11-736 Seminar on Endangered Languages

Alan W Black, Robert Frederking, Laura Tomokiyo

http://www.cs.cmu.edu/~ref/sel/

Language Technologies Institute
Fall 2017
[Intro: see class announcement]
Administrivia

- Please register!
- Slides will be available on website
- Regular meeting time?
- Readings, presentations; 14 weeks, incl. today
  - For next week, read handout from “Language Death” (David Crystal)
    - Individual projects in Tech for Endangered Languages
- Everyone signed in?
What are Endangered Languages?

[From here on, REF’s version, drawn largely from “Language Death” (Crystal)]

• The size of the problem
  – How do you count languages?
    • How do you count Arabic? Flemish?
  – How do you estimate unknown languages?
  – How do you evaluate the status of a small language?
    • Not just size (Zuni)
Normal languages are odd (to us)

- Natural situation of a language is *oral*
  - No dictionary, spell check, books, or even *name*
- Natural situation is *local*
  - No writing or telephones: if you don’t live nearby, your dialect will drift
- Standard national languages are a recent thing
  - More obvious in Europe/Africa/Asia
- Other languages/dialects are always political
**Real minority languages:**

**Spanish vs. Apache:**

<table>
<thead>
<tr>
<th>Language</th>
<th>Population</th>
</tr>
</thead>
<tbody>
<tr>
<td>CHINESE, MANDARIN</td>
<td>332,000,000</td>
</tr>
<tr>
<td>SPANISH</td>
<td>885,000,000</td>
</tr>
<tr>
<td>ENGLISH</td>
<td>322,000,000</td>
</tr>
<tr>
<td>Arabic? [if all local versions included together]</td>
<td>300,000,000</td>
</tr>
<tr>
<td>HAITIAN CREOLE FRENCH</td>
<td>737,200</td>
</tr>
<tr>
<td>ADYGHE (Russia)</td>
<td>148,530</td>
</tr>
<tr>
<td>NAVAJO</td>
<td>120,000</td>
</tr>
<tr>
<td>NYORE (Kenya)</td>
<td>12,693</td>
</tr>
<tr>
<td>APACHE, WESTERN</td>
<td>6413</td>
</tr>
<tr>
<td>ZUNI</td>
<td>775</td>
</tr>
<tr>
<td>TLINGIT</td>
<td>4</td>
</tr>
<tr>
<td>AFRO-SEMINOLE CREOLE</td>
<td>0</td>
</tr>
</tbody>
</table>

(http://www.sil.org/ftp/ethnolog13/langpop99.zip)
North Park, Pittsburgh

The name, Wahdo-Gwas, is Seneca for Rising from the Water. The Seneca Nation is one of the Haudenosaunee (Iroquois) original 5 nations (today 6 nations).

The Seneca Nation (Onödow’aga) are known as the “The Great Hill People” and are one of the Nations that lived and hunted in Western Pennsylvania prior to the European’s arrival. For the most part the Seneca were allies of the British during the American Revolution. Seneca’s lived in temporary rock shelters, and other dwellings while hunting and traveling over the land and vast waterway systems in the North East. Longhouses were used in the permanent settlements. Longhouses were occupied for about 20 years before moving on to a new location. Moving the settlements on a regular cycle was because of the Seneca’s knowledge of soil depletion, game migration patterns and changes. It was understood that the land had to be given time to replenish itself for continued and future renewal for the next seven generations. Farming, hunting, fishing, and gathering of food were how life...
Why should we care?

• We need diversity
• Languages express identity
  – [REF/German]
• Languages are repositories of history
  – And Culture
• Languages contribute to sum of human knowledge
  – Ethnobiology, technology
• Languages are interesting
And it's not hopeless

- Welsh
- Maori
- Hawaiian
- Cornish (may have been extinct)
Description vs. Documentation vs. Revitalization

• Linguistic Description
• Language Documentation
  – “Sleeping” languages: Miami
• Language Revitalization
What can LT do about it?

• Build MT systems?
LingWear for the Information Warrior

New Ideas

• The pre-development of appropriate interlingua representations for domains of interest facilitates generation into a new language within two weeks.

• The development of new MT engines (e.g. learnable transfer rules) and improved multi-engine integration supports rapid deployment of MT for a new language with scarce resources.

• Gisting and summarization in the source language followed by MT is better than vice versa.

Impact

• Allow military and relief organizations to converse in limited domains of interest with the local population in an area of conflict and/or disaster

• Allow military and other operatives in the field to assimilate foreign language information they encounter on-the-move

• Rapidly port and deploy the technology into new languages with scarce resources

Schedule


• 9/2001: Port to second language complete.

• 9/2002: Port to third language complete.

Carnegie Mellon University School of Computer Science: A. Waibel, L. Levin, A. Lavie, R. Frederking
Diplomat:
Rapid-deployment, wearable, speech-to-speech translation

• Develop a new language in weeks, graceful improvement for months or years.
• Speech in, speech out.
• Use user interaction to cope with errors.
• Human factors of working with computer naïve users.
• Languages: English and {Croatian, Haitian Creole, Spanish, some Korean, (Arabic)}
AVENUE:
Low Cost MT for Minor Languages

- Speakers of electronically underrepresented languages can participate in the information age.
- Policy makers can access ideas, viewpoints, and information from the developing world.
- MT for unforeseen translation needs: e.g., humanitarian aid.
- Documentation and preservation of endangered languages.
What can LT do about it?

• Build MT systems?
  – AVENUE experience: the community didn’t want state-of-the-art MT, but project had been funded to build MT!

• Haitian Creole was difficult, but still not small enough to be endangered!

• New language technology methods that the community can use *themselves* to save their language?
What can LT do about it?

• Three of the main ways languages die:
  – children don’t learn them;  [REF/German]
  – teenagers forget them;
  – young adults forget them

• Spelling checkers, SMS, etc.

• Navajo.org [Dene] (but in English!)
  – Winnebago [Hochunk] website errors

• Etc.?
Only the beginning...