Indigenous languages and technology
Challenges, standards, & tools for small language communities

Craig Cornelius, Ph.D.
Senior Software Engineer
International Engineering (I18N)
Google, Inc.
Outline

Technology and effects on indigenous language

Case study: Cherokee

Tech support for written and unwritten languages

Example: Aikuma

What tech companies can and can’t do

Actions that can be done by indigenous speakers

Conclusions
Technology is rapidly expanding around the world!
Tech enables communication for everyone
Tech’s positive impact for indigenous peoples

+ Enables communication
+ Provides access to information
+ Promotes education & literacy
+ Grows economic opportunities
Tech’s potential negatives for indigenous languages

- Media from dominant cultures
  - Print, radio, television, video, games
- Education in dominant language
- New concepts: imported words
- Reduced perceived value
- Less young < -- > elder interaction
Tech opens the floodgates to overwhelming content

Major world languages

- English
- French
- Spanish
- Russian
- Arabic
- Hindi
- Chinese
- Indonesian
- Italian
- Portuguese
- Myanmar
- Languages of Indigenous Peoples

etc.
Case study: Cherokee language

- Began unique writing system in 1820s
- Literacy grows quickly
- Printing press and typesets in 1830s
- Newspapers, books, educational materials
- Typewriters for Cherokee
- Immersion schools, 2001
- Cherokee Nation establishes Translation Department, 2008
GWY and computer tech

- Cherokee script in Unicode (1999)
- With tech companies: interfaces, keyboards, fonts, etc.
- New for concepts, “spam” (ᏤᎣᏰ), “email” = “lightning paper” (ᏣᎳᎩ)
Translators at work
Cherokee

November 19, 2012

December 20, 2012

Microsoft announces Windows 8 support for Cherokee, a native American language

Cherokee Is Now An Official Language In iOS
What is available now from Tech for languages

Character support:
- Unicode
- Input, keyboards
- Fonts

Tools and online services
- Media: blogs, social networks, video services
- Access to the internet
- Language-specific tools and archives

Frameworks for developing content and applications
Support for written languages:

Unicode: standard for writing systems*
- Any computer, operating system, and programming language
- For mobile and web-based
- First published in 1992
- Over 120,000 characters
- 129 modern & historic scripts …
- Common data for almost 200 languages (CLDR)
Computer input in any written language

Virtual keyboards & input tools for many languages:
- Alphabetic input
- Complex writing systems
- Ideographic systems

Options include:
- Web-based & soft keyboards
- Handwriting recognition
- Phonetic and character-based
Fonts: avoiding “tofu”

The font determines actual shape of characters on the screen or printed.

Most fonts cover a few scripts, but not all.

Available for all scripts block in Unicode

Special modified fonts (“encodings”) have been used for many languages

** Some devices prevent font installation
More support for languages

- Videos and captioning
- Crowdsourcing for translation
- User interface from users
- Tweets organized by language
Crowdsourcing for user interfaces in apps

Select Your Language

Afrikaans  Euskara beta
Bahasa Indonesia  Filipino
Bahasa Melayu  Français (Canada)
Bosanski beta  Français (France)
Català  Gaeldige beta
Čeština  Galego beta
Cymraeg  한국어
Dansk  Hrvatski
Deutsch  Íslenska beta
Eesti beta  Italiano
English (Pirate) beta  Kiswahili beta
English (UK)  Latviešu beta
English (Upside Down)  Lietuvių
English (US)  Magyar
Español  Nederlands
Español (España)  日本語
Esperanto beta  Norsk (bokmål)

Norsk (ny) beta  Polski
Português (Brasil)  Português (Portugal)
Română  Русский
Shqip beta  سری‌لانکایی
een

Welcome to Translations
The Translations application by Facebook allows translator into different languages. Join our community of translators everyone, everywhere, in all languages.

Facebook in your Language
Facebook will soon be available in your language. Stay tuned for updates on what locales are supported and how you can participate in the translation and voting process.
<table>
<thead>
<tr>
<th>Language</th>
<th>Users</th>
<th>Tweets</th>
<th>Top User</th>
<th>Tweets</th>
<th>First Tweet</th>
</tr>
</thead>
<tbody>
<tr>
<td>Euskara</td>
<td>17052</td>
<td>8930528</td>
<td>berria</td>
<td>95353</td>
<td>eastigarraga</td>
</tr>
<tr>
<td>Kiswahili</td>
<td>1296</td>
<td>6359198</td>
<td>MariaSTsehai</td>
<td>104252</td>
<td>issamichuzi</td>
</tr>
<tr>
<td>Kreyòl Ayisyen</td>
<td>14270</td>
<td>5238855</td>
<td>amour109</td>
<td>80595</td>
<td>tichrist</td>
</tr>
<tr>
<td>Cymraeg</td>
<td>14249</td>
<td>4680735</td>
<td>newyyddcymraeg</td>
<td>92301</td>
<td>meigwilym</td>
</tr>
<tr>
<td>Kapampangan</td>
<td>1379</td>
<td>2157062</td>
<td>keeyttguevarra</td>
<td>21818</td>
<td>desperada</td>
</tr>
<tr>
<td>Gaeilge</td>
<td>8023</td>
<td>1208870</td>
<td>Tuigim</td>
<td>67906</td>
<td>imeall</td>
</tr>
<tr>
<td>Frysk</td>
<td>2667</td>
<td>821667</td>
<td>omropfytsban</td>
<td>83379</td>
<td>eetweetje</td>
</tr>
<tr>
<td>Setswana</td>
<td>314</td>
<td>763816</td>
<td>sesutho</td>
<td>51189</td>
<td>WameDre</td>
</tr>
<tr>
<td>Asturianu</td>
<td>771</td>
<td>479357</td>
<td>iyangc</td>
<td>30782</td>
<td>Pingarates</td>
</tr>
<tr>
<td>Hausa</td>
<td>1331</td>
<td>409232</td>
<td>bbchausa</td>
<td>37671</td>
<td>mojaam</td>
</tr>
<tr>
<td>Yorùbá</td>
<td>2239</td>
<td>278701</td>
<td>yobamoodua</td>
<td>7388</td>
<td>kojere</td>
</tr>
<tr>
<td>Ikinyarwanda</td>
<td>289</td>
<td>249671</td>
<td>TweetRwanda</td>
<td>39897</td>
<td>kwitob</td>
</tr>
<tr>
<td>Soomaaliga</td>
<td>558</td>
<td>237739</td>
<td>Weedhsan</td>
<td>17356</td>
<td>HaPpYMaXaMeD</td>
</tr>
<tr>
<td>Gàidhlig</td>
<td>1126</td>
<td>207922</td>
<td>sconewt</td>
<td>27303</td>
<td>Seumas</td>
</tr>
</tbody>
</table>
How to speak the Irish language
Online communities and blogs

[Screenshot of a Facebook group for learning Shan language]
A worldwide collaboration to strengthen endangered languages

Our catalogue contains information on 3393 languages

Explore Language Map
Find existing language content

Locate and use web pages, applications, and services available in the user’s language of choice.

A few examples:
- Find sites about educational opportunities written in the Navajo language
- Provide list of mobile apps that have a user interface in the Oriya language
- Find services in my city for native Mayan speakers
- Located social media pages in Urdu
- Find Tweets written in Choctaw

Tech challenges: how to determine the language from text.
Tech frameworks for 3rd party developers

Keyboard development tools for operating systems, mobile devices and on the web. For example:

- iOS, Android, etc.
- Windows, OS X, Linux
- Web-based HTML / Javascript
Example: keyboard tools for 3rd party developers

Custom keyboards can be created.

Word lists can be added for suggestions.
Anyone can create new content in any language

Build web pages, web & mobile apps with script/language.

- Business, non-profit, family, personal interest, etc.
- Create audio/video with spoken and written text, including closed captions
- Label these with standard language tags for discovery

```html
<html lang="fj">
  ...
</html>
```

Tech challenges: keyboards, fonts, methods to label
Identify, create, and join language communities

Tools to create and discover language-based social media, discussion groups, community pages, chat rooms.

Examples:

- Locate social networks in Myanmar minority languages
- Find other speakers of Hausa
- Set up a discussion group on Gurindji
- Edit, comment, interact with my Vanuatu language communities

Tech challenges: Typing
Supporting languages that have no written form

Video & audio can be captured and shared

Educational materials can be created

Mobile platforms enable sharing and audio communication

**Challenge:**

How to make technology usable for non-literate users?
Case study: Aikuma for language documentation

"Aikuma: a free Android app that helps people record, share, and translate stories in the world's unwritten languages."

Created by Steven Bird, University of Melbourne

http://www.aikuma.org/
Aikuma: record, replay, respeak, translate

Hold the green button to listen to a phrase

Use the red button to record a translation.
What Aikuma does:

Set language: ISO 639 or by name
Add speakers
Record audio
Respeak the recording
Add metadata
Translate
Share
Internet usage is not universal (yet)!

Internet users in 2012 as a percentage of a country's population
“Access” will happen

Google[x] Project Loon: WiFi via balloon

Facebook’s Connectivity Lab: “drones, satellites and lasers”

Cellular networks continue to grow in coverage and speed
What tech companies probably cannot do

1. Provide user interfaces for products in all 7000 human languages
2. Build 100% accurate language detection for written text in all languages
3. Add machine translation for all the world's languages
4. Build speech recognition for most languages
5. Provide text to speech for most languages
6. Support all variations of language use, such as vocabulary / grammar. e.g., Mexican vs. Nicaraguan Spanish.
Don’t expect user interfaces in every language

Many companies support 50 - 150 user interface languages in major products. However:

- Translation is expensive
- Maintenance is a continuing cost
- Many languages are not standardized
- Few committed & organized translators
- Small impact for potential users

General approach: companies work to support at least one language spoken by most people, e.g., Spanish in Central / South America, Filipino in the Philippines.
Don’t expect machine translation for every language

Why not?

Translation is much harder than having a dictionary.

Translate needs massive amounts of data:

- Millions of words in parallel text
- Many samples of common usage
Support voice on more languages?

Perhaps:
- Support for human annotation / transcription tools for media such as video and audio

Probably not:
- Detect the language from audio or video
- Reliable speech to text (voice recognition)
- Automated text to speech
Why not speech recognition for every language

Similar to machine translation, speech recognition needs massive amounts of audio training data:

- A large sample of all common words and phrases
- From multiple speakers (500 or more)

Also, accents vary widely and context is extremely important.
Language variants: Arabic as an example
Things language communities should avoid:

Don’t use a “hacked font” (font-encoding):
- Text can’t be show without a special font. Fails on mobile devices.
- Search and other text processing fails.

Don’t create new writing systems.
- In most cases, an existing Unicode script can be applied.
- Be careful with novel use of diacritics. Many fonts may not be able to render them correctly.

Don’t give up!

Google
What indigenous communities can do...

Create content of all kinds & declare the language

Use your language in all communications: text, mail, audio, video, etc.

Establish and use language communities. Engage new users!

Use video to teach. Add closed captioning.

Encourage developers of input tools, fonts, applications, etc.
Build online tools

Dictionaries, grammars, etc.

Navajo Language Resources

Our Online Navajo Language Materials

- **Navajo Vocabulary:** List of vocabulary words in the Navajo language, with comparison to English.
- **Navajo Pronunciation Guide:** How to pronounce Navajo words.
- **Navajo Animal Words:** Illustrated glossary of animal words in the Navajo language.

---

Chino nyenga chino kotama, chino simudza musoro chawana.
— Shona proverb
Form communities and tell stories!

Share stories with family, friends, and strangers in your own language.

HAVE YOU SPOKEN YOUR TREASURE LANGUAGE TODAY?

Cherokee, Tigrigna, Lu Mien, Hawaiian, Welsh, Quechua, Warlpiri, Xhosa, Yoruba, Inupiaq
Encourage developers to think beyond the big languages

Find developers to support additional languages, e.g., games.

Increase awareness of the needs for: input methods, fonts, applications, etc.

Encourage and use standards-based tools that can use language data plugins.

Reward developers who support your languages.
Advice: Be persistent!

Use the language: the key to preserving it!

Help users understand how to write using the new tools.

Be proud of the language. Advocate for its public use!
Conclusions: Technology and Indigenous Languages

Great potential for indigenous languages

Dominant languages can be overwhelming

There is general support for text, audio, video, internet access

Advanced language tools are much harder

Indigenous communities can apply tech to help preserve and extend their languages & culture
Thank you
Mobile-only Internet Users (2010/2011)

- USA: 25%
- UK: 22%
- Egypt: 70%
- Brazil: 33%
- South Africa: 57%
- Russia: 19%
- China: 38%
- Indonesia: 44%
- India: 59%
- Thailand: 32%

*no data available*
Technology can also *dis*-connect us