Keyman: High Fidelity Text Input for All Languages

Solving a real problem

All Unicode writing systems have more than one way to encode a given string. We will use Khmer as an example. These two strings render identically:

ខ្មែរ /kmae/ **'Khmer'**

Standard keyboard layouts allow users to enter either sequence

ខ	◌ុ	ម	◌ែ	រ
81	D2	98	C2	9A

→ ខ្មែរ ✅

ខ	◌ែ	◌ុ	ម	រ
81	C2	D2	98	9A

→ ខ្មែរ ❌

What's the Big Deal?

Inconsistent sequences lead to real world problems!

- Search returns **different results**
- Text processors such as text-to-speech **stumble** on unexpected sequences.
- Dictionaries **sort wrongly**
- Malicious actors can **fool users**

Many minority language users encounter these issues all the time.

What is Keyman?

Keyman is a free and open source software keyboard engine.

Users can create their own keyboard layouts. Over 1,000 languages are already supported.

Keyman automatically transforms invalid sequences into valid sequences.

Keyman runs on Android, iOS, macOS, Linux, Windows and web.

Keyman also does text prediction with lexical models

Some words have many possible encodings

13 ways to encode ស្ទីប /sɔɔp/ 'to detect'

	◌ី	◌ើ	◌ប		✅
9F	CA	BE	94		
	◌ី	◌ែ	◌ី	◌ប	❌
9F	CA	C1	B8	94	
	◌ី	◌ី	◌ែ	◌ប	❌
9F	CA	B8	C1	94	
	◌ុ	◌ើ	◌ប		❌
9F	BB	BE	94		
	◌ើ	◌ុ	◌ប		❌
9F	BE	BB	94		

	◌ុ	◌ែ	◌ី	◌ប	❌
9F	BB	C1	B8	94	
	◌ុ	◌ី	◌ែ	◌ប	❌
9F	BB	B8	C1	94	
	◌ី	◌ុ	◌ែ	◌ប	❌
9F	B8	BB	C1	94	
	◌ី	◌ែ	◌ុ	◌ប	❌
9F	B8	C1	BB	94	
	◌ែ	◌ុ	◌ី	◌ប	❌
9F	C1	BB	B8	94	

	◌ែ	◌ី	◌ុ	◌ប	❌
9F	C1	B8	BB	94	
	◌ॆ	◌ॆ	◌ॆ	◌ॆ	❌
9F	C9	C1	B8	94	
	◌ॆ	◌ॆ	◌ॆ	◌ॆ	❌
9F	C9	B8	C1	94	

Yes, only #1 is correct but all these are easily typed on a standard keyboard!

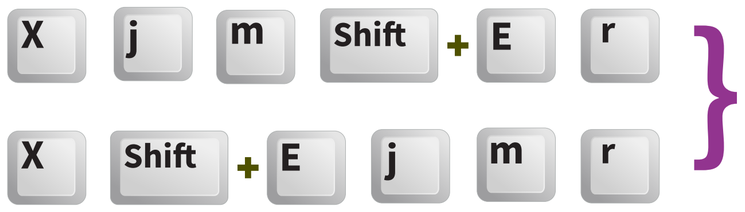
How does Keyman help solve this problem?

Keyman transforms input with rules as users type. Invalid sequences are transformed before they reach the document.

This reduces end user training requirements, improves data quality, reduces the risk of fraud, and best of all, it's completely free.

Keyman's lexical models also help with spelling consistency for less resourced language groups.

Did you know many language users employ specialized typing agencies to type text in their own language, because this is all just too hard?



ខ	◌ុ	ម	◌ែ	រ
81	D2	98	C2	9A

What about Normalization?

Normalization allows multiple sequences to be treated identically. This is great! But normalization rules are not complete and can never be extended because of **stability rules.**

Furthermore, many programs do not support normalization.

Avoiding the problem at input time simplifies the data quality problem.

Keyman
keyman.com

SIL International

References

[1] Open Forum of Cambodia. How to Type Khmer Unicode, Version 1.0:7–14, 2004.
[2] Solá, J. Issues in Khmer Unicode 4.0. Open Forum of Cambodia, Version 2.0:6-7, 2004. Retrieved from https://sourceforge.net/projects/khmer/
[3] Khmer Generation Panel. Association for computing machinery. Proposal for Khmer Script Root Zone Label Generation Rules, Version 1.5:15, 2016.
[4] Sok, Makara. Phonological Principles and Automatic Phonemic and Phonetic Transcription of Khmer Words. Master's Thesis: 35, 2016. Retrieved from: http://inter.payap.ac.th/wp-content/uploads/linguistics_students/Makaras-Thesis.pdf
[5] Zheng, Xudong. Phishing with Unicode Domains. April 14, 2017 Online: https://www.xudongz.com/blog/2017/idn-phishing/

All Khmer Unicode references are U+17##.
We've used the final byte here to reduce clutter!