Localization in minority language contexts

Kevin Scannell
Saint Louis University
21 May 2013

Endangered Languages

- More than 7000 languages spoken today
- Almost 2500 are endangered (UNESCO)
- 50% expected to die out in the next 100 years
- 90% or more in certain areas (N.Amer, AUS)

Technology and Revitalization

- Tech often plays a role in revitalization efforts
- Language documentation
- Language learning software/resources
- Social media
- Software localization

Some perspective

“The internet and digital world cannot save us. They cannot save Indigenous languages. Of course these things have benefits but they are not the Messiah. We don't need another website or DVD or multi-media application, these are short term, quick fix solutions. What we really need is sustainable initiatives, to create opportunities for Indigenous language users to communicate with each other in their native tongue. To get people speaking again.” John Hobson, University of Sydney, 16 May 2013

http://au.artshub.com/au/news-article/news/arts/digital-not-always-the-answer-195370

Obstacles, I

- No access to computers or the Internet
- Prohibitively expensive
- Language may have no writing system at all
- Limited literacy
- L2 learners of heritage language
- Special fonts or keyboards needed

Obstacles, II

- May have learned to use computer in English
- Lack of computing terminology in the language
- Limited content in the language (chicken&egg)
- Lack of commercial interest

Menu vs. Food

- Software interfaces are great, but...
- Also need Wikipedia articles
- Blogs
- Tweets
- Cat memes
- ...

Indigenous Tweets

- http://indigenoustweets.com/
- Tracks everyone using an indigenous or minority language on Twitter
- 35 language at launch, March 2011
- 146 languages at present
- Basque with ~4M tweets from 17k users
- 33 languages with a lone tweeter

Irish language (Gaeilge)

- First official language of Rep. of Ireland
- Taught in Irish schools
- > 1M claim competency in language
- Official Irish speaking regions: Gaeltacht
- Probably less than 20,000 L1 speakers
- “Definitely endangered” according to UNESCO

An Atypical Minority Language

- Support from cross-border language body
- Long written tradition; standardized spelling
- Excellent (bilingual) dictionaries
- Terminology committee; http://focal.ie/
- Official EU language since 2007
- Many skilled professional translators

My role

- Began learning the language in the 90's
- No web, no TG4, no streaming radio
- “Social media” = email list Gaeilge-A
- In 1999-2000 I created GaelSpell
- Began translating the software I used regularly
- Linux command-line tools, vi, etc.

Free/Open Source Strategy

- Essential for all indigenous/minority languages
- Reuse translations project to project
- Sense of “community ownership”
- Software is free of cost
- Lack of interest from commercial vendors
- Always danger of “the plugged being pulled”

Goal: Fully-localized desktop

- Firefox, 100% translated since 0.8 in 2004
- OpenOffice/LibreOffice, 100% since 2005
- Kubuntu, core system 100% translated
- KDE Applications, about 70% translated
- More than 2.5 million words translated
- Entirely volunteer effort; no grants

“Accessories” too

- Spelling and grammar checker (more later)
- Online dictionaries
- Thesaurus (http://borel.slu.edu/lsg/)
- Speech synthesis (http://abair.ie/)
- Survey at http://scriobh.ie/

Buíochas

- Séamus Ó Ciardhuáin, Ciarán Ó Bréartúin, Seanán Ó Coistín, Iarla Mac Aodha Bhuí, Marion Gunn, Mícheál Ó Meachair, Panu Höglund, Peadar Ó Gúilín, Sean V. Kelley, Gabriel Beecham, Sean Burke, Pat Folan, Alastair McKinstry, Brian Ó Broin, Justin McCubbin, Ciarán Ó Súilleabháin

Marketing

- The biggest drawback of open source
- Uptake of Irish language software is very low
- Anxiety about unfamiliar technical terms
- Bureaucracy to get installed in schools
- Mostly a lack of awareness of what's available!

nascanna.com

- Created by Dúrud Teoranta
- Support from Foras na Gaeilge
- Survey of everything that's been localized
- Install instructions
- How to help with software localization

Workflow

- Virtually everything done via PO files
- Either “natively”, when apps use gettext
- Or translate-toolkit for Mozilla, OpenOffice, etc
- Or ad hoc roundtrip scripts for everything else
- Translators use their favorite PO editors

Parallel Corpus

- http://borel.slu.edu/corpas/
- 36.5M words in 1.4M aligned pairs
- Software translations, glossaries, terminology
- Government documents, literature, ...

PO Compendia

- Some parallel corpus material under copyright
- TM from subset of software translations only
- Free to download as PO or TMX
- https://sourceforge.net/projects/gaeilge/

Spelling Reform

- An Caighdeán Oifigiúil; late 1940's
- Some of the best translations from the 1930's
- An Caighdeánaitheoir
- http://github.com/kscanne/caighdean
- Treats it as a problem in statistical MT
- Used by New En-Ir and RIA dictionaries

QA Process

- Product-dependent
- Mozilla; dedicated platform testers (more later)
- Manual proofreading
- Series of automated quality checks...

Consistency checks

- Verify that identical or nearly identical source strings (ignore difference in caps, etc.) are translated the same way
- Can be applied within a product, or across a family of related products (e.g. KDE apps)
- Sometimes distinct translations are needed, in which case...

Known ambiguities

- When legitimate inconsistencies are found, they are added to a DB of known ambiguities
- Translators get a warning when one of these messages appears
- For Irish: “Directory” (comhadlann vs. eolaire), “Player” (imreoir vs. seinnteoir), “Tab” (táb vs. cluaisín), 200+ others

Deprecated terms

- Irish terminology has been a work-in-progress
- Also the biggest obstacle for new localizers
- Check all translations against big database
- Uses simple regex pattern matching

Untranslatables

- Data-driven approach
- Generates DB of tokens that appear on source and target sides in QA'd translations more than, say, 75% of the time
- Then warn localizer if one of these terms appears in a source string but not in the target

An Gramadóir

- Robust spelling and grammar checker
- Combination of statistical and rule-based
- http://borel.slu.edu/gramadoir/form.html

Terminology Development

- Visit: http://goo.gl/FE8IY
- Goes back to Gaelic-L list in early 1990's
- “software”
- “spell checker”
- “download”
- “website”
- “homepage”
- “browser”

An Coiste Téarmaíochta

- Official terms in all domains
- Disseminated via http://focal.ie/
- Quite responsive to email requests for terms

Divergences

- The Coiste seem inclined towards cognates
- We prefer “cáipéis” over “doiciméad”
- “tosaigh” over “lainseáil” (launch)
- “luchtaigh” over “lódáil” (load)
- “comhéadan” over “tosach” (front-end)
- We do use “brabhsálaí” (not uncontroversially)

Future Prospects

- Maintain!
- Grow the community
- Users → testers → proofreaders → localizers
- More mobile apps
- Games
- Strengthen usage in gaelscoileanna
- http://scratch.mit.edu/

Other languages

- Scottish Gaelic (Michael Bauer)
- Haitian Creole (Jean Came Poulard)
- Chichewa (Edmond Kachale)
- 7000 others!