The State of Open Source Resources for Low Resource Languages

Richard Littauer
Saarland University
Saarbrücken, Germany
richard.littauer@gmail.com

Abstract

1. Introduction

Of the world's roughly 7000 languages, only a small number are present on the World Wide Web, or present in digital form. The majority of technological infrastructure that first world nations use and depend upon has been built with English, and serves English speakers. There are a few languages with large populations and state-backing which have a large foothold on the web. For instance, China is estimated to have more users of the internet than Japan, India, and the United States combined [1]. The majority of these users will be speakers of Standard Mandarin, and not English; the operating systems and infrastructural backbone will still depend upon English-language originating software, but the user interface will be in Chinese. Some thirty languages (**?**) are estimated to be at this level. The majority of the world's languages, however, have no significant user presence, and are not present on the World Wide Web.
There are various metrics would can be used to assess linguistic health in the digital sphere.

- It is spoken by living fluent speakers, including second-language learners.
- It is spoken by living, first-language learners.
- It is productive in it's morphology, growing in vocabulary, and not frozen in time.
- It is recorded in some form, including audio files.
- It has a writing system.
- It has a writing system that is used by modern speakers to record their own language.
- It has a writing system that can be used on a computer.
- The electronic writing system does not require excessive installation.
- All normal characters are available in Unicode.
- There is a growing corpus of written documents in the language.
- There are users who consistently use the language digitally.
- There is a formalized spelling system.
- There is a Bible translation.
- There are non-electronic documents.
- There is a dictionary.
- There is a machine-readable corpus.
- It is used on modern social media; Twitter, and so on.
- There is a Wikipedia entry.
- There are spellcheckers.
- There are syntactic tools.
- There are machine learning algorithms based on the language.
- There are speech-to-text or text-to-speech systems developed for the language.

- Definition of computational linguistics, and linguistic tooling - Code as it pertains to lrl - state of the field linguistically - State of the field computationally - Lack of sharing code or storing it usefully, due to factors: funding, academic cycle, inability, scope, lack of knowledge of domain - Some shared code
In this paper, I will talk about:
- Open source code - Longevity of linguistic scholarship and work - Data, rights, liability, and privacy - Funding - Institutional bottleneck - Linguistic colonialism - Ethical and moral concerns for military usage - Ethical and moral concerns for big business usage - Open Source work currently available - Case study on GitHub, SourceForge, some archival sites (UPenn, Max Planck, DFKI) - Case study using endangered-languages repository - Clean up resource - Add all listed resources in issues - Contact and create the LSA CELP Technology Subcommittee - Clone all SourceForge repositories - Rename to low-resource-languages - "List quality" - "the pages and subpages are often dead" - Get diagnostics on the state of the links I've found: - What percentage have been updated when - Downloaded, etc. - Review Excel results - Peer-to-peer solution for sharing code - Stub out example - Build a web searcher for automatically getting and sharing code Further Work: - Open source data repositories (touch on) - Working with Ethnologue Conculusion

2. Open Source code

What it is, the history of it in Computational Linguistics and elsewhere, and various incentive models for using open source methods

[1] `http://chinapower.csis.org/web-connectedness/`http://chinapower.csis.org/web-connectedness/

3. Data and privacy

Whether it makes sense to decouple code from data, especially in cases of low resource languages, where sparse data may be naturally enriched with annotation schemas and hard to separate out from the tools being used. In such cases, how do we as a community, researchers as providers, and developers as consumers, deal with licensing, privacy, and proprietary data? Does it make sense to provide links to code that can be used institutionally or commercially without also allowing for things like royalties for usage, or proper licensing for data? Bound up in this are also ethical concerns - well studied in theoretical field linguistics - about the language users themselves not wishing for their data to be used in certain ways.

4. Funding

IARPA and DARPA both are involved with low resource languages and both of them may have their own institutional values that are probably at ends with independent researchers, commercial consumers, and language communities. Does working on sparse data openly bring along with it ethical or moral concerns; if so, how can these be adequately explained, breached, and talked about? How can they be worked around or be part of the conversation? Note that DARPA and the like also use humanitarian reasons as their primary stated aim for work on sparse languages, which may be contrary to their military needs. There is already an extensive literature on moral uses of data – I could summarize that, and apply it specifically to low resource languages, which is something I do not think has yet been published.

5. Digital Permanence and Storage

Universities and institutions have short timelines and are largely dependent on specific, allocated, and thus finite funding. What other models are there for data storage? What concerns are there?

6. Choosing Repositories

Longer term plans for open source repositories; GitHub is useful currently, but it also a business, and as such its aims may not be aligned with its users. I would like to talk about building a database of open source repositories on a secure, permanent, peer-to-peer network. This is something I am actively involved in professionally (I currently work at IPFS, which is building such a network). I would like to talk about linguistic and scientific applications of using versioned, p2p, and distributed systems for storing both open source code related to low resource languages as well as language data.

7. Language Specific Needs

- Disambiguate low-resource language, minoirty languages, endangered languages, and sparse languages (among others) are used often synonymously, but are distinct and come along with different stakeholders and communities, which means different values, methods, and goals. - A review of low resource language resources and their target communities and languages, in general; a state of the field for the issue. - Specific examples of cross-language applicability of an open source coding library (such as NLTK, or more specifically, family-related usage of parsers or MT models), and what that says about the incentives and use cases for open source libraries.

8. Example Use Case

I propose a study of RichardLitt/endangered-languages: - It's uses (specifically) - Current considerations in it's planning - reception - User evaluations from other open source scientists - Future goals

9. Tool

Build a web-application tool for serving a decentralized data store for endangered language tools and data
Example:
I have already put a subset of repositories listed on endangered-languages into IPFS, a p2p resource for storing and disseminating data in a decentralized and persistent fashion.
Process:
1. ‘cat‘ the endangered-languages README.md, then ‘grep‘ for ‘/.*(//github.com/.*?/[a-zA-Z0-9-]*).*/‘ (all github.com repos). 2. Output list into separate file. 3. ‘awk‘ the first few repos, until a random divider, and clone the git repos: ‘awk ’1;/kuromoji-server/exit’ ../githublist.md — xargs -n1 git clone‘ 4. ‘ipfs add -r repos‘ 5. ‘ipfs pin add repos‘