဘူးဒေါက်တာ။ နည်းနည်းလမ်းလျှောက်ရင်ဘဲပင်ပန်းနေပြီ။အရင်တုန်း ကဆို ကိလိုများများလျှောက်နိုင်တယ်။	ก็ไม่เหนื่อย แค่ขยับเดินนิดหน่อยก็ดูเหนื่อย แล้ว เมื่อก่อนเดินเป็นกิโลก็ยังไหว		
9.	ဆရာဝန်	ပြီးတော့ အခြားဘာလက္ခဏာတွေရှိသေးလဲခင်ဗျာ။	แล้วมีอาการอะไรอีกไหมครับ		
10.	ဧည့်သည်(လူနာ)	စဉ်းစားလို့မရဘူးဒေါက်တာ။	นึกไม่ออกกะหมอน		

# Corpus Creation

**Corpus size: 87,525 sentence pairs**

**Myanmar-Thai parallel text corpus**

**Text about medical conversation**

**80,625 sentence pairs without any tag**

**6,900 sentence pairs with tag**

```
1 pd.DataFrame(dataset_dict["my_dataset"]["translation"])
```

Python

	src	tgt
0	အ နာ ပတ် လည် ကို သ န့် ရှင်း ရေး လုပ် ပါ	ท่า ความ สะอาด รีม ขอบ แผล
1	ကျွန် တော် ပင် လယ် ထဲ မှာ ရေ ဆော့ နေ ခဲ့ တယ် ။	ผม เล่น น้ำ ใน ทะเล ครับ
2	ဓာတ် မ တ ည့် မှု ကြောင့် နှာ ခေါင်း ပေါက် ထဲ က မြူး ကပ် လွှာ ရောင် ရမ်း ခြင်း ။	โรค จมูก อักเสบ จาก ภูมิ แพ้
3	ဒီ လို လက္ခဏာ ဖြစ် နေ တာ 2 ရက် လောက် တော့ ရှိ ပါ ပြီ	มี อาการ เมื่อ 2 วัน ที่ แล้ว
4	သူ့ ပါး ပြင် မှာ ဖောင်း ရောင် ပြီး နာ ကျင် နေ တယ် ။	แก้ม ของ เขา บวม และ เจ็บ
...	...	...
80620	စီး က ရက် သောက် တာ	สูบ บุหรี่
80621	နီ နေ တဲ့ ဒဏ် ရာ ကို အ ခြေ အ နေ ကြ ည့် လိုက် ရ သ ဖြင့်	ดู จาก ลักษณะ แผล บูน แดง
80622	နှိပ် လိုက် တာ နဲ့ နာ လွန်း လွန်း ကို ကျွန် တော့် တုန် သွား တယ် ။	กด แล้ว เจ็บ มา ผม ချပ်
80623	လက် ထိပ် က အောက် ဆီ ဂျင် က လည်း ကောင်း ပါ တယ် ။	ค่า ความ อิ่ม ตัว ออกซิเจน จาก ปลาย นิ้ว มือ ยัง ดี อยู่
80624	တော် တော် နာ တယ်	เจ็บ มาก เลย

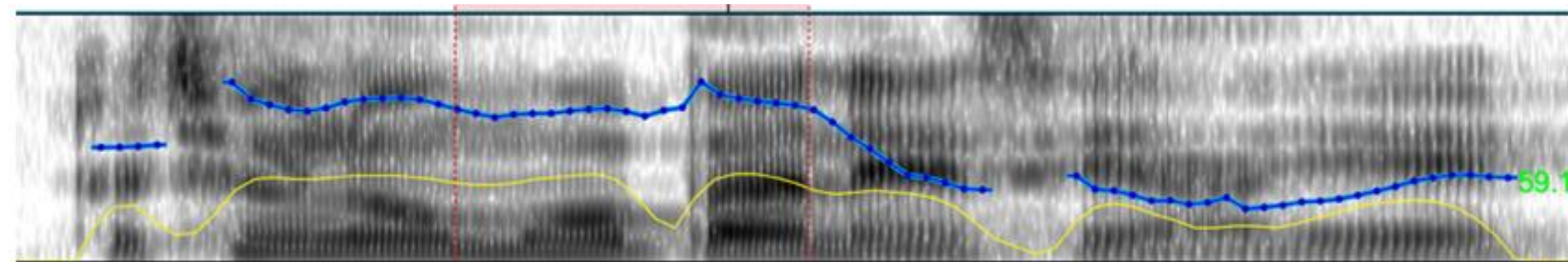
80625 rows × 2 columns

# Speech corpus

## Why Speech ? Why sound recording is needed ?

The sound records are use to analysis of Phonology

Actual voice as acoustic →



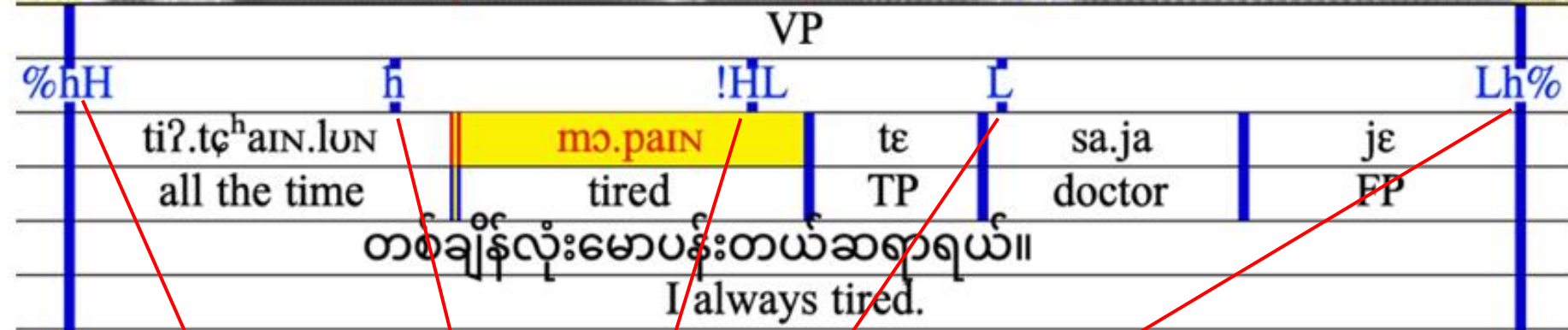
Tone tier (from analysis) →

Phoneme →

Word tier →

Text →

Translation →



Manual tag for MT >>

|%hH| တစ်ချိန်လုံး |h| မောပန်း |!HL| တယ် |L| ဆရာရယ် |Lh%|



TOTAL  
87,525 sentence pairs

80,625 sentence pairs

အ နာ ပတ် လည် ကို သ န် ရှင်း ရေး လုပ် ပါ

6,900 sentence pairs

|%hH| တစ်ချိန်လုံး |h| မောပန်း |!HL| တယ် |L| ဆရာရယ် |Lh%|

```
1 pd.DataFrame(dataset_dict["my_dataset"]["translation"])
```

Python

	src	tgt
0	အ နာ ပတ် လည် ကို သ န် ရှင်း ရေး လုပ် ပါ	ทำ ความ สะอาด ร่ม ขอบ แผล
1	ကျွန်တော် ပင် လယ် ထဲ မှာ ရေ ဆော့ နေ ခဲ့ တယ် ။	ผม เล่น น้ำ ใน ทะเล ครับ
2	ဓာတ် မ တ ည့် မှု ကြောင့် နှာ ခေါင်း ပေါက် ထဲ က မြူး ကပ် လွှာ ရောင် ရမ်း ခြင်း ။	โรค จมูก อักเสบ จาก ภูมิ แพ้
3	ဒီ လို လက္ခဏာ ဖြစ် နေ တာ 2 ရက် လောက် တော့ ရှိ ပါ ပြီ	มี อาการ เมื่อ 2 วัน ที่ แล้
4	သူ့ ပါး ပြင် မှာ မောင်း ရောင် ပြီး နှာ ကျင် နေ တယ် ။	แก้ม ของ เขา บวม และ เจ็บ
...	...	...
80620	စီး က ရက် သောက် တာ	สูบ บุหรี่
80621	နို့ နေ တဲ့ ဒဏ် ရာ ကို အ ခြေ အ နေ ကြ ည့် လိုက် ရ သ ပြီ ၊	ดู จาก ลักษณะ แผล บวมแดง
80622	နှိပ် လိုက် တာ နဲ့ နှာ လွန်း လွန်း ကို ကျွန်တော့် တုန့် သွား တယ် ။	กด แล้ เจ็บ มา ผม จึง ขยับ
80623	လက် ထိပ် က အောက် ဆီ ဂျင် က လည်း ကောင်း ပါ တယ် ။	คำ ความ ชံน ตัว ออกซิเจน จาก ปลาย นิ้ว มีอ ย့် ดี
80624	တော် တော် နာ တယ်	เจ็บ มาก เลย

80625 rows x 2 columns

```
1 pd.DataFrame(dataset_dict["my_dataset"]["translation"])
```

Python

	src	tgt
0	%H  မ နက်  H  တောင်  L  မ  H  လင်း  LH  သေး  H  ပါ ဘူး  L  ။  %h  လူ တွေ  h  လည်း  HL  မ  H  နို့  LH  သေး  H  ပါ ဘူး  L%  ။	เข้า มีด เลຍ คะ ยัง ไม่ มี ใคร ตื่น เลຍ คะ
1	%h  နှာ ခေါင်း  H  ပေါက်  HL  ဝိတ်  L  နေ  H  သ  h  လို  L  မျိုး ဘဲ  H  ခင်  h  ဗျာ  L%	เหมือน ร จมูก ปัด เลຍ ครับ
2	%H  စက် ရုံ  h  က  H  နှာ  L  ခေါင်း စည်း  H  ကို  L%	พวก หน้ากาก ที่ โรง งาน แยก
3	%h  ဒီ အ  H  ခါ မှာ  h  တော့  H  ဗီ  LH  တာ  L  မင်  HL  သောက်  H  ဖို့ မှာ  h  လိုက်  H  မယ် ခင် ဗျာ  LH%  ။	ครั้งนี้ หมอ จะ สั่ง วิตามิน
4	%h  ဘ ကြီး မ သေ  HL  ချာ ဘူး  H  ဖြစ်  LH  နေ  H  တယ်  L%  ။	ลุง ไม่ มันใจ นะ
...	...	...
6895	%h  သန္ဓေ တား ဆေး  H  တွေ သောက်  HL  လား ခင် ဗျာ  L%	กิน ยาคุม กำเนิด ไหม ครับ
6896	%h  ဒီ ရက် နောက် ပိုင်း  HL  လူ နာ  L  က နေ သွား  h  တိုက် နည်း  H  မှန် အောင် တိုက်  H  ပေး ပါ  L%	หลังจาก ที่ คน ไหม่ ลอง ไป ပြော ပုံ ဝယ် ရွက်
6897	%LH  မှန် ပြောင်း မ ကြ ည့် ခင် လိုက် နာ လုပ် ဆောင် ရ မဲ့  H  အ ချက် တွေ ရှိ လို့  L  နား ထောင် ပေး  h  ပါ  L%  ။	เดี๋ยว เข็ญ รับ ฟัง เรื่อง การ ปฏิบัติ ตัว ก่อน มา
6898	%H  သူ့ မျက် နှာ အ ရောင် က  H  ဘယ် လို ဖြစ် နေ ပြီ လဲ  L%  ။	สีหน้า เขา เป็น อย่างไร
6899	%H  အ ကယ် လို့  H  ရော  L  ဂါ အ ခံ ရှိ နေ  L  ရင် အဲ့ ဒါ  L  ကို ဦး  H  ဆုံး ထိ န်း ရ ပါ  L  မယ်  L%  ။	หาก มี โรคประจำตัว ควร รับ การรักษา จน

6900 rows x 2 columns



# Corpus as Data preprocessing

	No. of sen.	No. of Wd.	No. of unique Wd.
Myanmar side	87,525	722,028	13,678
Thai side	87,525	616,695	10,203

	First dataset	Second dataset
	Baseline model data / sentence pairs	Finetuned model data / sentence pairs with ToBI tag on Myanmar side
Train	64,500	5,520
Validation	8,062	1,380
Test	8,063	0
<b>Total</b> each dataset	<b>80,625</b>	<b>6,900</b>
<b>Total</b> sentence pairs on corpus created	<b>87,525</b>	

Derive from  
phonological analysis

`#additional_special_tokens = ["|H|", "|L|", "|h|", "|hL|", "|hH|", "|HL|", "|LH|", "|%H|", "|%h|",  
"|%hH|", "|%L|", "|%LH|", "|H%|", "|L%|", "|h%|", "|hH%|", "|LH%|"]`

`#additional_special_tokens = sorted(additional_special_tokens, key=len, reverse=True)`

# Scoring & translation analysis on MT

Model	BLEU	Meteor	chrF	TER <sup>35</sup>
Baseline model	25.05	0.43	66.72	182.70
Finetuned-model-withToBI	<b>32.52</b>	<b>0.50</b>	<b>70.02</b>	<b>142.69</b>
Finetuned-model-without ToBI	28.81	0.44	66.79	174.10

<b>7.11b</b>	<b>Sentence type</b>	Yes-No Question
	<b>Source</b>	အ ရမ်း အန္တရာယ် များ လား ခင် ဗျာ ။
	<i>[meaning]</i>	<i>Is this so dangerous?</i>
	<b>Reference</b>	เป็น อันตราย มาก ไหม
	Baseline	ใช่ สูง มาก เลย นะ ครับ
	Finetuned-withToBI	อันตราย มาก ไหม ครับ
	Finetuned-without ToBI	เยอะ ไหม ครับ

the conclusion from the question that “How will ToBI help improve better translation?”  
I found that the translation is better because some words usually have same intonation

ə ... ɔ̌ /ma.....bu/ 'no/ not'

[h.....L] tone event

ə|h| ... ɔ̌|L|

By detecting these markers, the surrounding words of those word with ToBI are sequenced better.

# Limitations and further suggestions

## Limitations

- Self creation corpus is time-consuming task and has a high-cost for generating various data. The reason is Machine translation need amount of data to run model and tagged data requires time to be tagged by hand. There is no app or software for prosodic or ToBI tagging like Part-Of-Speech for now. Therefore, tagging ToBI in detail may not provide enough data If tagged data with all detailed. So I chose to rough tagged in the data that is, not using all the symbols that have been proposed on chapter 6.

## Further suggestions

- Tagged data can be more complex such text|POS|ToBI|etc. due to linguistic feature addition. This supposed to give better translation.
- Tagging linguistic features on a large data should be operated by software or any application such POS-tagger, because tagging each word, sentence, or text on dataset are admit that it is labouring and time-consuming task. For instance, ToBI tagging with application can be operate by detecting range of fundamental frequency in audio files then annotate at least tonal event such H, h, L.

# Thank you

---

[tanintornlimp@gmail.com](mailto:tanintornlimp@gmail.com)