

Innovating Myanmar's Future with NLP/AI

Ye Kyaw Thu
LU Lab., Myanmar
NECTEC, Thailand
Email: ykt.nlp.ai@gmail.com

*26 Nov 2024
@GUSTO Innovation
Bootcamp 2024, Yangon,
Myanmar*

Outlines

1. Who Am I?
2. R&D in AI (Language acquisition)
3. R&D in NLP
4. Hacking Lipidipika
5. Key Takeaways

Who Am I?

- NLP/AI Researcher
- Especially focus on under resourced languages such as Burmese, Khmer, Thai, sign language, Braille
- Corpus development, Machine Translation (TM), Automatic Speech Recognition (ASR), Text to Speech (TTS), Language Acquisition for Robot, Speech Translation, etc.
- Homepage: <https://sites.google.com/site/yekyawthunlp/>
- GitHub: <https://github.com/ye-kyaw-thu>

Who Am I?



- Doctor of Science, GITS, Waseda University, Tokyo, Japan (26th October 2011)
- Master of Science, GITS, Waseda University, Tokyo, Japan (15th March 2006)
- Graduation of One Year Special Japanese Language Course, The Japanese Language School of the International Student Institute (国際学友会日本語学校), Tokyo, Japan (12th March 2004)
- Bachelor of Science (Physics), Dagon University, North Dagon, Myanmar (15th December 2000)
- International Advanced Diploma in Computer Studies, NCC Education, UK (13th December 1999)
- International Diploma in Computer Studies, NCC Education, UK (13th June 1998)

Who Am I? (@Waseda Univ, Tokyo, Japan)



- Research Associate (Apr 2009 - Mar 2012)
- Lab က
မဟာတန်းကျောင်းသားတွေရဲ့ thesis
ကံကြကည်ပေးရ
- ကုယ့်ကျောင်းရဲ့ event တွေမှ
ကူညီရ (Entrance exam, interview,
Freshers welcome, workshop etc.)
- Attending weekly faculty meetings
- Writing meeting notes in Japanese

Fig. Okuma Auditorium of Waseda Campus

Who Am I? (@NICT, Kyoto, Japan)



Fig. National Institute of Information and Communications Technology (NICT), Kyoto, Japan
(Photo: 10 Feb 2014)

- Researcher (Apr 2012 to Mar 2016)
- VoiceTra (Voice to Voice Translation Project)
ပရောဂျက်မှ
မြန်မာစာကို
ထည့်လိုပြန့်
ခြုံဆောင်
- Machine Translation R&D

Who Am I? (@NICT, Kyoto, Japan)



- ကိုယ်တိုင် editing
လုပ်ပေးခဲ့တဲ့ flyer ပါ
 - Link:
<https://voicetra.nict.go.jp/en/>

Who Am I? (@OPU, Okayama, Japan)



- Researcher (Sept 2016 to Mar 2018)
- Advising undergrad/master's students
- Language acquisition for robot

Fig. Okayama Prefectural University (OPU), Okayama, Japan

Who Am I? (@OPU, Okayama, Japan)

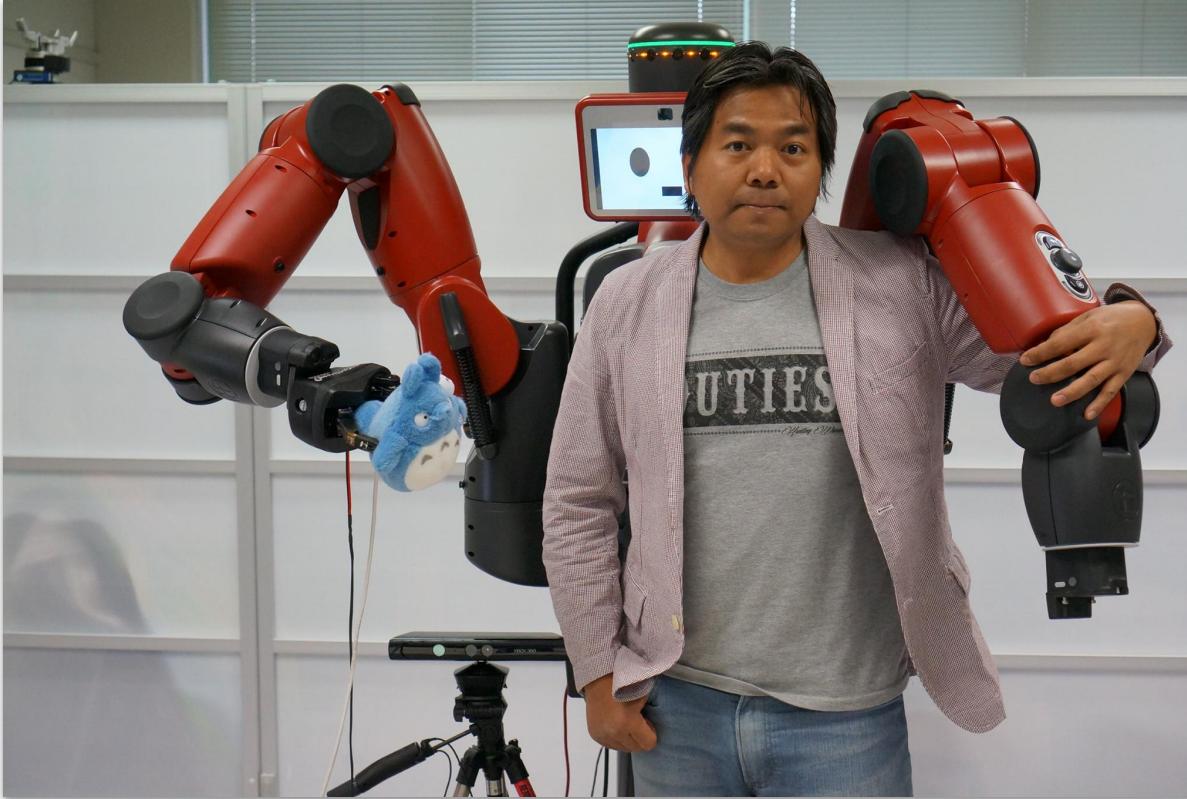


Fig. Ye with Baxter Robot at AI Lab., Okayama Prefectural University (OPU)

Who Am I? (@NECTEC, Khlong Luang, Thailand)



- Visiting Professor
(from Jan 2019 to Present)
- Medical domain projects
- Supervising students at various institutions, including AU, KU, KMITL, SIIT

Fig. A photo with KMITL internship students at NECTEC office

Who Am I? (@LU Lab., Pyin Oo Lwin, Myanmar)



- Lab Leader at LU Lab., Myanmar
- 2019 မေလမှ
စတည်ထောင်
- Internship Camp
- R&Ds with master's
and doctoral
students

Fig. A group photo with 2019 internship students at LU Lab.

Who Am I?

The screenshot shows the GitHub profile of user 'ye-kyaw-thu'. The profile features a circular profile picture with a drawing of a face. The user has 486 followers and is following 2 people. Their email is listed as yktntp@gmail.com. The 'Pinned' section contains six projects:

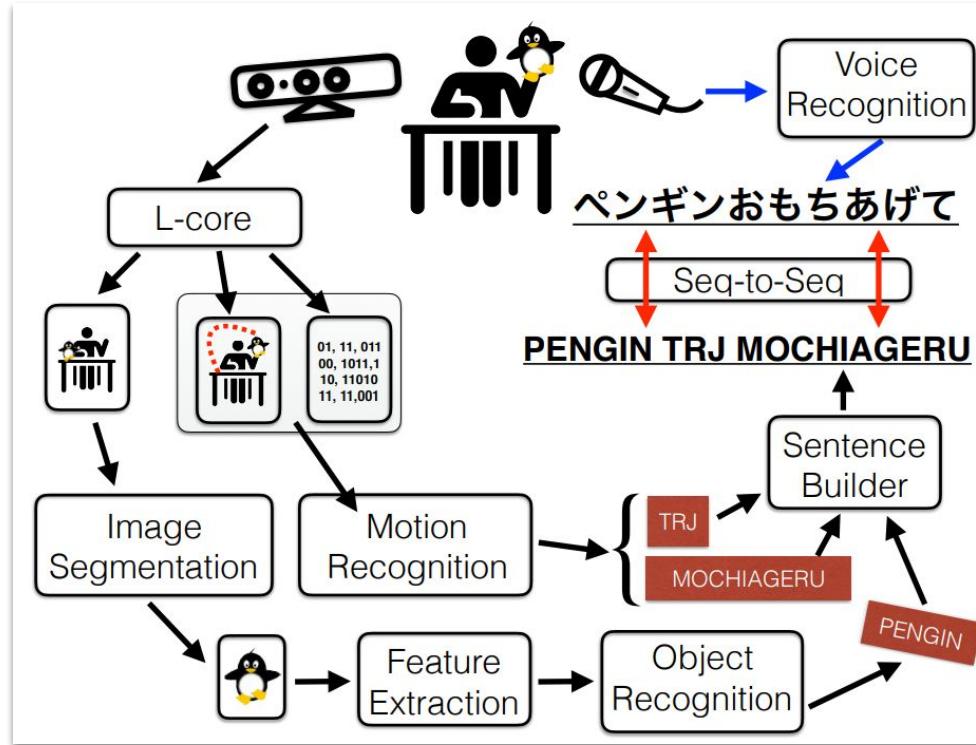
- myParaphrase** (Public) - Paraphrase Dataset for Burmese (Myanmar Language). A Jupyter Notebook with 4 stars.
- myUDTree** (Public) - Universal Dependency Tree for Myanmar Lan. 10 stars, 1 fork.
- myWord** (Public) - syllable, word and phrase segmenter for Burmese (Myanmar language). Python code with 48 stars, 8 forks.
- papers** (Public) - Some published papers. 10 stars, 1 fork.
- myPOS** (Public) - myPOS (Myanmar Part-of-Speech) Corpus for Myanmar NLP Research and Developments. Python code with 68 stars, 14 forks.
- myG2P** (Public) - Myanmar (Burmese) Language Grapheme to Dictionary for speech recognition (ASR) and. Perl code with 53 stars, 9 forks.

Below the pinned projects, it says '819 contributions in the last year' with a timeline from Dec to Nov.

Fig. Check this GitHub page

- လုပ်ခွဲတဲ့ NLP R&D
အလုပ်တွေနဲ့
ပတ်သက်တဲ့
ဒေတာ၊ မောဒယ်၊
ပရိုဂရမ်တွေကို
ရှုနှင်သမျှ
ရှုပေးထားပါတယ်
- GitHub:
<https://github.com/ye-kyaw-thu>

R&D in AI (Language acquisition)



- စက်ရှပ်ကို voice command တချို့ပေးပြီး ခိုင်းတာကိုလုပ်တတ်အောင် သင်ပေးတဲ့ ပရောဂျက်
- ကိုယ်ကျမ်းကျင်တဲ့ Machine translation ကို အသုံးချခဲ့

Fig. Overview of robot language acquisition (an example with conceptual structure “PENGIN TRJ MOCHIAGERU” and Japanese syllable sequences “ペンギンおもちあげて” (Move up penguin))

R&D in AI (Language acquisition)



(a)
Lunch-
box



(b)
Piggy-
bank



(c)
Emo



(d)
Box



(e)
Goldfish



(f)
Accessory-
box



(g)
Cup



(h)
Penguin



(i)
Pikachyuu



(j)
Totoro

- Object ၁၀ မျိုး သုံးခဲ့
- lunch-box,
piggybank, emo, box,
goldfish, accessory
box, cup, penguin,
Pikachyuu and
Totoro
- အရောင်၊
ပုံသဏ္ဌန်အမျိုးမျိုး

Fig. Ten objects

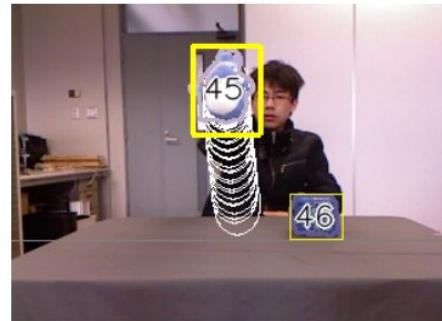
R&D in AI (Language acquisition)



(a) Move-close-to



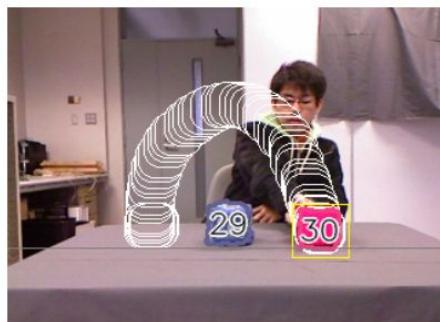
(b) Move-away



(c) Move-up



(d) Move-onto



(e) Move-over



(f) Move-circle

Fig. Ten motions

R&D in AI (Language acquisition)

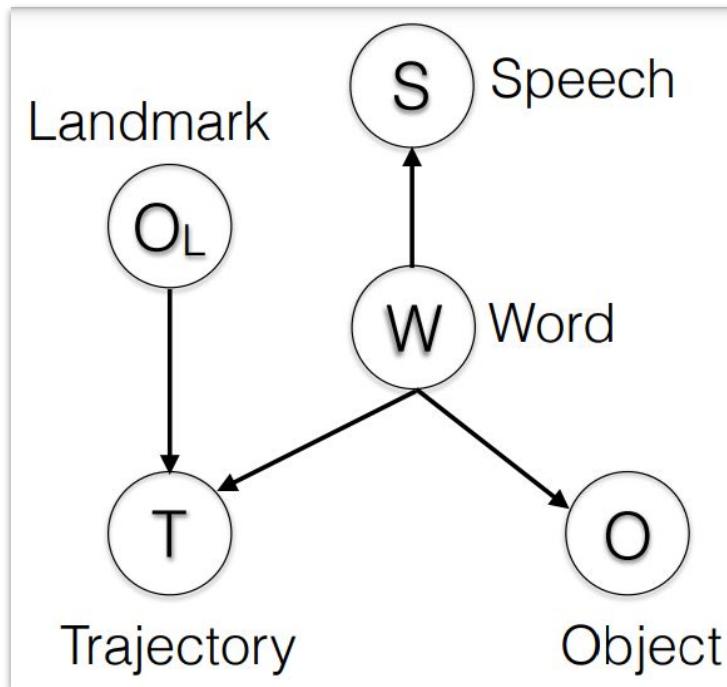


Fig. Graphical model of a lexicon containing words referring to objects and motions

- Only seven grammar patterns in total as follows:
 1. OBJECT LND OBJECT TRJ HANASU
 2. OBJECT TRJ MAWASU
 3. OBJECT TRJ MOCHIAGERU
 4. OBJECT TRJ OBJECT LND CHIKAZUKERU
 5. OBJECT TRJ OBJECT LND HANASU
 6. OBJECT TRJ OBJECT LND NOSERU
 7. OBJECT TRJ OBJECT LND TOBIKOESASERU

R&D in AI (Language acquisition)

BENTOO LND	Ⅲ	べんとお	Ⅲ	0.310345	0.0274354	...	
BENTOO LND	Ⅲ	べんとおに	Ⅲ	1	0.252406	0.566667	...
BENTOO LND	Ⅲ	ぺんとお	Ⅲ	0.13171	0.0102134	...	
BENTOO LND	Ⅲ	ぺんとおに	Ⅲ	0.852853	0.0939635	...	
BENTOO TRJ	Ⅲ	べんとお	Ⅲ	0.482759	0.349605	...	
BENTOO TRJ	Ⅲ	べんとおお	Ⅲ	1	0.349605	0.5625	...

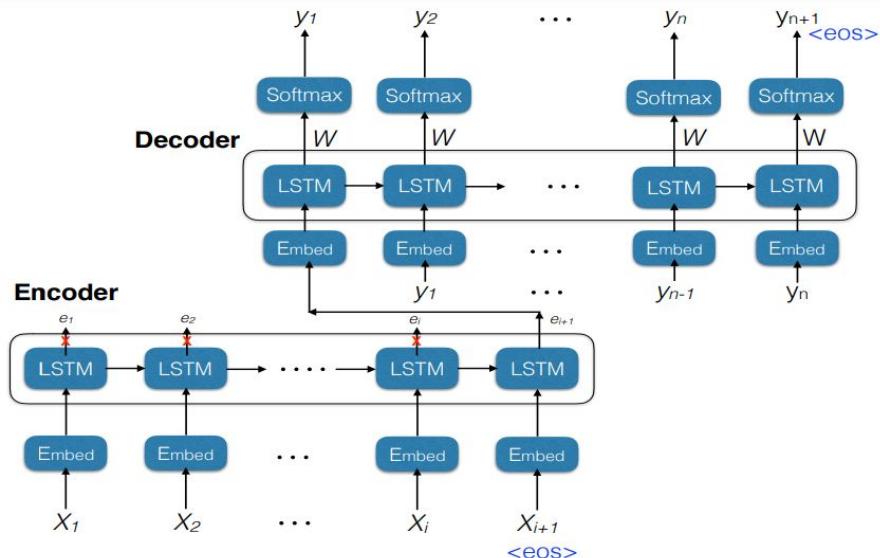
Fig. Phrase table of PBSMT (Phrase based Statistical Machine Translation)

R&D in AI (Language acquisition)

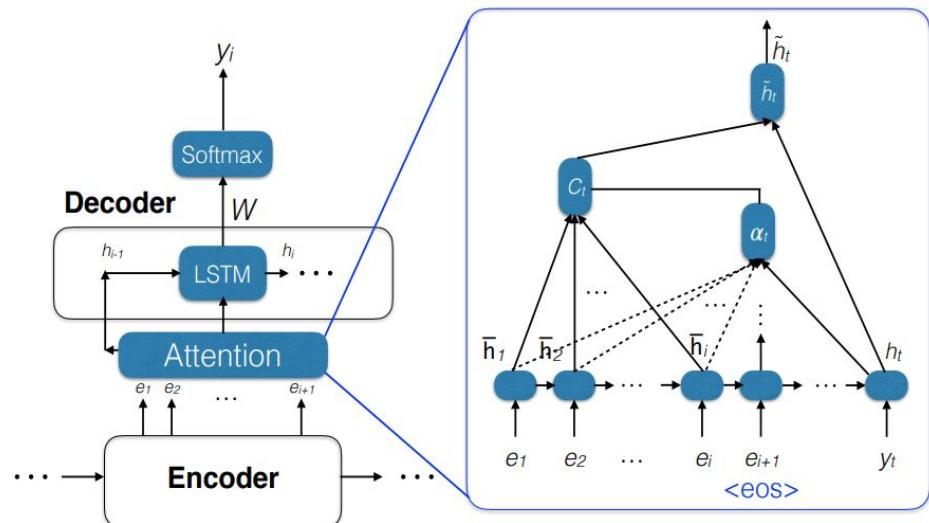
BENTOO LND [X] Ⅲ べんとお [X] Ⅲ 0.310345 0.0274354 ...
BENTOO LND [X] Ⅲ べんとおに [X] Ⅲ 1 0.252406 0.566667 ...
BENTOO LND [X] Ⅲ ぺんとお [X] Ⅲ 0.126784 0.0102134 ...
BENTOO LND [X] Ⅲ ぺんとおに [X] Ⅲ 0.850649 0.0939635 ...
BENTOO TRJ [X] Ⅲ べんとお [X] Ⅲ 0.482759 0.349605 ...
BENTOO TRJ [X] Ⅲ べんとおお [X] Ⅲ 1 0.349605 0.5625 ...

Fig. Phrase table of HPBSMT (Hierarchical Phrase based Statistical Machine Translation)

R&D in AI (Language acquisition)



(a) Encoder-Decoder translation model



(b) Encoder-Decoder with attention model

Fig. Sequence-to-Sequence learning model

R&D in AI (Language acquisition)

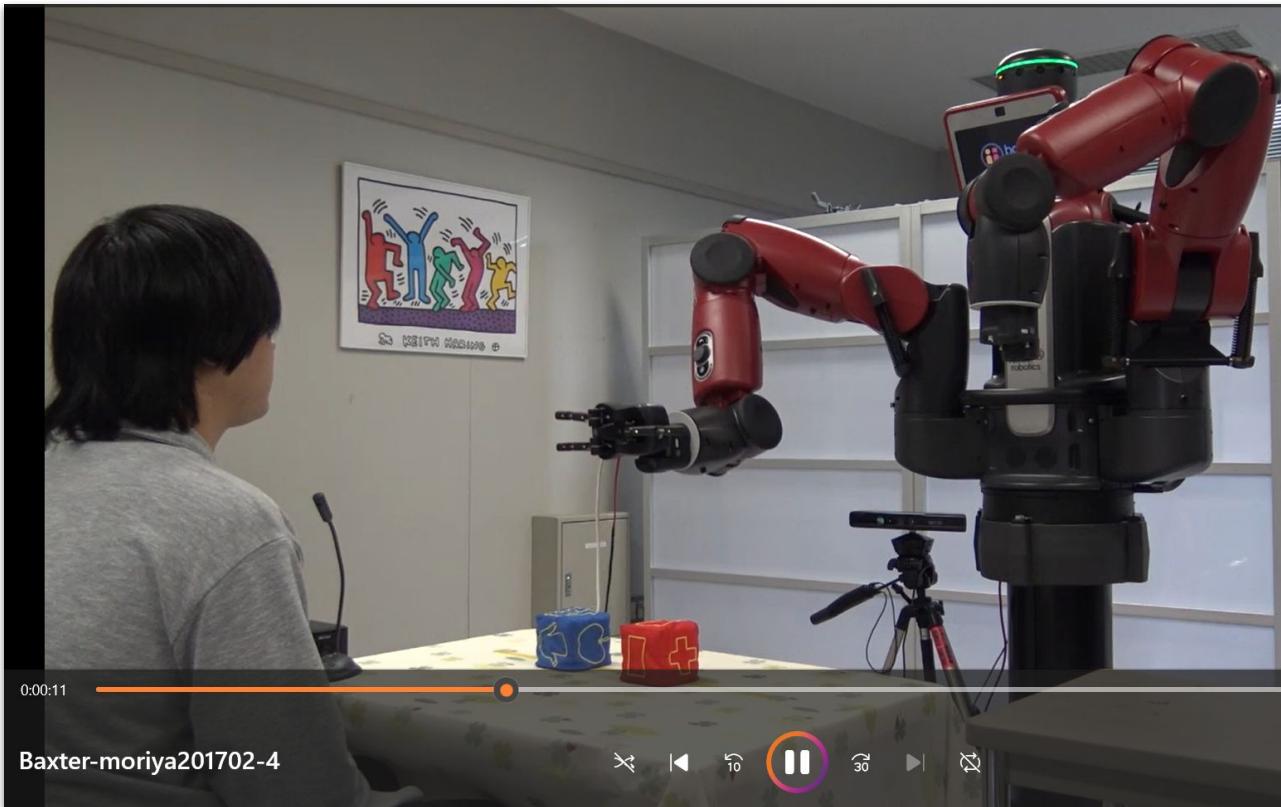


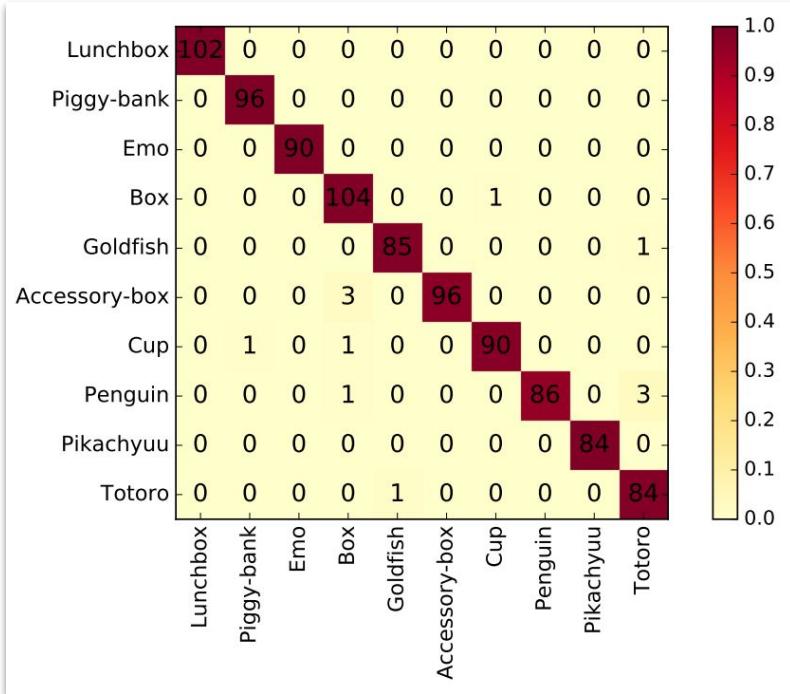
Fig. Training with Baxter robot

R&D in AI (Language acquisition)



Fig. An example of ASR error on “Move-up Totoro” (Totoro o mochiagete) sentence together with related image

R&D in AI (Language acquisition)



- CAFFE deep learning framework with open trained network model of IMAGENET is used for image features extraction from segmented object images.
- The extracted features are used for training object recognition. Although we trained several unsupervised classifier such as Complex Tree, KNN, we selected to use highest accuracy classifier Gaussian-kernel SVM for object classification.

Fig. Confusion matrix of object recognition by Gaussian Kernel SVM classification

R&D in AI (Language acquisition)

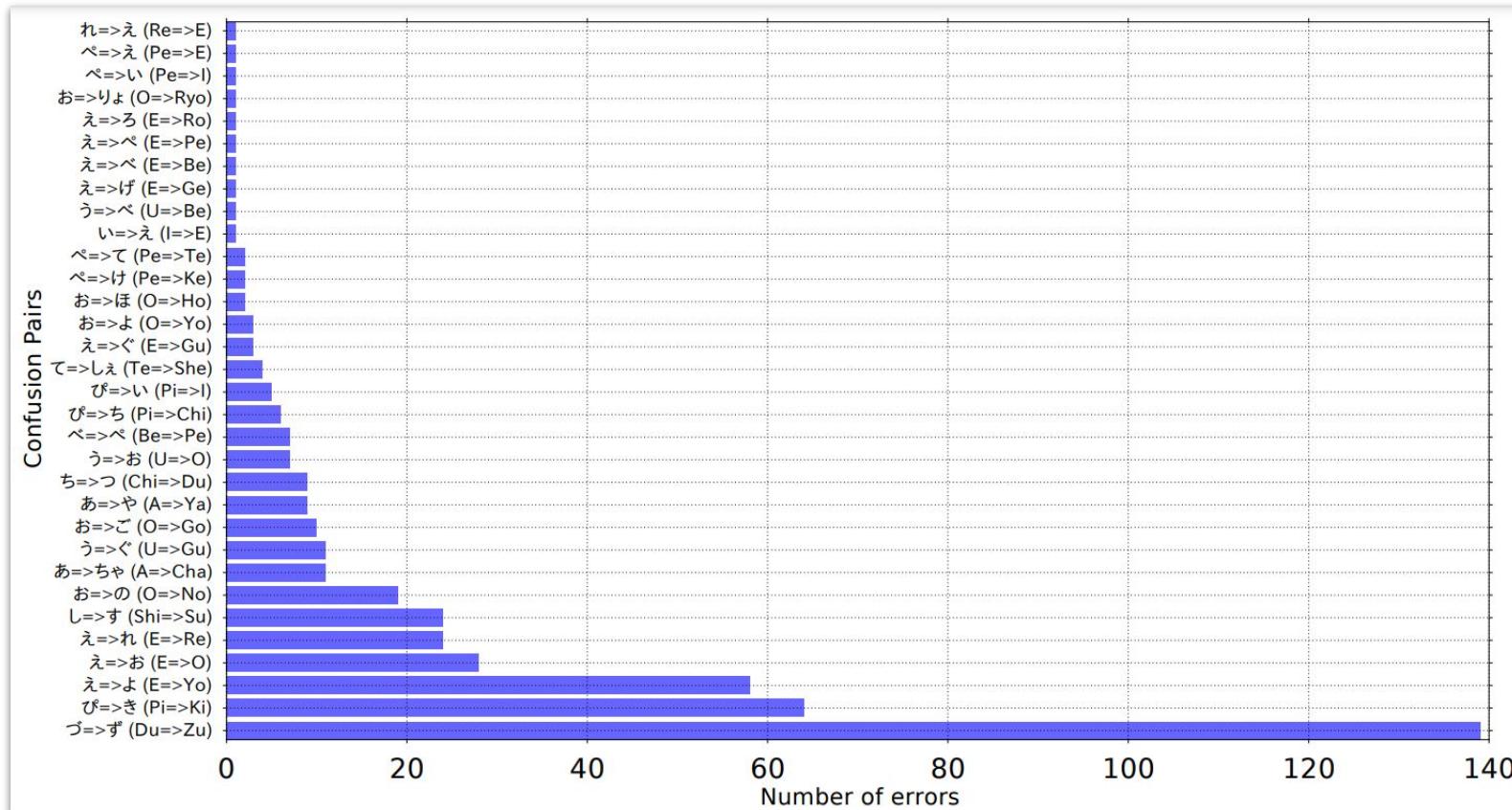


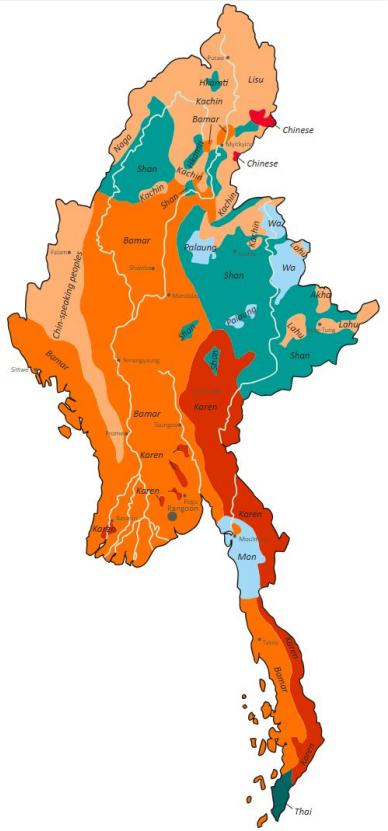
Fig. Confusion pairs of Japanese Automatic Speech Recognition (ASR)

R&D in AI (Language acquisition)

MT Methods	CS-to-SS			SS-to-CS		
	Baseline	+ASR Recog	+Object Recog	Baseline	+ASR Recog	+Object Recog
PBSMT	79.68%	71.36%	77.61%	46.69%	46.91%	44.88%
HPBSMT	79.68%	70.71%	77.83%	46.69%	43.50%	45.10%
Encoder-Decoder	100.00%	81.93%	99.22%	100.00%	99.28%	100.00%
Attention	92.53%	77.41%	88.14%	100.00%	95.98%	100.00%

Table. BLEU scores for machine translation between conceptual structure (CS) and syllable sequences (SS) (+ ASR Recog denotes the result with automatic speech recognition, + Object Recog denotes the result with Object recognition)

R&D in NLP (Motivation)



- Approximately a hundred languages spoken in Myanmar
- Burmese is the official language and spoken by two third of the population
- Languages spoken by ethnic minorities represent six language families: Sino-Tibetan, Austro-Asiatic, Tai-Kadai, Indo-European, Austronesian, and Hmong-Mien
- Reference:
https://en.wikipedia.org/wiki/Languages_of_Myanmar

Fig. Ethnolinguistic map of Myanmar (1972, Wikipedia)

R&D in NLP (Motivation)

rk: ကလေချွဲ ထိ ဘေးလုံး ကျောက် နှိမ်ရေး ||
my: ကောင်လေး တွေ့ ဘေးလုံး ကန် နေတယ် ||
("Boys are playing football" in English)

bm: သူ သစ်ပင် တွေ့ ပန်းပင် တွေ့ စိုက်တယ် ||
po: ငွေ့ ဆို့ သောင်းမွှဲး ဖုံး ကပ်းမွှဲး ဖုံး ||
English: He planted trees and flowers.

dw: ကောနသား ကြောန်း မှန်းမှန် သွား ဟူယ် ||
my: ကောင်လေး ကျောင်း မှန်မှန် တက် တယ် ||
("The boy goes to school regularly" in English)

bk : ငါ မောလင်း နိုင်ငံခြား သော မယ် ||
my : ကျွန်တော် မန်က်ဖြန် နိုင်ငံခြား သွား မယ် ||
("I will go foreign tomorrow ." in English)

Fig. Example sentences of Rakhine, Pao, Dawei and Beik languages

R&D in NLP (Motivation)

my: ମେଲାଃଧିପଃ ମତଃ କ୍ଷେତ୍ରବନ୍ଦୀ ॥
 kc: Ma ni gasup taw nga ma ai .
 rw: CVMRE RI GVSØP MÈ .

Fig. Example sentences of Kayah, Kachin and Rawang

R&D in NLP (Motivation)

Burmese	ကခေယင်ဆရွှေ့ညွှန့်ချုပ်ကဲတတ္ထဒနပါးမယရလဝသသယာင့်အ
Sanskrit	ကခေယင်ဆရွှေ့ညွှန့်ချုပ်ကဲတတ္ထဒနပါးမယရလဝရမသသယာင့်အ
Arakanese	ကခေယင်ဆရွှေ့ညွှန့်ချုပ်ကဲတတ္ထဒနပါးမယရလဝသသယာင့်အ
Mon	ကခေယင်ဆရွှေ့ညွှန့်ချုပ်ကဲတတ္ထဒနပါးမယရလဝသသယာင့်အ၏မြှု
Thai Mon	ကခေယင်ဆရွှေ့ညွှန့်ချုပ်ကဲတတ္ထဒနပါးမယရလဝသသယာင့်အ၏မြှုအမြှု
Letalanyah	ကခေယင်ဆရွှေ့ညွှန့်ချုပ်ကဲတတ္ထဒနပါးမယရလဝသသယာင့်အ၏မြှုအမြှု
S'gaw	ကခေယင်ဆရွှေ့ညွှန့်ချုပ်ကဲတတ္ထဒနပါးမယရလဝသသယာင့်အ၏မြှု
E Pwo	ကခေယင်ဆရွှေ့ညွှန့်ချုပ်ကဲတတ္ထဒနပါးမယရလဝသသယာင့်အ၏မြှု
W Pwo	ကခေယင်ဆရွှေ့ညွှန့်ချုပ်ကဲတတ္ထဒနပါးမယရလဝသသယာင့်အ၏မြှု
Pa'O	ကခေယင်ဆရွှေ့ညွှန့်ချုပ်ကဲတတ္ထဒနပါးမယရလဝသသယာင့်အ၏မြှု
Geba	ကခေယင်ဆရွှေ့ညွှန့်ချုပ်ကဲတတ္ထဒနပါးမယရလဝသသယာင့်အ၏မြှု
Karen Ni	ကခေယင်ဆရွှေ့ညွှန့်ချုပ်ကဲတတ္ထဒနပါးမယရလဝသသယာင့်အ၏မြှု
Rumai Palaung	ကခေယင်ဆရွှေ့ညွှန့်ချုပ်ကဲတတ္ထဒနပါးမယရလဝဝသသယာင့်အ၏မြှု
Pale Palaung	ကခေယင်ဆရွှေ့ညွှန့်ချုပ်ကဲတတ္ထဒနပါးမယရလဝဘအဝါ
Shwe Palaung	ကခေယင်ဆရွှေ့ညွှန့်ချုပ်ကဲတတ္ထဒနပါးမယရလဝဘအဝါ
Asho	ကခေယင်ဆရွှေ့ညွှန့်ချုပ်ကဲတတ္ထဒနပါးမယရလဝဘအဝါ
Khamee	ကခေယင်ဆရွှေ့ညွှန့်ချုပ်ကဲတတ္ထဒနပါးမယရလဝဘအဝါ
Moken	ကခေယင်ဆရွှေ့ညွှန့်ချုပ်ကဲတတ္ထဒနပါးမယရလဝဘအဝါ
Shan	ရခင်သရုတ်တအပေါ်မယရလဝရှိက ဇော်ဗော
Shan Pali	ရခင်သရုတ်တအပေါ်မယရလဝရှိက ဇော်ဗော
Shan Ni	ကရကရင်သရုတ်တအပေါ်မယရလဝရှိက ဇော်ဗော
Khamti	ကရကရင်သရုတ်တအပေါ်မယရလဝရှိက ဇော်ဗော

- Font Development for Ethnic Languages
- Source: from Ben Mitchell

R&D in NLP (Corpus building, myPOS)

- 1 ဒီ/adj ဆေး/n ၊/ppm ၁၀၀/num ရာခိုင်နှုန်း/n ဆေးဘက်ဝင်/adj အပင်/n များ/part မှ/ppm ဖော်စပ်/v ထား/part တဲ့/part ဖွစ်/v တယ်/ppm ။/punc
- 2 အသစ်/n စား/v ထား/part တဲ့/part ဆူယ်တဲ့/n ၊/ppm အသီး/n စား/v နေ့/part ပါ/part ပေါ့/part ။/punc
- 3 မှ/part ကျွန်းမာ/v လျှင်/conj နတ်/n | ဆရာ/n ထံ/ppm မေးမြန်း/v ၍၏/conj သက်ဆိုင်ရာ/n နတ်/n တို့/part အား/ppm ပူဇော်ပသာ/v ရာ/part သည်/ppm ။/punc
- 4 ပေဟိုင်/n | ပုယျာဗုံး/n ။/punc
- 5 နဝမှ/adj အိပ်မက်/n ကောသလာ/n | မင်း/n | အိပ်မက်/n ရဲ/num နက်ရှိုင်း/adj ကျယ်ဝန်း/adj ဘေး/part ရေကန်/n ၍၏း/adj တစ်/tn ခု/part တွင်/ppm သတ္တဝါ/n တို့/part ဆင်း/v ၍၏/conj ရေဘော်/v ၍၏/part ၏/ppm ။/punc
- 6 အပြင်ပန်း/n ၍၏ုံး/v ရင်/conj ခက်/adj သလို/part ထင်း/v ရာ/part ပေမယ့်/conj တကယ့်/adj လက်တွေ့/n အခြေအနေး/n ၊/ppm တွေ့/part အဲဒီ/pron လို့/ppm မှ/part ဟုတ်/v ပါ/part ဘူး/part ။/punc
- 7 8/fw bit/fw ပုံရိပ်/n တစ်/tn ခု/part သည်/ppm 256/fw color/fw သို့မဟုတ်/conj gray/fw scale/fw များ/part တို့/ppm အထောက်အကူး/n ၍၏့/v သည်/ppm ။/punc

Fig. Example sentences from the myPOS (Version 3.0) corpus

- Github link: <https://github.com/ye-kyaw-thu/myPOS/tree/master/corpus-ver-3.0>

R&D in NLP (Corpus building, myPOS)

Unit	myPOS (ver. 1.0)	Ext-1: my-zh	Ext-2: my-ko	Ext-3: ASEAN-MT my	myPOS (ver. 3.0)
Sentences	11,000	10,000	10,052	12,144	43,196
Words	239,598	103,909	106,864	114,134	564,505
Average Words/Sentence	21.78	10.17	10.64	9.40	13.07

Table. Statistical information of myPOS version 1.0 to 3.0

- Corpus development က အချိန်ပေးရတယ်
- မှတ်မံသလောက် myPOS အတွက် ဆိုရင် 2016 က စ လုပ်ဖြစ်ခဲ့တယ်
- 2017 မှာ version 1.0 release လုပ်နိုင်တယ်
- 2020 မှာမှ version 3.0 ကို release လုပ်နိုင်ခဲ့တယ်

R&D in NLP (Corpus building, myUDTree)

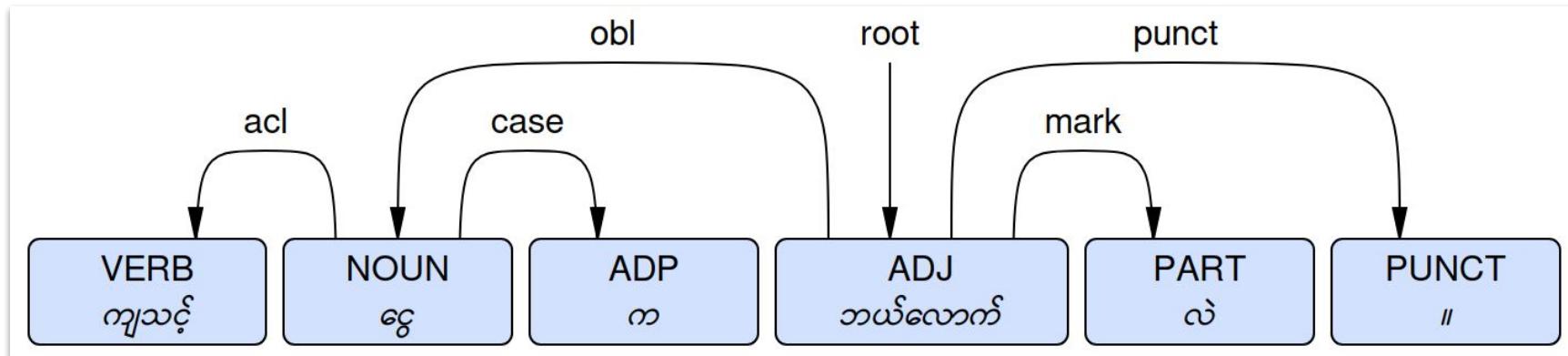


Fig. Dependency tree for the example sentence: "ကျောင့်ငွေ က ဘယ်လောက်ပဲ"

- Universal Dependencies (UD) Corpus is a type of corpus that is annotated according to the grammar rule of the respective languages with Part-of-Speech (POS), morphological features, and syntactic dependencies.
- Github link: <https://github.com/ye-kyaw-thu/myUDTree>

R&D in NLP (sylbreak, released in 2017)

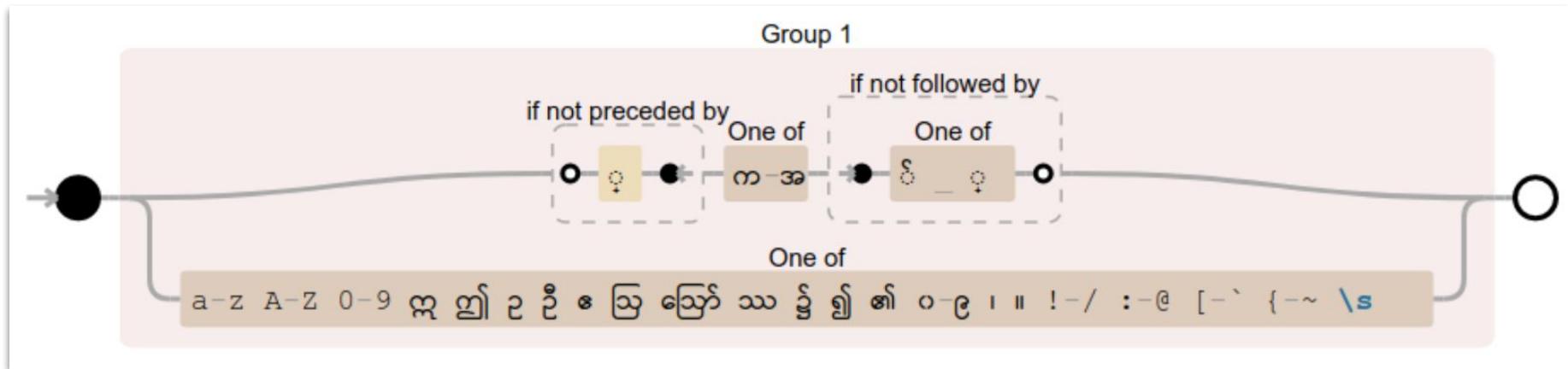


Fig. Visualization of syllable breaking regular expression (RE) for Myanmar

- If you use shell (`sylbreak.sh`), perl (`sylbreak.pl`) and python (`sylbreak.py`) scripts, no need to make installation.
 - Github link: <https://github.com/ye-kyaw-thu/sylbreak>

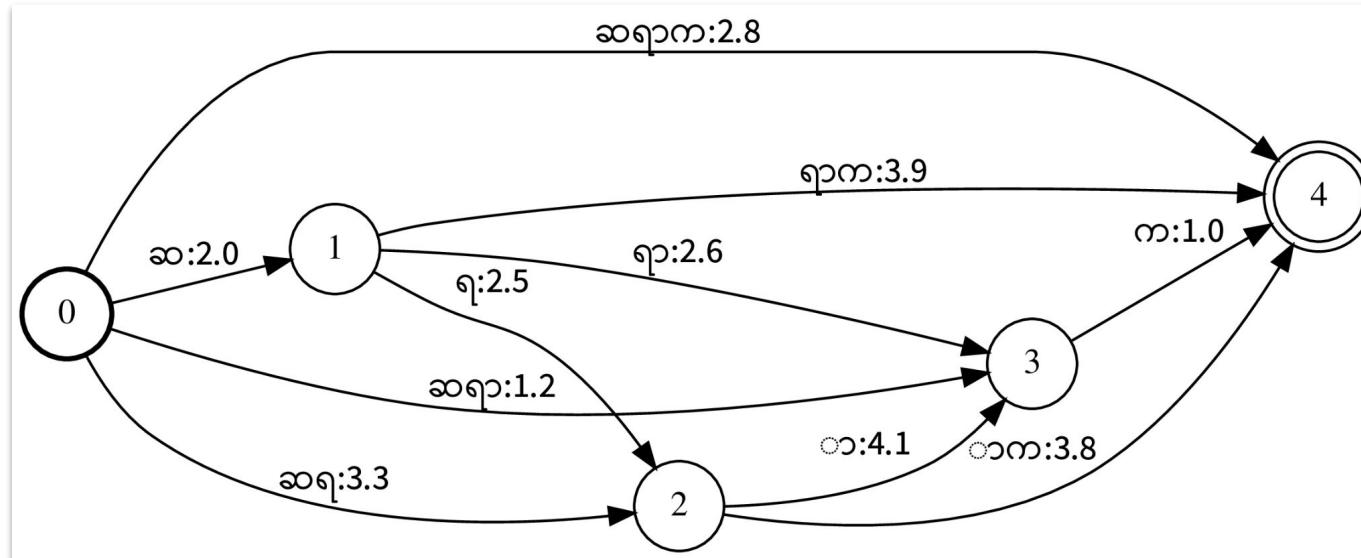
R&D in NLP (sylbreak)

Language	Unsegmented Sentences	Syllable Segmented Sentences
Burmese	ခွင့်လွတ်ပါ။	ခွင့် လွတ် ပါ။
Shan	ပုံ,ချောင်းခလား။	ပုံ, ချောင်း ခလား။
S'gaw Karen	ဝံသံစူးအီ၏.	ဝံ သံ စူး အီ၏.
Pwo Karen	ဖျိုခံယောဆံး.	ဖျို ခံ ယော ဆံး .
Pa'O	ခွင်လွတ်ဟဲ့င်း	ခွင်္ လွတ်္ ဟဲ့င်း
Mon	သူးအခေါင်ညီ	သူး အ ခေါင် ညီ

Fig. Some example of syllable breaking results

R&D in NLP (Word Segmentation, myWord)

myWORD Segmentation Tool

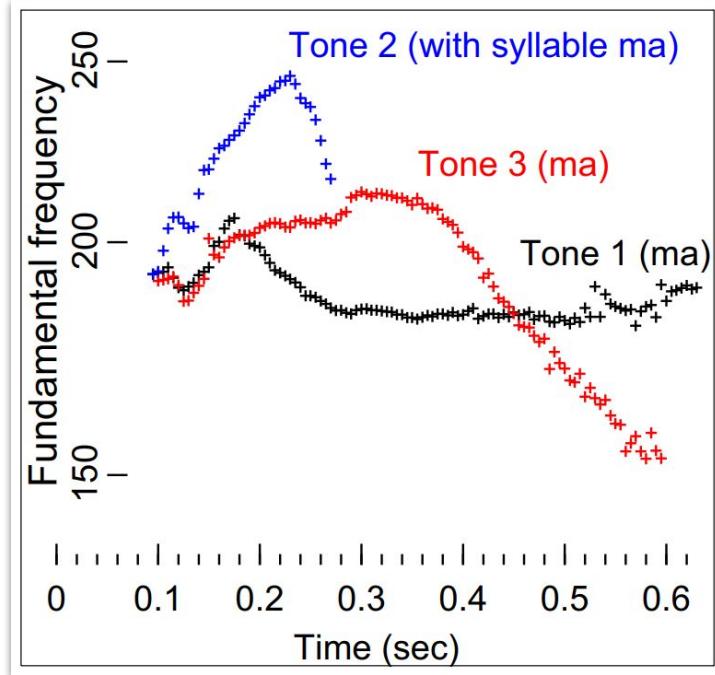


- Unigram
- Bigram
- Viterbi

Fig. Word segmentation as graph for the input sentence "ဆရာတ"

Github link: <https://github.com/ye-kyaw-thu/myWord>

R&D in NLP (Speech corpus building at NICT)



- Myanmar language (Burmese) is a tonal language
- Tone is carried by syllable and is featured by both fundamental frequency and duration of syllable

Fig. An example of three tones of Myanmar syllable Ma (“မ”)

- Paper: https://www.isca-archive.org/interspeech_2015/thu15_interspeech.pdf

R&D in NLP (speech corpus building at NICT)



- Speech corpus building (Read speech), 3.59 hrs of female, 3.35 hrs of male
- Used Marantz recording device
- Recording with maximum decibel (dB) is -12
- Training with HMM based Ossian Speech synthesizing tool

Fig. Recording room environment and the device

R&D in NLP (HMM based TTS, InterSpeech 2015)

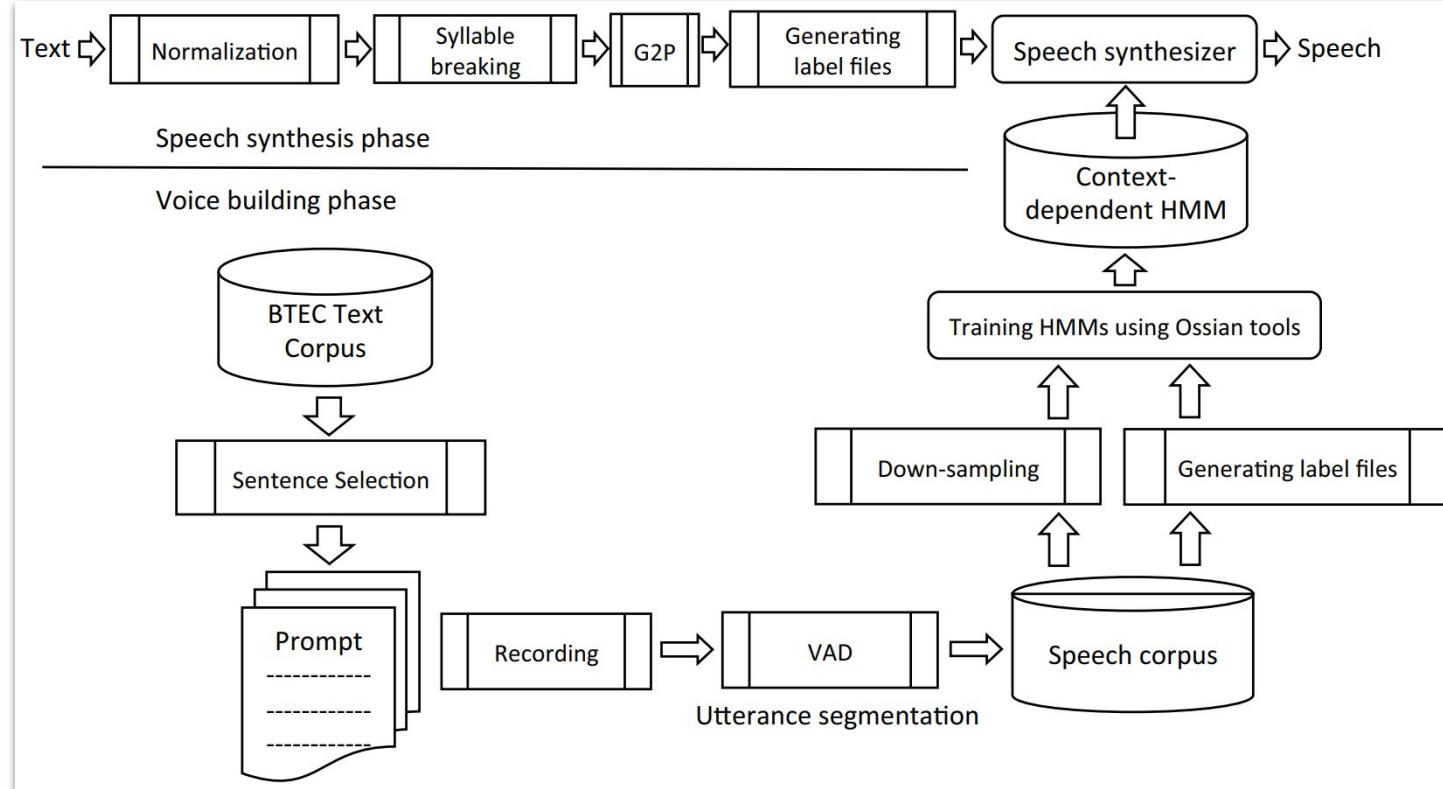


Fig. The 1st Applicable TTS System for Myanmar Language

R&D in NLP (myMediCon at LU Lab.)

LREC-COLING 2024

Turin, Italy, 20-25 May 2024

myMediCon: End-to-End Burmese Automatic Speech Recognition for Medical Conversations

Hay Man Htun^{*}, Ye Kyaw Thu[†], Hutchatai Chanlekha[‡], Kotaro Funakoshi[§], Thepchai Supnithi[†]

^{*}Department of Electrical Engineering, Kasetsart University, 50 Ngamwongwan Road, Lat Yao, Chatuchak, Bangkok 10900, Thailand

[†]Language and Semantic Technology Research Team (LST), Artificial Intelligence Research Group (AINRG),

National Electronics and Computer Technology Center (NECTEC), 112 Phahonyothin Road, Klong Nueng, Klong Luang, Pathumthani 12120, Thailand

[‡]Department of Computer Engineering, Kasetsart University, 50 Ngamwongwan Road, Lat Yao, Chatuchak, Bangkok 10900, Thailand

[§]Department of Information and Communication Technology, Tokyo Institute of Technology, 4259 Nagatsuta-cho, Midori-ku, Yokohama, Kanagawa 226-8503, Japan



東京工業大學
Tokyo Institute of Technology



R&D in NLP (Myanmar sign language corpus building)



Fig. Sign language corpus building, Example annotation with ELAN software

- MSL4Emergency
(Myanmar Sign language Corpus for the Emergency Domain)
- Github link:
<https://github.com/ye-kyaw-thu/MSL4Emergency>
- Shared 558 Myanmar sign language videos

R&D in NLP (Myanmar sign language corpus building)

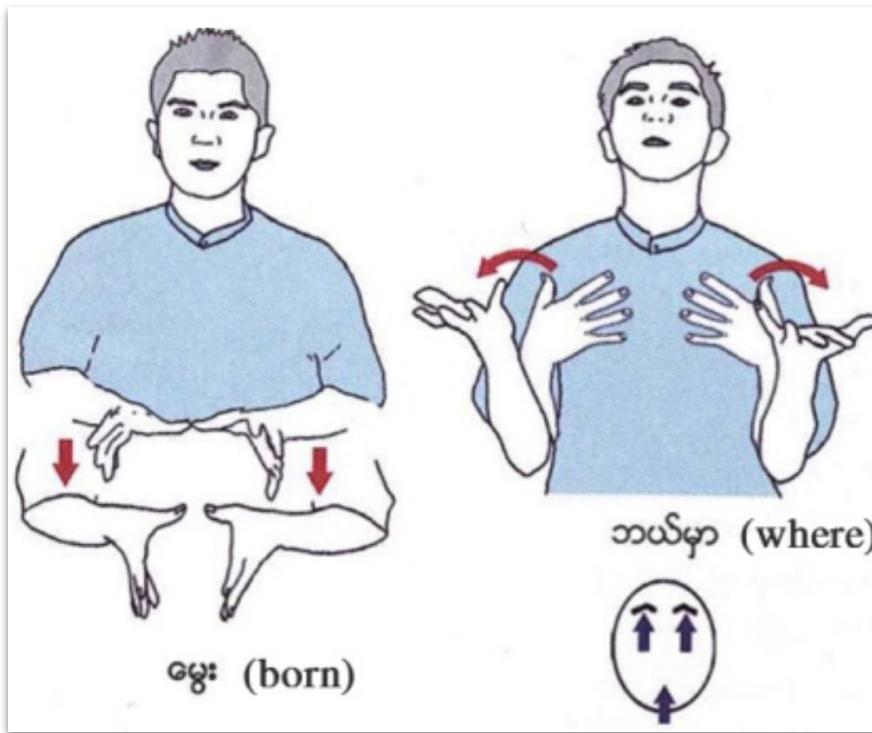


Fig. An example of non manual feature

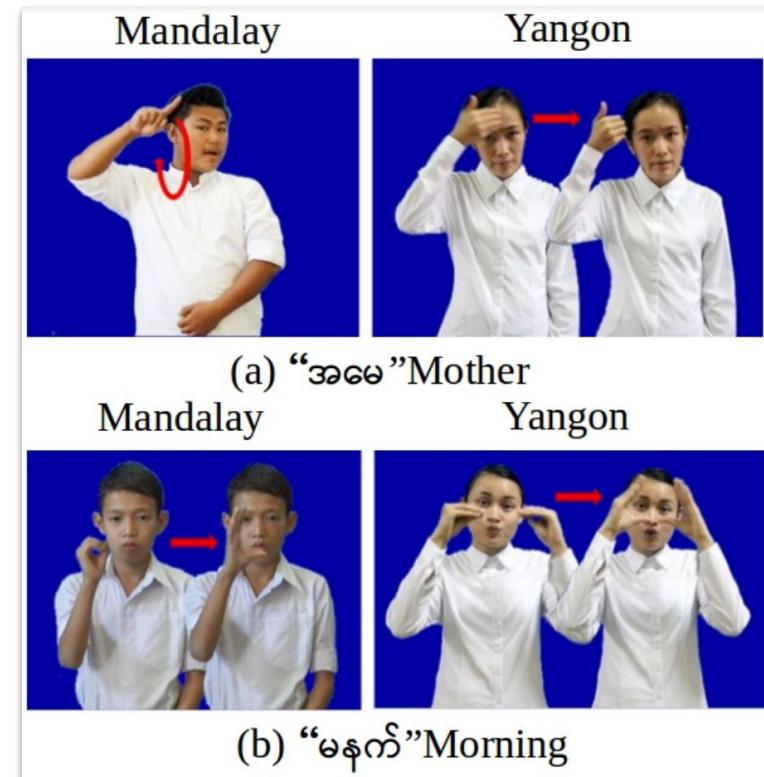


Fig. Sign language dialects

R&D in NLP (Myanmar Fingerspelling)

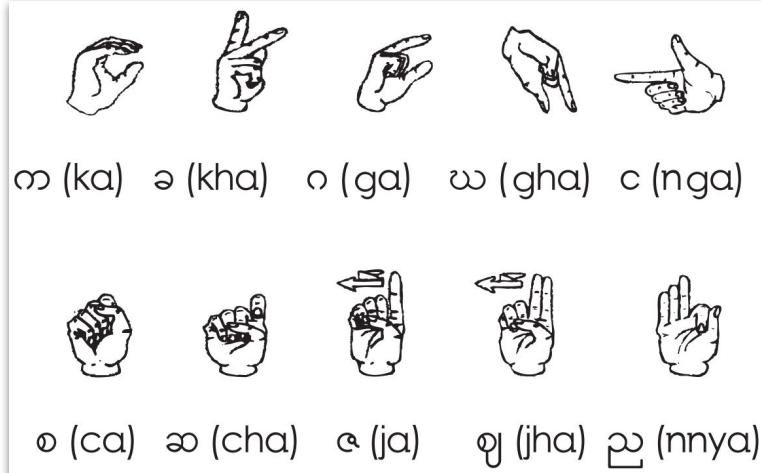


Fig. Ka to Nya of
Myanmar fingerspelling

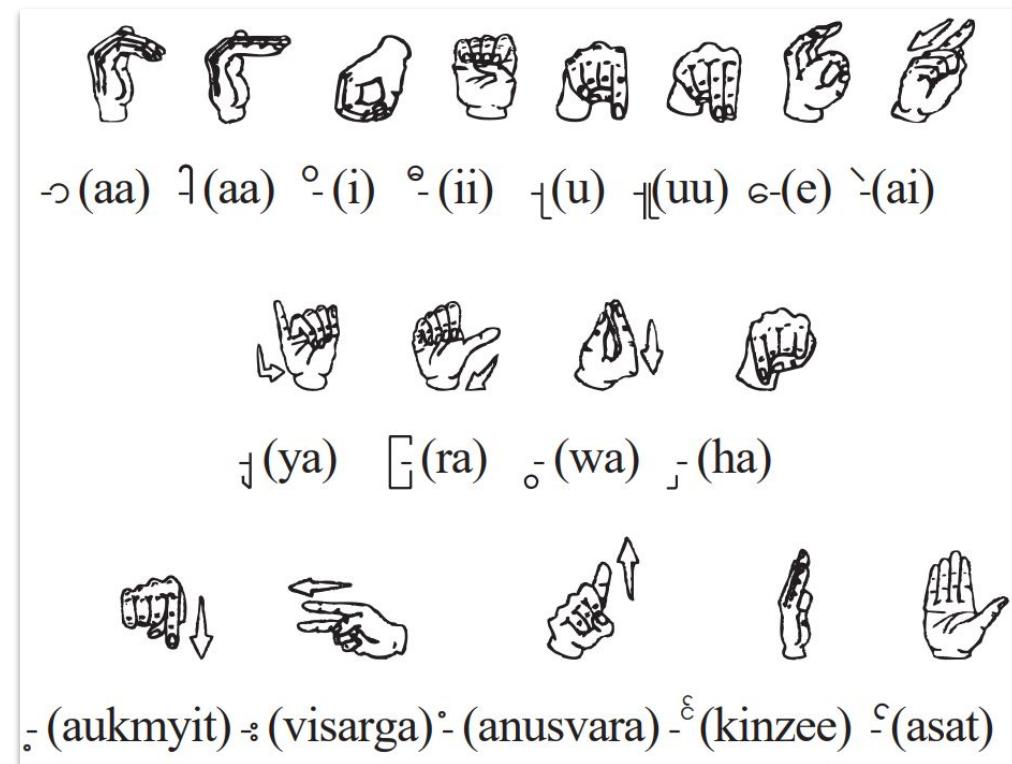
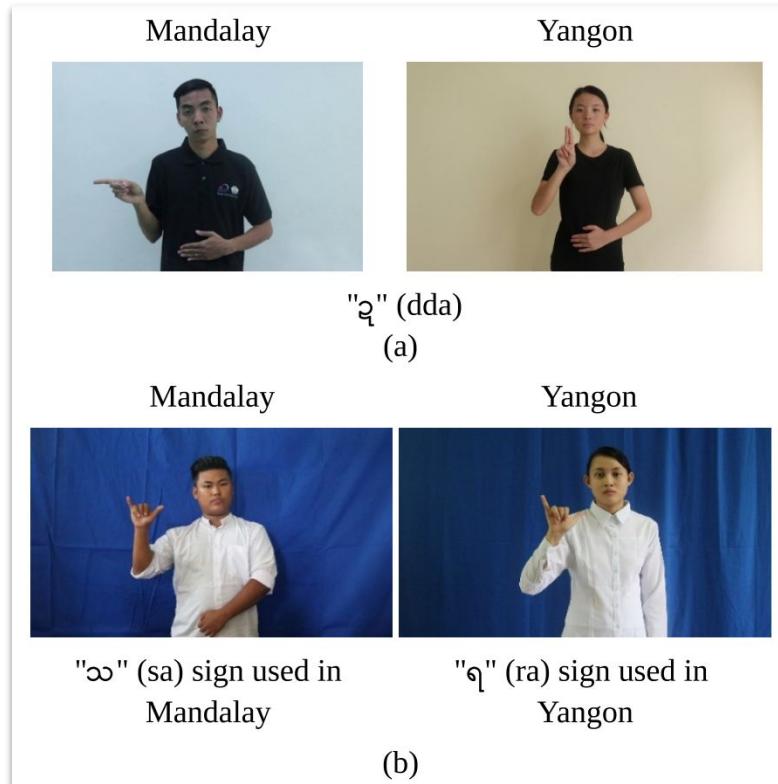


Fig. Vowels, consonant signs, kinzee etc.

R&D in NLP (Myanmar Fingerspelling)



Alphabet Fingerspelling	A (a)	B (b)	C (c)	D (d)	E (e)
BDA					
ASL					

Fig. BDA (British Deaf Association) and ASL (American Sign Language) fingerspelling alphabets “A (a)” to “E (e)”

Fig. Yangon vs Mandalay

R&D in NLP (SignWriting for Myanmar sign language)



Fig. SignWriting inventor
Valerie Sutton

- Sutton SignWriting is 50 years old in 2024
- Internationally-used writing system for writing the movements of all sign languages
- The writing system can be applied to writing any sign language in the world because it writes body movement
- There are 40 to 60 countries writing their sign languages using SignWriting
- Link: <https://www.signwriting.org/>

R&D in NLP (SignWriting for Myanmar sign language)

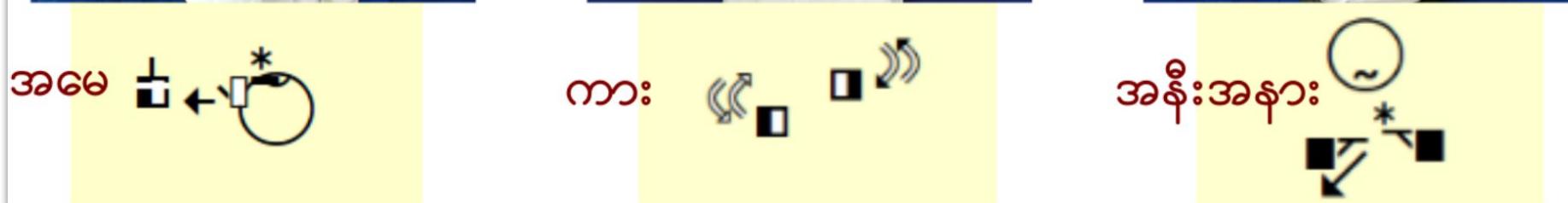


Fig. SignWriting transcription

R&D in NLP (Myanmar Braille corpus building)

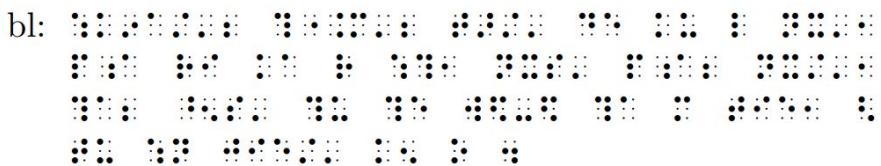
my: ပွင့် ဖူး ကြ ချိန် အေ မှန်။ ဝ သာတ္ထ လေ ချို့ ဖျိန်း။ (“Budding and blooming on time, Blowing the gentle breeze of the rainy season.” in English)

bl:



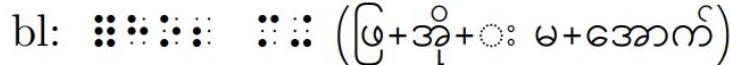
my: ကျောင်း သခံမ်း တွင် ဒု ကူ လ နှင့် ပါ ရို ကာ ရ သေ့ နစ် ပါး နှင့် သား ဖြစ် သူ သူ ဝဏ္ဏ သာ မ တို့ အ တူ နေ ထိုင် ကြ ၏။ (“The two hermits named Dukula and Parika lived with their son, SuwunnaSama in the hermitage.” in English)

bl:



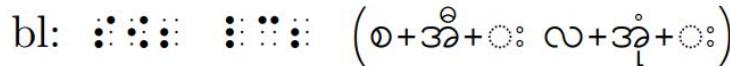
my: ဖြီး မောက် (ဖ+ဗို+ဗို+ဗို+ဗိုး ဝေ+မ+ဝေ+က+ို)

bl:



my: စည်း လုံး (စ+ညု+်+ဗိုး လ+ံုံု+ဗို+ဗိုး)

bl:



my: ရို သေ (ရ+ို့+ဗို ဝေ+သ)

bl:

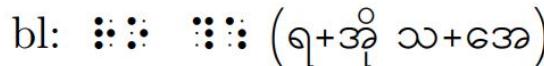


Fig. Mu-Haung Myanmar Braille

Fig. Mu-Thit Myanmar Braille

R&D in NLP (myOCR)

Font Name	Text Image	Font Name	Text Image
Burmese Handwriting Style 04	နည်းပညာကဏ္ဍ	Kamjing	နည်းပညာကဏ္ဍ
Myanmar Ayar3	နည်းပညာကဏ္ဍ	Z01-UMoe	နည်းပညာကဏ္ဍ
Z03-Press	နည်းပညာကဏ္ဍ	Z09-LatYaySat	နည်းပညာကဏ္ဍ
Masterpiece Spring Revolution		Masterpiece Uni Type	နည်းပညာကဏ္ဍ
Myanmar Chatulight	နည်းပညာကဏ္ဍ	Myanmar Phiskel	နည်းပညာကဏ္ဍ
Myanmar Sanpya	နည်းပညာကဏ္ဍ	Myanmar Yin Mar	နည်းပညာကဏ္ဍ
NKSSmart3	နည်းပညာကဏ္ဍ	Pyidaungsu	နည်းပညာကဏ္ဍ

Fig. Synthetic images of နည်းပညာကဏ္ဍ ("technology sector" in English) with different fonts

R&D in NLP (myOCR)

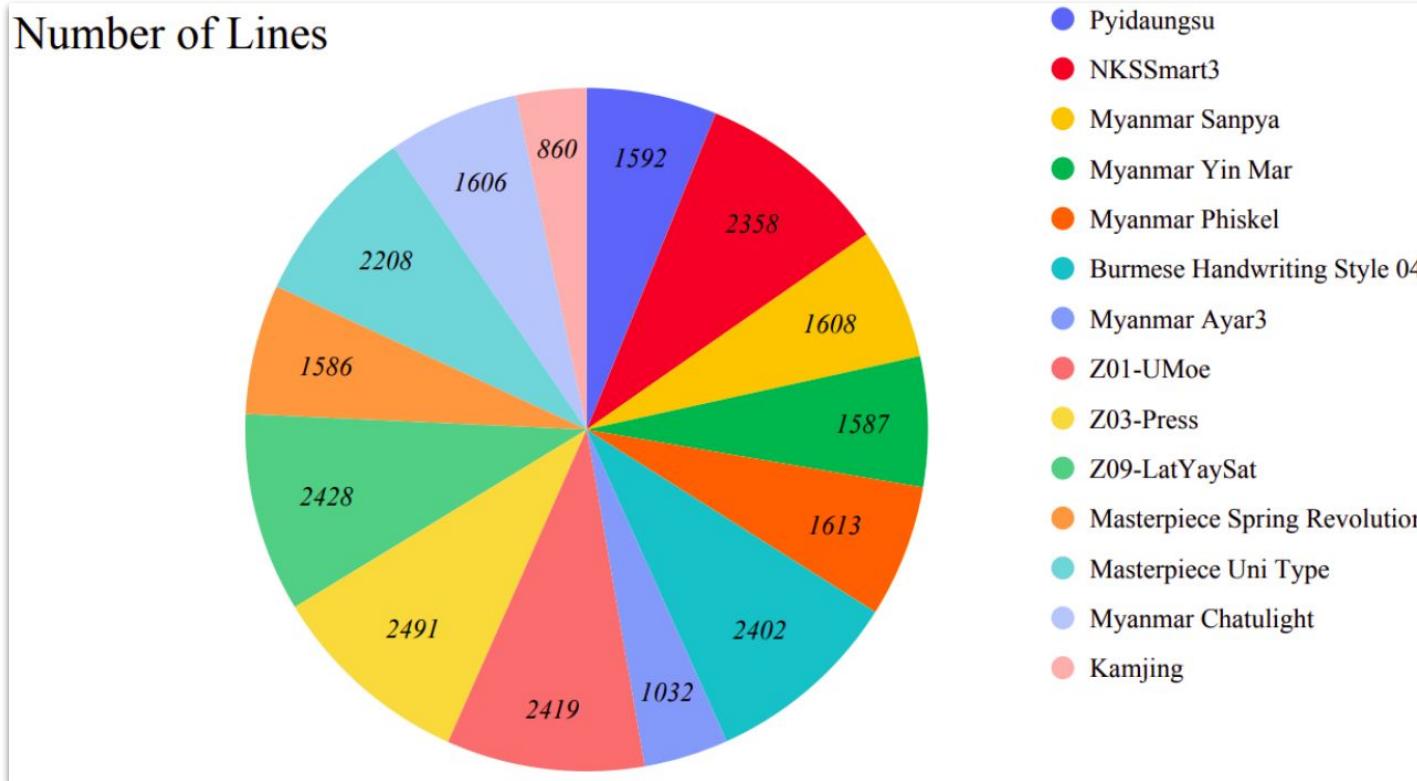


Fig. Number of lines of each font used in the synthetic images (used 14 fonts)

R&D in NLP (myOCR)

	Images	Words	Characters
Train	18,052	30,008	1,884,829
Valid	5,802	9,475	599,906
Test	1,936	15,975	201,764
Total	25,790	55,458	2,686,499

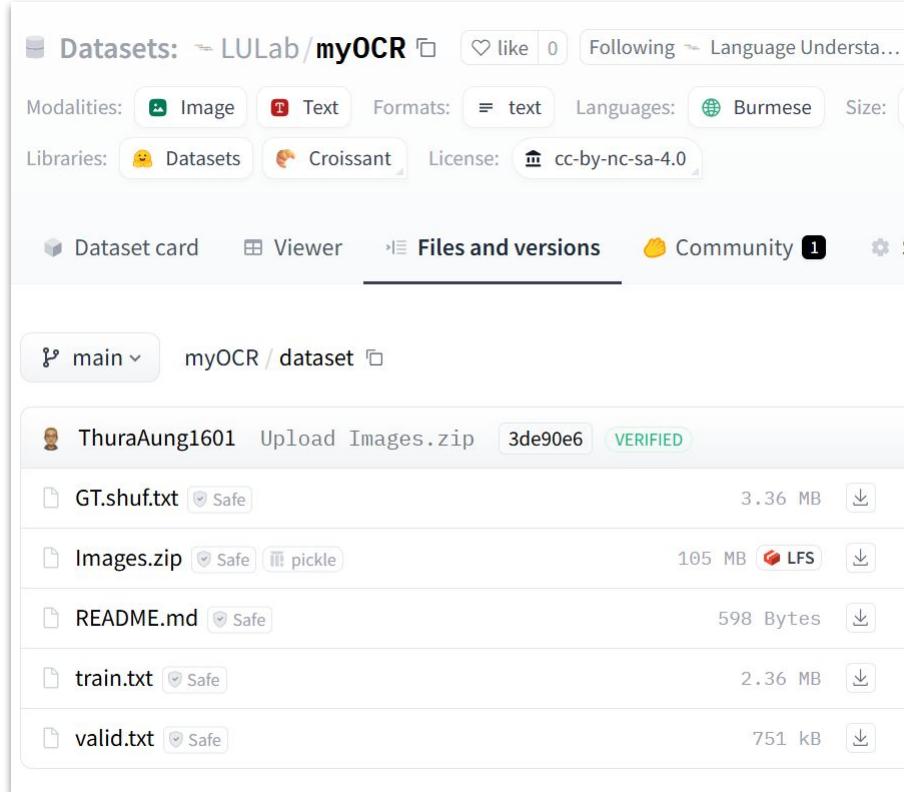
Table. myOCR image dataset partitioning for the experiments

R&D in NLP (myOCR)

	Iteration	3,000			6,000			9,000		
		Hidden State	64	128	256	64	128	256	64	128
None	Base	N/A	N/A	N/A	30.20	23.92	35.47	23.46	18.94	19.15
	+N-gram	N/A	N/A	N/A	14.65	10.03	17.25	9.42	7.16	7.40
	+SymSpell	N/A	N/A	N/A	13.98	9.81	15.95	9.49	7.53	7.65
	+BiLSTM-S2S	N/A	N/A	N/A	15.07	10.83	19.60	10.20	7.54	7.54
	+Transformer-S2S	N/A	N/A	N/A	9.89	6.37	12.95	6.12	3.57	3.24
	+mT5-base	N/A	N/A	N/A	32.19	30.98	33.60	30.74	30.09	30.57
	+mBART-50	N/A	N/A	N/A	11.24	11.01	11.69	10.94	10.57	10.63
+BiLSTM	Base	15.87	17.32	12.30	10.74	11.05	9.69	10.04	10.11	9.18
	+N-gram	5.25	6.43	2.98	2.33	2.58	1.79	1.98	1.99	1.59
	+SymSpell	5.26	6.06	3.17	2.40	2.54	1.83	2.04	1.99	1.53
	+BiLSTM-S2S	7.05	6.56	5.31	4.58	4.27	3.83	4.10	4.05	3.71
	+Transformer-S2S	3.65	3.35	1.69	1.32	1.36	0.83	1.05	1.00	0.66
	+mT5-base	30.41	30.68	29.73	29.56	29.75	29.51	29.69	29.61	29.54
	+mBART-50	10.73	10.87	10.64	10.54	10.61	10.62	10.60	10.61	10.64

Table. Word Error Rate (WER) scores for each model
 (mainly focus on post-OCR error correction)

R&D in NLP (myOCR)



- myOCR paper reading at iSAI-NLP 2024, Pattaya, Thailand
- Hugging Face link:
<https://huggingface.co/datasets/LULab/myOCR>
- GitHub link:
<https://github.com/ye-kyaw-thu/myOCR>
- Work-in-progress

Fig. Image and text dataset information of myOCR version 1.0

Hacking the Lipidipika

- Yaw Atwinwun U Phoe Hlaing (1830-1883)
- Civil servant, best known for his treatise, Rājadhammasaṅgaha (ရာဇဓမ္မသင်ဟ)
- An accomplished writer and wrote a number of important treatises throughout his lifetime, on politics, mathematics and Buddhist philosophy
- He compiled Lipidipika (1869)



Fig. Minister of the interior of Yaw
51

Hacking the Lipidipika

- “ကမ္မာ လုံးသည်” ဟူသော အယူအဆဘက်မှ ရဲ့စွာ ရပ်တည်ခဲ့ (မင်းယုဝေ၊ ပထမ မြန်မာများ)
- ကန္ဒာင်မင်းသားနဲ့ တွဲဖက်ညီ
- ခေတ်သစ်ထူထောင်လို
- မြန်မာနိုင်ငံ ခေတ်မီ တုံးတက်ရေး အစွမ်းကုန်ဆောင်ရွက်ခဲ့



Fig. Group of Myanmar Ambassadors in Paris

Hacking the Lipidipika

လိပိဒီပိကာ ကျမ်းသည် မေ့စ် ဥပဒေကို အခြေခံထားလျက် မြန်မာ ဘာသာဖြင့် ကြေးနှစ်ရိုက်နည်းကို တီထွင် ပြုစုသောကျမ်း ဖြစ်သည်။ ထိုကျမ်းကို မန္တလေးမြို့မြို့ ရုံးစိုက်သော ပြီတိသူ ကိုယ်စားလှယ် မေဂျာ မက္ကာမေဟန်က အက်လိပ်ဘာသာ ပြန်ဆိုသည်။ မြန်မာမင်းပိုင် နေပြည်တော်အတွင်း ကြေးနှစ် သွယ်တန်း ဆက်သွယ်ပြီးနောက် ကြေးနှစ် ဆက်သွယ်ရေး အလုပ်သမားတို့ အကျိုးငှာ ပြုစွင်း ဖြစ်သည် ဆိုသည်။

Fig. About Lipidipika (source: မင်းယုဝေ၊ ပထမ မြန်မာများ)

Hacking the Lipidipika

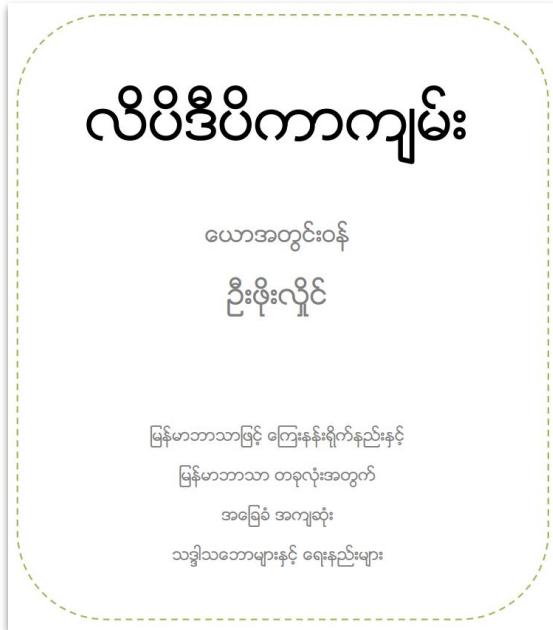


Fig. Book number 2, (Typed by CleanText), 16 pages proposal written by U Phoe Hlaing

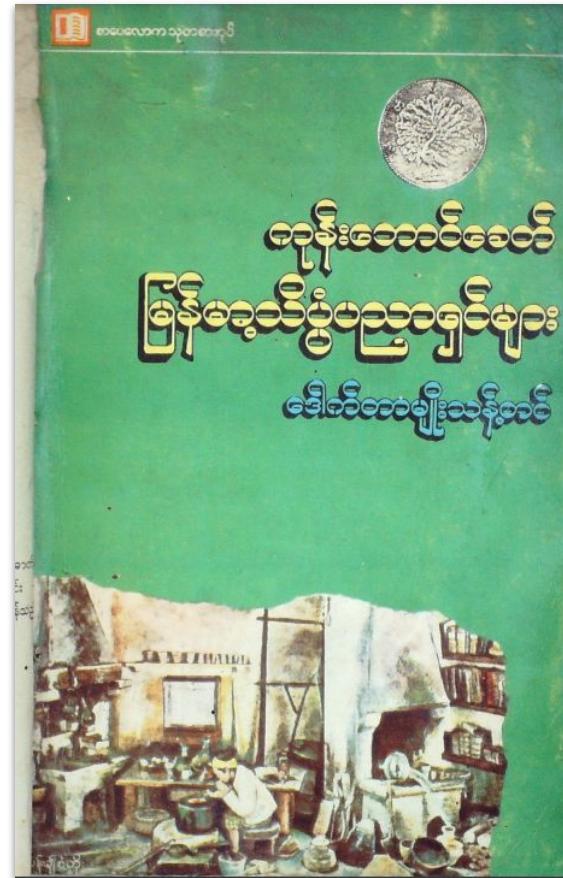


Fig. Book number 3

Hacking the Lipidipika

က	ခ	ဂ	ယ	ဗ
စ	ဆ	ဇ	ဈ	ဉာဏ်
ဋ	၅	၂	၁	၃
၈	၈	၃	၇	၏
၆	၆	၅	၄	၆
၃	၄	၂	၀	၃
၃	၆	၅	၁	၃

Fig. 35 consonants

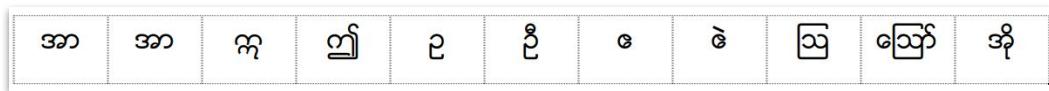


Fig. 11 independent vowels



Fig. 10 dependent vowels

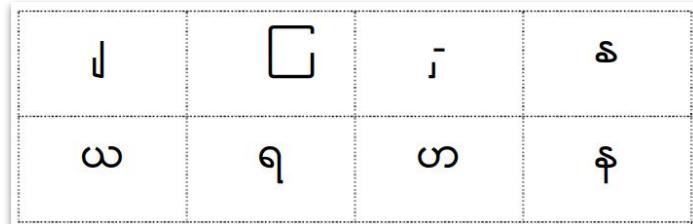


Fig. 4 consonant signs

Hacking the Lipidipika

ကအ ကအာ ကလ ကၢ် ကည ကီ ကဇ ကဲ ကသ ကၢ်သိ ကအိ

ବ	ତ	ର	କ
ଠ	ଥ	ଙ	ଖ
ଘ	ଘ	ସ	ଶ
୦	୨	୮	୭
୩	୪	୯	୬
୫	୬	୮	୫
୮	୭		
ଅଁ	ଅୟଃ	ଅ	

၅၁။ အကျဉ်းရေးရာနိုက် (အ)သရကို ရေးသော်မည်းတွင် အပါအဝင်ထား၍ ရေးသောကြောင့် စာလုံးပုံ အားမရှိ။
အကျဉ်းရေးရာနိုက်ကား သရေးပုံတမျိုး၊ မျိုးရေးပုံ တမျိုးစီသာ ဖြစ်သောကြောင့် မျိုးနောက်
အသံထွက်စေလိုသည့်အတိုင်း သရကို ရေးထားရသည်။

၅၂။ နိဂုဟိတ် ကပ်၍ သည်းခံ အကွရာ ရေးရာနိုက် (သ)အကွရာ၊ (အ)အကွရာ၊ (ည)အကွရာ၊ ဝိသဇ္ဇနီ၊ (ခ)အကွရာ၊ (အ)အကွရာ၊ နိဂုဟိတ် ဤသို့အစဉ်အတိုင်း ရေးရာ၏ ပုံကား-

သအည်းခအံ

ଗ୍ରୀକ୍ ଆଧୁନିକାତ୍ମିକିଣିଙ୍କ ବେଳେ ଯୁଦ୍ଧରେ ପରିବର୍ତ୍ତନ ଆବଶ୍ୟକ ହେଲା ।

Fig. 70 paragraphs, containing rules for encoding of Myanmar text

Hacking the Lipidipika

စအ၊ ကအား၊ ပရှုံ၊ နအနား၊ တအနား၊ လျေပ၊ စ၊ သို့၊ စအ၊ ရေး၊ ကရှုံး၊ အ၊ သံး၊
စအ၊ ရေး၊ တအို့၊ စအ၊ ကအား၊ ပရှုံ၊ နအနား၊ တအနား၊ ရအနာ၊ လအမား၊ တဝအင်၊ တအိုင်၊
စအိုက၊ လျေပ၊ ရအနာ၊ မဟအ၊ စအ၊ ရဝ္မာ၊ အ၊ ရအပ၊ ရအပ၊ လျေပ၊ ဆြုင်၊ ရအနာ၊ ရဟကူး၊
သအညာ၊ မယအား၊ ကအို့၊ ၁၂၃၊ ချေ၊ နဟအစ၊ အ၊ တဝတင်း၊ ပရှုံး၊ ပရေး၊ သွေး၊ လျေပ၊
ဆြုင်၊ ကရအ၊ ရအ၊ မအညာ၊ ခအတာ၊ လဟဇ၊ အ၊ ဆအနာ၊ တဝအင်၊ လအညား၊ မအညာ၊
ရဝ္မာ၊ မအညာ၊ မဟရအ၊ လျေပ၊ ဆြုင်၊ ပရှုံး၊ ပရေး၊ သအညာ၊ မယအား၊ ကအို့၊ မအ၊
ပရအတာ၊ တအင်၊ လဟယြုက၊ လအိုက၊ ရအ၊ မအညာ။

Fig. Example encoding of Lipidipika (source: လိပ်ဒိပ်ကာကျမ်း၊
နီးဘုံးလိုင်)

Hacking the Lipidipika

စ၊ ကား၊ ပြော၊ နှိုး၊ တန်း၊ လုပ်၊ စေ၊ သူ၊ စာ၊ ရေး၊ ကြီး၊ အ၊ သုံး၊ စာ၊ ရေး တို့၊ စ၊ ကား၊ ပြော၊ နှိုး၊ တန်း၊ ရန်၊ လမ်း၊ တွင်၊ တိုင်၊ စိုက်၊ လုပ်၊ ရန်၊ မှ၊ စ၊ ၍၊ အ၊ ရပ်၊ ရပ်၊ လုပ်၊ ဆောင်၊ ရန်၊ ၍၊ သည်၊ များ၊ ကို၊ ၁၂၃၁၊ ခု၊ နှစ်၊ အ၊ တွင်း၊ ပြီး၊ ပြော၊ အောင်၊ လုပ်၊ ဆောင်၊ ကြာ၊ ရ၊ မည်၊ ခတ်၊ လျှော့၊ အ၊ ဆန်၊ တွင်၊ လည်း၊ မည်၊ ရွှေ့၊ မည်၊ မျှ၊ လုပ်၊ ဆောင်၊ ပြီး၊ ပြော၊ သည်၊ များ၊ ကို၊ မ၊ ပြတ်၊ တင်၊ လျှောက်၊ လိုက်၊ ရ၊ မည်။

Fig. Original Myanmar text (source: လိပ်ဒီပိကာကျမ်း၊ ၉းဘုံးလိုင်)

Hacking the Lipidipika

၆၇။ မီးဖြူ။ မီးဝါ အစရှိသော မီးမျိုးစုတို့ဖြင့်ရင်း၊ အလံပျိုးစုတို့ဖြင့်ရင်း စကားပြောရန် ဤကတ်အိုး နှစ်ဦးတို့ဖြင့် စကားပြောဆို နိုင်သကဲ့သို့ အမှတ်အသားထား၍ ပြောဆို နိုင်ကုန်၏။

၆၈။ စကားပြောရန် ကတ်အိုး လုပ်ဆောင်ခြင်းသည်လည်း အမျိုးမျိုး ရှိကုန်၏။ အခါး။ ကတ်အိုး၏ တီးခတ်သော အချက်ဖြင့် စာလုံးကို မှတ်ကုန်၏။ အခါး။ကတ်အိုးစာလုံး ဟူ၍ အမှတ်သညာ ပေးသော အရေးအဆွဲးတို့ဖြင့် လုပ်ဆောင်ကုန်၏။ အခါး။ကတ်အိုး၏ သံလိုက်ကို အိမ်မြောင်ကဲ့သို့ထား၍ လုပ်ဆောင်ကုန်၏။

၆၉။ ဤသို့ အစရှိသည်ဖြင့် အထူးထူး အထွေထွေ လုပ်ဆောင်ကြေားသော်လည်း ယခုအခါ နိုင်ငံတော်တွင် အီးအစ ပင့်မ ဖြစ်သေ့့ကြောင့် သိသာ ထင်ရှားလွယ်အောင် ကတ်အိုး မျက်နှာပြင်၏ စာလုံးပုံထား၍ အိမ်မြောင်နှင့် တည့်စေသော ကတ်အိုးဖြင့် လုပ်ဆောင်စေတော်မူ၏။

Fig. Paragraph 67 to 69, explanation for practical implementation

Hacking the Lipidipika

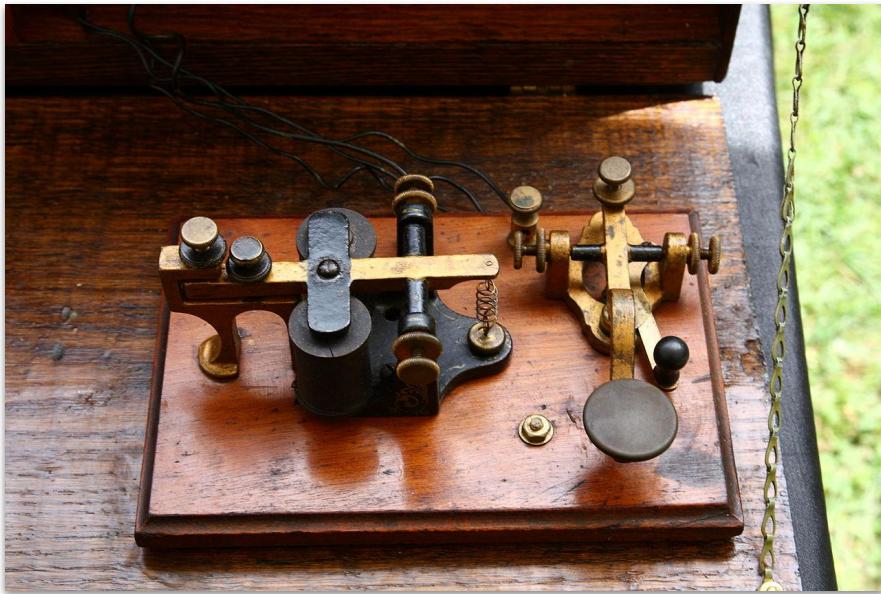


Fig. Morse key and sounder (source: Wikipedia)

International Morse Code

1. The length of a dot is one unit.
2. A dash is three units.
3. The space between parts of the same letter is one unit.
4. The space between letters is three units.
5. The space between words is seven units.

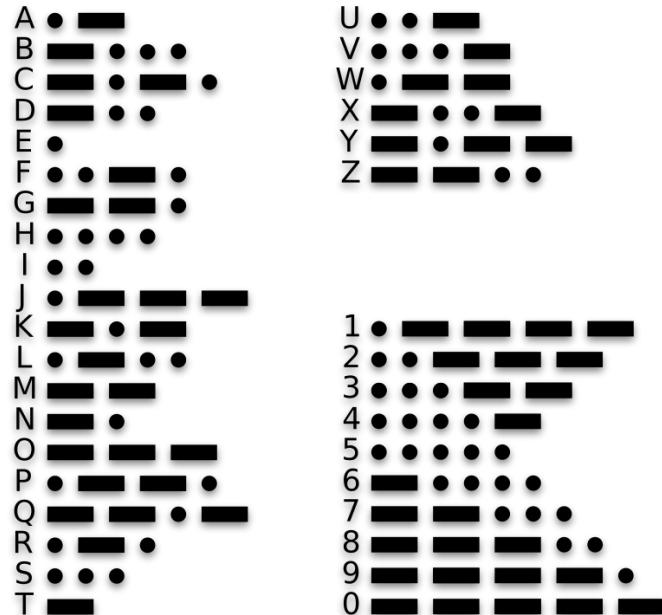


Fig. International Morse code (source: Wikipedia)

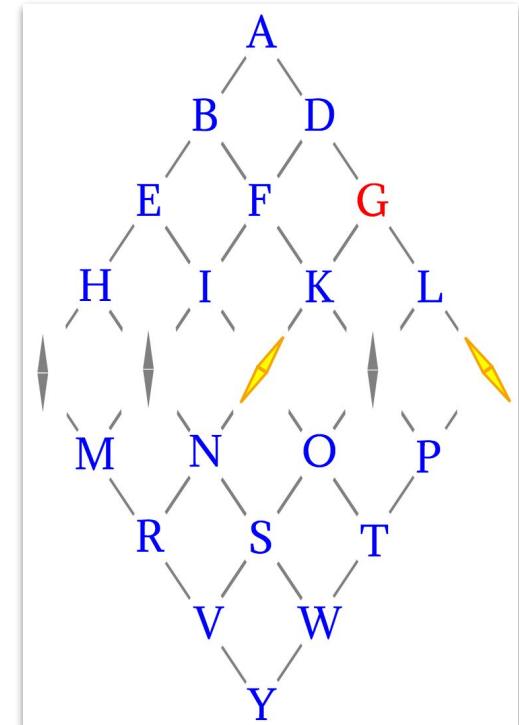
Hacking the Lipidipika



Fig. Hughes telegraph, an early (1855) teleprinter



Fig. Five needle Cooke and Wheatstone telegraph
(source: PowerHouse Collection)



Hacking the Lipidipika

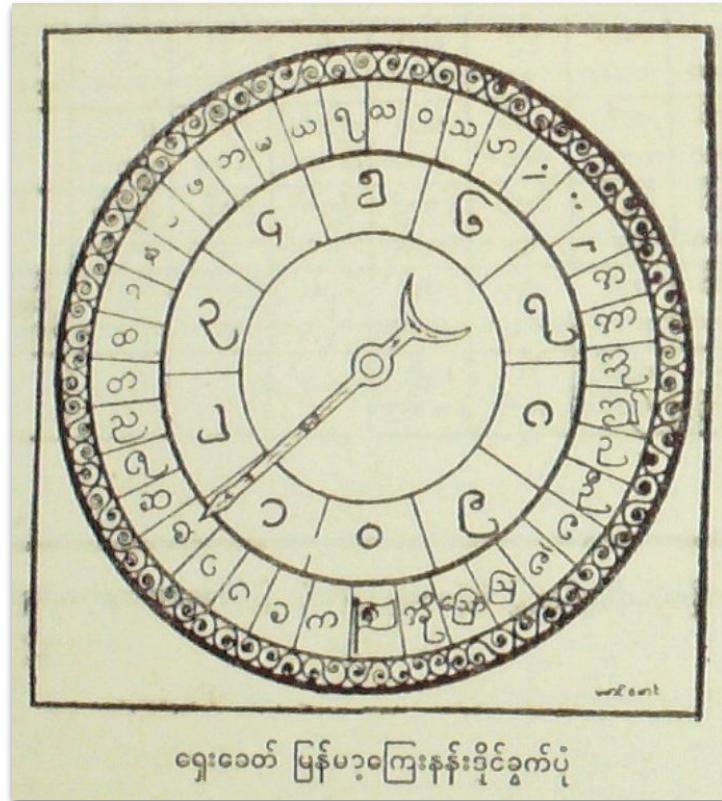


Fig. Hints from “ကုန်းဘောင်ခေတ် မြန်မာသိပ္ပံပညာရှင်များ”

Hacking the Lipidipika

ထိစဉ်ကြေးနှစ်းဖြင့် သတင်းပေးပို့ရာတွင် နည်းစနစ် သုံးမျိုး ရှိနေ၏။ ပထမနည်းမှာ ဓာတ်အိုးကိုတီးခတ်၍ အသံဖြင့်အချက်ပြု ဆက် သွယ်သော ‘မော့စ်’၏ နည်းဖြစ်၏။ ဒုတိယနည်းမှာ၊ တဖက်မှ ပေးပို့ သော စကားကို လက်ခံသည့်စက်တွင်ပါသော စွဲ။ ပေါ်၍ အစက်များ ရေးထိုးစေသော ‘စတိမ်းဟီး’၏ နည်းဖြစ်သည်။ တတ္ထာ နည်းမှာ ဓာတ်အိုးမျက်နှာပြင် ခိုင်ခွဲက်၍ စာလုံးများရေးသားထားပြီး စာလုံးကို အိမ်မြောင်လက်တံဖြင့် တည့်၍ ညွှန်းသော ‘ကုတ်နှင်းဝိုင်တုံး’ တို့၏ နည်းဖြစ်၏။။

Fig. Hints from “ကုန်းဘောင်ခေတ် မြန်မာသိပ္ပံပညာရှင်များ”

Hacking the Lipidipika

ဤနည်းသုံးမျိုးအနက် တတိယနည်းဖြစ်သော ကွဲတန်ငိုစတုံးနည်းသည် သိသာထင်ရှား၍ လူတိုင်းအလွယ်တကူ သုံးနိုင်သောနည်းဖြစ်သဖြင့် မြန်မာတို့ လက်ခံကျင့်သုံးရန် မင်းတန်းမင်းက ဆုံးပြတ်သည်။ ဤနည်းကို ရွှေးစုံ ကြီးမားသော အေက်အခဲတရပ်နှင့် ရင်ဆိုင်ခဲ့ရ၏။ မြန်မာတို့ စာရေးရာတွင် သရ ၁၁ လုံး၊ ပျေည်း ၃၅ လုံး ပေါင်း၄၆ လုံးသော မူလ အကွဲရာကို အခြေခံသည်အပြင် ရေးနည်းကို ကျဉ်းစေရန် အသုံးပြုသည့် ပျေည်းနှင့်ကပ်သော သရ ၁၁ လုံး၊ ပုံစံပြောင်းသည့် ပျေည်း ၅ လုံး စသည်ဖြင့် စုစုပေါင်း ၆၂ လုံး ရှိရာ စာလုံးတိုင်းကို ၆၇၈ကို ရေးရှု အိမ်ပြောင်လက်တံ့ဖြင့် တည်ညွှန်းမည်ဆိုပါက လုန်စာ ရှုပ်ထွေးဖွယ်ရာ ရှုမည်ဖြစ်၏။ ဤအေက်အခဲအား မြန်မာပညာရှိ ယော အတွင်းဝန် ဦးဘုံးလှုံးက လိပ်ဒိုကာ ခေါ် မြန်မာ ကြေးနှုန်းစာ ရေးနည်းကို တိထွင်ပေးခြင်းဖြင့် ဖြေရှင်းနိုင်ခဲ့ပေသည်။ ၅၂



Fig. Hints from “ကုန်းဘောင်ခေတ်
မြန်မာသီပိုပညာရှင်များ”

Fig. ABC machine

Hacking the Lipidipika

```
60 def process_text(line, break_pattern):
61     """Processes the text line by line for syllable breaking and
62     conversion."""
63     # Step 1: Substitute ဗ with စ and ဃ with ဃ
64     line = line.replace("ဗ", "စ").replace("ဃ", "ဃ").replace("ၢ", "ၣ")
65
66     # Step 2: Delete ။, ၌, and other symbols
67     line = re.sub(r"[\u2019!-/:@[-`{~-}]", "", line)
68
69     # Step 3: Apply syllable segmentation
70     syllables = break_syllables(line, break_pattern).split(" ")
71
72     # Step 4: Substitute Myanmar consonants
73     consonant_substitutions = {
74         "ာ": "ာ", "ု": "ု", "ိ": "ိ", "ီ": "ီ", "ုာ": "ာ", "ုု": "ု", "ို": "ိ", "ီု": "ီ",
75         "ေ": "ေ", "ဲ": "ဲ", "ဳ": "ဳ", "ော": "ာ", "ေု": "ု", "ေိ": "ိ", "ေီ": "ီ", "ောာ": "ာ", "ောု": "ု", "ောိ": "ိ", "ောီ": "ီ", "ော့": "့", "ေား": "း", "ော္": "္", "ော်": "်", "ောျ": "ျ", "ောွ": "ွ", "ောှ": "ှ", "ောြ": "ြ", "ောွှ": "ွှ", "ောြှ": "ြှ", "ောြွ": "ြွ", "ောြွှ": "ြွှ"},
```

Fig. Coding of Lipidipika encoding (used Python programming language)

Hacking the Lipidipika

The screenshot shows a web-based application for encoding Burmese text into a telegraph code. The interface is split into two main sections: 'Input' on the left and 'Output' on the right.

Input: This section contains a text area with the following Burmese text:
စကားပြောန်းတန်းလုပ်စေသူတရေးကြီးအသုံးတရေးတို့စကားပြောန်းတန်းရန်
လမ်းတွင်တိုင်စိက်လုပ်ရန်မှစ၍အရပ်ရပ်လုပ်ဆောင်ရန်ရှိသည်များကိုဘုရားချုံစွဲ
အတွင်းပြီးပြောအင်လုပ်ဆောင်ကြရမည်ခတ်လေ့အဆန်တွင်လည်း
မည်ရွှေမည်မှုလုပ်ဆောင်ပြီးပြောသည်များကိုမပြတ်တင်လျှောက်လိုက်ရမည်။

Below the text area is a 'Choose File' button with the placeholder 'No file chosen'. At the bottom are two blue buttons: 'Encode' and 'Clean'.

Output: This section displays the encoded output text in a large text area:
စ ကအား ပရအဗြာ န်း တန်း လောပ် စ။ သို့ စအာ ရေး ကရှုံး အ သ
ုံး စအာ ရေး တအို့ စ ကအား ပရအဗြာ န်း တန်း ရန် လမ်း တွင် တက္ကာ
င် စက္ကာောက် လောပ် ရန် မှ စ ရွှေး အ ရပ် ရပ် လောပ် ဆုံးရှုံး ရန် ရှုံး
သည် မျှအား ကက္ကာော ၁ ၂ ၁ ခု နှစ် အ တွင်း ပရှုံး ပရအေး အယ်း
လောပ် ဆုံးရှုံး ကရအ ရ မည် ခတ် လှေ အ ဆန် တွင် လည်း မည် ရွှေး
မည် မျှ လောပ် ဆုံးရှုံး ပရှုံး ပရအေး သည် မျှအား ကက္ကာော မ ပရအတ်
တင် လွှုံးက် လက္ကာောက် ရ မည်

At the bottom of the output area is a blue 'Download Output' button.

Fig. Lipidipika encoding prototype

Key Takeaways

- I hope you learned something from the R&D work I introduced
- In short, corpus building and NLP R&D are crucial for our languages and our country
- Yes, we must follow current R&D trends, but it is equally important to establish your own identity
- Study previous work thoroughly and conduct literature reviews before starting your own research
- When you begin something, dive deep into it

Thank you!

Q&A

Ye Kyaw Thu
LU Lab., Myanmar
NECTEC, Thailand

Email: ykt.nlp.ai@gmail.com

FYI

- Plan to share Lipidipika code when I finished some more works