



Fixing Burmese

Dealing with Zawgyi

Shane Carr, Jeremy Hoffman, Luke Swartz
42nd Internationalization & Unicode Conference
September 12, 2018 Santa Clara, CA

Outline

- What is Zawgyi? Why should we care?
- Google Search
 - What was broken... and how we fixed it
 - Deep dive: Zawgyi Detection & Conversion
- What's Next / A Path to Unicode?

Myanmar: Country and Script

■ July 10, 2017, 2:00 PM PDT

Myanmar

54 million people \cong Spain

676,578 km² \cong Texas

Per capita GDP \cong India/Pakistan
(4X since 2000!)

**Fastest growing internet market
in Asia** (especially since ~2012
reforms)



PHOTOGRAPHER: TAYLOR WEIDMAN/BLOOMBERG

The Unprecedented Explosion of
Smartphones in Myanmar

What is Zawgyi?

Alice writes a message on Zawgyi device:



ကျေးဇူးပြုပြီးကျွန်တော့်ကိုကူညီပါ

...and Bob can't read it on a Unicode device:



ော်းူးျ□□□ပီ
းက်ြန⊕ေတ
ာ့့့ိုကူညီပါ

WTF?



Unicode: 0x1000 - 0x109f

Legend: Burmese Shan* Mon Sanskrit and Pali S'gaw Karen Western Pwo Karen Eastern Pwo Karen Geba Karen Kayah Rumai Palaung

80 code points for Burmese language

Used by other Myanmar script languages

	0	1	2	3	4	5	6	7	8	9	A	B	C	D	E	F
U+100x	က	ခ	ဂ	ဃ	င	စ	ဆ	ဇ	ဈ	ည	ဉ	ဋ	ဌ	ဍ	ဎ	ဏ
U+101x	တ	ထ	ဒ	ဓ	န	ပ	ဖ	ဗ	ဘ	မ	ယ	ရ	လ	ဝ	သ	ဟ
U+102x	ဠ	အ	ဓ	ဆ	ဋ	ဌ	ဍ	ဎ	ဏ	ဏ	ဝေ	သြ	ါ	ာ	ာ	ါ
U+103x	။	ေ	ဲ	့	ံ	ံ	ံ	ံ	ံ	ံ	ံ	ံ	ံ	ံ	ံ	ံ
U+104x	ံ	ံ	ံ	ံ	ံ	ံ	ံ	ံ	ံ	ံ	ံ	ံ	ံ	ံ	ံ	ံ
U+105x	ံ	ံ	ံ	ံ	ံ	ံ	ံ	ံ	ံ	ံ	ံ	ံ	ံ	ံ	ံ	ံ
U+106x	ံ	ံ	ံ	ံ	ံ	ံ	ံ	ံ	ံ	ံ	ံ	ံ	ံ	ံ	ံ	ံ
U+107x	ံ	ံ	ံ	ံ	ံ	ံ	ံ	ံ	ံ	ံ	ံ	ံ	ံ	ံ	ံ	ံ
U+108x	ံ	ံ	ံ	ံ	ံ	ံ	ံ	ံ	ံ	ံ	ံ	ံ	ံ	ံ	ံ	ံ
U+109x	ံ	ံ	ံ	ံ	ံ	ံ	ံ	ံ	ံ	ံ	ံ	ံ	ံ	ံ	ံ	ံ

Zawgyi-One encoding: 0x1000 - 0x109f

	0	1	2	3	4	5	6	7	8	9	A	B	C	D	E	F
U+100x	က	ခ	ဂ	ဃ	င	စ	ဆ	ဇ	ဈ	ည	ည	ဋ	ဌ	ဍ	ဎ	ဏ
U+101x	တ	ထ	ဒ	ဓ	န	ပ	ဖ	ဗ	ဘ	မ	ယ	ရ	လ	ဝ	သ	တ
U+102x	ဠ	အ	--	အ	ဦ	ဥ	ဦ	ဧ	--	ဩ	ဩ	၂	၁	၀	၀	၀
U+103x	၂	၆	'	၂	၂	--	၀	၀	၀	၀	၂	၂	၀	၂	--	--
U+104x	၀	၀	၂	၃	၄	၅	၆	၇	၈	၉	၀	၂	၃	၄	၅	၆
U+105x	--	--	--	--	--	--	--	--	--	၂	--	--	--	--	--	--
U+106x	က	ခ	ဂ	ဃ	င	စ	ဆ	ဇ	ဈ	ည	ည	ဋ	ဌ	ဍ	ဎ	ဏ
U+107x	က	တ	တ	ထ	ထ	ဒ	ဓ	န	ပ	ဖ	ဗ	ဘ	မ	ယ	ရ	လ
U+108x	၂	၂	၂	၂	၂	၀	၀	၂	၂	၂	၀	၀	၀	၀	၀	၀
U+109x	၇	၈	၉	၀	၀	၀	၀	၀	--	--					--	--

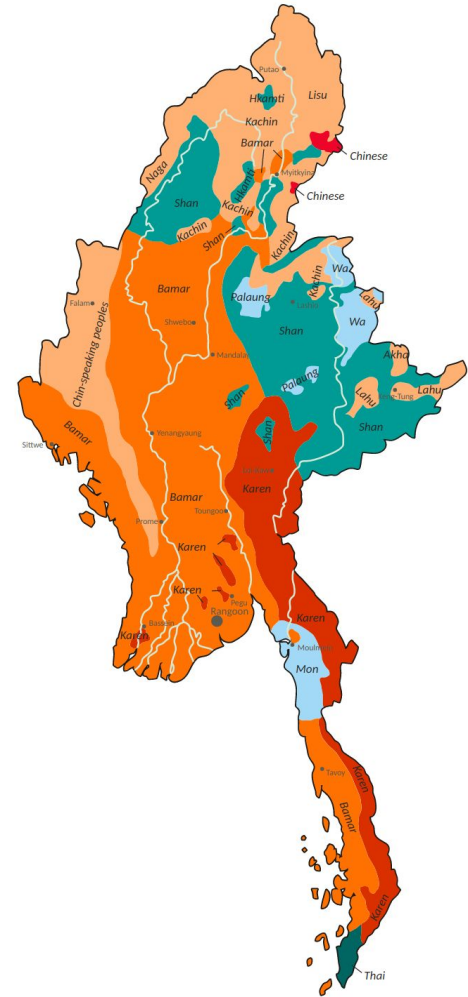
All 160 needed for Burmese language!

8 code points vs. 1 in Unicode!

Multiethnic, Multilingual Myanmar

- 68% Burmese (Bamar): 32M
- 9% Shan (Tai): 3.2M
- 7% Karen (several varieties) 2.6M
- 4% Arakanese (Rakhine): 2.4M
- 3% Chin and Kachin (Jingpho): 1.7M
- 2% Mon: 750K
- 1% Rohingya: 400K
- 6% other

...mostly using Myanmar script



Problems with Zawgyi

- Not a different encoding → **“pretends” to be Unicode**
- Co-opts Myanmar script codepoints → **cannot encode non-Burmese languages**
- Many ways to produce the ~same visual → **hard to type, search, sort, etc.**
- Large installed base/momentum → **hard for users to adopt Unicode**

Burmese & Google Search

Google Search for "မြန်မာနိုင်ငံ" (the country Myanmar) in mid 2017 (1 year ago)

Zawgyi font and query:

Unicode font and query:

မြန်မာနိုင်ငံ - Google တွေ့ရပါ

မြန်မာနိုင်ငံ

အားလုံး ဝီဒီယို ပုံရိပ်ရိပ် မြေပုံများ နောက်ထပ် အကိုးအကား ကိုရိုယာများ

ရလဒ် ၁,၂၄,၀၀၀ ခန့် (၀.၄၅ စက်ကန့်)

မြန်မာနိုင်ငံ အတုကုရာဇဝင်ပုံစံပြ

မြန်မာနိုင်ငံ အတုကုရာဇဝင်ပုံစံပြ ပုံစံကို ဖြည့်စွက်ရန် သတင်းပို့ပါ

မြန်မာနိုင်ငံ - အသိပညာပဟုသတ
aungmyintmyat408.blogspot.com/2015/10/blog-post_35.html

မြန်မာနိုင်ငံ လုံးရေသည် ၅၁.၄ သန်းကျော်သာရှိ...
www.phothutaw.com

မြန်မာနိုင်ငံ၏ စီးပွားရေးအကောင်းဘက်သို့ ...
www.dawnmanhon.com/2017/02/blog-post_902.html

UI strings are garbled!

Better image results

Text results are less reputable

မြန်မာနိုင်ငံ - Google တွေ့ရပါ

မြန်မာနိုင်ငံ

အားလုံး ဝီဒီယို ပုံရိပ်ရိပ် မြေပုံများ နောက်ထပ် အကိုးအကား ကိုရိုယာများ

ရလဒ် ၂,၇၃၀,၀၀၀ ခန့် (၀.၁၈ စက်ကန့်)

မြန်မာနိုင်ငံ - ဝီကီပီးဒီးယား
https://my.wikipedia.org/wiki/မြန်မာနိုင်ငံ

Category:မြန်မာနိုင်ငံ - ဝီကီပီးဒီးယား
https://my.wikipedia.org/wiki/Category:မြန်မာနိုင်ငံ

မြန်မာနိုင်ငံ အတွက်ရုပ်ပုံများ

အိန္ဒိယစစ်တပ် အကြီးအကဲ မြန်မာနိုင်ငံ ငှ...
burmese.voanews.com/a/3875188.html

Number of results: Unicode: 3 million Zawgyi: 100 thousand

Wikipedia is first result

Less relevant images

Goal

Google's mission: to organize the world's information and make it universally accessible and useful.

Unicode and Zawgyi content should be easy to find and read, no matter if the user has Unicode fonts, Zawgyi fonts, or both.

What was broken in Google Search

1. Indexing documents

2. Search queries

3. Search Engine Results Page (SERP)

What was broken in Google Search... and how we fixed it

1. Indexing documents

We detected Zawgyi with regular expressions. Poor precision/recall.

Text outside the Burmese Unicode range was incorrectly considered Zawgyi.
Documents in Mon or Shan would be mangled!

Fix: new probabilistic Zawgyi detector/converter.

2. Search queries

3. Search Engine Results Page (SERP)

What was broken in Google Search... and how we fixed it

1. Indexing documents

2. Search queries

Mostly in Zawgyi... which doesn't match the Unicode we index.

Fix: apply Zawgyi detector/converter, with hedge for uncertainty.

3. Search Engine Results Page (SERP)

What was broken in Google Search... and how we fixed it

1. Indexing documents

2. Search queries

3. Search Engine Results Page (SERP)

Entire SERP is in Unicode: UI elements, snippets of search results.

On devices with Zawgyi font, Unicode strings are garbled.

Fix: deploy Web Font on SERP so that Unicode appears correctly.

Live demo of Burmese Google search today

Zawgyi query for "Myanmar"

Unicode query for "Myanmar"

Deep Dive: Zawgyi Detection/Conversion

How to test if a string is Zawgyi or Unicode?

It's a hard problem: There is no perfect signal for detecting Unicode vs Zawgyi. Requires probabilistic models.

Option 1: Heuristics

- For example, handwritten regular expression rules
- Tends to overpredict Zawgyi, especially with Shan/Mon/Karen
- Does not work well with mixed content (yes/no classification only)

Option 2: Machine Learning

- Uses sample data and teaches an algorithm to predict

Machine Learning to Detect Zawgyi

Models evaluated:

Our Pick

1. Markov chain

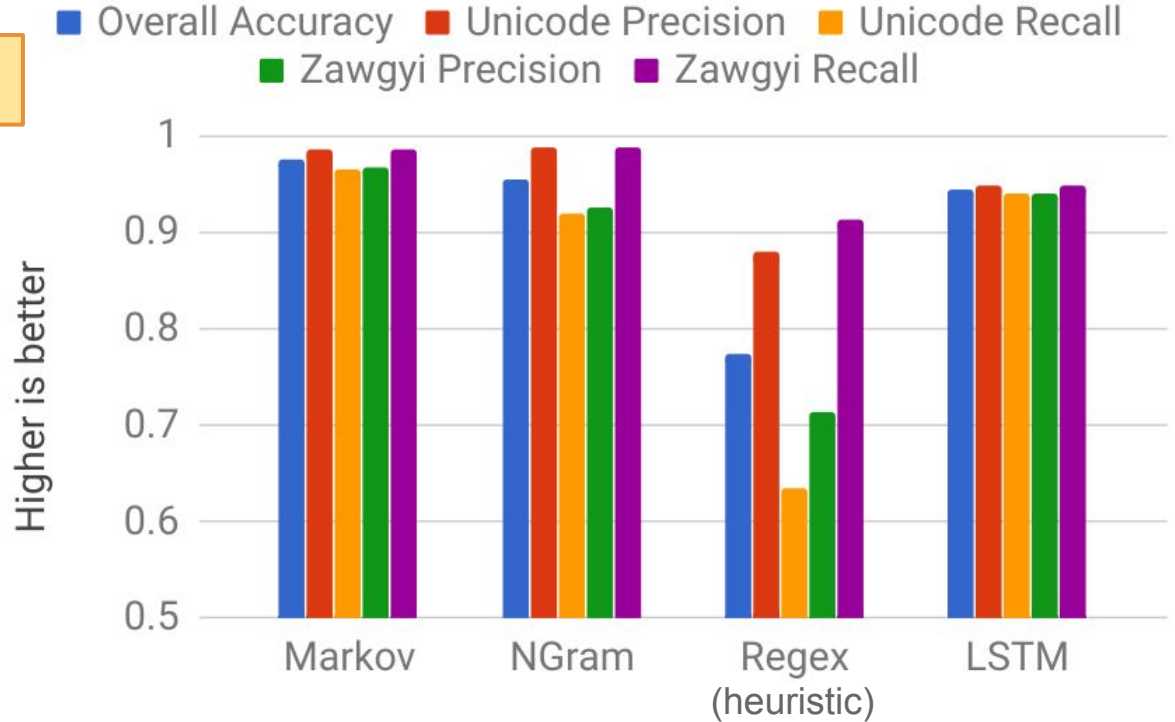
- Transition probabilities

2. NGram (2-gram)

- Likelihood of code point pairs

3. LSTM

- Neural network

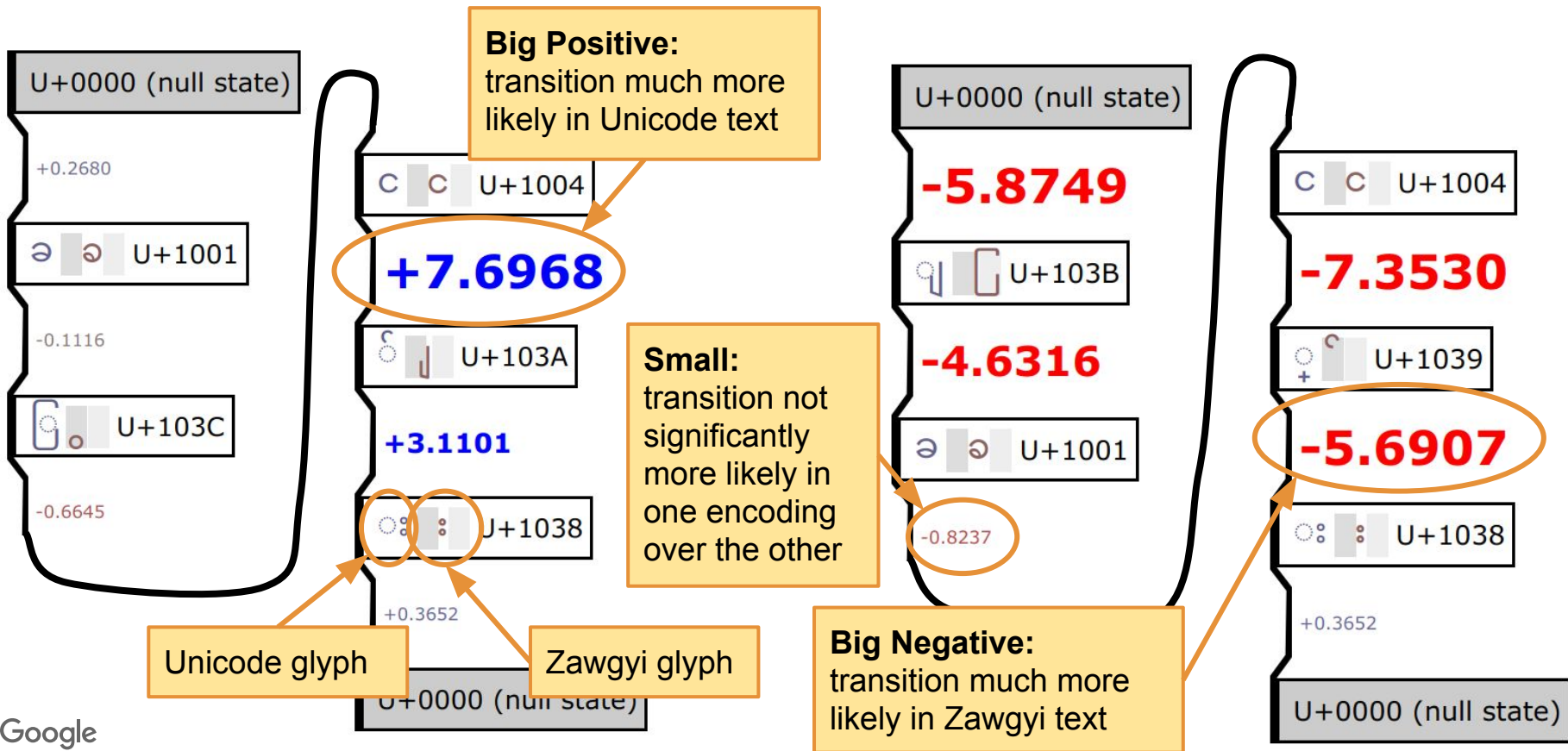




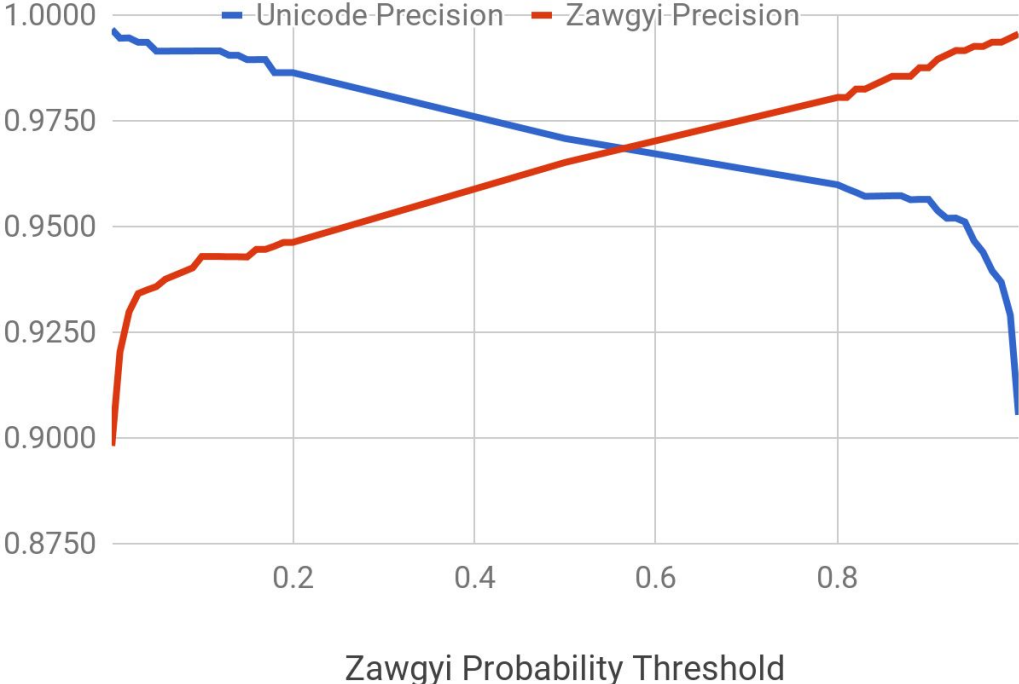
Score: +10.6639
Zawgyi Probability: 0.0000



Score: -24.0087
Zawgyi Probability: 1.0000

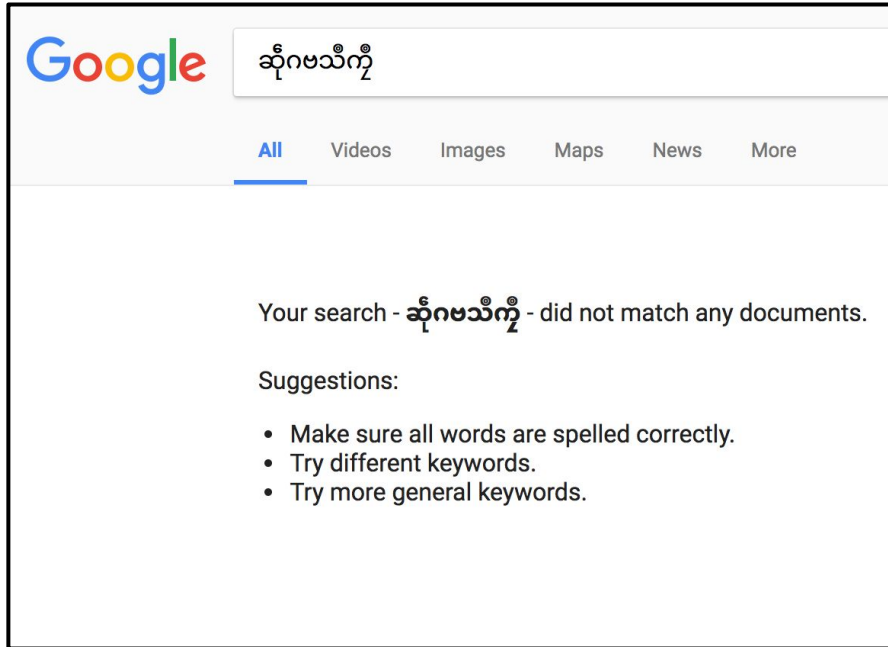


Classifier precision on Burmese queries



Mon Language Before/After

September 13, 2017



Google ဆိုဂဗသီကို

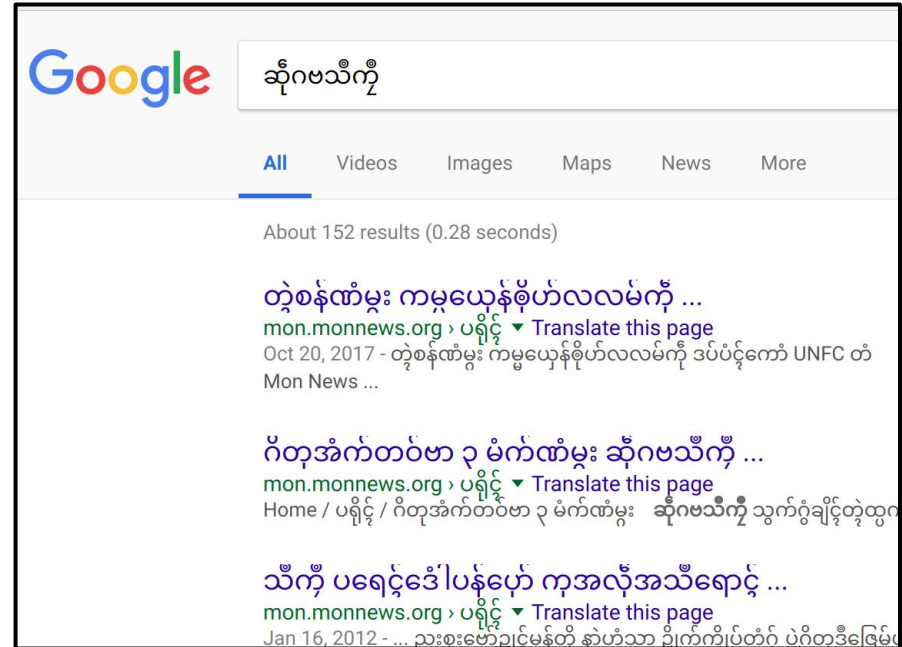
All Videos Images Maps News More

Your search - ဆိုဂဗသီကို - did not match any documents.

Suggestions:

- Make sure all words are spelled correctly.
- Try different keywords.
- Try more general keywords.

August 30, 2018



Google ဆိုဂဗသီကို

All Videos Images Maps News More

About 152 results (0.28 seconds)

တုံ့စနစ်ထံမှ: ကမ္ဘာ့ယေန်စိုဟ်လလမ်ကို ...
mon.monnews.org › ပရိုင့် Translate this page
Oct 20, 2017 - တုံ့စနစ်ထံမှ: ကမ္ဘာ့ယေန်စိုဟ်လလမ်ကို ဒပ်ပိုင့်ကောံ UNFC တံ
Mon News ...

ဂိတုအက်တဝ်ဗာ ၃ မိက်ထံမှ: ဆိုဂဗသီကို ...
mon.monnews.org › ပရိုင့် Translate this page
Home / ပရိုင့် / ဂိတုအက်တဝ်ဗာ ၃ မိက်ထံမှ: ဆိုဂဗသီကို သွက်ဝိုချိုင့်တွဲထွက်

သီကို ပရေင့်ဒေံပန်ဟော် ကုအလိုအသီရောင့် ...
mon.monnews.org › ပရိုင့် Translate this page
Jan 16, 2012 - ... ညးစးဗော့ၤၤမိန်တံ နာ်ဟံသာ ဝါက်ကိပ်တံဂ် ပံဂိတုဒီၤခြေမ်

Conversion, too!

Z → U and U → Z

CLDR transliteration rules. Why?

1. Standardized syntax
2. One source of truth: any client of transliteration uses the same rule file
3. Ecosystem: ICU4C+ICU4J

Also available as an auto-generated standalone script in Java and JavaScript (no ICU dependency)

```
82
83 # hatoh
84 [\u103D|\u1087] → \u103E ;
85 \u1088 → \u103E \u102F ;
86 \u1089 → \u103E \u1030 ;
87
88 # Vowels
89 \u1033 → \u102F ;
90 \u1034 → \u1030 ;
91
92 # asat
93 \u1039 → \u103A ;
94
95 # lower dot
96 [\u1094\u1095] → \u1037 ;
97
98 # Special cases for 1025 vs 1009;
99 \u1025 \u1039 → \u1009 \u103a;
00 \u1025 \u1061 → \u1009 \u1039 \u1001;
01 \u1025 \u1062 → \u1009 \u1039 \u1002;
02 \u1025 \u1065 → \u1009 \u1039 \u1005;
03 \u1025 \u1068 → \u1009 \u1039 \u1007;
```


Our library is open-source!

github.com/googlei18n/myanmar-tools

Available in package repositories.

[Live Demo](#)

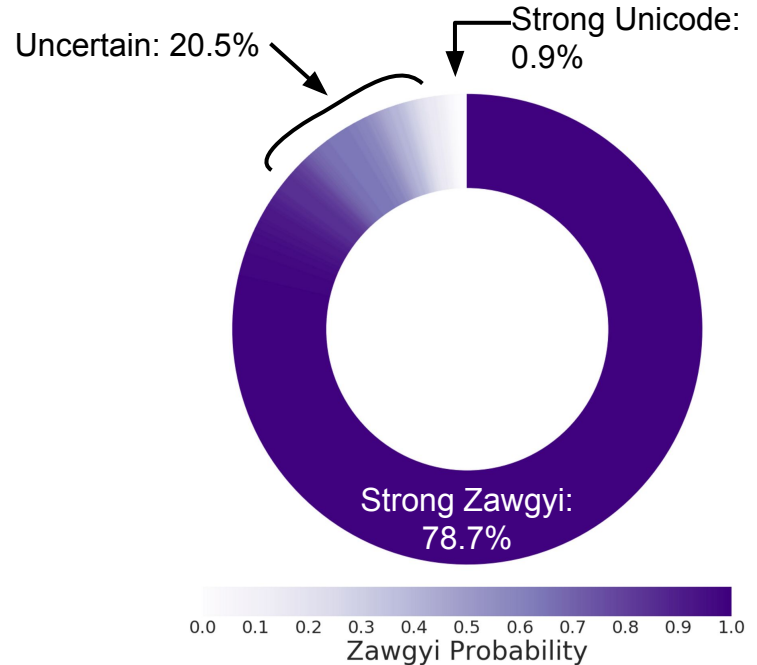
What about ambiguous queries?

Classifier is uncertain for some queries.

Fortunately, most ambiguous queries are unchanged when we run our Zawgyi-to-Unicode converter!

For the remaining queries (3%), we can try both the unconverted and converted version of the query.

Incoming Google Search Queries by Probability of Zawgyi Encoding, as of mid 2017



What's Next?

Work on an app that shows user-generated content?

- **Detect & Convert Zawgyi**

- Open source tools: github.com/googlei18n/myanmar-tools
- At indexing and query/input time

- **Use Web Font for UI messages**

- Noto Sans Myanmar is free/open source google.com/get/noto
- Noto Sans Myanmar UI Webfont at fonts.google.com/earlyaccess

A path to full Unicode?

- Shorter term: **defragment Zawgyi support in Android**
- Longer term: **two-way communication between Zawgyi & Unicode devices**
- Long Term Challenge: **more human than technical**

Short Term: Defragment Android Zawgyi

- Currently a mess
 - various implementations, fonts, locale codes used by different OEMs
 - Unicode-compliant apps look garbled when phone in Zawgyi mode
- **Make any Unicode app UI “just work”** (even if phone is in “Zawgyi mode”)
- **Standardize locale codes**
 - should make it easier for apps to know what mode a user is in
 - makes it easier to switch back & forth
- **Compatibility library APIs**

Google ○ detection/conversion for content, etc.

2-way communication between Zawgyi & Unicode devices

- Most urgent long-term problem (due to network effect)
- No clear solution yet (invisible characters? detect-and-convert?)
- Would need to be cross-platform, backwards- and forwards-compatible
- Let us know if have ideas / want to contribute!
 - lswartz@google.com

Long Term Challenge: More Human than Technical

- Zawgyi has huge installed base/mindshare
- Conflation of fonts & input methods
 - need “Zawgyi style” Unicode input
- Under-appreciation of Zawgyi’s problems
 - especially effect on non-Burmese languages

မေးခွန်းများနှင့်အဖြေများ

(Questions & answers)