Supporting 1,000+ Languages?

Language Technology at Scale

Craig Cornelius, Luke Swartz, Daan van Esch
42nd Internationalization & Unicode Conference
September 12, 2018 Santa Clara, CA

Our Goals Today

- **Many millions** of people around the world speak a **native language** that is **not well-supported** by information technology products

- Due to **resource constraints**, companies have always had to make **hard decisions** on **which languages to support** at what levels

- We'll discuss our work towards **scaling** support to a **large number of languages**

- Our goal: **stimulate discussion** on how everyone here can help large and small languages alike **cross the digital divide**

The World's Linguistic Diversity

- 7B+ people in the world speak **7,000+ languages**
  - **~1,300** with **>100K speakers** (according to *Ethnologue*)

- Yes, many languages are endangered...but lots are **very much alive and kicking**
  - Hundreds of languages have a **Wikipedia**, many more used on **social media platforms**
  - Yet more languages **used everyday, but primarily in spoken form** → voice technology?

- **Technology** can help languages continue to **thrive**
  - Making it **easier to type** in Santali (7M speakers, India) helped volunteers **create** a Wikipedia
  - ~50% of web content in English → **machine translation** can help break down language barriers
  - Academic collaboration: machine learning for **speech recognition** in **Indigenous languages**
  - Seeing languages **cross the digital divide** makes many communities proud

Trends

- "Next Billion Users" coming online
  - ITU/UNESCO report *State of Broadband: Broadband catalyzing sustainable development*
  - End of **2016**: **3.2 billion mobile** subscribers with broadband internet access
  - Forecast: another **additional 2.6 billion** subscribers by **2022**
  - An average of **~1.1M new** mobile broadband users **every day for six years**
- Most new users are concentrated in areas with **high linguistic diversity**
  - Many will be **more comfortable in native language** than any second language
  - Google/KPMG report: **90%** of users coming online in **India** in the **next 5 years** will **be Indian-language users**, typically won't speak English
- Users will expect technology to **support their language**
  - Increasingly common product feedback from Google users, e.g. in Gboard

Which Languages Exactly to Support?

- Depends on your product use case and target markets

- Does your product rely on language in its **spoken** or **written** form?
  - If you're expecting people to speak to your product, expect more languages
  - “Arabic” and “Chinese” are more than just “ar” and “zh”

- Would it be acceptable/normal for a user to use a **second language**?
  - For example, many users in Africa prefer to search in English/French…
  - ...even if they speak a different language at home → domain differences impact usage!

- **User Interface** (UI) localization vs. **Content**
  - A word processing product may want to let users create content in any languages they like, even if the UI is not localized → still requires fonts, rendering, line-breaking etc.

Language Technology at Scale

Overall Philosophy

- Build in **i18n from the start**
- Create **forward-looking roadmaps** with **market analyses**
- Use repeatable **processes** across languages
- **Drive down investment** needed per language
  - Create **push-button automatic** infrastructure, **reuse** resources, **R&D** for **low-resource** scenarios
- Build dashboards to track everything, **automatically**

Fonts & Rendering across the Scripts of the World

- Encoding: Unicode
- Fonts: Noto google.com/get/noto
- Text Rendering: HarfBuzz harfbuzz.org

Google

Enabling Input in the World's Languages

- Keyboards: Gboard, the Google Keyboard
    - **450+ language varieties** supported on Android today (up from ~220 at IUC 41)
    - We think it makes **sense to go to 1,000+** eventually
    - But we'll also have to rely on **crowdsourcing** and **user-contributed dictionaries**

- Speech Processing
    - **Tremendous growth** in voice usage in markets like India
    - Many users are now "**voice-first**", strongly prefer speaking to typing
    - **119 varieties** supported by Google's Speech Recognition systems today

- Handwriting
    - **Going strong** in a number of languages with **complex scripts**, like Chinese or Malayalam

Finding Training Data

- Many online text resources available across **thousands of languages**

- **Web crawls** let you find **more data**, using **language identification** tools trained on labelled text

Open Resource	Type of data	# of distinct language varieties
Tatoeba	Sentences	313
Wikipedia	Sentences	570
UDHR	Sentences	549
Bibles.org	Sentences	923
JW.org	Sentences	882
An Crúbadán	Wordlists	2,500
Unilex	Wordlists	998
PanLex	Wordlists	5,700

Details: Manasa Prasad, Theresa Breiner and Daan van Esch, "Mining Training Data for Language Modeling across the World's Languages", in *Proceedings of the 6th International Workshop on Spoken Language Technologies for Under-resourced Languages* (Gurgaon, India, 2018)

Unilex: Unicode Lexicon for all Languages

- Unicode Consortium project
  - github.com/unicode-org/unilex

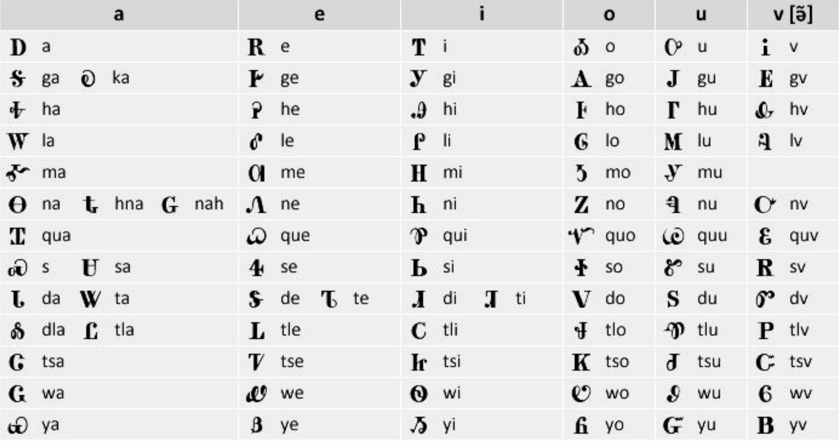
- **Open-source** data for 1,000+ languages on
  - word frequency
  - (some) pronunciation
  - (some) hyphenation
  - ...and more

- Data mostly from open-source **web crawler**:
  - github.com/googlei18n/corpuscrawler

Now for some in-depth case studies...

Study #1: ᏣᎳᎩ, Tsa La Gi, Cherokee

- Cherokee syllabary by Sequoyah in 1820s
  - Adopted by Cherokee Nation, literacy soars
- First printing press in 1828 (Georgia) - standard font
  - Cherokee Phoenix published starting 1828, re-established in Oklahoma
- Typewriters and Selectric type ball
- Encoded fonts developed
- Unicode 3.0 in 2000 (84 code points)



a	e	i	o	u	v [ə̃]
Ꭰ a	Ꭱ e	Ꭲ i	Ꭳ o	Ꭴ u	Ꭵ v
Ꭶ ga Ꭷ ka	Ꭸ ge	Ꭹ gi	Ꭺ go	Ꭻ gu	Ꭼ gv
Ꭽ ha	Ꭾ he	Ꭿ hi	Ꮀ ho	Ꮁ hu	Ꮂ hv
Ꮃ la	Ꮄ le	Ꮅ li	Ꮆ lo	Ꮇ lu	Ꮈ lv
Ꮉ ma	Ꮊ me	Ꮋ mi	Ꮌ mo	Ꮍ mu	
Ꮎ na Ꮏ hna Ꮐ nah	Ꮑ ne	Ꮒ ni	Ꮓ no	Ꮔ nu	Ꮕ nv
Ꮖ qua	Ꮗ que	Ꮘ qui	Ꮙ quo	Ꮚ quu	Ꮛ quv
Ꮜ s Ꮝ sa	Ꮞ se	Ꮟ si	Ꮠ so	Ꮡ su	Ꮢ sv
Ꮣ da Ꮤ ta	Ꮥ de Ꮦ te	Ꮧ di Ꮨ ti	Ꮩ do	Ꮪ du	Ꮫ dv
Ꮬ dla Ꮭ tla	Ꮮ tle	Ꮯ tli	Ꮰ tlo	Ꮱ tlu	Ꮲ tlv
Ꮳ tsa	Ꮴ tse	Ꮵ tsi	Ꮶ tso	Ꮷ tsu	Ꮸ tsv
Ꮹ wa	Ꮺ we	Ꮻ wi	Ꮼ wo	Ꮽ wu	Ꮾ wv
Ꮿ ya	Ᏸ ye	Ᏹ yi	Ᏺ yo	Ᏻ yu	Ᏼ yv

Cherokee language: Renewal efforts

- Cherokee is “endangered”: in Oklahoma “definitely”; in North Carolina “severely”
  - Fewer than 5% of children raised with Cherokee language, fewer than 20K speakers
  - *“No one under 40 speaking Cherokee”*

- Now, Cherokee language programs at immersion schools, high school, North Eastern College (Talequah, OK), Eastern Band, other organizations

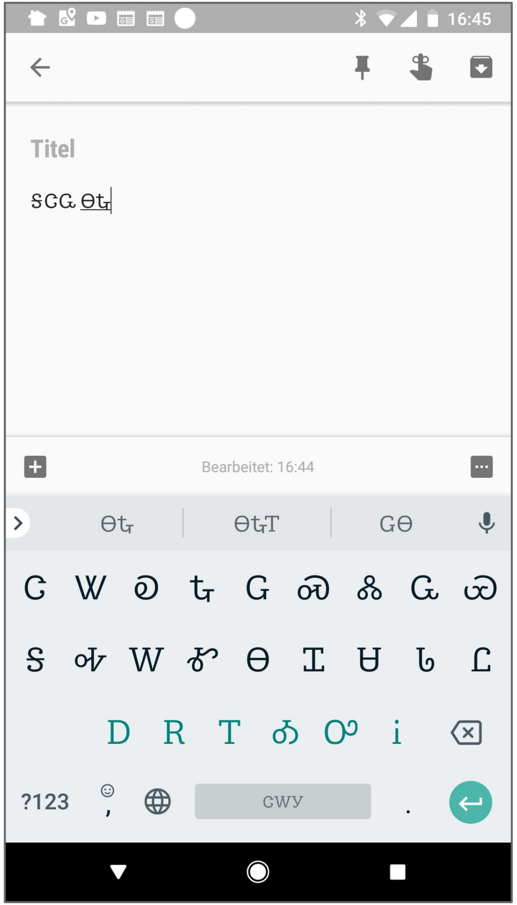
- Cherokee syllabary used in local signage and official documents

- A chance meeting in 2009 --> Cherokee Nation and Google

Cherokee and the internet

CHR: 84 characters, fonts available, keyboard layout defined

- September 2010: Cherokee keyboard and font on iOS
- March 2011: google.com/webhp?hl=chr
    - Virtual keyboards in Google Input Tools (phonetic and syllabic)
- November 2012: Gmail localized
- Noto Sans Cherokee font
    - On Chrome and Chromebooks
    - May 2015: Cherokee font on Android 5.0*
- 2015: Cherokee lower case added to Unicode 8.0
- 2018: Syllabic Android Gboard



Titel

ᏕᏣᏩ ᎾᏧ

Bearbeitet: 16:44

- Some vendors and carriers

Study #2: Pular / Fulani and Adlam script

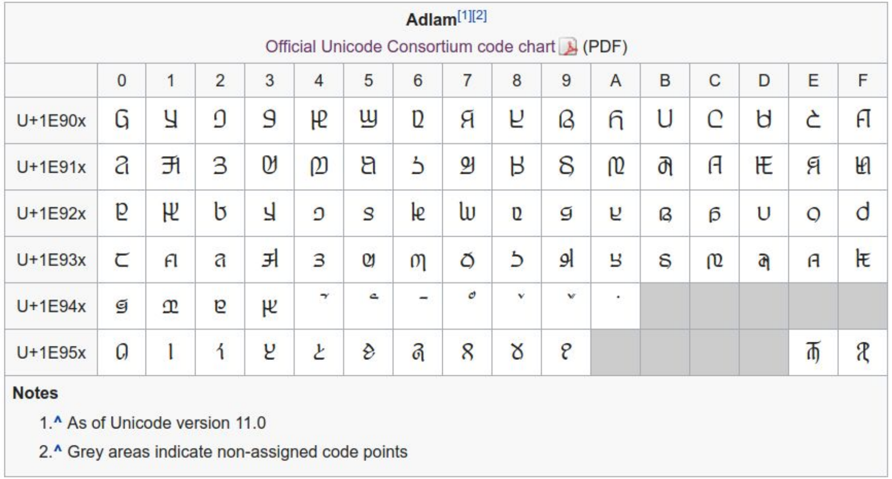
- Pular: language of Fulani people, across the Sahel (Africa)

- Adlam: alphabet invented by Ibrahima and Abdoulaye in 1989
    - Adlam is very phonetic for Pular, easy to learn, read, write. More natural that Latin or Arabic script
    - An alphabet of 28 letters, including 5 vowels

- Spreading across 20+ countries - a powerful tool for literacy
    - Up to 40 million speakers of the language - many potential users as cell phone support expands

Adlam, standards, & implementation

- 2016: Adlam added to Unicode 9.0
- Nov. 2016: Atlantic article on Adlam
- March 2017: Talks@Google by Abdoulaye and Ibrahima
- Google support today
  - Noto Sans Adlam font
  - Google Input Tools layout for Chrome
  - Gboard and Adlam font on Android 8



Adlam[1][2]

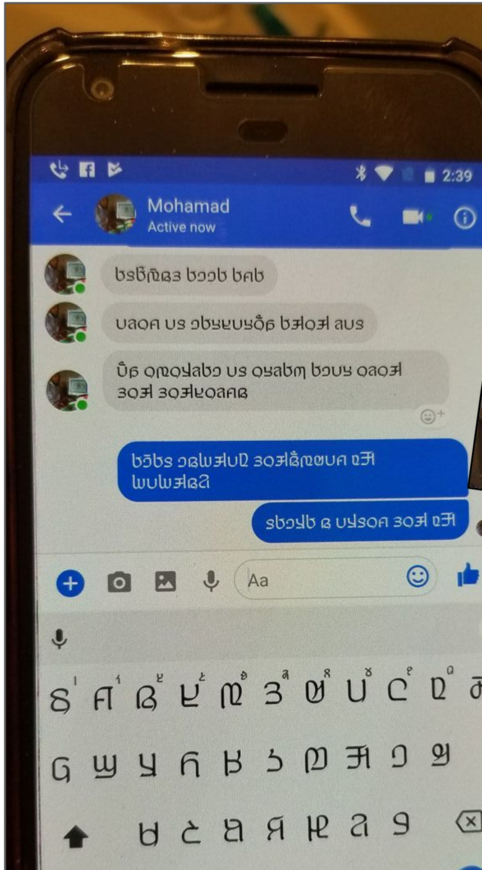
Official Unicode Consortium code chart (PDF)

	0	1	2	3	4	5	6	7	8	9	A	B	C	D	E	F
U+1E90x	𞤀	𞤁	𞤂	𞤃	𞤄	𞤅	𞤆	𞤇	𞤈	𞤉	𞤊	𞤋	𞤌	𞤍	𞤎	𞤏
U+1E91x	𞤐	𞤑	𞤒	𞤓	𞤔	𞤕	𞤖	𞤗	𞤘	𞤙	𞤚	𞤛	𞤜	𞤝	𞤞	𞤟
U+1E92x	𞤠	𞤡	𞤢	𞤣	𞤤	𞤥	𞤦	𞤧	𞤨	𞤩	𞤪	𞤫	𞤬	𞤭	𞤮	𞤯
U+1E93x	𞤰	𞤱	𞤲	𞤳	𞤴	𞤵	𞤶	𞤷	𞤸	𞤹	𞤺	𞤻	𞤼	𞤽	𞤾	𞤿
U+1E94x	𞥀	𞥁	𞥂	𞥃	𞥄	𞥅	𞥆	𞥇	𞥈	𞥉	𞥊					
U+1E95x	𞥐	𞥑	𞥒	𞥓	𞥔	𞥕	𞥖	𞥗	𞥘	𞥙					𞥞	𞥟

Notes

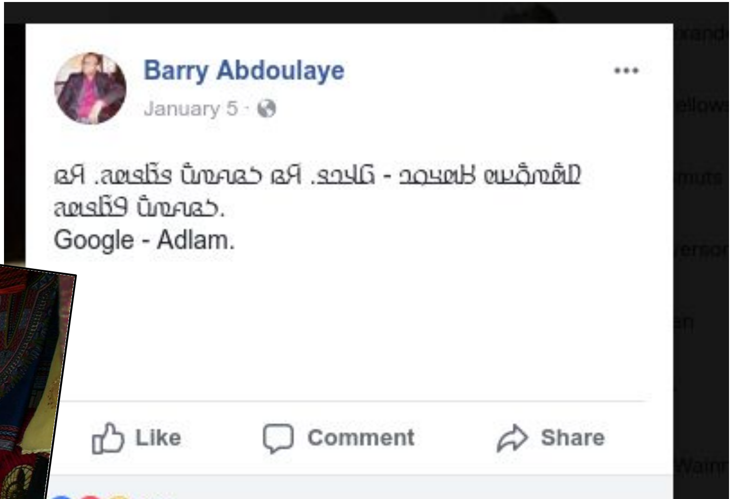
1. ^ As of Unicode version 11.0
2. ^ Grey areas indicate non-assigned code points

Adlam today



Mohamad

Active now



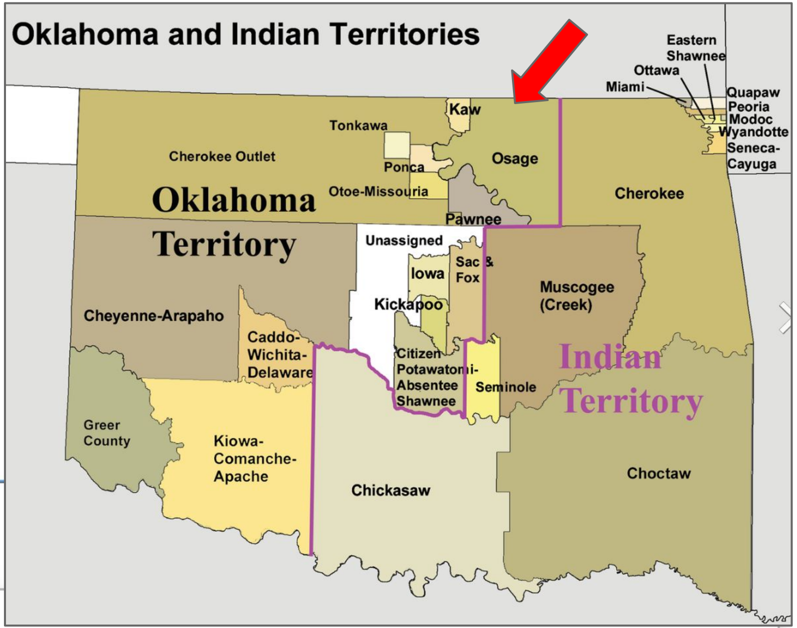
Barry Abdoulaye

January 5 ·

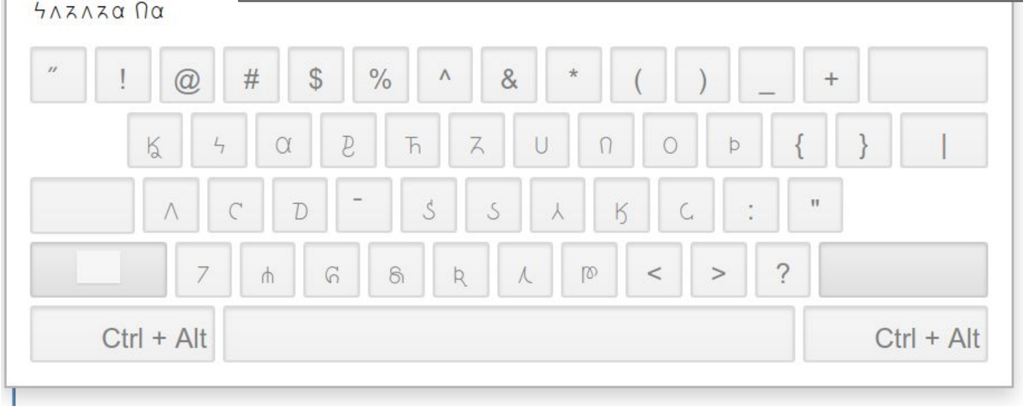
Google - Adlam.

Study #3: Osage

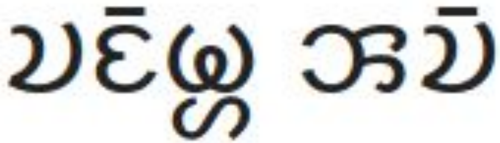
- Last native speaker died in 2012
  - About 200 second language speakers
- Osage Nation’s language programs
  - Osage in Unicode 9, 2016
  - Google Input Tools keyboard
  - Google’s Noto Osage font in Android 9.0



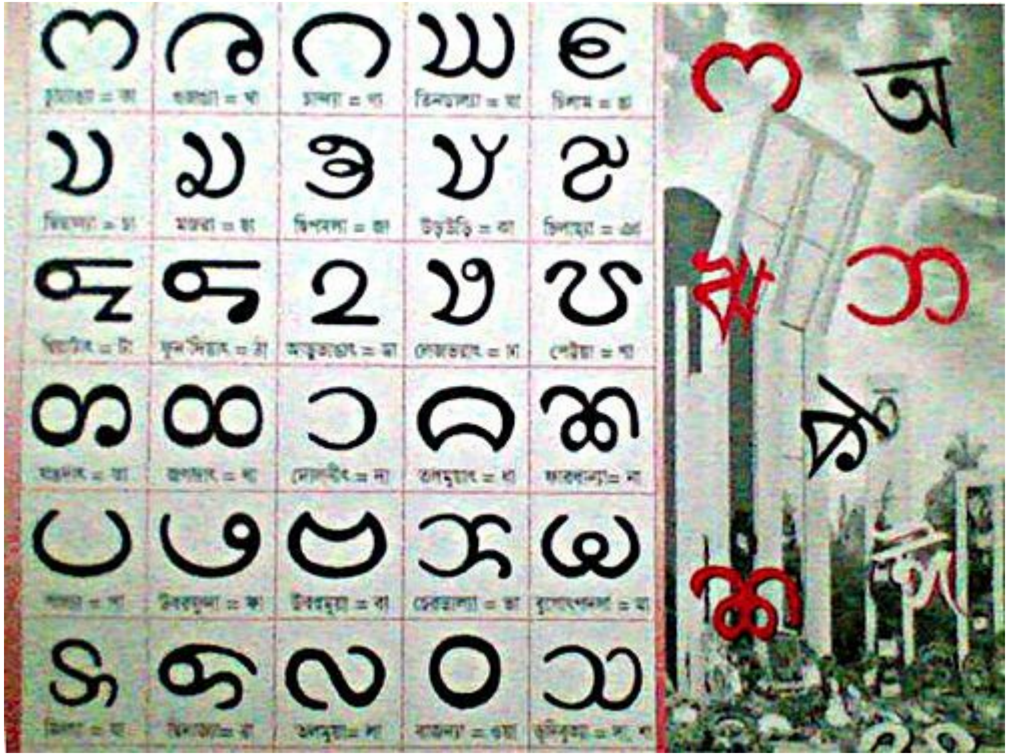
𐓏𐓘𐓻𐓘𐓻𐓟



Study #4: Chakma (Bangladesh/India)



- Spoken in India and Bangladesh
  - About 300K in Assam / Tripura
  - About 330K in eastern Bangladesh
- Ancient abugida, revived in 20th century
  - 2016: Unicode 9.0 & RigengUni font
  - 2017: Noto Sans Chakma, Google Input Tools keyboard
  - 2018: Font added to Android 9



Thank you!

"I'm in love with this innovation. I have always wanted to download a local keyboard." (Ewe speaker and Gboard beta tester, Ghana, ~6M speakers)

"I am grateful for contributing on this keyboard testing as my local language is now accessible on the Internet." (Lango speaker and beta tester, Uganda, ~1.5M speakers)

"Please add the Mizo language." (Feature request from India, ~800K speakers)