OSS & LRLs

Richard Littauer
Colloquium for an MSc in Computational Linguistics
Saarland University

Abstract

- Number of languages: 7000.
- Number of languages used (for real) online: Around 30.
- Why? Tools are hard to make.
- So… why not use open source to help?

Abstract

- What are low resource languages?
- What is digital presence?
- What is open source?
- What open source for LRLs exists?

Abstract

- Case studies: Gaelic and Naskapi
- Code resource: RichardLitt/low-resource-languages
- Discussion

Abstract

Contributions:

- First published analysis of FLOSS for LRLS;
- New description of the only database of solely open source resources
- State-of-the-field for Gaelic, and some field work on Naskapi
- First suggestion of using the decentralized web for LRLs

Outline

- Introduction
- Low Resource Languages (LRL)
- Language Resources
- Open Source Software (OSS)
- OSS for LRLs
- Gaelic and Naskapi
- Best Practice Recommendations
- Discussion
- Future work

Introduction

- Half of the world's 7000 will die by 2100.
- Less than 5% have a significant digital presence.
- English is everywhere.
- Tools are expensive.
- Let's share resources to make the process easier.

Low Resource Languages

Terminology!

- Endangered, dormant, extinct, historic, constructed and revitalized
- Official, de facto, de jure, majority, and minority
- Low, sparse, non-central, and under resourced languages; incident, source, and target
- Language vitality?

GIDS

- Fishman 1991
- Domains
- Transmission
- 1-8 levels, with 1 being stable

GIDS

- Lewis 2009
- Directionality
- More granularity
- More levels (0 (English), 9, 10 (Revitalized; extinct))

UNESCO

- UNESCO 2003~
- Nine metrics, with 0-5 ratings each.
  - Transmission, number of speakers, proportion of speakers, trends, new domains, materials, government support, attitudes, documentation.
  - "Languages cannot be assessed simply by adding the numbers; we therefore suggest such simple addition not be done.”

LEI

- Lee and Van Way 2016 — Catalogue of Endangered Languages (Google)

- Intergenerational transmission, absolute number of speakers, speaker number trends (whether increasing or decreasing), and domains of use.

- "Let’s just put a number on it.” (Paraphrase)

Digital Presence

- Kornai 2013
- Thriving, Vital, Heritage, and Still
  - Demographics, prestige, the identity function of the language, the level of software support, and Wikipedia presence for a language
  - Also the Digital Language Support project

Digital Presence

- Gibson 2016: Emergent and Latent

- Soria 2017: 15 different factors, each with a scale

Language Resources

- Corpora: Written, audio, bilingual, annotated corpora, &c.
- Code: Codecs, language identification, parsers, spell checkers, tokenizers, lemmatizers, POS taggers, NER, TTS, MT, &c.
- Aggregators: CLDR, ELP, Ethnologue, Glottolog, Omniglot, ODIN, OLAC, Wikipedia, WALS

Language Resources

- BLARK (Krauwer 98, ELRA)
- LREs (2010+, LREC)
- 4400 entries; 133 LRLs with 400~ entries

Who makes Language Resources?

- Linguists
- Computational Linguists
- Language communities
- Missionaries
- Nonprofits
- Enterprise
- DARPA

Open Source Software

- FLOSS: Free libre open source software
- OSI: Free redistribution; source code; derivative works; no discrimination; non-restrictive; license distribution; integrity and patchability.
- Licenses: MIT, Apache, BSD, GPL, CC, Unlicense
- GitHub.
- Permanence and relevance, rot, and storage
- Funding?

OSS for LRLs

- Mapping linguistic coördinates
- OSS in the Aggregators
- Linked Open Data
- Cross-linguistic Projection
- An OSS database for LRLs

Mapping linguistic coördinates

- Garrette and Baldridge 2013: low-resource-pos-tagging-2014
- Gawne and Ring 2016: Using CartoDB, Google Maps, and TileMill
- Open, closed, and semi-closed resources

OSS in the Aggregators

- Most resource aggregators don’t have code (ELP, Glottolog, Omniglot, WALS)
- Or there’s a paywall (CLARIN)
- Or no links to code (LRE Map)

OSS in the Aggregators

- Some have limited resources (CLDR, OLAC, LDC)
- Or lots (RNLD, EMELD, ACL Wiki) but without good coverage or clarity.
- In general: hundreds, not thousands of resources.

Linked Open Data

- OWLG and the LLOD
- SPARQL
- Most resources are available elsewhere, anyway.

Multilingual Projection

- LinGO (Bender 2016)
- Agic 2015: "If all you have is a bit of the Bible: Learning POS taggers for truly low-resource languages.”
- NLTK
- No metrics on use, hard to track forks

low-resource-languages

- GitHub repository using Markdown links
- 500~ links, 30~ specific languages
- 19 contributors, 166 stars
- Crowd-sourced, collaborative, and curated
- Easily editable and searchable

Gaelic

- 60k speakers, 30k literate
- Unesco: Definitely endangered. Ethnologue: Provincial. LEI: Threatened. Kornai: Living.
- Lots of academic research; a language Bórd; and some small groups working on Gaelic HLT.

Gaelic

- Several language packs at the OS level for Ubuntu and Windows input;
- A large Wikipedia
- A Hunspell checker
- A large Crúbadán corpus (1,541,302 words and 17,308 documents)
- A large Indigenous Tweets corpus (>.5m words)
- Some open source. But not much.

Gaelic

- No language technology courses for Gaelic in Scottish Universities
- Unlikely to see more growth without concerted lobbying
- The best resource for finding OSS for Gaelic is arguably my list.

Naskapi

- 1000~ speakers, most bilingual in English (and French and Montagnais)
- An SIL outpost, who’ve translated a Bible and other resources
- UNESCO: Vulnerable. Ethnologue: 4 (Educational). LEI: Nothing. Kornai: Dead.

Naskapi

- No language packs
- No Hunspell
- No primary texts online.
- 2k words in the Crúbadán Corpus
- ODIN has one IGT reference

Naskapi

- Keyman used for the script.
- Naskapi Development Council works with SIL, but most of the information is in FLEx and on Webonary.
- Literacy seems to be limited in young speakers to Roman characters and spoken domains.
- There’s work to be done.

Best Practice Recs

- Choose a license: MIT. (Scannell suggests GPL.)
- Choose repositories: GitHub, but also decentralize it.
- Sans platform: Use IPFS and the decentralized web. Offline first development may also be crucial.
- Market your code.

Discussion

- Is digital presence necessary? It’s unclear.
- Is FLOSS ethical? It’s unclear.
- Is data privacy a concern for language technologists? For data, yes. For code, it probably shouldn’t be.
- Should we open source? Definitively yes. It allows access, growth, and ecosystem development.

Future Work

- Make a database actually *linking* FLOSS code for LRLs.
- Extend Kornai's metrics.
- Tease our language diversity and typological relations for cross-lingual projection.
- Metrics for code in LREC and ACL papers.
- Develop a p2p storage system for linguistics code
- Developing resources for Naskapi and Gaelic.
- Taking control go the ISO 649-3 standard.

Thanks

Questions?