

Innovation in Three Months (Potential R&D Topics for Interns)

Meeting #1

Ye Kyaw Thu
Lab Leader, LU Lab., Myanmar

Overview

1. Designing an Educational Programming Language
2. Myanmar Text Readability Scoring: Formulas & Evaluation
3. Creating a Myanmar SQuAD Dataset for Multilingual Question Answering
4. Humor Detection in Myanmar Text: Binary & Multiclass Classification
5. Optimizing Myanmar Keyboard Layouts via Character Frequency Analysis
6. Automatic Speech Recognition (ASR) for Myanmar: Daily-Use Phrases
7. Text-to-Speech (TTS) Synthesis for Myanmar Language

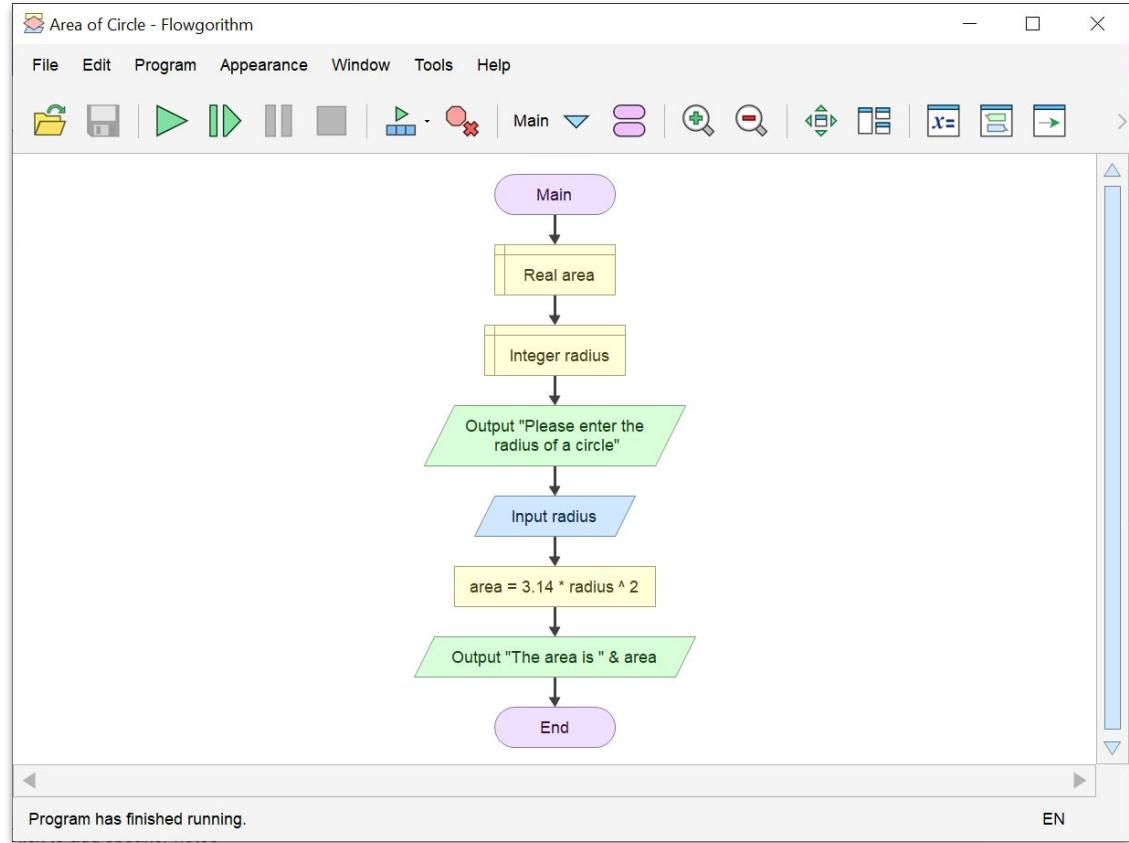
1. Designing an Educational Programming Language

Programming languages designed for **educational purposes** play a critical role in introducing learners to **computational thinking**, **problem-solving**, and **software development**. The landscape includes both visual and textual languages, each with unique strengths for different age groups and learning objectives.

- Visual Languages: Flowgorithm, Scratch, Alice
- Textual Languages: Pascal, Python, Ruby, Lua
- Specialized Edu Languages: Logo, Mama, RoboMind, CircuitPython

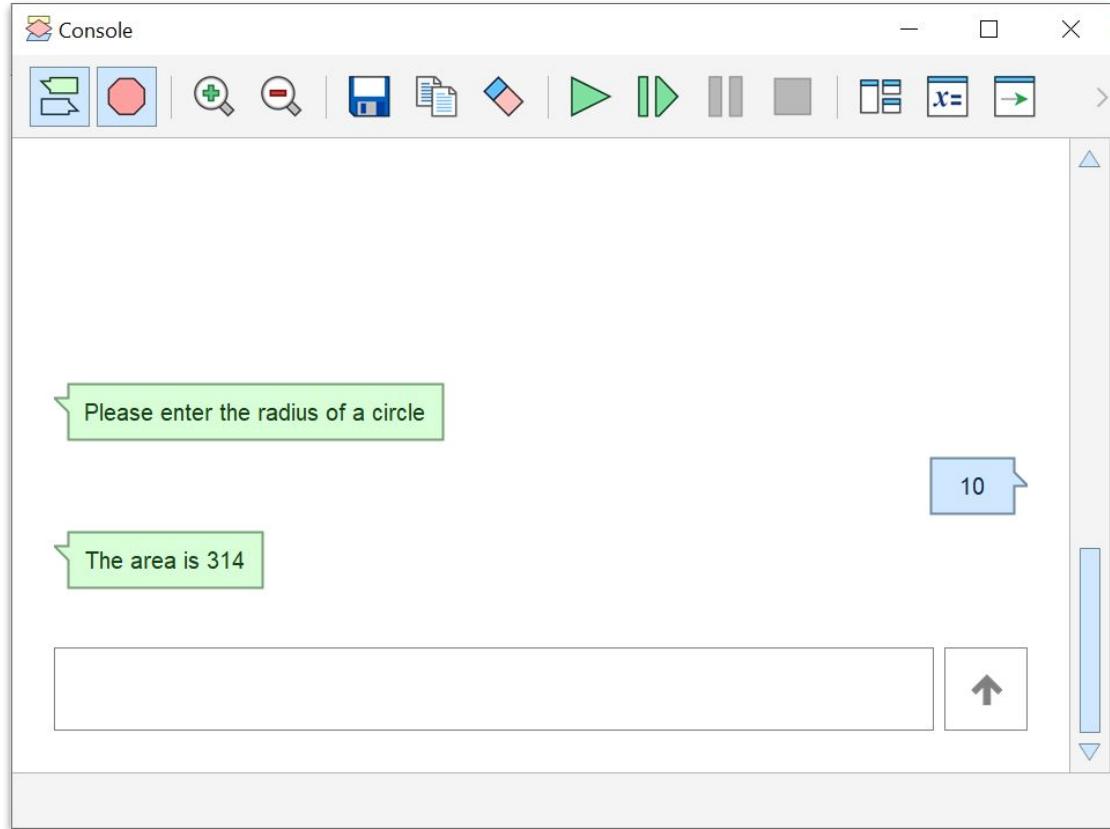
1. Educational Programming Language (Flowgorithm)

- အလယ်တန်း၊
အထက်တန်း
ကျောင်းသားတွေအ
တွက်
သင့်တော်တယ်။
Algorithm စဉ်းစားပြီး
Flowchart ဆပြီး Run
ခိုင်းတဲ့ ပုံစံပါ။
- ကလေးတွေအတွက်
က ခက်တယ်။



1. Educational Programming Language (Flowgorithm)

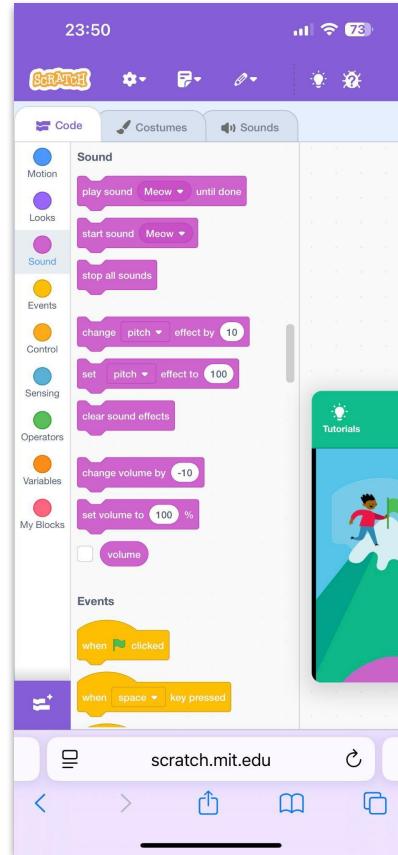
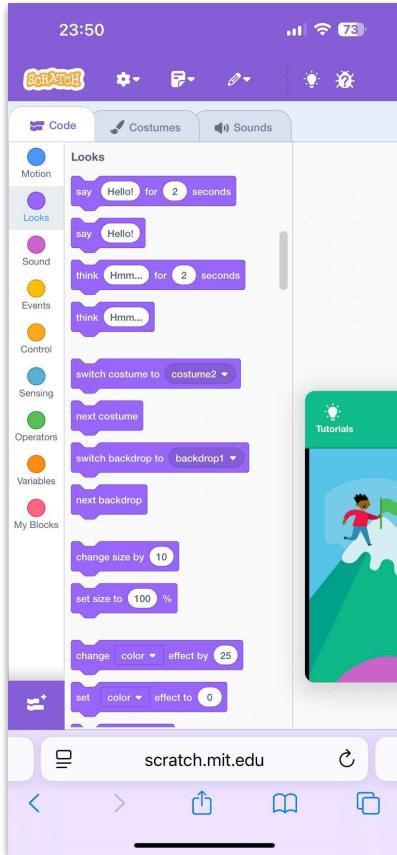
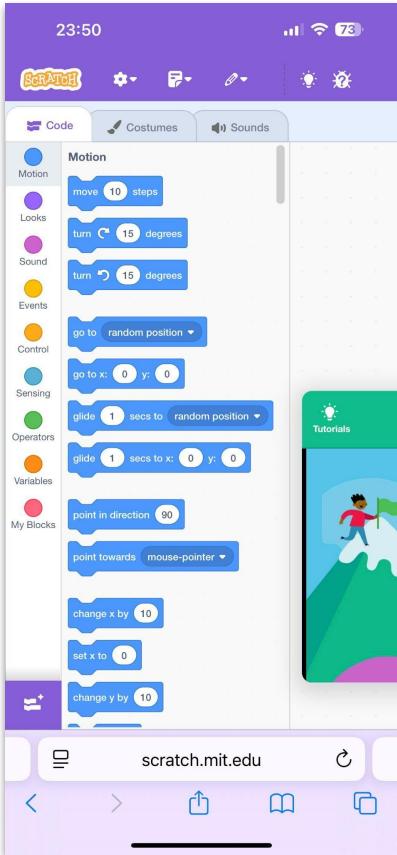
- ရွှေ့က flowchart ကို
Run လိုက်ရင်
မြင်ရမယ့် Screen
- စက်ဝိုင်းတစ်ခုပဲ
ရေးယာကို radius
ရုံကထည့်ပြုး
တွက်ခွင်းတဲ့
flowchart ပါ။



1. Educational Programming Language (Flowgorithm)

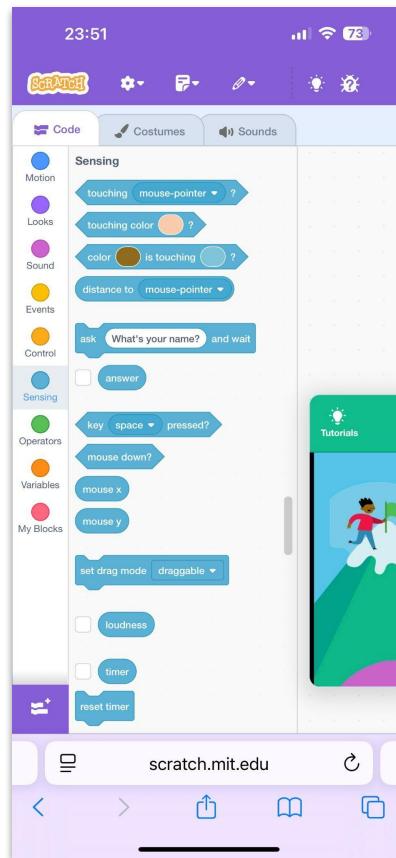
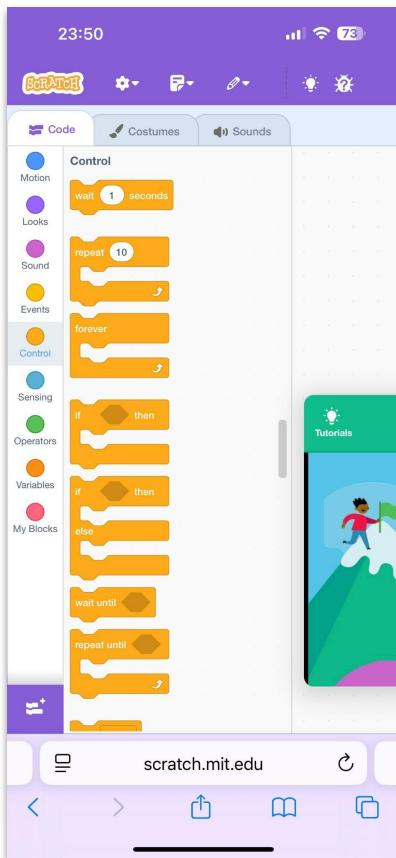
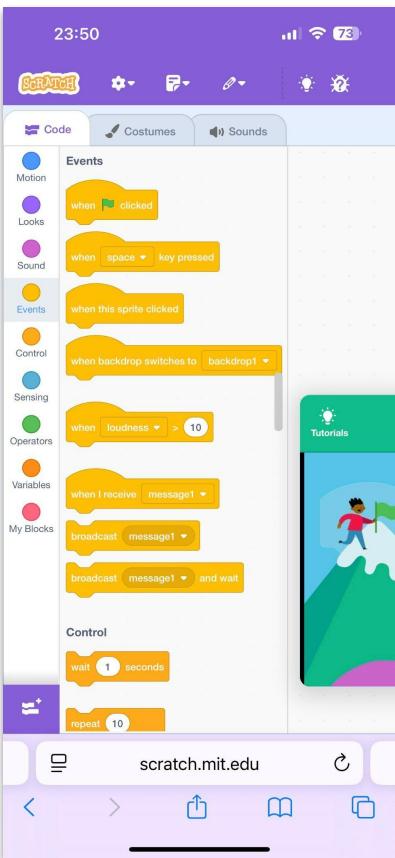
```
1  <?xml version="1.0"?>
2  <flowgorithm fileversion="1.0">
3    <attributes>
4      <attribute name="name" value="Area of a Circle"/>
5      <attribute name="authors" value="Devin Cook"/>
6      <attribute name="about" value="This example inputs the radius of a circle and
7        outputs the area."/>
8      <attribute name="saved" value="8/1/2014 11:17:52 PM"/>
9    </attributes>
10   <function name="Main" type="None" variable="">
11     <parameters/>
12     <body>
13       <declare variables="area" type="Real"/>
14       <declare variables="radius" type="Integer"/>
15       <output expression="Please enter the radius of a circle"/>
16       <input variable="radius"/>
17       <assign variable="area" expression="3.14 * radius ^ 2"/>
18       <output expression="The area is &amp; #39; &amp; #39; &amp; #39; area"/>
19     </body>
20   </function>
21 </flowgorithm>
```

1. Educational Programming Language (Scratch)



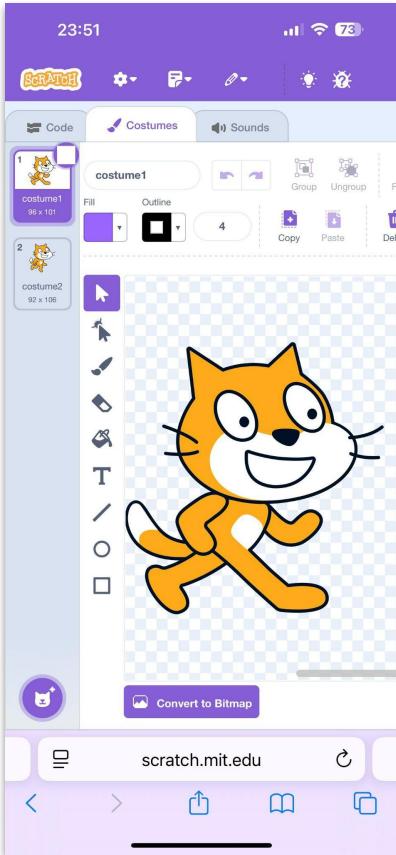
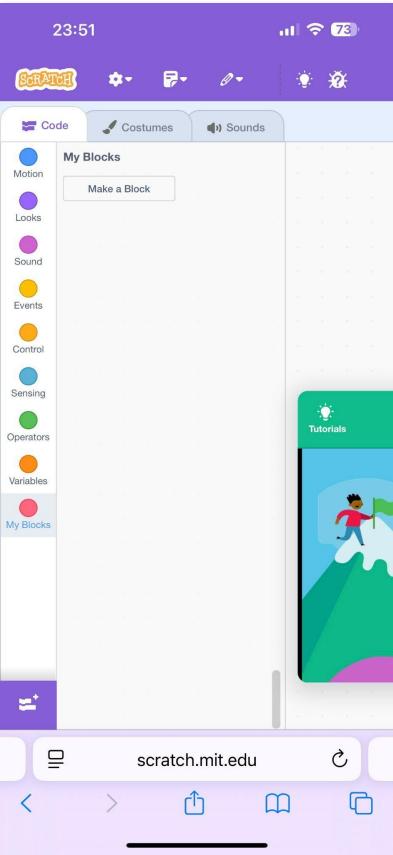
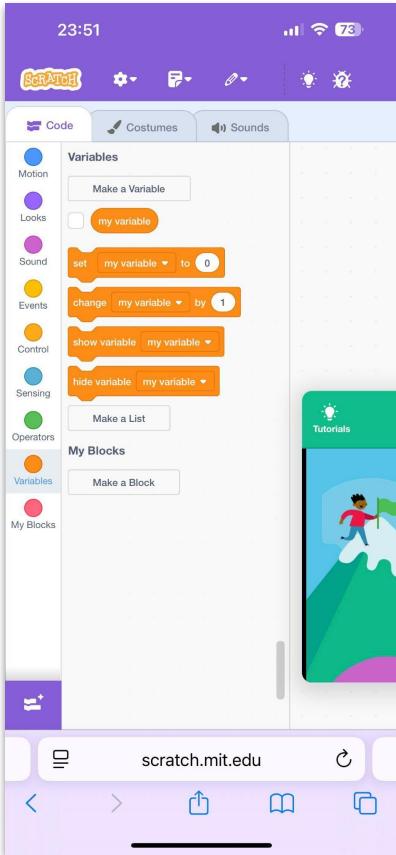
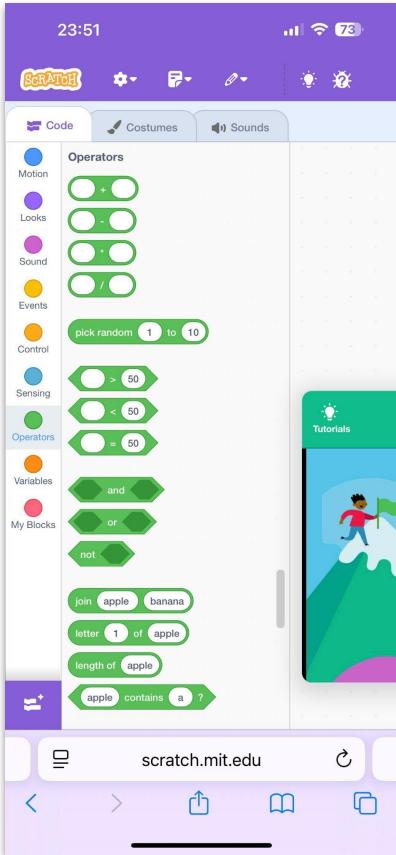
- ဘလောက်တံ့ဌေးတောက်ကွန်ပြု၍ တောက်စန်း
drag and drop လုပ်ပြီး code လုပ်သွားတဲ့ ပုံစံပါ။
- ခုမြင်နေရတာက motion, looks နဲ့ sound နဲ့ဆိုင်တဲ့ ဘလောက်တံ့ဌေးတွေ

1. Educational Programming Language (Scratch)

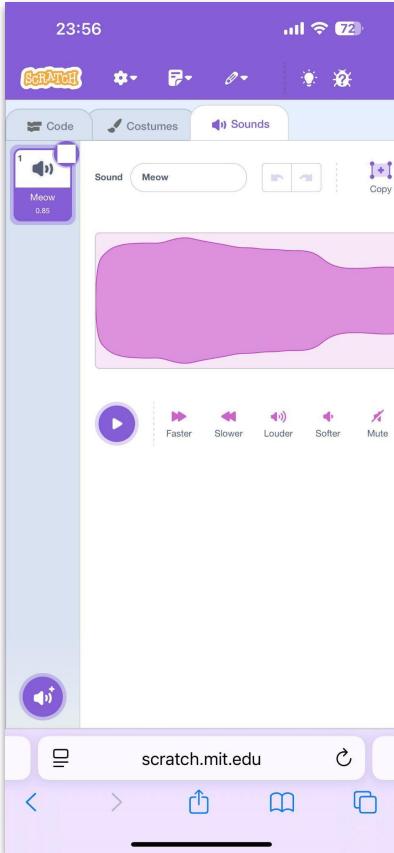


- Events, control ቁር
sensing ቁር የሚጠናውን
ቈለያကတဲ፡ጭዎን
- Event እስተካገኘ
ዋයን አጋጣጥ ይሞላ
እስተካገኘ
ቈተም ማተኞች በ
ጠልቅ ቁጥር ተፈጸማል
space bar ቁጥር ተፈጸማል
በሆነ ማረጋገጫ
- Control ጉዳዚ ይመለከት
if ተስተካክል
wait ተስተካክል
የuser የሚቀርቡትን
ልቦታ የሚመለከት

1. Educational Programming Language (Scratch)



1. Educational Programming Language (Scratch)

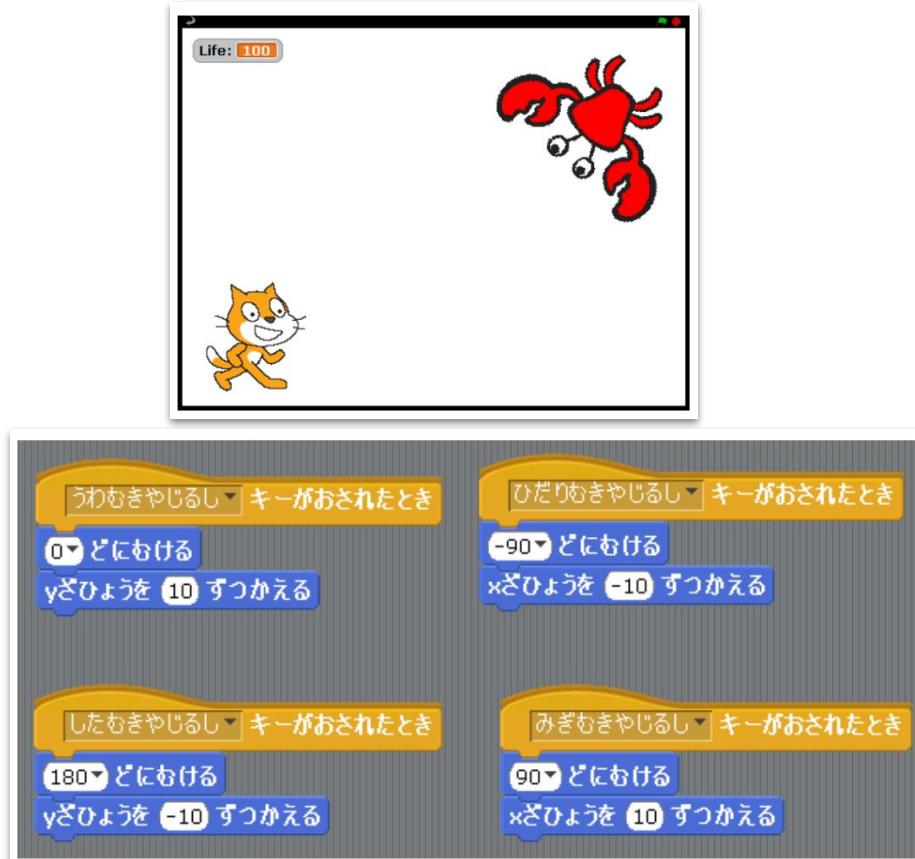


- Default အနေနဲ့ ကြောင်အော်သံက ပါပြီးသား
- ကိုယ့်အသံကို record လုပ်ထားပြီးတော့လည်း coding မှာ ခေါ်သံးလို့ရတယ်
- Story ရေးတာမျိုးဆိုရင် background music အနေနဲ့ play လုပ်တာမျိုးကိုလည်း လုပ်လို့ရတယ်
- Scratch ကတော့ တော့တော်လေး လုပ်လိုက်မှုကိုလည်း ပြောနိုင်တယ်
- ဒါ MIT ရဲ့ Scratch project က အောင်မြင်ပါတယ်

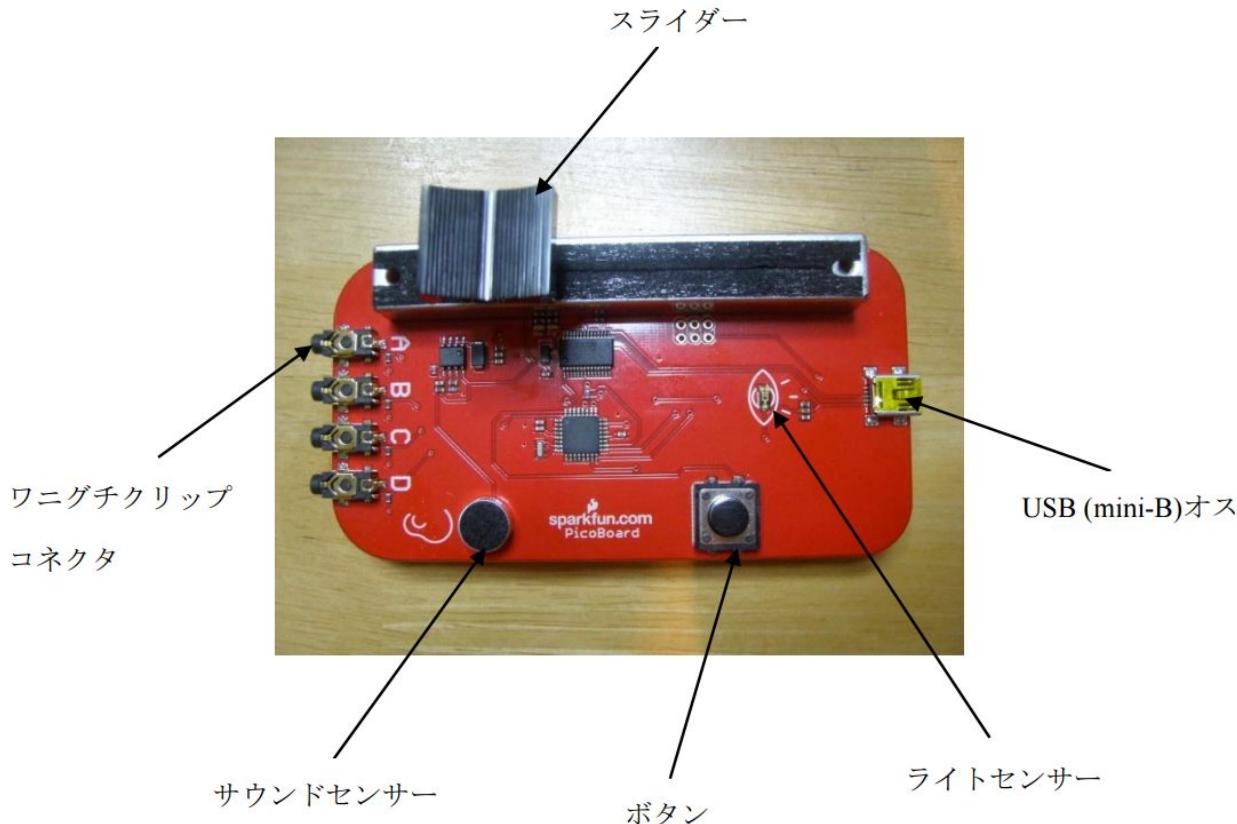
1. Educational Programming Language (Scratch)



1. Educational Programming Language (Scratch)

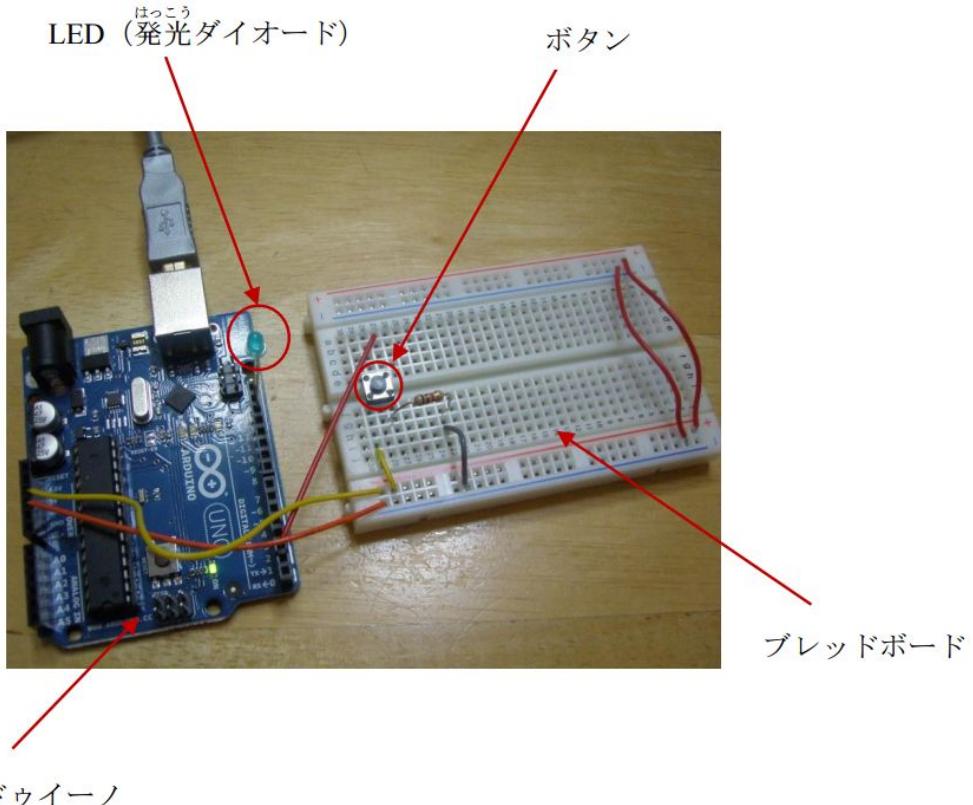


1. Educational Programming Language (Scratch)



- Pico board, Arduino board စတဲ့ physical hardware
ပေါ်နဲ့လည်း
ပို့စွဲလည်း
ချိတ်ဆက်ပြုး
Scratch coding
လုပ်လို့ ရတယ်

1. Educational Programming Language (Scratch)



Waseda Univ. မှ Research
Associate အဖြစ် အလုပ်လုပ်စဉ်က
ဂျပန် အလယ်တန်း
ကျောင်းသားတွေကို Summer
Seminar မှာ သင်ပေးခဲ့တယ်။
အမှတ်တရ တစ်ခု။ (July 31,
2011)

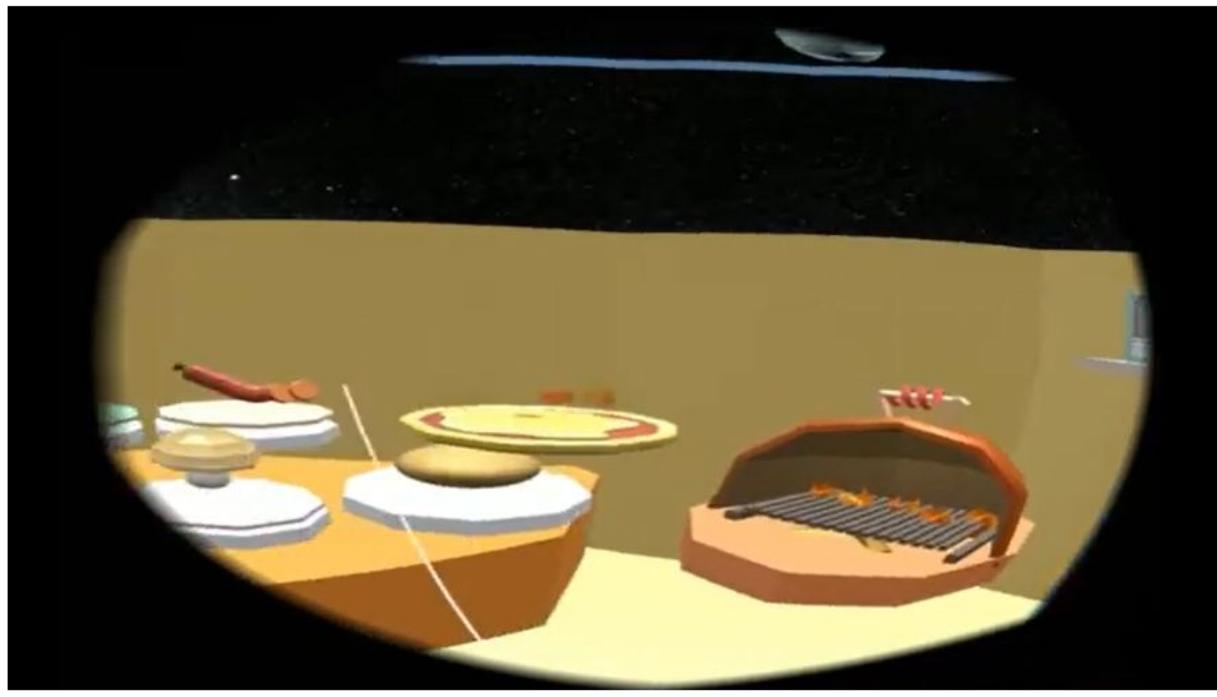
わせだだいがく
早稲田大学

こくさいじょうほうつうしんけんきゅうか
国際情報通信研究科 じょしゅ
助手

Ye Kyaw Thu (イエ チョウ トウ)

メール : wasedakuma@gmail.com

1. Educational Programming Language (Alice)



A free gift from **Carnegie Mellon University**

ORACLE



1. Educational Programming Language (Alice)



<https://www.youtube.com/watch?v=yM83so-EmaQ>

VR ပါ support
လုပ်တယ်။

ဗာရင်း 2 နှုပ္ပ
စမ်းကြည့်ဖွေးတယ်
။

ဗားရဲင်း 3 မှ
OOP ကိုပါ
support
လုပ်တယ်လို့
သာရာ။

1. Educational Programming Language (Alice)

The screenshot shows the homepage of www.alice.org. On the left, a dark sidebar lists navigation links: About Alice, Get Alice, Resources, Community, News, Research, and Support. The main content area has a yellow background. It features the Alice logo at the top left. Two large sections are displayed side-by-side. The left section is for Alice 3, which includes a white Alice logo icon above the text: "Alice 3 has all of the features that have made Alice an exciting and creative first programming experience with an added emphasis on object-oriented concepts." A blue "Get Alice 3" button is at the bottom. The right section is for Alice 2, which includes a white Alice logo icon above the text: "Alice 2 has a proven record as great tool for learning logical and computational thinking skills and the fundamental principles of programming." A blue "Get Alice 2" button is at the bottom. The URL "www.alice.org" is visible in the browser's address bar.

About Alice

Get Alice

Resources

Community

News

Research

Support

www.alice.org

Alice 3

Alice 2

Alice 3 has all of the features that have made Alice an exciting and creative first programming experience with an added emphasis on object-oriented concepts.

Get Alice 3

Alice 2 has a proven record as great tool for learning logical and computational thinking skills and the fundamental principles of programming.

Get Alice 2

1. Educational Programming Language (Alice)



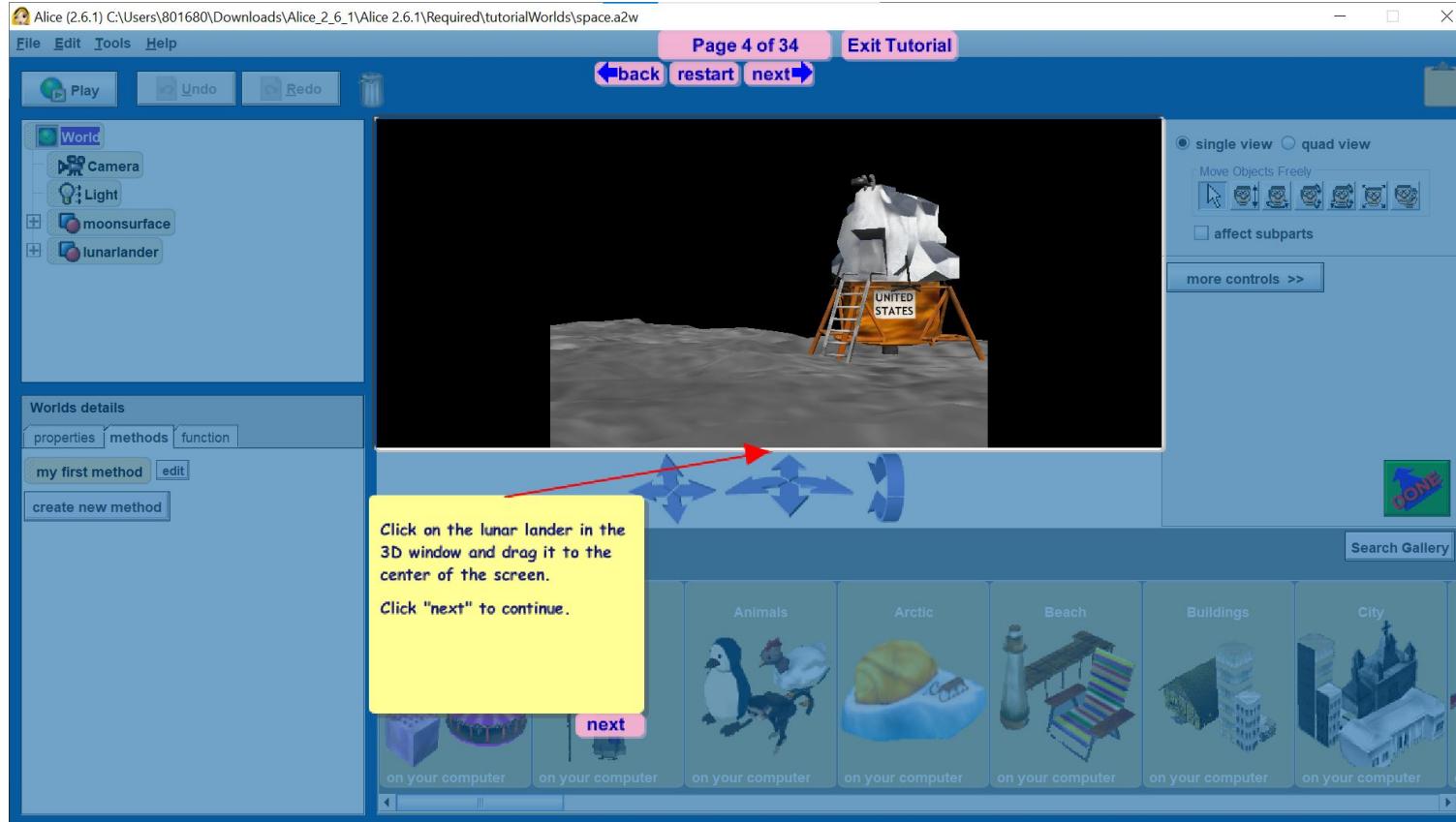
a free gift to you from

Carnegie Mellon

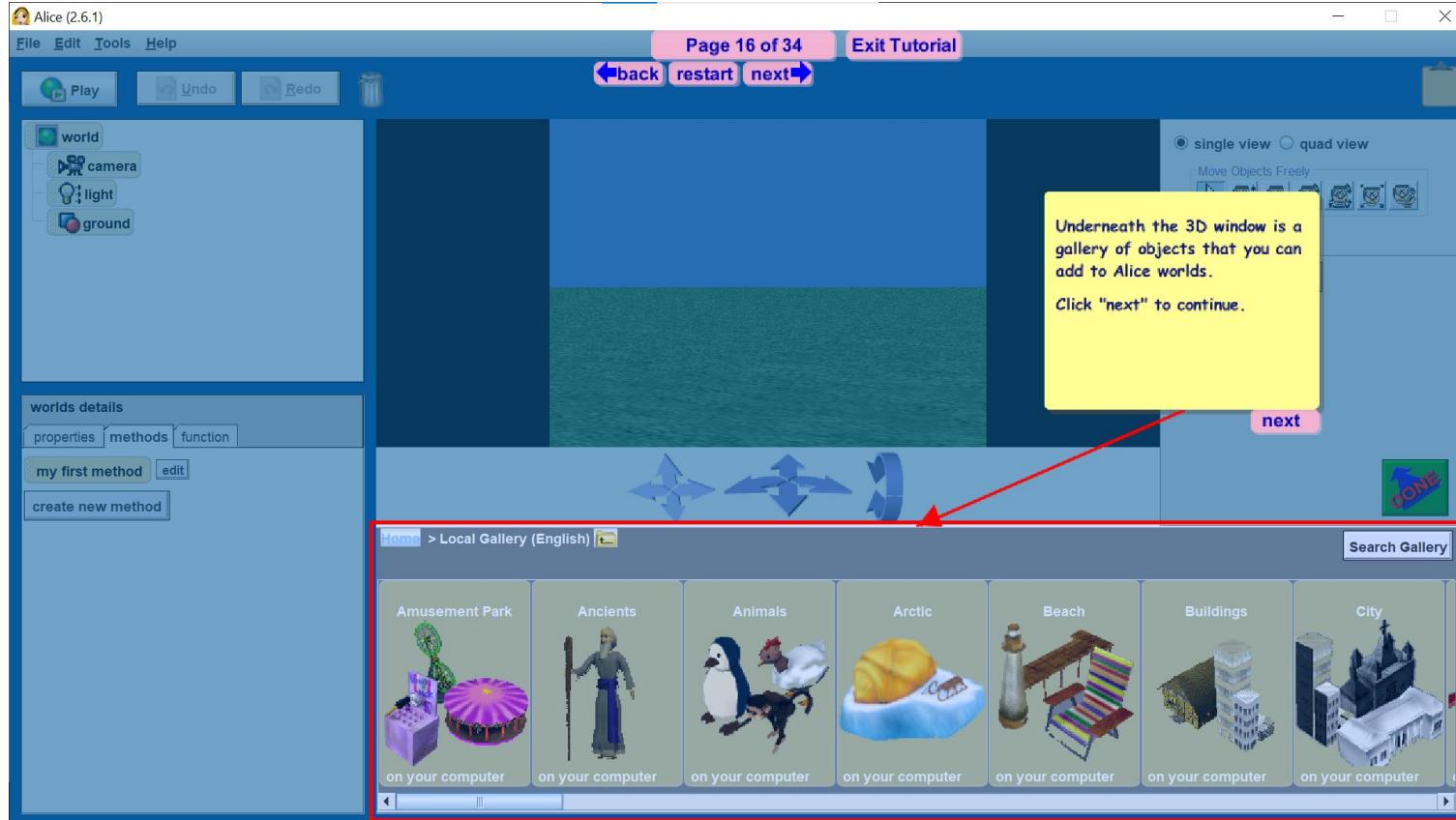
Loading . . .

version: 2.6.1

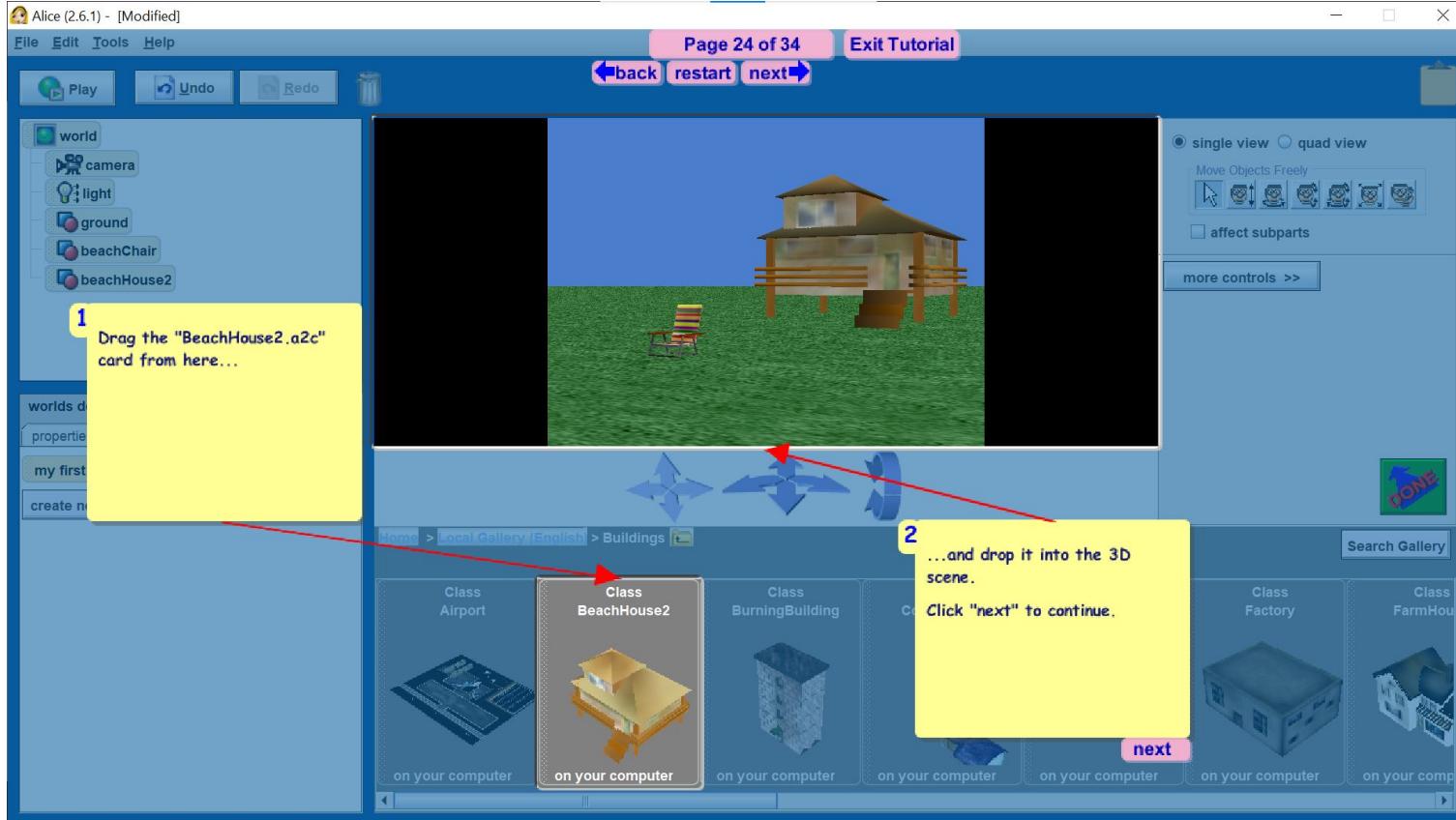
1. Educational Programming Language (Alice)



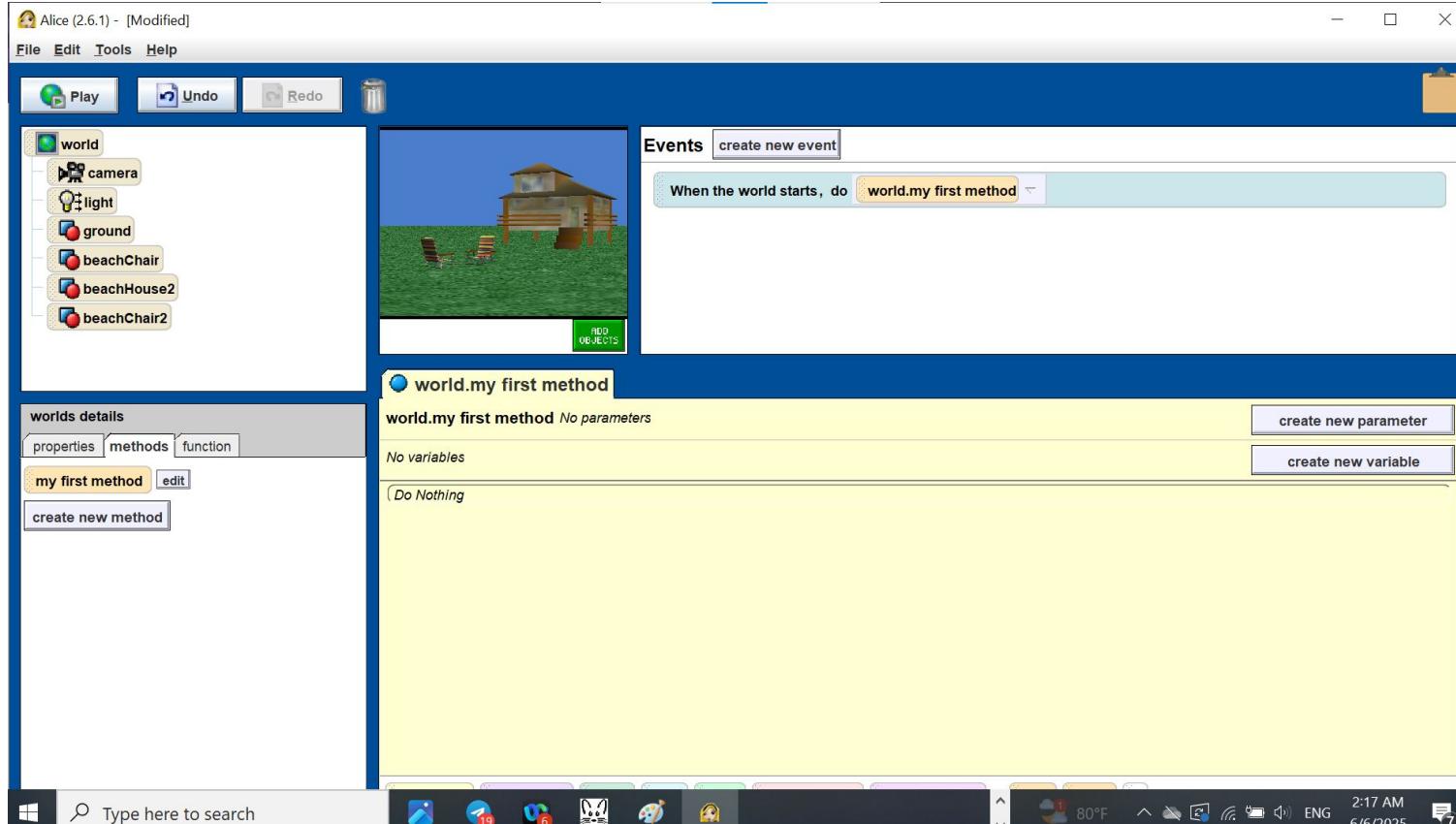
1. Educational Programming Language (Alice)



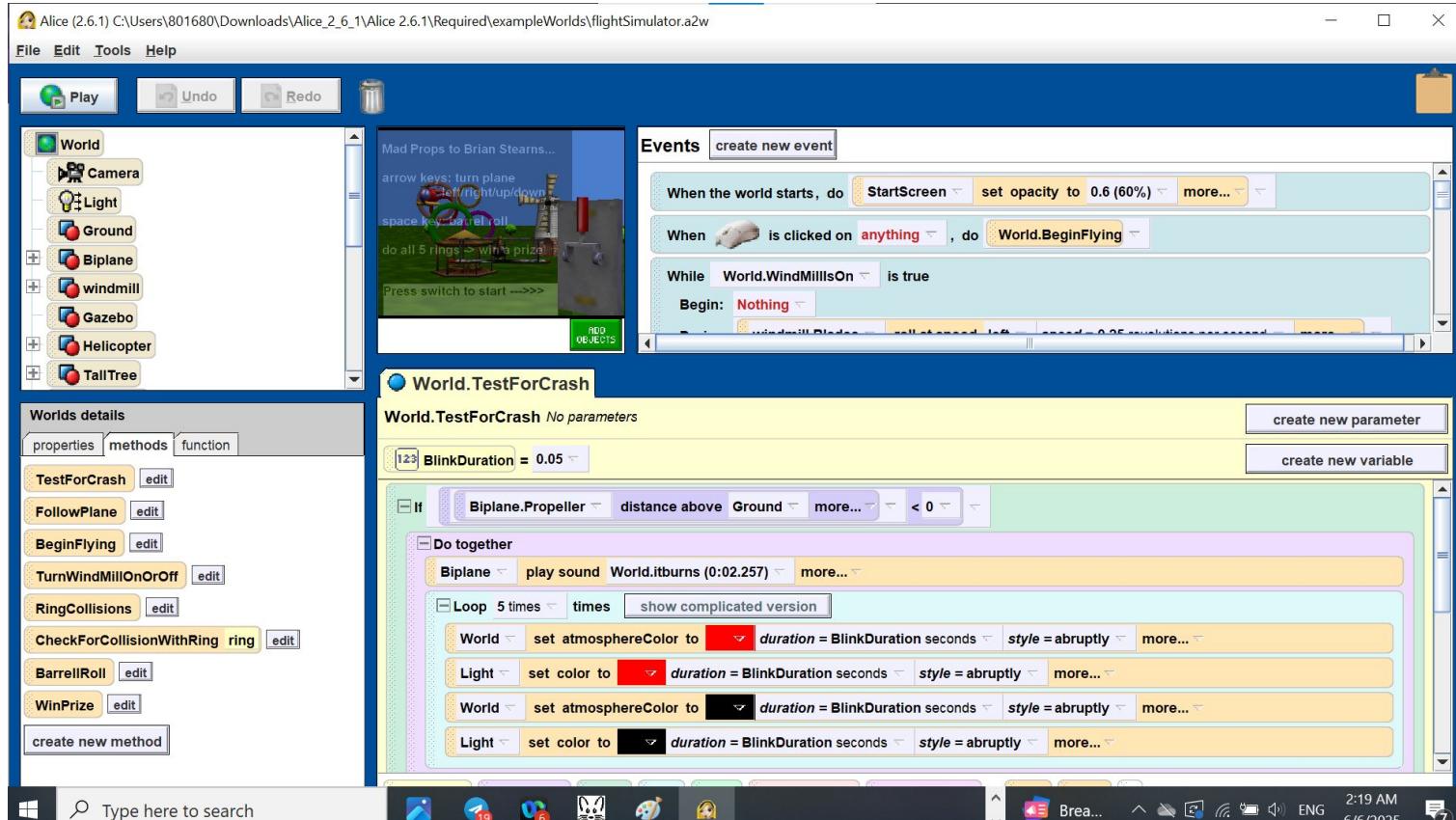
1. Educational Programming Language (Alice)



1. Educational Programming Language (Alice)



1. Educational Programming Language (Alice)



1. Educational Programming Language (Alice)



ကိုယ်ဆောထားတာကို
ပီးသိရှိကြထားပြီး ပြန် play
လုပ်ချင်းလို့ ရတယ်။

Screen capture လုပ်လို့ ရတယ်။

အခြေခံအားဖြင့် Alice က
3D Game Play Environment ပါ။

1. Educational Programming Language (Representar)

A Language Based on Two Relations between Symbols

Agustín Rafael Martínez
Universidad de Buenos Aires
Argentina
agustin@dc.uba.ar

Abstract

We present a language with all the power of abstraction and the simplicity of two fundamental relations: substitution and categorization. With a graphic symbol representing each one of them, we created a playful visual programming environment aimed at teaching with high expressive power. This environment includes tools to inspect the program execution and a console to try visual expressions. This is achieved without resorting to text, since the symbols are user-defined drawings. To address complex problems, the language offers another set of tools to define text-based programs. Here we show a functional prototype of our rule-based, general-purpose declarative programming language.

CCS Concepts: • Software and its engineering → General programming languages; • Social and professional topics → Computing education.

Keywords: general-purpose programming languages, visual programming, computer science education

The approaches to these problems are oriented in two identifiable directions. There are projects that decide to maintain simplicity at the cost of having a limited power of abstraction [27] [14] [7]. On the other hand, there are projects that decide to increase the expressive power of visual languages [10] [32], and what is gained in expressiveness is lost in complexity. Faced with this situation, teachers must choose the tool to introduce their students to programming and then identify the instance or instances in which the language change is necessary.

In this work we present a different solution. The proposal is not to change the language, but only to change the tools. Our language can work in a visual way with a high power of abstraction. We can define categories and parameterize any sequence of symbols, no matter what kind they are. When we need to move towards models where text is more appropriate to express them, we can change the tools but continue programming with the same language.

The name of the language is *Representar*, which in Spanish means ‘to represent’. The name highlights the vision of the

ဒီ စာတမ်း ဖတ်ပါ။

1. Substitution
2. Categorization

ဆိတဲ့ concept
နစ်ခေါ်ကိုပဲ
အငြော်ထားတဲ့
programming
language တစ်ခု

1. Educational Programming Language (Representar)



... substitute by ...

- Formal Grammar တော့ရဲ့ rewriting လိုပ်ဆရာက နားလည်တယ်



... is inside of ...

- အုပ်စွဲဖွဲ့တာ၊
အမျိုးအစားသတ်မှတ်တာလို့ မြင်လို့
ရတယ်

1. Educational Programming Language (Representar)

Números Naturales

The screenshot shows a programming environment with a vertical toolbar on the left containing icons for addition (+), subtraction (-), multiplication (×), division (÷), and modulus (%). Above the toolbar is a row of 12 icons representing various actions or objects. Below this are three rows of 4x4 grids. The first grid contains numbers 1-5 and hand icons. The second grid contains numbers 1-5 and hand icons. The third grid contains numbers 1-5 and hand icons. At the bottom, there are three sequences of icons: 1) Up arrow, Up arrow, Eyes icon, 2; 2) Up arrow, Up arrow, Eyes icon, Up arrow; 3) Down arrow, 5, Eyes icon, 4.

- မျက်လုံးပုန်းက evaluate လုပ်တာ။
တနည်းအားဖြင့်
ရလဒ်သိမဟုတ
output ကို ကြည့်တာ။

1. Educational Programming Language (Representar)

Representar

Números Naturales

3 + 1 = 4

1. Educational Programming Language (Representar)

- ရှေ့မှာ ရှင်းပြထားတာကို အခြေခံပြီး natural number တွေ အားလုံးကို ပေါင်းလုံး ရှုပို့ဆိုရင် ဘယ်လို့ လုပ်ရင် ရနိုင်မလဲ ဆုံးတာကို
စဉ်းစားကြည့်ရအောင်

1. Educational Programming Language (Representar)



1. Educational Programming Language (Representar)

Core Components:

- Digits (d): 0 to 9
- Non-Zero Digits (d-0): 1 to 9
- Number (N): A sequence of digits not starting with 0
- List of Digits: [d]

Two Rule Types:

- Substitution: Transforms one pattern into another ($A \rightarrow B$), e.g., $\uparrow 2 \rightarrow 3$
- Categorization: Classifies symbols into broader sets, e.g., 2 in d-0, d in [d]

Arrows for Operations:

- \uparrow = Increment (Count Up)
- \downarrow = Decrement (Count Down)

1. Educational Programming Language (Representar)

Addition Rules (Selected):

- $0 + N \rightarrow N$
- $d-0 + d-0 \rightarrow \downarrow d-0 + UP\ d-0$
- $N + d-0 \rightarrow \uparrow d-0 + N$
- $d-0 + N \rightarrow (\downarrow d-0) + (\uparrow N)$
- $N\ d + N\ d \rightarrow (N + N)\ 0 + d + d$

Wildcard Rules:

- $(*) \rightarrow *$ (Generic pattern)

Visual Semantics:

- $()$ = grouped expression
-  = final result/output

1. Educational Programming Language (Representar)

$89+90 = 179$ ကို ဘယ်လို တွက်သွားသလဲ ဆိုရင် အောက်ပါအတိုင်း

Step-by-step:

1. Break into digits:

$$89 = 8 \ 9, \ 90 = 9 \ 0 \rightarrow N \ d + N \ d$$

2. Apply Rule:

$$(N + N) \ 0 + d + d$$

3. Add digits:

$$9 + 0 \rightarrow 9, \ 8 + 9 \rightarrow \downarrow 8 + \uparrow 9 \rightarrow 7 + 10$$

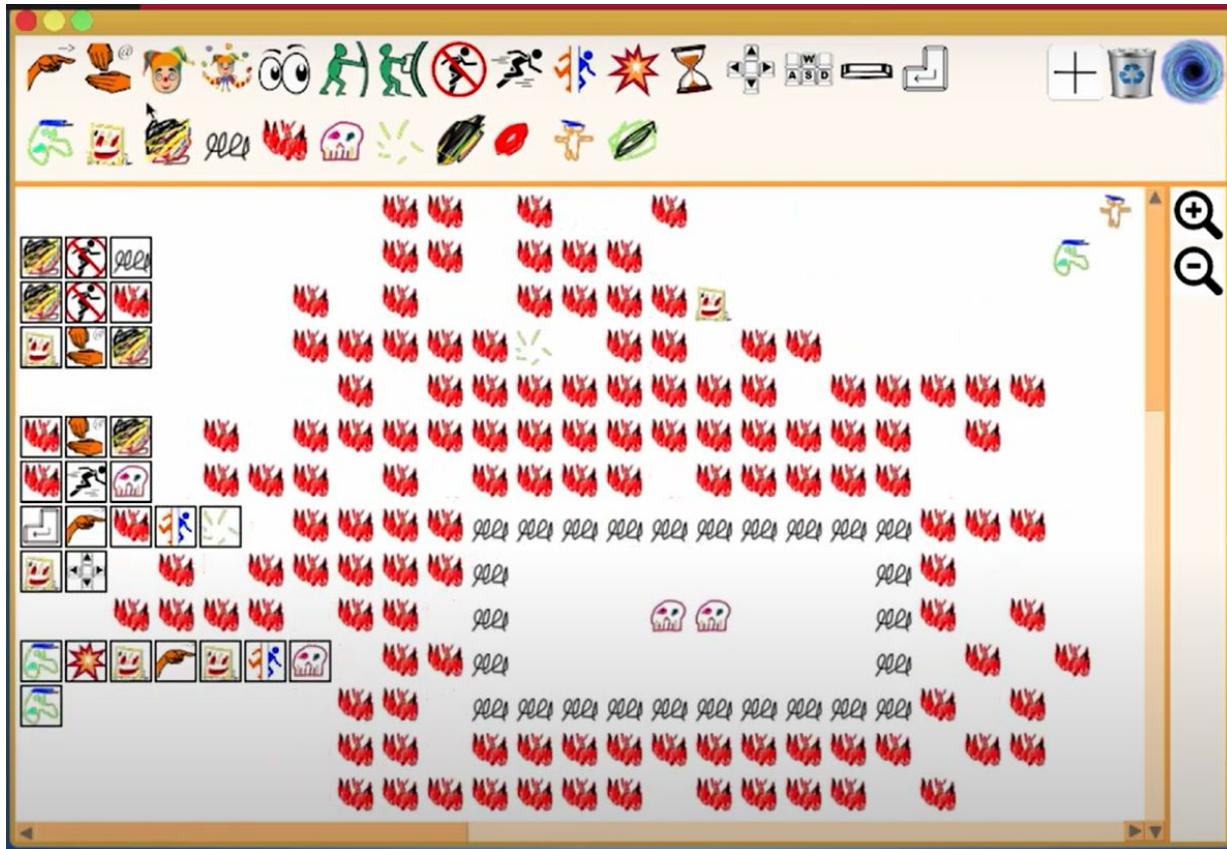
4. Handle carry:

$$10 \rightarrow \text{split as } 0 \text{ and } \uparrow 1$$

5. Final Assembly:

$$\rightarrow 1 \ 7 \ 9 \rightarrow \text{Result: EYE 179}$$

1. Educational Programming Language (Representar)



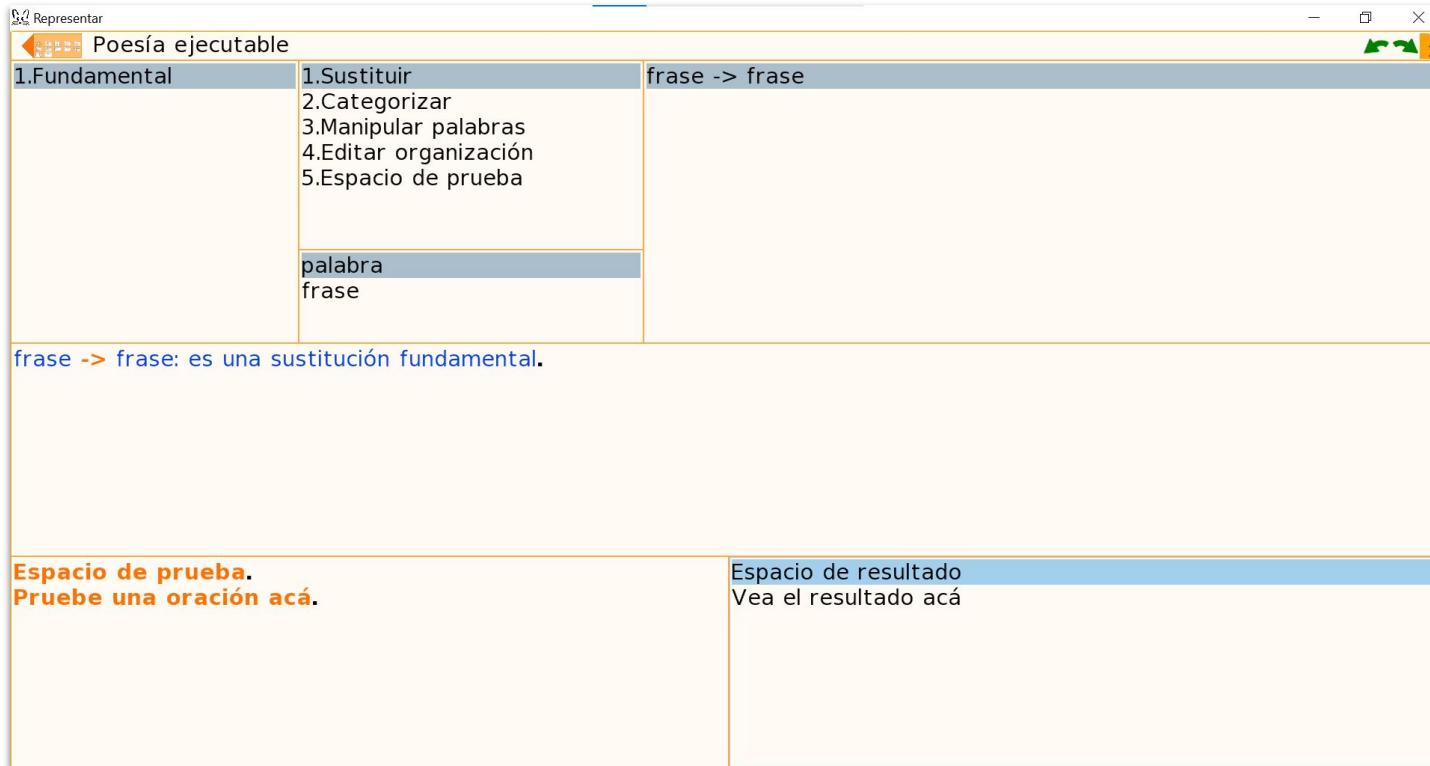
- Symbol တွေကို
ကြိုက်သလဲ
ရေးဆုံးနိုင်ပြီး
- Rule ပြင်ပြားတာနဲ့
တခြား simulation
တောလည်း
အမျှိုးမျှိုး လုပ်လို့
ရနိုင်တယ်
- ကလေးတွေ အခြေခံ
problem solving
ကိစ္စတွေကို
မြင်သွားလိမ့်မယ်

1. Educational Programming Language (Representar)

- အလွယ်ဆုံးကတော့ smalltalk နဲ့ ရေးထားတာကို ငါတို့က တွေး
programming language တစ်ခုခုနဲ့ အထားထုံးရေးတာပဲ
- ဒါပေမဲ့ အဲဒါပဲ လုပ်တယ် ဆိုရင် development အပိုင်းပဲ ဖြစ်ပြီးတော့
သူတေသန အပိုင်းက မပါသေးဘူး
- အဲဒါကြောင့် research အနေနဲ့ ဘယ်လို contribution လုပ်လို့ ရနိုင်မလဲ
ဆိုတာကို တပတ်လုံး စဉ်းစားတာ၊ တိုင်ပင်တာ လုပ်စေချင်တယ် (ဥပမာ
substitution နဲ့ categorization အပြင် တွေး rule ထပ်ဖြည့်တာမျိုး)
- သို့မဟုတ် Symbol (ပုံ) တွေပဲမဟုတ်ပဲ text နဲ့ကော့ (ဥပမာ မြန်မာစာ)
ရေးပြီး ဘယ်လို programming language အသစ်တမျိုး ထွင်လို့ ရနိုင်မလဲ
လေ့လာစေချင်တယ်

1. Educational Programming Language (Representar)

- Text based programming environment



2. Myanmar Text Readability Scoring

Readability scoring measures how easy or difficult a text is to understand. It is crucial for:

- **Education:** Ensuring textbooks match students' reading levels.
- **Content Creation:** Helping writers tailor content for different audiences (e.g., news, legal documents, websites).
- **Language Learning:** Assisting second-language learners with appropriately leveled texts.
- **Accessibility:** Making information accessible to people with varying literacy levels.
- **Government & Healthcare:** Simplifying public documents for better comprehension.

2. Myanmar Text Readability Scoring

- ရုပ်နစ်အတွက် Readability score တိုင်းပေးထဲ online website တစ်ခု ဖြစ်တဲ့ jReadability က လေ့လာကြည့်ပါ
- Link:
<https://jreadability.net/sys/en>

The screenshot shows a web browser window with the URL jreadability.net/sys/en. The title bar reads "jReadability Japanese Text Readability Measurement System". The main content area has a text input box containing Japanese text: "LGBTQの学生 9割ハラスメントなど経験 NPO法人が調査". Below the text input are several checkboxes: "Generate text details" (checked), "Generate vocab list" (checked), "Use advanced visualization" (unchecked), "Remove parentheses and text inside" (checked), and "Remove ruby characters in Aozora-Bunko text" (unchecked). At the bottom, there are buttons for "Run", "Clear", and "Reset". The footer of the page includes the copyright notice "© 2013-2024 Jae-ho Lee and Yoichiro Hasebe".

2. Myanmar Text Readability Scoring

- Run လိုက်ရင် Readability score ကတ်ကပေးလိမ့်မယ်
- Score: 1.82
- Input လုပ်ထားတဲ့
စာတဲ့မှာပါတဲ့
စာကြောင်းအရေအတွက်၊
စာလုံးအရေအတွက်၊
စာလုံးအမျိုးအစားအရေအတွက်
စာကြောင်းတစ်ကြောင်းထဲမှာ
ပါဝင်တဲ့ ပျမ်းမျှစာလုံး စာတဲ့
အချက်အလက်တွေကိုပါ
ထုတဲ့ပြပေးတယ်

jR jReadability Japanese Text Readability Measurement System

Text info	
Enable 'vocab list' to show the total number of morpheme types.	
Text Readability Level *	Lower Advanced Difficult
Readability Score	1.82
Total Num of Sentences	3
Total Num of Word Tokens	104
Total Num of Lexical Types	63
Total Num of Characters (incl. symbols and white spaces)	167
Num of Words per Sentence	34.67

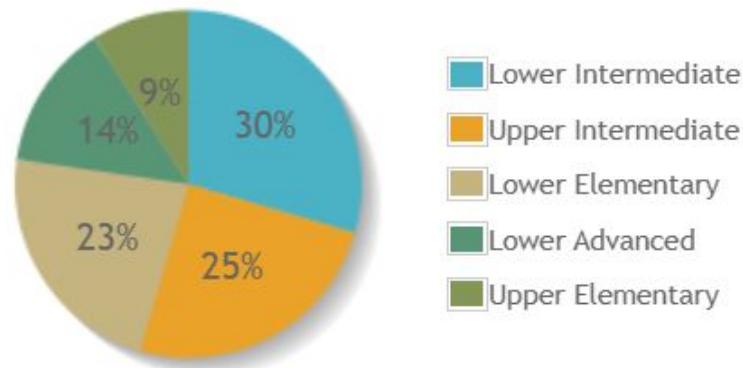
* The level represents the readability for learners of the Japanese language, not for native speakers of Japanese.

2. Myanmar Text Readability Scoring

Vocab level distribution

Words that are not listed in the system dictionary are excluded.

Lower Intermediate	13
Upper Intermediate	11
Lower Elementary	10
Lower Advanced	6
Upper Elementary	4

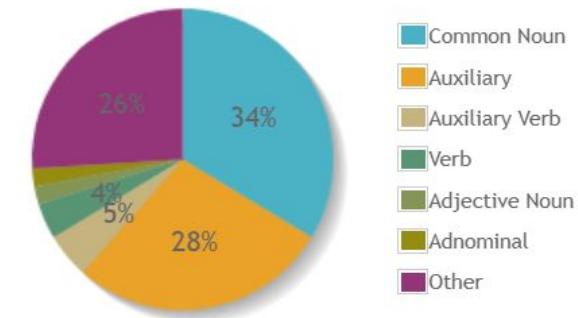


2. Myanmar Text Readability Scoring

POS-type distribution

Symbol are excluded.

Common Noun	35
Auxiliary	29
Auxiliary Verb	5
Verb	4
Adjective Noun	2
Adnominal	2
Other	27



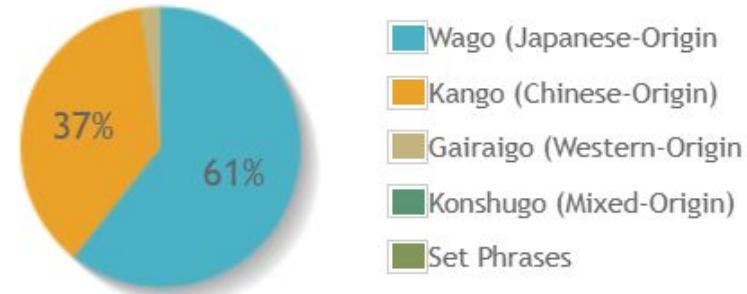
- မြန်မာစာအတွက် ဆိပ်ရင်တွေ LU Lab. ရဲ့ myPOS ရှိတယ်။ Ver. 3 က စာငြောင်းရေ စုစုပေါင်း 43,196 ရှိတယ်။
- Link: <https://github.com/ye-kyaw-thu/myPOS/tree/master/corpus-ver-3.0>

2. Myanmar Text Readability Scoring

Word-type distribution

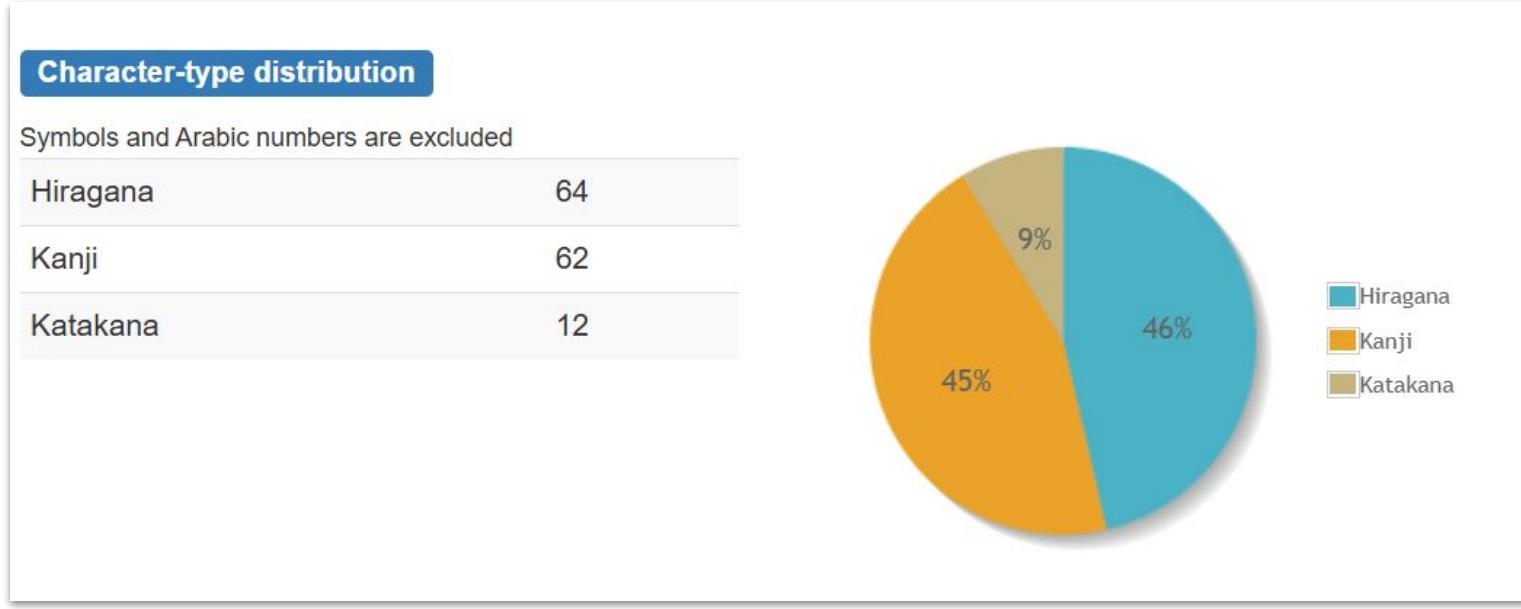
Set phrases include patterns like *arigato*

Wago (Japanese-Origin)	54
Kango (Chinese-Origin)	33
Gairaigo (Western-Origin)	2
Konshugo (Mixed-Origin)	0
Set Phrases	0



- မြန်မာစာအတွက် က word-type ကို ဘယ်လို ခဲ့ကြမှုလဲ
- ဥပမာ အသံစံနှင့်မှာ ပါတဲ့ စာလုံး၊ မွေးစားစကားလုံး၊ ဘမ်းစကား၊ ပါဌီ၊ အဂ်လိုပ် စာလုံး

2. Myanmar Text Readability Scoring



- မြန်မာစာအတွက် က word အဖြစ်သတ်မှတ်ရတာက ခေါ်တာမိုလို ဝဏ္ဏ (Syllable) ကစွဲကို စဉ်းစားဖို့ လိုအပ်မလား?! သို့မဟုတ် compound word ကိုထည့်တွက်ကြမလား
- ရခိုင်၊ ထားဝယ်၊ ဘိတ်၊ အင်းသား စတဲ့ spoken dialogue ကိစ္စတွေကိုကော...

2. Myanmar Text Readability Scoring

Total num of sentences: 3 Num of words per sentence: 34.67

Click on colored word shows definition

Download (CSV: Shift-JIS) | Download (CSV: UTF-8)

Levels of content words: Lwr-Elm Upr-Elm Lwr-Int Upr-Int Lwr-Adv Upr-Adv

Levels of functional expressions: Elementary Lower-Int Intermediate

1	LGBTQ の 学生 9 割 ハラスメントなど 経験 N P O 法人 が 調査
2	LGBTQ など 性的 少数 者の 若者を 対象に した 調査 で、この 1 年に 学校で 困難 や ハラスメントを 経験 した 中高生は 9 割 に のぼり、その うち 6 割 超 が 教職員 が 要因 に なっ て いる こと が、N P O 法人 の 調査 結果 で 判明 した。
3	学校 で 性 の 多様 性 について 学ぶ 内容 に 誤り や 差別 的な 発言 が ある こと も 浮かんだ。

- အခု jReadability က လုပ်ပြထားသလိုမျိုး မြန်မာစာအတွက် Visualization လုပ်ပေးနိုင်ရင် ပုံပြီး interactive ဖြစ်လိမ်မယ်။
- User တွေ မြန်မာစာကို လေ့လာတဲ့ နိုင်ငံခြားသားတွေအတွက် ပို အဆင်ပြေလိမ့်မယ်။

2. Myanmar Text Readability Scoring

Num of word tokens: 104 Num of word types: 63

Click on colored word shows definition

Sort key Apperance Yomi POS Freq Vocab Level

Appearance	Lemma	Yomi	POS	Freq	%	Surface	Level	Level
1	LGBTQ		名詞-普通名詞-一般	2	1.92	LGBTQ (2)		
2	の	ノ	助詞-格助詞	4	3.85	の (4)		
3	学生	ガクセイ	名詞-普通名詞-一般	1	0.96	学生 (1)	1	初級前半
4			空白	2	1.92	(2)		
5	9	キュウ	名詞-数詞	2	1.92	9 (2)	1	初級前半
6	割	ワリ	名詞-普通名詞-助数詞可能	3	2.88	割 (3)		

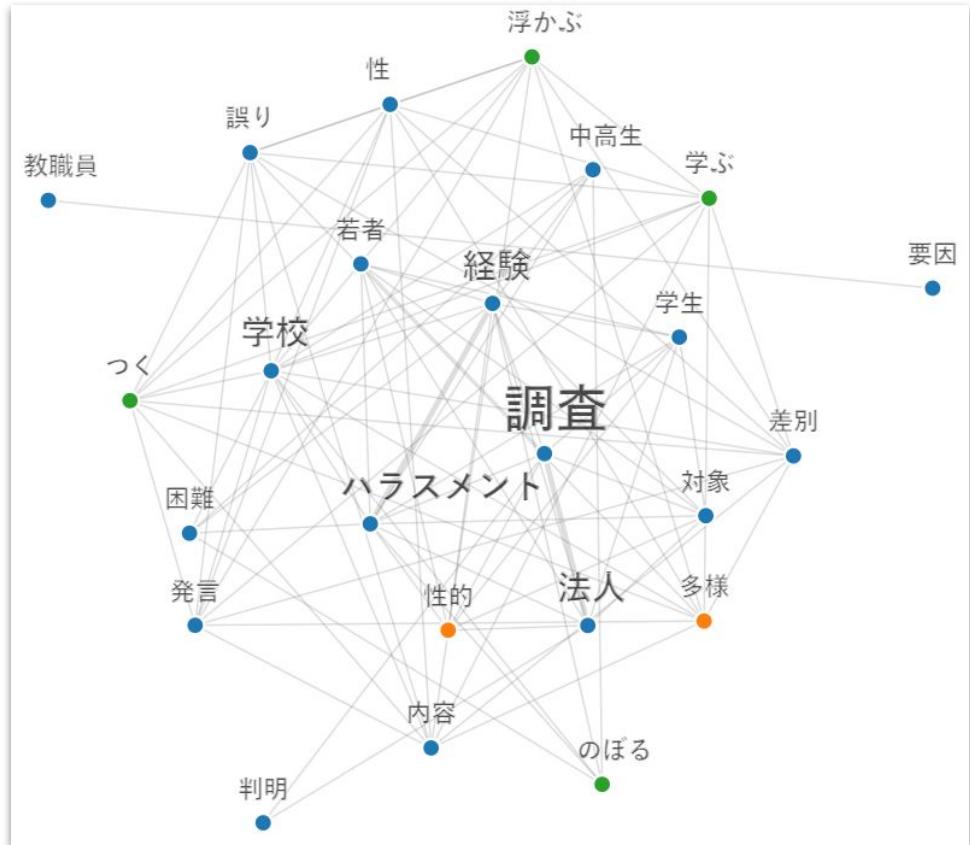
2. Myanmar Text Readability Scoring



- Facility အနေနဲ့ Word Cloud ပါ ဖြည့်ကြမယ်။

2. Myanmar Text Readability Scoring

- Word Network ကတော့
စာလုံးတွေအကြေား
ချိတ်ဆက်မှုက
visualization လုပ်ပြော

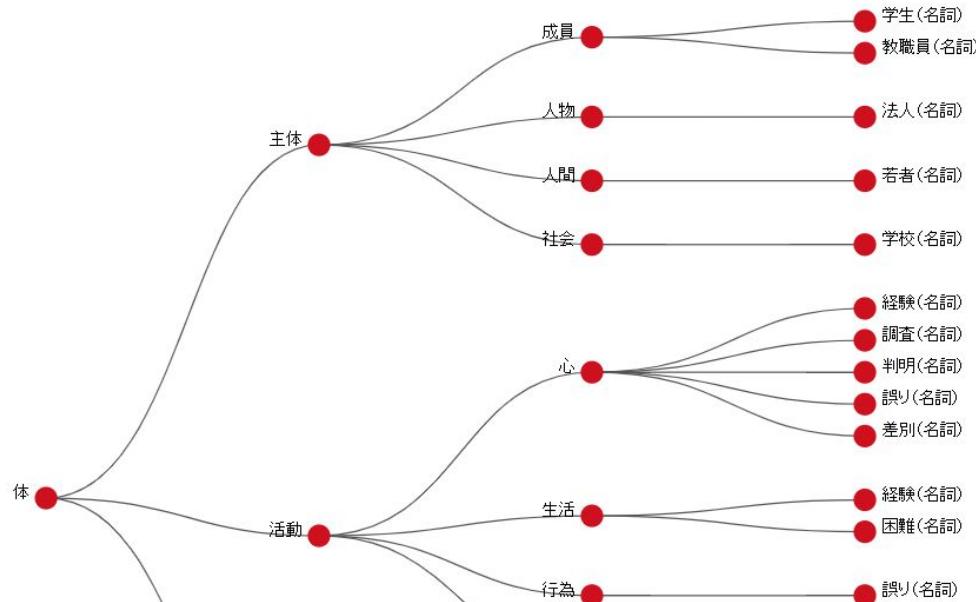


2. Myanmar Text Readability Scoring

The following hierarchies are drawn using words (the most frequent 100 words) in text in accordance with "分類語彙表" (table of

● 体 ● 用 ● 相

Click on a word shows dictio



- "分類語彙表" (table of vocabulary by semantic categories)
● ဘာသာဇာု
ပညာရင်တွေအတွက်လ
သုံး အသံးဝင်တယ်
- TTS (Text to Speech)
လည်း လုပ်ပေးနိုင်တယ်

2. Myanmar Text Readability Scoring

- Hasebe, Yoichiro and Lee, Jae-Ho (2015) 'Introducing a Readability Evaluation System for Japanese Language Education' Proceedings of the 6th International Conference on Computer Assisted Systems for Teaching & Learning Japanese (CASTEL/J), pp.19-22.
- <https://jreadability.net/file/hasebe-lee-2015-castelj.pdf>
- ဒီစာတမ်းကို ဖတ်ပြီး လွှဲလာကြပါ
- <https://github.com/joshdavham/jreadability>
- ဒေတာကိုလည်း ပြင်ကြရမယ်။ မြန်မာစာအတွက် ဘယ်လို ဖော်မြှုလာနဲ့ Readability score တွက် ကြမလဲ ဆုတာကိုလည်း စဉ်းစားကြရမယ်။

2. Myanmar Text Readability Scoring

INTRODUCING A READABILITY EVALUATION SYSTEM FOR JAPANESE LANGUAGE EDUCATION

Yoichiro Hasebe (長谷部陽一郎), Doshisha University
Jae-Ho Lee (李在鎬), University of Tsukuba

Abstract: This study introduces a readability evaluation system that was developed to support educators and learners of Japanese (available at <http://jreadability.net>). The system analyzes input text and estimates its readability, using a formula based on a regression analysis of data collected from 100 language textbooks and the balanced corpus of contemporary written Japanese (BCCWJ). In addition to scoring text in six-level categories, the system has rich functionalities that are implemented to support teachers and learners in carrying out various reading-related activities efficiently and effectively. Furthermore, feedback collected on an earlier version of the system is discussed, which confirms the usefulness of our method of evaluating text readability, while suggesting the necessity for further improvements.

Keyword: readability, text analysis, web application, Japanese education, BCCWJ

- ဆရာနဲ့ ဒီခေါင်းစဉ်ကို လပ်မယ့် ကျောင်းသားတွေလည်း စာတမ်းကောင်း တစ်စောင် ရေးနှင့်အောင် ကြုံးစားကြရအောင်။

2. Myanmar Text Readability Scoring

Model

```
readability = {mean number of words per sentence} * -0.056
            + {percentage of kango} * -0.126
            + {percentage of wago} * -0.042
            + {percentage of verbs} * -0.145
            + {percentage of particles} * -0.044
            + 11.724
```



* "kango" (漢語) means Japanese word of Chinese origin while "wago" (和語) means native Japanese word.

- မြန်မာစာ မလုပ်ဖူးသေးရင်တော့ လွယ်တယ်လို့ပဲ မြင်လိမ့်မယ်။ စာလုံး ဖြတ်ရေးတာ မဟုတ်လို့ ပြီးတော့ စံသတ်မှတ်ချကတွေ မရှိ၊ corpus မရှိ စတာတွေနဲ့ ငါတို့ ပြင်ရမှာတွေရှိတယ်...

2. Myanmar Text Readability Scoring

```
from jreadability import compute_readability

# "Good morning! The weather is nice today."
text = 'おはようございます！今日は天気がいいですね。'

score = compute_readability(text)

print(score) # 6.438000000000001
```



- သုံးလအတွင်းမှာ မီရင် Library အနေနဲ့ပါ ထုတ်ချင်တယ်

2. Myanmar Text Readability Scoring

Readability scores

Level	Readability score range
Upper-advanced	[0.5, 1.5)
Lower-advanced	[1.5, 2.5)
Upper-intermediate	[2.5, 3.5)
Lower-intermediate	[3.5, 4.5)
Upper-elementary	[4.5, 5.5)
Lower-elementary	[5.5, 6.5)

- တကယ်တမ်း ဒီလိုမျိုး score ထောက်မှတ်ဖိုက မြန်မာစာပညာရှင်တရာ့ဝန်လည်း တုပောင်နိုင်ရင်ပိုကောင်းတယ်။
- အနုသုံးဆုံးတော့ user study လုပ်ကြရအောင်...
- အုက္ခယာ ရအောင် ဆရာ ဥပမာအနေနဲ့ ဂျပန်စာကို ပြသွားပေမဲ့ အဂါလပ်စာ၊ ကိုရီးယား စတုံး တွေး
သာသာစကားတွေအတွက် readability score
ဘယ်လိုတွေကိုကြသလဲ
ဆိုတာကိုလည်း ရှာဖွေလေ့လာကြပါ

3. Creating a Myanmar SQuAD Dataset

The screenshot shows the SQuAD Explorer interface at rajpurkar.github.io/SQuAD-explorer/explore/1.1/dev/. The main title is "SQuAD2.0: The Stanford Question Answering Dataset". Below it, there are two main sections: "Explore SQuAD" and "Visualize Model Predictions".

Explore SQuAD: This section displays the "Version 1.1, dev set" with several buttons for categories: Super_Bowl_50, Warsaw, Normans, Nikola_Tesla, Computational_complexity_theory, Teacher, Martin_Luther, and another Computational_complexity_theory button.

Visualize Model Predictions: This section shows predictions from the "r-net+ (ensemble) (Microsoft Research Asia)". The predictions are displayed in a grid of colored boxes:

Row 1	Super_Bowl_50	Warsaw	Normans	Nikola_Tesla
Row 2	Computational_complexity_theory	Teacher	Martin_Luther	Southern_California
Row 3	Sky_(United_Kingdom)	Victoria_(Australia)	Huguenot	Steam_engine
Row 4	1973_oil_crisis	Apollo_program	European_Union_law	Oxygen
Row 5	Amazon_rainforest			

- Version 1.1 ကို စိတ်ဝင်စားထဲ
- Link: <https://rajpurkar.github.io/SQuAD-explorer/explore/1.1/dev/>

3. Creating a Myanmar SQuAD Dataset

- SQuAD dataset အတွက်က ဒီစာတမ်းကို refer လုပ်ပါ။ နာမည်ကြီးတဲ့ ခုံတေပါ။

The screenshot shows a research paper titled "SQuAD: 100,000+ Questions for Machine Comprehension of Text" by Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. The paper was submitted on June 16, 2016, and last revised on October 11, 2016. The page includes a summary, a question-and-answer section, and a conclusion. The arXiv logo is at the top left, and there are navigation links like 'CODE', 'Notebook', and social sharing icons.

arXiv > cs > arXiv:1606.05250

Computer Science > Computation and Language

[Submitted on 16 Jun 2016 (v1), last revised 11 Oct 2016 (this version, v3)]

SQuAD: 100,000+ Questions for Machine Comprehension of Text

[CODE](#) [Notebook](#)

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, Percy Liang

Ask the author(s) a question! :)

[Ask](#)

powered by CatalyzeX

We present the Stanford Question Answering Dataset (SQuAD), a new reading comprehension dataset consisting of 100,000+ questions posed by crowdworkers on a set of Wikipedia articles, where the answer to each question is a segment of text from the corresponding reading passage. We analyze the dataset to understand the types of reasoning required to answer the questions, leaning heavily on dependency and constituency trees. We build a strong logistic regression model, which achieves an F1 score of 51.0%, a significant improvement over a simple baseline (20%). However, human performance (86.8%) is much higher, indicating that the dataset presents a good challenge problem for future research.

The dataset is freely available at [this URL](https://www.semanticscience.org/resource/SQuAD)

3. Creating a Myanmar SQuAD Dataset

- ဒီစာတမ်းကို refer လုပ်ပါ

The screenshot shows a red header bar with the arXiv logo and navigation links for 'cs' and 'arXiv:1912.05200v2'. The main content area has a light gray background. At the top left, it says 'Computer Science > Computation and Language'. Below that, a small note indicates the submission and revision dates. The title 'Automatic Spanish Translation of the SQuAD Dataset for Multilingual Question Answering' is prominently displayed. Under the title are several interactive icons: 'CODE', 'Notebook', 'Copy', 'Bookmark', 'Bell', and 'Share'. Below the title, the authors' names are listed: Casimiro Pio Carrino, Marta R. Costa-jussà, José A. R. Fonollosa. A text input field with placeholder 'Ask the author(s) a question! :)' and a blue 'Ask' button are shown. At the bottom, a note states 'powered by CatalyzeX'. The main text body discusses the development of a multilingual question answering system, specifically focusing on translating the English SQuAD dataset into Spanish and training Spanish QA models on it.

arXiv > cs > arXiv:1912.05200v2

Computer Science > Computation and Language

[Submitted on 11 Dec 2019 (v1), last revised 12 Dec 2019 (this version, v2)]

Automatic Spanish Translation of the SQuAD Dataset for Multilingual Question Answering

[CODE](#) [Notebook](#) [Copy](#) [Bookmark](#) [Bell](#) [Share](#)

Casimiro Pio Carrino, Marta R. Costa-jussà, José A. R. Fonollosa

Ask the author(s) a question! :)

Ask

powered by [CatalyzeX](#)

Recently, multilingual question answering became a crucial research topic, and it is receiving increased interest in the NLP community. However, the unavailability of large-scale datasets makes it challenging to train multilingual QA systems with performance comparable to the English ones. In this work, we develop the Translate Align Retrieve (TAR) method to automatically translate the Stanford Question Answering Dataset (SQuAD) v1.1 to Spanish. We then used this dataset to train Spanish QA systems by fine-tuning a Multilingual-BERT model. Finally, we evaluated our QA models with the recently proposed MLQA and XQuAD benchmarks for cross-lingual Extractive QA. Experimental results show that our models outperform the previous Multilingual-BERT baselines achieving the new state-of-the-art value of 68.1 F1 points on the Spanish MLQA corpus and 77.6 F1 and 61.8 Exact Match points on the Spanish XQuAD corpus. The resulting, synthetically generated SQuAD-es v1.1 corpora, with almost 100% of data contained in the original English version, to the best of our knowledge, is the first large-scale QA training resource for Spanish.

3. Creating a Myanmar SQuAD Dataset

- ဒီစာတမ်းကိုလည်း refer လုပ်ပါ။ ၂၀၂၅ မေလမှာ တင်ထားတဲ့ preprint ပါ။

The screenshot shows a red header bar with the arXiv logo and navigation links for 'Search' and 'Help'. Below it, a white content area with a grey header bar. The header bar contains the category 'Computer Science > Computation and Language'. The main title 'IndicSQuAD: A Comprehensive Multilingual Question Answering Dataset for Indic Languages' is displayed prominently. Below the title, the authors' names are listed: Sharvi Endait, Ruturaj Ghatare, Aditya Kulkarni, Rajlaxmi Patil, Raviraj Joshi. The abstract begins with a paragraph about the rapid progress in question-answering systems and the lack of representation for Indic languages. It describes the creation of IndicSQuAD, a dataset derived from SQuAD, covering nine major Indic languages. The text discusses the challenges of maintaining linguistic fidelity and accurate answer-span alignment across diverse languages, the development of training and test sets, and the evaluation of models using monolingual and multilingual BERT variants. It concludes by mentioning potential future directions like expanding to more languages and incorporating multimodal data, and provides a link to the dataset and models.

arXiv > cs > arXiv:2505.03688

Computer Science > Computation and Language

[Submitted on 6 May 2025 (v1), last revised 13 May 2025 (this version, v2)]

IndicSQuAD: A Comprehensive Multilingual Question Answering Dataset for Indic Languages

Sharvi Endait, Ruturaj Ghatare, Aditya Kulkarni, Rajlaxmi Patil, Raviraj Joshi

The rapid progress in question-answering (QA) systems has predominantly benefited high-resource languages, leaving Indic languages largely underrepresented despite their vast native speaker base. In this paper, we present IndicSQuAD, a comprehensive multi-lingual extractive QA dataset covering nine major Indic languages, systematically derived from the SQuAD dataset. Building on previous work with MahaSQuAD for Marathi, our approach adapts and extends translation techniques to maintain high linguistic fidelity and accurate answer-span alignment across diverse languages. IndicSQuAD comprises extensive training, validation, and test sets for each language, providing a robust foundation for model development. We evaluate baseline performances using language-specific monolingual BERT models and the multilingual MuRIL-BERT. The results indicate some challenges inherent in low-resource settings. Moreover, our experiments suggest potential directions for future work, including expanding to additional languages, developing domain-specific datasets, and incorporating multimodal data. The dataset and models are publicly shared at [this https URL](https://https://)

3. Creating a Myanmar SQuAD Dataset

- ပုဂ္ဂိုလင်းလင်း ပြောရရင်တော့ မြန်မာစာကိုလည်း multilingual Q&A မှာပါ။
- SQuAD data, model:
<https://www.kaggle.com/datasets/stanfordu/stanford-question-answering-dataset?resource=download>
- Plain text:
https://huggingface.co/datasets/rajpurkar/squad/tree/main/plain_text

3. Creating a Myanmar SQuAD Dataset

Teacher

The Stanford Question Answering Dataset

The role of teacher is often formal and ongoing, carried out at a school or other place of formal education. In many countries, a person who wishes to become a teacher must first obtain specified professional qualifications or credentials from a university or college. These professional qualifications may include the study of **pedagogy**, the **science of teaching**. Teachers, like other professionals, may have to continue their education after they qualify, a process known as continuing professional development. Teachers may use a **lesson plan** to facilitate student learning, providing a course of study which is called the **curriculum**.

What is a course of study called?

Ground Truth Answers: the curriculum. curriculum curriculum

What is another name to describe the science of teaching?

Ground Truth Answers: pedagogy pedagogy formal education

Where do most teachers get their credentials from?

Ground Truth Answers: university or college. university university or college

What can a teacher use to help students learn?

Ground Truth Answers: lesson plan lesson plan lesson plan

Where is a teacher most likely to be teaching at?

Ground Truth Answers: school school school

3. Creating a Myanmar SQuAD Dataset

Doctor_Who

The Stanford Question Answering Dataset

Doctor Who is a British science-fiction television programme produced by the BBC since 1963. The programme depicts the adventures of the Doctor, a Time Lord—a space and time-travelling humanoid alien. He explores the universe in his TARDIS, a sentient time-travelling space ship. Its exterior appears as a blue British police box, which was a common sight in Britain in 1963 when the series first aired. Accompanied by companions, the Doctor combats a variety of foes, while working to save civilisations and help people in need.

Who is the producer of Doctor Who?

Ground Truth Answers: BBC BBC BBC

What year did Doctor Who first show on TV?

Ground Truth Answers: 1963 1963 1963

What is Doctor Who's space ship called?

Ground Truth Answers: TARDIS TARDIS TARDIS

What does the outside of the Tardis resemble?

Ground Truth Answers: a blue British police box a blue British police box blue British police box

What type/genre of TV show is Doctor Who?

Ground Truth Answers: science-fiction science-fiction science-fiction

3. Creating a Myanmar SQuAD Dataset

Prime_number

The Stanford Question Answering Dataset

A prime number (or a prime) is a natural number greater than 1 that has no positive divisors other than 1 and itself. A natural number greater than 1 that is not a prime number is called a composite number. For example, 5 is prime because 1 and 5 are its only positive integer factors, whereas 6 is composite because it has the divisors 2 and 3 in addition to 1 and 6. The fundamental theorem of arithmetic establishes the central role of primes in number theory: any integer greater than 1 can be expressed as a product of primes that is unique up to ordering. The uniqueness in this theorem requires excluding 1 as a prime because one can include arbitrarily many instances of 1 in any factorization, e.g., $3, 1 \cdot 3, 1 \cdot 1 \cdot 3$, etc. are all valid factorizations of 3.

What is the only divisor besides 1 that a prime number can have?

Ground Truth Answers: itself | itself | itself | itself | itself

What are numbers greater than 1 that can be divided by 3 or more numbers called?

Ground Truth Answers: composite number | composite number | composite number | primes

What theorem defines the main role of primes in number theory?

Ground Truth Answers: The fundamental theorem of arithmetic | fundamental theorem of arithmetic | arithmetic | fundamental theorem of arithmetic | fundamental theorem of arithmetic

Any number larger than 1 can be represented as a product of what?

Ground Truth Answers: a product of primes | product of primes that is unique up to ordering | primes | primes | primes that is unique up to ordering

3. Creating a Myanmar SQuAD Dataset

In meteorology, precipitation is any product of the condensation of atmospheric water vapor that falls under **gravity**. The main forms of precipitation include drizzle, rain, sleet, snow, **graupel** and hail... Precipitation forms as smaller droplets coalesce via collision with other rain drops or ice crystals **within a cloud**. Short, intense periods of rain in scattered locations are called "showers".

What causes precipitation to fall?
gravity

What is another main form of precipitation besides drizzle, rain, snow, sleet and hail?
graupel

Where do water droplets collide with ice crystals to form precipitation?
within a cloud

- Reading Comprehension (RC), or the ability to read text and then answer questions about it, is a challenging task for machines, requiring both understanding of natural language and knowledge about the world. Consider the question "**what causes precipitation to fall?**"

3. Creating a Myanmar SQuAD Dataset

Predictions by r-net+ (ensemble) (Microsoft Research Asia)

Article EM: 84.3 F1: 91.2

Apollo_program

The Stanford Question Answering Dataset

The Apollo program, also known as Project Apollo, was the third United States human spaceflight program carried out by the National Aeronautics and Space Administration (NASA), which accomplished landing the first humans on the Moon from 1969 to 1972. First conceived during Dwight D. Eisenhower's administration as a three-man spacecraft to follow the one-man Project Mercury which put the first Americans in space, Apollo was later dedicated to President John F. Kennedy's national goal of "landing a man on the Moon and returning him safely to the Earth" by the end of the 1960s, which he proposed in a May 25, 1961, address to Congress. Project Mercury was followed by the two-man Project Gemini (1962–66). The first manned flight of Apollo was in 1968.

What project put the first Americans into space?

Ground Truth Answers: Project Mercury | spacecraft | Project Mercury | Apollo | Project Apollo
Prediction: Project Mercury

What program was created to carry out these projects and missions?

Ground Truth Answers: National Aeronautics and Space Administration (NASA) | National Aeronautics and Space Administration (NASA) | Apollo | Project Mercury | Apollo program
Prediction: Apollo program

3. Creating a Myanmar SQuAD Dataset

Predictions by SLQA+ (ensemble) (Alibaba iDST NLP)

Article EM: 83.9 F1: 89.6

Apollo_program

The Stanford Question Answering Dataset

The Apollo program, also known as Project Apollo, was the third United States human spaceflight program carried out by the National Aeronautics and Space Administration (NASA), which accomplished landing the first humans on the Moon from 1969 to 1972. First conceived during Dwight D. Eisenhower's administration as a three-man spacecraft to follow the one-man Project Mercury which put the first Americans in space, Apollo was later dedicated to President John F. Kennedy's national goal of "landing a man on the Moon and returning him safely to the Earth" by the end of the 1960s, which he proposed in a May 25, 1961, address to Congress. Project Mercury was followed by the two-man Project Gemini (1962–66). The first manned flight of Apollo was in 1968.

What project put the first Americans into space?

Ground Truth Answers: Project Mercury | spacecraft | Project Mercury | Apollo | Project Apollo

Prediction: Project Mercury

What program was created to carry out these projects and missions?

Ground Truth Answers: National Aeronautics and Space Administration (NASA) | National Aeronautics and Space Administration (NASA) | Apollo | Project Mercury | Apollo program

Prediction: Apollo program

3. Creating a Myanmar SQuAD Dataset

Predictions by BERT (ensemble) (Google AI Language)

Article EM: 88.0 F1: 92.8

Apollo_program

The Stanford Question Answering Dataset

The Apollo program, also known as Project Apollo, was the third United States human spaceflight program carried out by the National Aeronautics and Space Administration (NASA), which accomplished landing the first humans on the Moon from 1969 to 1972. First conceived during Dwight D. Eisenhower's administration as a three-man spacecraft to follow the one-man Project Mercury which put the first Americans in space, Apollo was later dedicated to President John F. Kennedy's national goal of "landing a man on the Moon and returning him safely to the Earth" by the end of the 1960s, which he proposed in a May 25, 1961, address to Congress. Project Mercury was followed by the two-man Project Gemini (1962–66). The first manned flight of Apollo was in 1968.

What project put the first Americans into space?

Ground Truth Answers: Project Mercury | spacecraft | Project Mercury | Apollo | Project Apollo

Prediction: Project Mercury

What program was created to carry out these projects and missions?

Ground Truth Answers: National Aeronautics and Space Administration (NASA) | National Aeronautics and Space Administration (NASA) | Apollo | Project Mercury | Apollo program

Prediction: The Apollo program

3. Creating a Myanmar SQuAD Dataset

- SQuAD dataset ကို NLLB, Gemini (i.e. LLM) တိန် အကြမ်း ဘာသာပြန်ချလိုက်ပြီးတော့ အဲဒါကိုမှ မြန်မာ-အဂံလွှပ် NMT တော့ baseline ထားပြီး လူကဗျာပြန်စစ်ပြန်ပြင်တာ ဖြည့်နှင့်သလောက် ဖြည့်လိုက်ပြီး ဘယ်လောက်ထဲ BLEU, CharF score တွေက တက်လာသလဲ ဆုတာကို လက်တွေလုပ်ကြည့်ကြမယ်။
- မြန်မာဘာသာ SQuAD data ကို LLM ရဲ့ understanding, reading comprehension တွေကို တိုင်းတာကြည့်နှင့်ရင် စာတမ်းအသစ် ရေးလုံး ရနိုင်တယ်
- ပြီးတော့ လက်ရှိရှိပြီးသား Microsoft, Alibaba, Google, Stanford တို့ရဲ့ English SQuAD ရလဒ်တွေနဲ့ နိုင်းယဉ်တာမျိုးလုပ်ကြည့်လိုရတယ်
- ကြံးစားပြီး group member တွေအချင်းချင်း တိုင်ပြုပြီး လုပ်သွားကြရင် သုံးလအတွင်း ရလဒ် တော့ထွေကဲလိမ့်မယ်လို့ မျှော်လင့်တယ်

4. Humor Detection in Myanmar Text

“မြို့မြင်”

“ဟော... တွေပြန်ပြီ၊ ဒီသေနာကျနဲ့”

“ငါ-ငါ ချစ်တယ်ပြောသားတယ်လေ၊ အဲ-အဲဒါ အဖြူ- အဖြူ သီ-သီချင်”

သူငယ်ချင်းတွေက မြင် ဘယ်လိုဖြမ်မလဲ၊ ဘာပြောမလဲဆိတာ
စောင့်ကြည့်ရင်း ရှိသတဲ့။ မြင်က

“ဘချို့”

“ဟင်... မြို့မြင်...”

“နင် ငါကို တကယ်ချစ်သလား”

“ချစ်-ချစ်ပါတယ်ဟာ”

“ငါအချို့ကို လိုချင်တယ်ပေါ့”

“အင်း...”

အဲဒီတော့ မြင် ပြောလိုက်တဲ့စကားကြောင့် ဘချို့ခဲ့များ မျက်လုံး
ပြေားသွားပြီး မြင်ရဲ့ သူငယ်ချင်းတွေ ပြုင်တူပြီးလိုက်ကြသတဲ့။ မြင်
ပြောလိုက်တဲ့စကားက

“ဒီမယ်ဘချို့... တောရကျောင်းက စော်ကြီးပြုပြင်ပြီးလို့ အဲဒီ
စော်ကြီး ဘုရားပွဲကျင်းပပြီး ပွဲပါတဲ့နှစ်ကျေရင် နင်ကိုချစ်မယ် ဘချို့”
“ဂျာ...”

ဘချို့ခဲ့များ (များ)ကိုတောင် ပိုအောင်မပြောနိုင်ဘဲ (များ)ဆိုပြီး
ငေးကြည့်ရင်း ကျွန်ုတ်သတဲ့။ အဲဒီအဖြစ် အဲဒီစကားက အဲဒီရွှာမှာ နာမည့်
ကြီးပြီး ရာဇ်ဝါတွင်သွားသတဲ့။

ဘချို့ကိုလည်း သူငယ်ချင်းတွေက နာမည့်ပြောင်းပေးကြသတဲ့။
တောရကျောင်း အချို့ကြီး...တဲ့။

- မြန်မာ့ကျေးလက် ဟာသမား (၁)၊ တွင်းကြီးသား တင်ဝင်းခြီး

4. Humor Detection in Myanmar Text

ဆီသွားဖို့ လက်မှတ်ရပြီးသား။ ဒီတော့ သေမှာ အမှန်ပဲ။ ဒါကြောင့် ဘုန်းကြီးတွေ အလုပ်မရှိတဲ့ရက်မှာ သေပါကွယ်။ အဖေတို့ကတော့ လူ လေးရဲ့ အသာချုပ် အရုံသင့်လုပ်ထားပြီးသား။”

အေသလို လူကြားထဲမှာ ပြောတော့ မောင်ဘအေးရော မသင်းမေပါ မျက်လုံးဖြူးနေကြသလို၊ အခြားသူ လေးငါးညီးလည်း မျက်လုံးဖြူးနေကြ တာပေါ့။

အဲဒီအချိန် မောင်ဘအေးရဲ့ မိခင်ပြောလိုက်တဲ့ စကားကြောင့် အေးလုံးဖြူးရယ်ကြပေမယ့် မောင်ဘအေးတို့ လင်မယားမှာတော့ ပိုပြီး မျက်လုံးဖြူးသွားသတဲ့။

မိခင်ဖြစ်သွားရဲ့ စကားက...

“လူလေးရယ်... မင်းမိန်းမ အမည်းကိုက သင်းမေတဲ့။ သင်းမေ ဆိုတာ သေမင်းကွယ်။ သေမင်း-သေမင်း”

အဲဒီအဖြစ်နဲ့ အဲဒီစကားက ရာဇ်ဝင်တွင်သွားသတဲ့။

- မြန်မာ့ကျေးလက် ဟာသများ (၁)၊
တွင်းကြုံးသား
တင်ဝင်းဦး

4. Humor Detection in Myanmar Text

အသလိပြာတော့ ကိုမြခင်က အရက်ချက်ကို မပေးဘဲ၊ လေးချက် အထိနှုပြီး သွားသွား သောက်နေသတဲ့။ အဲဒီတော့ ကိုအေးမောင်က ပုလင်း လက်ကျွဲထဲက အရက်ကို ယူပြီးမေ့မယ်အလုပ် ကိုမြခင်က...

“ဖွတ်ထဲဖွတ်နှတ်၊ မန္တုတ်လောက်၊ ဖွတ်မန္တုတ်လောက် ဖွတ်လာ၊ လာတဲ့ဖွတ်နဲ့ ဖွတ်နဲ့ပေါင်း... ဖွတ်...”

အသလိပိုလိုက်တော့ ကိုအေးမောင်က...

“ဒါ ဘယ်သွဲပြာတာလဲလဲ။ ဖွတ်ဆိုတာ”

အဲဒီတော့ ကိုမြခင်က...

“ဟ - ဒီအနား မင်းနဲ့ငါပရှိတာ၊ ငါက ပြောတဲ့လဲဆိုတော့ မင့်ကို စောင်းပြောတာပေါ့ ကျ”

အဲဒီရက်ကတည်းက ကိုအေးမောင်ဟာ ကိုမြခင်ဆီက အရက်ကို ကပ်မသောက်ရဲတော့ဘူးတဲ့။ ကိုမြခင် ရွတ်လိုက်တဲ့စကားကြောင့် ကိုအေးမောင်လည်း နာမည်ကြီး သွားသတဲ့။

အဲဒီစကားက ကိုအေးမောင်ကိုမြင်တိုင်း “ဖွတ်ထဲဖွတ်နှတ်၊ မန္တုတ်လောက်၊ ဖွတ်မန္တုတ်လောက်ဖွတ်လာ၊ လာထည့် ဖွတ်နဲ့ဖွတ်နဲ့ပေါင်း ဖွတ်” လိုအုပြီး ခေါ်ကြသတဲ့။ မှတ်ကရော...”

- မြန်မာ့ကျေးလက်
ဟာသများ (၁)၊
တွင်းကြုံသား
တင်ဝင်းဦး

4. Humor Detection in Myanmar Text

အိပ်ဆုံးများ

ဆရာမ = ‘နတ်ပြည်ကို သွားချင်တဲ့ကလေးတွေ လက်ထောင်ပြပါ။’ အေရန်မှုလွှဲ၍ ကျွန်ကလေးများအားလုံး လက်ထောင်ပြကြ၏။

ဆရာမ = ‘အေရန်... သားက နတ်ပြည် မသွားချင်လို့လား။’

အေရန် = ‘ဟူတ်ပါတယ် ဆရာမ။ အိမ်ကို တန်းတန်း မတ်မတ် ပြန်ခဲ့ရမယ်လို့ အမေက မှာထားလို့ပါ’



- နှစ်သက်စရာ ဟာသများ၊ ဇုဒေး

4. Humor Detection in Myanmar Text

လုပ်ကြောင်းမှာ

- | | |
|------|---|
| သား | = ‘ဒီညတော့ သား စာမကည့်နိုင်တော့ဘူး၊ အရမ်းပင်ပန်းနေပြီ။’ |
| ဖခင် | = ‘အလုပ်ပင်ပင်ပန်းပန်း လုပ်ရလို့ဆိုပြီး သေတဲ့ လူဆိုတာ မရှိသေးပါဘူးကွာ။’ |
| သား | = ‘ဒီတော့... သားက ပထမဆုံးဖြစ်အောင် လုပ်ကည့်ရမှာလား။’ |



- နှစ်သက်စရာ ဟာသများ၊ အင်အေး

4. Humor Detection in Myanmar Text

အကျင့် ပြောမှတ်သူ

ကျောင်းဆေးခန်း ဆရာဝန်က သူ၏ ကျောင်းမိဘဖြစ်သူအား
ပြောပြနေ၏။

‘အစ်မကြီးသမီး မျက်မှန် တပ်ရလိမ့်မယ် ထင်တယ်။’

ထိုအခါ ကျောင်းသူ၏ မိခင်က...

‘ဘယ်လိုလုပ်ပြီး ပြောနိုင်တာလဲဟင်’ ဟု မေးလိုက်လေသည်။

ထိုအခါ ကျောင်းဆေးခန်း ဆရာဝန်က...

‘အစ်မကြီးသမီး ပြေတင်းပေါက်ကနေ ဝင်လာတာကိုကြည့်ပြီး
ပြောတာပါ’ ဟု ပြောလိုက်လေသတည်း။



- နှစ်သက်စရာ ဟာသများ၊ အင်အေး

4. Humor Detection in Myanmar Text

အနောက်တိုင်းဟာသ ကိုးမျိုး [ပြင်ဆင်ရန်]

- ၁။ မြတ်ရာမှ နိမ့်ရာသို့ လျှောကျခြင်း (Bathos)
- ၂။ သရုပ်ပျက် (Caricature)
- ၃။ ဥပမာမြောက် (Comic Simile)
- ၄။ ပကာသနကို ဖောက်ထွင်းပစ်ခြင်း (Debunking)
- ၅။ သဏ္ဌာန်လုပ် သရုပ်ဖော် (Impersonation)
- ၆။ အသံထွက် ဆင်တူ (Pun)
- ၇။ ဝေနညာဏ် (Witticism)
- ၈။ ခေတ်ပြောင် (Satire)
- ၉။ စကားထာ စကားဂုဏ် (Riddle)

- ၁။ အနန္တာအနန္တငါးပါးကို ပြက်ရယ်ပြခြင်း (Blasphemy)
- ၂။ ကျောင်းသားအလွှဲဖြေ (Howlers or Boners)
- ၃။ ခနိုးခနဲ့ (Irony)
- ၄။ နာခေါင်းသွေးထွက် (Practical Joke)
- ၅။ ကလေးရွှေ့ (Nonsense and Nursery Tales)
- ၆။ စာဖျက် (Parody)

4. Humor Detection in Myanmar Text

- အရှင်ဆုံး ဒေတာကို ဘယ်လိုပံ့စွန့်စုံကြမလဲ ဆိုတာကို စဉ်းစားကြရလိမ့်မယ်
- အငြော်ထဲက ဟာသ၊ ရုပ်ရှင်ထဲက ဟာသ၊ Facebook ထဲက ဟာသ၊
စာအပ်ထဲက ဟာသ
- စာကြောင်း တစ်ကြောင်းချင်းစီ စုဖို့ ခက်လိမ့်မယ်
- အခြေခံအားဖြင့်က စာကြောင်း တစ်ကြောင်းနဲ့ အထက်ပဲ
- Label or annotation က ဘယ်လို လုပ်ကြမလဲ
- Humor Classification/Detection က NLP/AI အနေနဲ့က မလွယ်တဲ့ အလပ်ပါ
- Language Understanding အပိုင်းအနေနဲ့ AI field အနေနဲ့လည်း အရေးကြီးတဲ့ R&D topic တစ်ခုပါ

4. Humor Detection in Myanmar Text

- Lab member သူရအောင်က ...
- Offensive
- Mildly Appropriate
- Neutral
- Wholesome ဆိုပြီး Class ခဲ့တာ တွေဖြံး

4. Humor Detection in Myanmar Text

- Irony (သရော် ခနဲ့)
- Hyperbole (အတိုင်းသယ)
- Self-deprecation (မိမိကိုယ် မိမိ ပြက်)
- Metaphorical (ရူပမေယာ ဆောင်)
- Positive (အကောင်းမျင်)
- Cultural Ref (ရှိုးရာ ပြန်ညွှန်း) ဆိုတာမျိုးတွေလည်း
တွေ့ပါတယ် ဆရာ

5. Optimizing Myanmar Keyboard Layouts

kKg ကီးဘုတ် ပိတ်ဆက် (Version 1)

ကျွန်တော်က မြန်မာစာလုံးတွေကို အရမ်းချစ်ပါတယ်၊ သို့သော် မြန်မာစာကို ရိုက်ဖို့အတွက် အဲဒီအခိုန်တုန်းက ရှိနေတဲ့ မြန်မာစာကီးဘုတ်လက်ကွက်တွေကိုကြည့်ပြီး၊ လွှာကျင့်ဖို့ အရမ်းကို အပျင်းကျခဲ့ပါတယ်။ ၂၀၀၄ ခုနှစ်လောက်ကပါ။ ကျွန်တော် ဂျပန်၊ တိုက္ခိမှာရှိတဲ့ Waseda တက္ကသိုလ်မှာ မဟာဘွဲ့အတန်းတက်နေစဉ်၊ မြန်မာစာနဲ့ ပတ်သက်တဲ့ UI (User Interface) သုတေသနတွေကို စလုပ်ဖြစ်တော့ မြန်မာစာကို ကွန်ပြု၍တာသုံးပြီး ရိုက်ဖို့လိုအပ်လာတဲ့ အခါ၊ kKg ကီးဘုတ်ကို စပြီးစဉ်းစားဖြစ်ခဲ့ပါတယ်။ ငါက QWERTY ကီးဘုတ်ကို သုံးပြီး အင်လိပ်လို ရိုက်တတ်နေတာပဲ၊ QWERTY ကီးဘုတ်အကွဲရာအစီအစဉ်အတိုင်း မြန်မာစာကိုကော mapping လုပ်လို မရနိုင်ဘူးလားလို့။ အဲဒါနဲ့ ဒီ kKg (ကခါ) ကီးဘုတ်ကို လုပ်ဖြစ်သွားခဲ့ပါတယ်။

5. Optimizing Myanmar Keyboard Layouts

kKg ကီးဘုတ် ခလုပ်နေရာချုပ် (keyboard layout) က မြန်မာစာ စာလုံးတွေရဲ့ အသံထွက်ပေါ်ကို အခိုက အခြေခံ ထားပါတယ်။ အင်လိပ်ကီးဘုတ် ရိုက်နေကျသူများက အကျဉ်းတဝ် ရှိပြီးသားဖြစ်တဲ့ Shift ကီးနဲ့တွဲပြီး အင်လိပ် စာလုံး အသေး၊ အကြီး ပြောင်းတဲ့ပုံစံကိုလဲ အသုံးပြုထားပါတယ်။ အဲဒီ အချက် ပျက်အပေါ်ကို အခြေခံထားပြီး တော့

- က ကို `k` ကီး၊ ခ ကို `K` (`shift + k`) ကီး။
- ဂ ကို `g` ကီး၊ ယ ကို `G` (`shift + g`) ကီး။
- ဓ ကို `s` ကီး၊ ဆ ကို `S` (`shift + s`) ကီး။
- ဇ ကို `z` ကီး၊ ဈ ကို `Z` (`shift + z`) ကီး။
- ဒ ကို `d` ကီး၊ ဓ ကို `D` (`shift + d`) ကီး။
- န ကို `n` ကီး၊ ဏ ကို `N` (`shift + n`) ကီး။
- ပ ကို `p` ကီး၊ ဖ ကို `P` (`shift + p`) ကီး။
- ဟ ကို `b` (`shift + b`) ကီး၊ ဘ ကို `b` ကီး။
- မ ကို `m` ကီး။
- ယ ကို (`shift + y`)၊ ရ ကို `y` ကီး။
- လ ကို `l` ကီး၊ ဋ ကို (`shift + l`) ကီး မှာ အသီးသီး နေရာချိန်ပါတယ်။

5. Optimizing Myanmar Keyboard Layouts

kKg ကီးဘုတ်မှာ အသံထွက်ပေါ်အခြေခံပြီး နေရာချတဲ့ အယူအဆအပြင် နောက်ထပ် အယူ
အဆတရုက မြန်မာစာလုံးတွေရဲ့ စာလုံးပုံသဏ္ဌာန်နဲ့ အဂ်လိပ်စာ စာလုံးပုံစံတွေရဲ့ဆင်တူမှာ
အသုံးပြုမှုနဲ့ နီးစပ်တဲ့ အပေါ်ကို အခြေခံတာပါ။
ဥပမာ။။

မြန်မာဗျည်း c ကို အဂ်လိပ်စာလုံး small c  ကီးမှာ နေရာချထားတာမျိုး၊

မြန်မာစာလုံး ဃဲ့ ဃဲ့ ကို အဂ်လိပ်စာလုံး full stop  ကီးမှာ နေရာချထားတာမျိုး၊

မြန်မာစာလုံး । (ပုံဒ်မ)ကို အဂ်လိပ်စာလုံး comma  ကီးမှာ နေရာချထားတာမျိုး၊

မြန်မာစာလုံး ၂၂ (ယပင်း)ကို အဂ်လိပ်စာလုံး small j  ကီးမှာ နေရာချထားတာမျိုးပါ။

5. Optimizing Myanmar Keyboard Layouts

၁	၉	၇	၈	၅
k	K	g	G	c
ၦ	၃၀	၆	၂၂	၂၂
s	S	z	Z	q
၂	၄	၂	၂	၂
v	X	V	~	N
၈	၈	၃	၈	၄
T	x	d	D	n
၁	၅	၂	၃	၈
p	P	B	b	m
၁	၄	၂	၁	၁
Y	y	I	w	t
	၃	၂၂	၁၁	
	h	L	a	

၁	၁	၁	၁
r	R	i	I
၁	၁	၁	၁
u	U	A	e
၁	၁	၁	၁
O	၁	.	>

- kKg ခဲ့ ပျည်း၊ သရာ keyboard mapping

5. Optimizing Myanmar Keyboard Layouts



5. Optimizing Myanmar Keyboard Layouts

```
key <AC01> { [ U1021, U1031, a, A ] }; // ၁၀၀၀
key <AC02> { [ U1005, U1006, s, S ] }; // ၁၁၀၀
key <AC03> { [ U1012, U1013, d, D ] }; // ၁၂၀၀
key <AC04> { [ U103A, U1039, f, F ] }; // ၁၃၀၀ padsint
key <AC05> { [ U1002, U1003, g, G ] }; // ၁၄၀၀
key <AC06> { [ U101F, U103E, h, H ] }; // ၁၅၀၀
key <AC07> { [ U103B, U103C, j, J ] }; // ၁၆၀၀
key <AC08> { [ U1000, U1001, k, K ] }; // ၁၇၀၀
key <AC09> { [ U101C, U1020, l, L ] }; // ၁၈၀၀
key <AC10> { [ U1038, U104B, semicolon, colon ] }; // ၁၉၀၀
key <AC11> { [ apostrophe, quotedbl ] }; // ' "
```

5. Optimizing Myanmar Keyboard Layouts

The screenshot shows the Keyman Developer software interface. On the left, the 'Welcome' screen features the KeymanDeveloper logo, the URL <https://keyman.com/developer>, and the version 'Version 17.0.330'. Below this, the SIL International logo and the text 'Created by SIL International' and 'Copyright © SIL International. All Rights Reserved.' are displayed. A large 'Get Started!' button is prominent. To the right, the 'Character Map' panel is open, showing a grid of characters from the Unicode range U+2800 - U+28FF. The grid is organized into three columns and four rows, with each cell containing a character or a dot. The first row contains '•', '•', and '•'. The second row contains '•', '•', and '•'. The third row contains '•', '•', and '•'. The fourth row contains '•', '•', and '•'. Below the grid, there are buttons for 'U+2800 - U+28FF' and 'Help'. A 'Project Manager' section is also visible, stating: 'The Project Manager allows you to manage all the files related to a keyboard layout in a single location.' and 'More help...'. The overall interface is clean and professional.

- Prototyping အတွက် Keyman Developer ကိုလည်း သုံးနိုင်တယ်

5. Optimizing Myanmar Keyboard Layouts

MY-AKKHARA:

A Romanization-based Burmese (Myanmar) Input Method

Chenchen Ding, Masao Utiyama, and Eiichiro Sumita

Advanced Translation Technology Laboratory,

Advanced Speech Translation Research and Development Promotion Center,
National Institute of Information and Communications Technology

3-5 Hikaridai, Seika-cho, Soraku-gun, Kyoto, 619-0289, Japan

{chenchen.ding, mutiyama, eiichiro.sumita}@nict.go.jp

Abstract

MY-AKKHARA is a method used to input Burmese texts encoded in the *Unicode* standard, based on commonly accepted Latin transcription. By using this method, arbitrary Burmese strings can be accurately inputted with 26 lowercase Latin letters. Meanwhile, the 26 uppercase Latin letters are designed as shortcuts of lowercase letter sequences. The frequency of Burmese characters is considered in MY-AKKHARA to realize an efficient keystroke distribution on a QWERTY

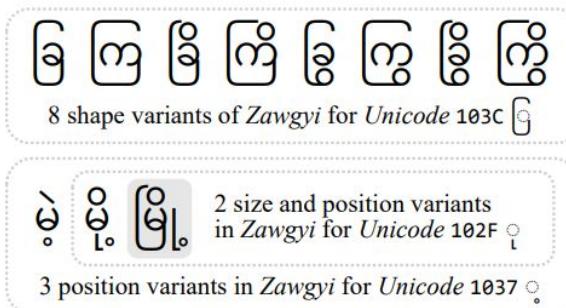


Figure 1: Shape, size, and position variants in Zawgyi for identical *Unicode* characters. The characters may

- Romanization-based approach ပါ
- ဆရာတေသနအရှင်စက်တဲ့မှာလည်း
စမ်းသုံးကြည့်ပေမဲ့
အရမ်း user-friendly
မဖြစ်ဘူးလို့
ခံစားရတယ်
မင်းတလည်း
စမ်းသုံးကြည့်ပြီး
လွှဲလာကပါ

5. Optimizing Myanmar Keyboard Layouts

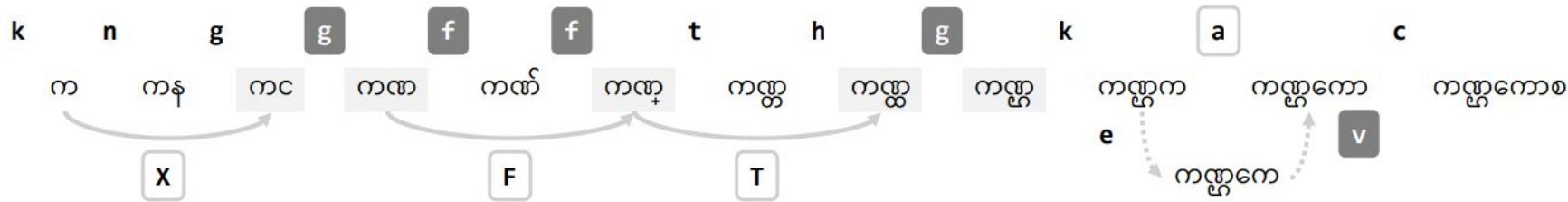


Figure 3: Example of the proposed input method. The top row is the typed Latin letters; the inputted Burmese string after each keystroke is presented in an increasing manner. Latin letters with frame are the shortcuts and those with dark background are special design that should be remembered by users. Burmese strings with gray background have a character alternation from their previous status. Although g and h are regarded as alternation operators, they are also part of the Romanization, i.e., the first g after n and h after t. The shortcuts mainly save the extra alternation by g, h and double keystroke (i.e., X, T, and F). Lowercase a is a shortcut for an extremely common character combination that can be inputted using ev.

- သုတေသန အနေနဲ့သွားမှုမိုလို ရိပြီးသား မြန်မာစာ စာရိုက်နည်းတွေကို
literature review အနေနဲ့လွှဲလာထားကြရလိမ့်မယ်

5. Optimizing Myanmar Keyboard Layouts



- Dvorak Layout ကို လွှဲလာပါ
- မြန်မာစာအတွက် Frequency based keyboard layout ကို proposal တင်ကြရင်ကော

5. Optimizing Myanmar Keyboard Layouts

- LU Lab မှာ ရှိတဲ့ monolingual corpus ကနေ frequency distribution ကို ထုတ်ကြည့်လို ရလိမ့်မယ်
- Field အနေနဲ့က HCI + NLP ဖြစ်သွားလိမ့်မယ်
- ဖြစ်နိုင်ရင် web interface, mobile phone keyboard prototype ထုတ်ချင်တယ်
- User study လုပ်ရလိမ့်မယ် အနည်းဆုံး user ၁၀ ယောက်လောက်နဲ့
- Baseline ကတော့ လက်ရှိ Myanmar3, PyiDaungSu keyboard layout
- Text input, Keyboard layout, UI/UX စိတ်ဝင်စားတဲ့ သူတွေ Group တစ်ခုပဲ့ပြီး ကြံးစားလုပ်ကြရင် သုံးလာအတွင်း ပြီးနှင့်တယ်၊ ရလဒေတော့ အနည်းဆုံး ထွက်လိမ့်မယ်

6. ASR

- ASR အနေနဲ့လည်း ပြန်မှတ်အတွက် သို့မဟုတ် သူတေသနအတွက်က လုပ်စရာတွေ အများကြီးပါပဲ
- အသံဒေတာကို ပြင်တာကနေ၊ Noise cleaning၊ မောဒယ်ဆောက၊ WER တို့ငါး စသည်ဖြင့် လုပ်ရတဲ့ engineering အလုပ်တွေက အများကြီးပါပဲ
- စိတ်ဝင်စားတဲ့ သူတွေနဲ့ အဖွဲ့တစ်ဖွဲ့ ဖွဲ့ပြီး သုံးလအတွင်း ပြီးမယ့် ခေါင်းစဉ် တစ်ခုကို လုပ်ချင်တယ်
- ဒုမ္မန်းကို သတ်မှတ်ချင်တယ်

7. TTS

- TTS လည်း ASR လိပါပဲ speech processing အပိုင်းမြှုပ်နည်းတွင် overlap ဖြစ်တဲ့ engineering အလုပ်တွေ ရှိပါတယ်
- ပုံမှန်က TTS ကတော့ professional native speaker (generally female) နဲ့ အသံဖမ်းရတယ်
- သို့သော လက်ရှိမှာ multiple-speaker TTS လည်း
လုပ်ကြပါတယ်
- ဒါ Intern3_2025 အတွက်က ASR ရော TTS ရောက် ဒီမိန်း
အတူတူ သတ်မှတ်ပြီး သွားမလားလို့ စဉ်းစားနေတယ်။ အဲဒါမှ
recording လုပ်ထားတဲ့ အသံဒေတာကုံ နှစ်ဖွဲ့စလုံး ရဲလုပ်ပြီး
သံးနိုင်တာမိုလို
- ပြီးတော့ ခုချွှန်မှာ ASR ရော TTS ရောက end-to-end အနေနဲ့
သွားနေကြတာမိုလို ...
- ASR-TTS ပုံကြောက်ကို ဘယ်ဘယ် slide ဘုံးပြီး

Q&A
(မေးစာရှင်ရုံးမေးပါ)