



Fixing Input Methods for Abugida Scripts

Elango Cheran

Unicode Technology Workshop

Nov 7, 2023

Big Ideas

Problem for many users

- **We overlook** the problem in S/SE Asian keyboard design
- **1.5 billion people** affected
- Some users cope, many don't

Problem for companies

- **Search and AI:** depends on user-generated content
- **Chat/social:** depends on people talking to each other

Simple idea to fix

Pull apart abugidas' vowels from consonants

Agenda

Problem & Background

Solution

Implementation in Tamil

Feedback

Future Work

Problem

"Input methods"?

"Abugida scripts"?

"Input methods"?

"Abugida scripts"?

Abugida

க்

k

கா

kaa

க

ka

கெ

ke

கொ

ko

C+V base shape

க

k

கா

kaa

க

ka

கெ

ke

கொ

ko

C+V combining shapes

க

k

கா

kaa

க

ka

கெ

ke

கொ

ko

Sound vs. shape

क → k

क

ka

Current keyboards

Gboard design

Based on Unicode encoding (how you write)

Keycaps	Unicode characters (C's, V's, and combining marks)
Key press on consonant	V keys flicker to display C+V letter
Consonant conjuncts	Needs long press on C key

What users say

“ I type **Tamil using English** with transliteration. It's **easier.** ”

What users say

“ The native Tamil keyboard works...
...you type the way you write...

...**I don't like it completely, but I don't
know why.**”

What are the stakes?

Previously...

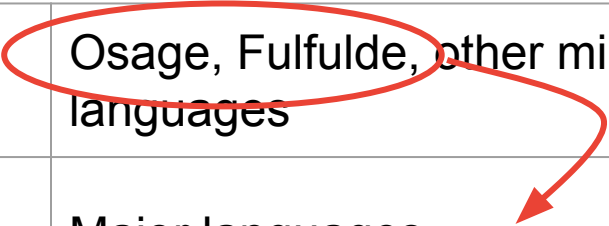
digitally disadvantaged



No keyboard	
Keyboard exists	

Previously...

No keyboard	Osage, Fulfulde, other minority languages
Keyboard exists	Major languages



No keyboard	Other minority languages
Keyboard exists	Major languages, Adlam, Osage

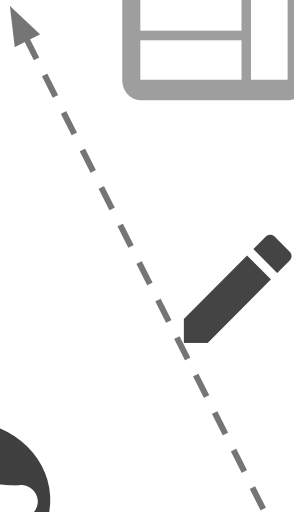
No keyboard	Other minority languages
Keyboard exists, hard to use	Abugida script languages
Keyboard is easy to use	Other major languages, Adlam, Osage

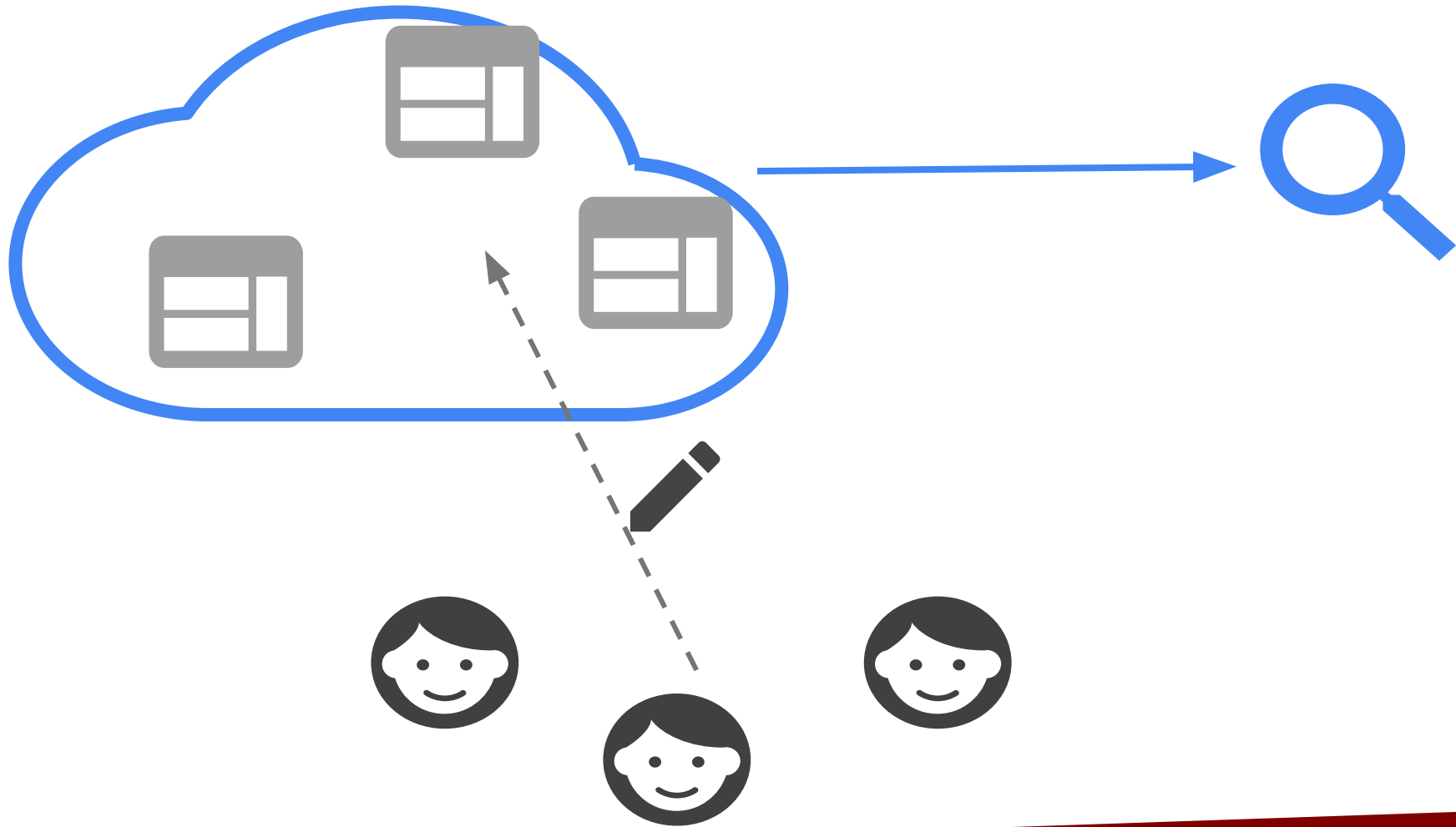
also
disadvantaged

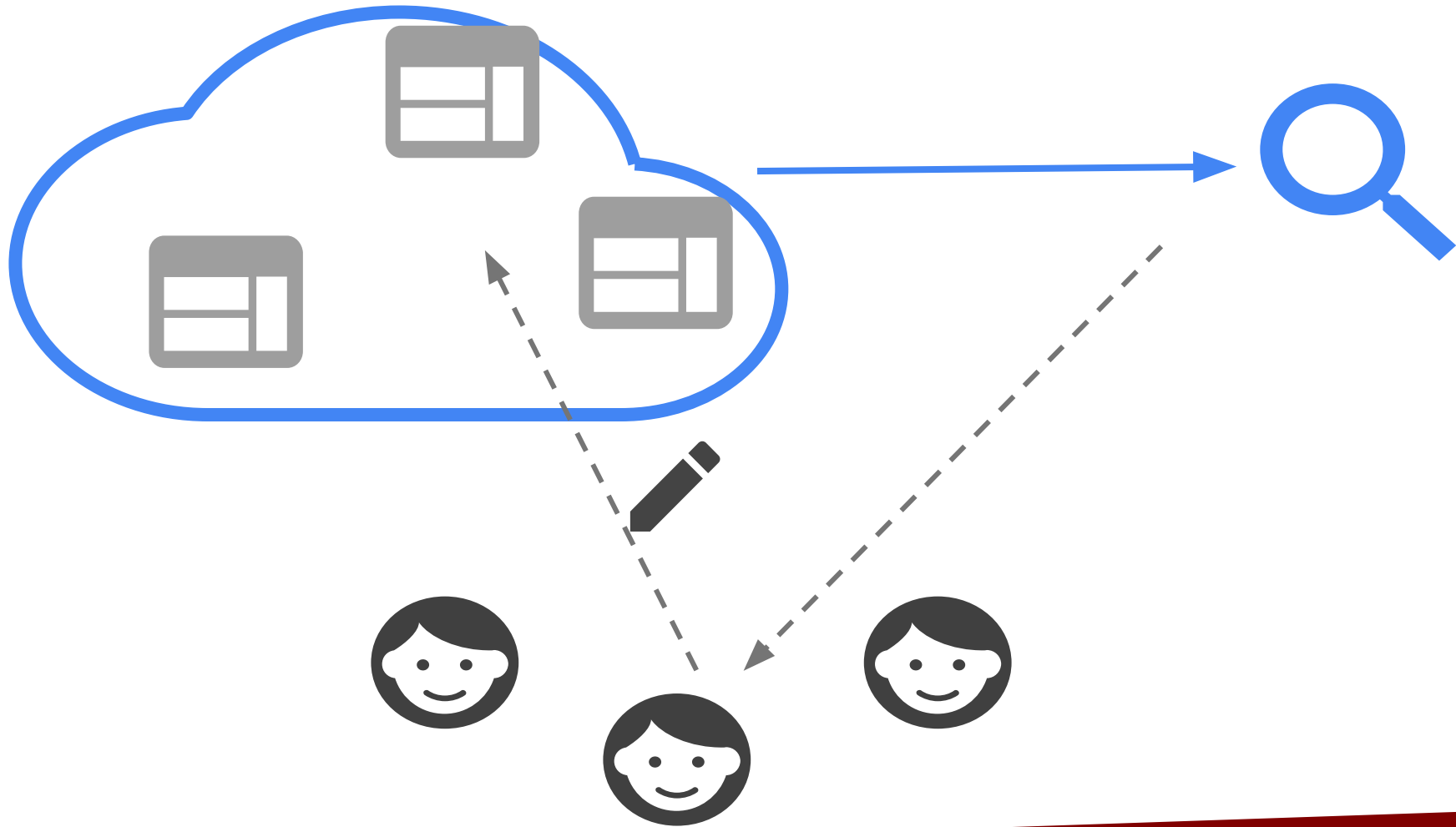


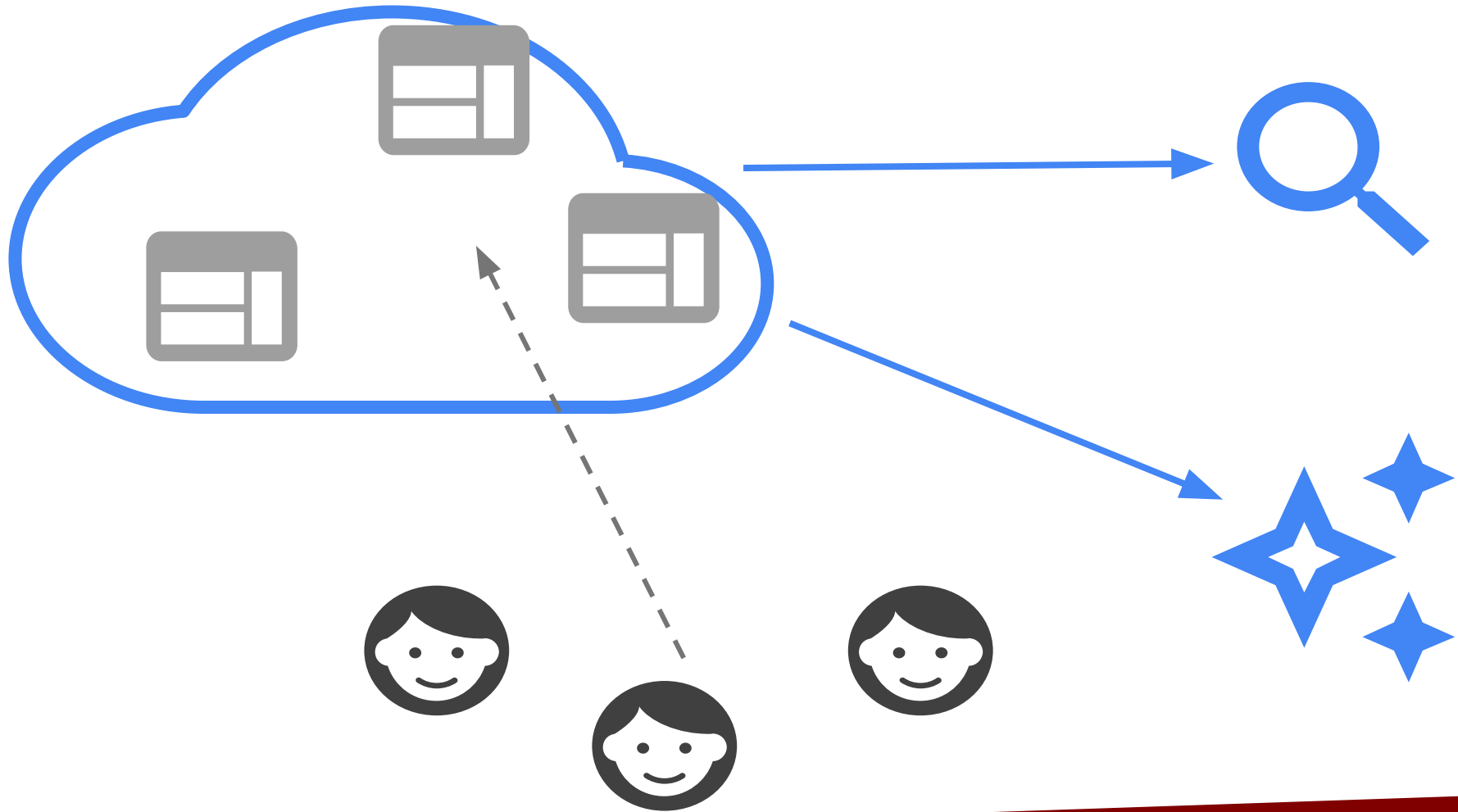


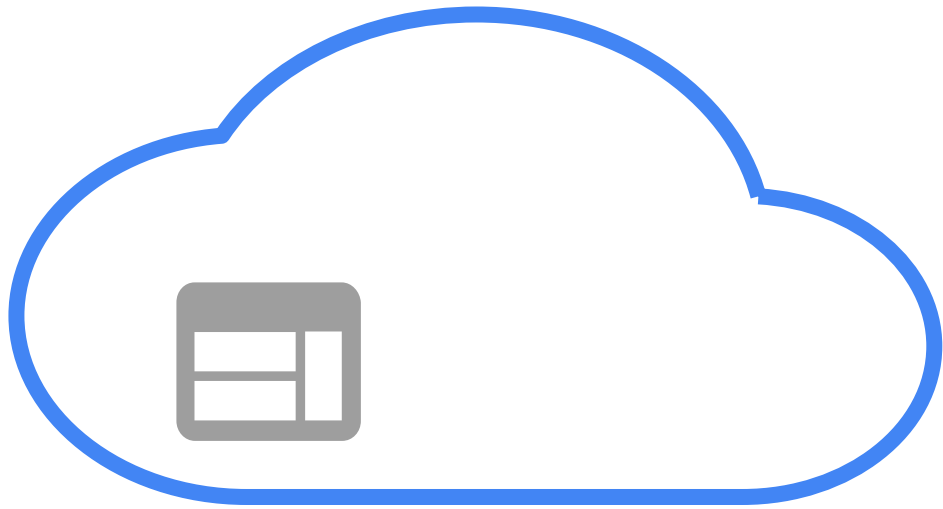
Users ↔ Content











Google and Microsoft Bet on 27-Year-Old Stanford Alum to Make AI Work For a Billion Users

Saritha Rai

Thu, 2 November 2023 at 4:00 am GMT-7 · 7-min read

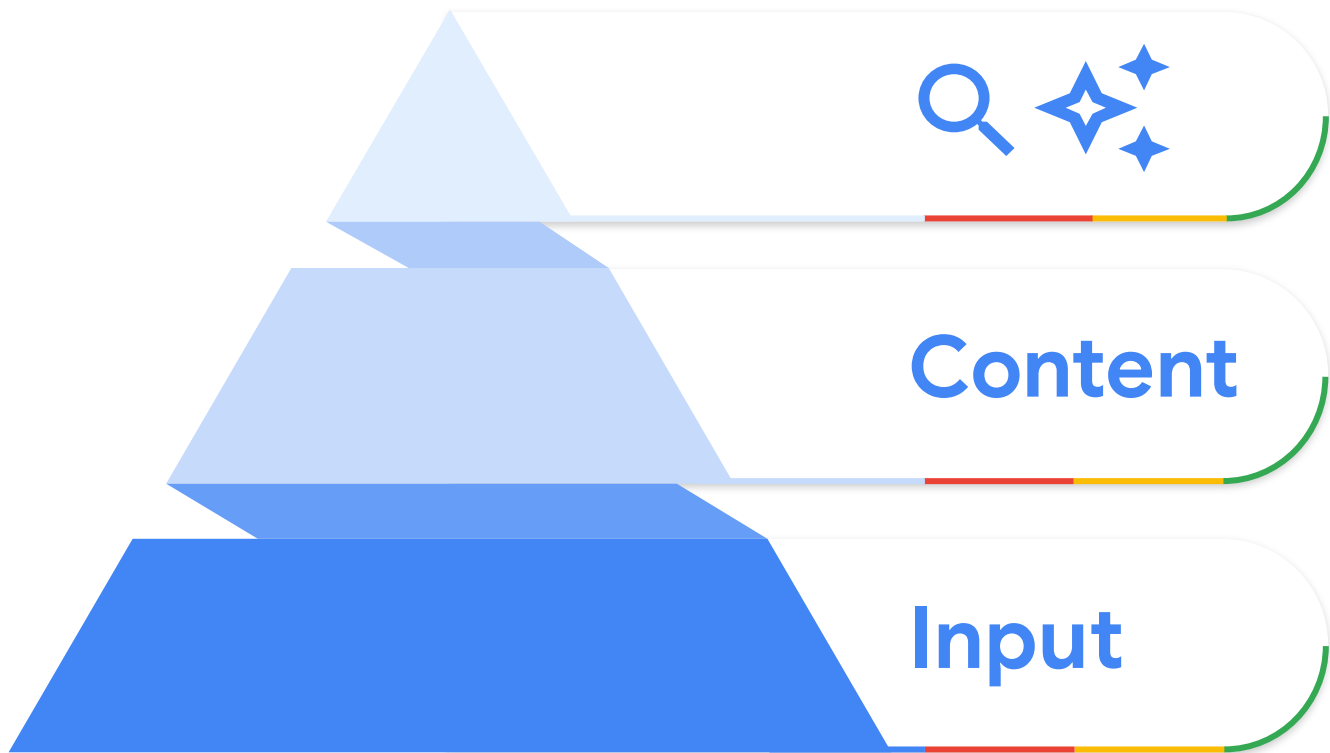
“ **Many AI services** have been disproportionately **developed with English-language internet data**, such as articles, books and social media posts. As a result, these AI models **poorly represent the diversity of languages** for internet users in other countries who are accessing AI-powered smartphones and apps faster than they’re learning English. ”

“ Many AI services have been disproportionately developed with English-language internet data, such as articles, books and social media posts. As a result, these AI models poorly represent the diversity of languages for internet **users** in other countries who **are accessing AI-powered smartphones and apps faster than they’re learning English.** ”

“ Nearly one billion such potential users live in India alone, as the government pushes for a rollout of AI tools in every sphere from healthcare to education to financial services.

”

“ When used **for South Asian languages, some large language models have been found to make up words and struggle with basic grammar.** There are also concerns these AI services may reflect a **more skewed view of other cultures.** ”



Fixing the problem

Consonants & Vowels

தமிழ் எழுத்துகள்

	அ	ஆ	இ	ஈ	உ
க்	க	கா	கி	கீ	கு
ங்	ங	நா	நி	நீ	நு
ச்	ச	சா	சி	சீ	சு
ஞ்	ஞ	ஞா	ஞி	ஞீ	ஞு
ட்	ட	டா	டி	டீ	டு
ண்	ண	ணா	ணி	ணீ	ணு
த்	த	தா	தி	தீ	து

க் + இ = கி

k + i = ki

Native speakers	Unicode
sounds	code points
how you speak	how you write
vowels, consonants	base consonant, combining mark, independent vowel
letter = consonant + vowel	grapheme cluster = base consonant + combining mark

Native speakers	Unicode
sounds	code points
how you speak	how you write
vowels, consonants	base consonant, combining mark, independent vowel
letter = consonant + vowel	grapheme cluster = base consonant + combining mark

Gboard



Phonemes = C & V sounds

Phonemes
have their
own keys

வணக்கம்

வ்	அ	ண்	அ	க்	க்	அ	ம்
v	a	n	a	k	k	a	m

What users say

“ This is a good
improvement. Keep it up. ”

creator of the iOS & macOS Tamil keyboards

What users say

“ This is faster than Gboard. I only have to know the vowel letters, and I don't have to remember how to write all the ways that the vowel signs are written on consonants. ”

What users say

“ This is faster than the iOS keyboard. I always type the key for the pulli sign because the iOS keyboard doesn't always add the pulli where it needs to be, which is confusing. ”

What users say

“This is how I expected that the Tamil keyboard would work when I bought my first phone! When I saw how the keyboard actually worked, **I was disappointed. There was no other option, so I started using what exists.**”

Backspace

Expectation: it is like an "undo" button

Backspace
on
phonemes

Current backspace problems

- **Usually:** removes code points
- **Sometimes:** removes letters (grapheme clusters)
- **Or both:** remove code points (backwards), remove grapheme clusters (forwards)
- **Overall:** not consistent or standardized

Backspace
on Gboard

Future work

Should work well

- Southern Indic
 - Tamil
 - Malayalam
 - Telugu
 - Kannada
 - Sinhala
- I plan to work on Malayalam next

Needs research

- North Indic - Devanagari (Hindi), Bangla, Marathi, etc.
 - Schwa deletion 🙄
 - Consonant conjuncts can still benefit
- SE Asian languages - Thai, Khmer, etc.
 - Tones in languages
 - How regular is the phonology?

New keyboard makers & keyboard maker news

- Translation Commons - keyboard initiative
 - Making keyboards for languages without
- CLDR Subcommittee on Keyboards
 - Spec coming out in a few months
- Keyman by SIL will support CLDR Keyboard spec
 - Will make it easy to define transform rules

Recap

Find me on the interwebs

<https://github.com/echeran/keyboards>

`elango@unicode.org`

Problem for many users

- **We overlook** the problem
- **1.5 billion people** affected
- Some users cope, many don't

Problem for companies

- **Search and AI:** depends on user-generated content
- **Chat/social:** depends on people talking to each other

Simple idea to fix

Pull apart abugidas' vowels from consonants

Problem for many users

- **We overlook** the problem
- **1.5 billion people** affected
- Some users cope, many don't

Educate and
advocate

Problem for companies

- **Search and AI:** depends on user-generated content
- **Chat/social:** depends on people talking to each other

Realize
needs & role

Simple idea to fix

Pull apart abugidas' vowels from consonants

Create
keyboards