# The State of Open Source Resources for Low Resource Languages

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

**Abstract**

## Contents



## 1. Introduction

At least half of the world's 7000 languages will be extinct this century (Grenoble, 2011, p. 27). 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. Tt is estimated that less than 5% of the world's langauges will be used online or have significant digital presence (Kornai, 2013).

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 - 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 Conclusion

## 2. Language Case Studies

### 2.1. Metrics

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.

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

- 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.

To get a better idea of hose these metrics can be implemented, we can look at several different languages and how they use code to further language development.

## 2.2. Scottish Gaelic

Scottish Gaelic is a Celtic language spoken mainly in the United Kingdom, which UNESCO defines as *definitely endangered* [2]. A large corpus compiled by the An Crubádán project is available online [3] (Scannell, 2007).

## 2.3. Naskapi

### 2.3.1. Language Background

Naskapi is a Cree language in the Algonquin family spoken in central Quebec (MacKenzie and Jancewicz, 1994), which UNESCO defines as *vulnerable* [4]. Virtually the entire population of around 800 Naskapi live within the reservation Kawawachikamach, around 10 miles from Schefferville, QC. Schefferville is only accessible by train or plane, and contains another local tribe called the Innu (which has more than 17,000 members, scattered among Quebec and Labrador[5]), who live on their own reservation and who speak Montagnais or Innu-aimun, a related language. The two languages are similar, and the Naskapi youth are often diglossic in Montagnais (but the Innu are often not) (MacKenzie, 1980).

The Naskapi speak English as a first or second language, while the Innu speak French (and some speak three or all four languages). They moved to Kawawachikamach in the 1960s, after initially being resettled in Schefferville in the early 1950s. Some of the elders still remember being a nomadic people who followed caribou and were raised in the bush. However, half of the population is under the age of 16, as the First Nations population is the largest growing population in Canada.[6]

All of the Naskapi speak their own language regularly, in all contexts. In the schools, there are Naskapi-only classes held until Grade 8 (Llewellyn and Ng-A-Fook, 2017). While there are a few social workers, teachers, and nurses who speak solely English, most jobs in Kawawachikamach are held by Naskapi. There has been a long tradition of missionaries, and almost all of the Naskapi are Protestant. At church, they use Montagnais hymnals and an Montagnais bible.

### 2.3.2. Literacy Developments

In recent years, the Naskapi Development Council, which works with translators provided by the local tribal council (called the Band), has produced a Naskapi to English bilingual dictionary in three volumes (MacKenzie and Jancewicz, 1994). This was produced by linguists from the Summer Institute of Linguistics, funded by Wycliffe Bible Translators [7].

Today, the SIL linguists are a team of six: two long term linguists, and two pairs of husband and wife pairs who are training how to work as bible translators in this community before moving on to working with other Cree communities in Canada. Naskapi does not have a complete bible. A new testament, started in the 70's, was recently published (Naskapi Development Corporation, 2007). Genesis, Exodus, and Psalms, have also been translated, and several children stories and books of oral legends from a an elder have been produced. The full-time translators are two people: a young woman in her mid-twenties, and an older gentleman of around 50 years of age. At times, elders also contribute to the bible translation effort by marking up their pre-publication drafts, which they then go over with the translators.

When there is a need to come up with a new term, the elders are consulted, and they agree on an appropriate translation. For instance, "grill" is translated as "metal-net". A grill is not a pre-existing word in Naskapi, but net is, and it is easy to imagine the metaphor of a grill on which you braise meat as being a metal net. However, these decisions are not replicated outside of the bible. Likewise, when there is a term which needs to be invented at the school, the teachers there decide on an appropriate term - for instance, for situations like Halloween, where "Frankenstein" may need to be translated into a local alternative. These decisions are largely one-off, although they may be used year to year, and informally recorded in their respective domains.

The linguists use the Fieldworks Language Explorer (FLEx) [8] to document new linguistic terms. FLEx was developed by SIL International, and provides linguists with an out-of-the-box solution for recording linguistics terms using interlinear glossed text. It is also open source, and available on GitHub [9]. Users can export as a PDF (among other file formats), or export words to an online interface

[2] http://www.unesco.org/languages-atlas/en/atlasmap/language-iso-gla.html

[3] http://crubadan.org/languages/gd

[4] http://www.unesco.org/culture/languages-atlas/en/atlasmap/language-id-2354.html

[5] https://en.wikipedia.org/wiki/Innu

[6] http://www12.statcan.gc.ca/census-recensement/2016/dp-pd/index-eng.cfm

[7] https://www.wycliffe.org/

[8] https://software.sil.org/fieldworks/

[9] https://github.com/sillsdev/FieldWorks

known as Webonary [10]. This allows language workers to automatically create a useable, free dictionary for members of the community.

Naskapi uses the Innuit syllabics spelling system (Comrie, 2013), as well as two other roman-based systems with only minor differences. For instance, a macron, such as û is used in place of a double *uu* to indicate vowel length. Computational writing using the syllabic system is possible by using Keyman [11], (free, open source software available on GitHub [12]) which must be installed manually on a computer. It allows a user to type roman letters which are converted to the right syllabic phrase, and is forgiving for phonemic variants. For instance, "ju", "chu", "tchu" and so on might all be interpreted and replaced by the appropriate syllabic.

Currently, the school has a computer lab with over a dozen computers, but no in-house computer technician. One of the Wycliffe translators needed to visit the school to check on Keyman updates, and the students are not regularly trained in how to set up Keyman on their own, or how to set it up on their phones or other portable devices. While Facebook and other online platforms are increasingly popular, the majority of talking takes place in Naskapi written in local characters, or in English.

#### 2.3.3. Computational Tools

There are no spell checkers, word lists, or large corpora available digitally except for the dictionary. As well as the SIL-sponsored Webonary, there is also work done by atlasling.ca, which is a Canadian government-backed venture, originally cofounded by MacKenzie, who also worked on the Naskapi dictionary [13]. This website also has some options for looking at languages, but does not seem to be updated by local translators from the community. It is sourced from the previously published dictionary, which the SIL linguists have indicated is not up to date and has insufficient English to Naskapi translations. These are insufficient because of the nature of Naskapi; a root word is used with a slot system, and any word which mentions water is included under the English heading. This makes translating something as simple as "the mug is red" difficult, as you need to know to look for "red" as a root word, and then to find the appropriate example from which you can extrapolate the correct form for translation.

There is a potentially large corpus of spoken language in Naskapi from the local radio station, but this is not linguistically digested. There does not appear to be any adult-level secular written corpora which could be utilized to jumpstart a corpus. The Band employs translators (who generally have other jobs - one this author interviewed was a band Councilman, one of four elected officials underneath the Chief) who may be able to provide bilingual texts in English, French, or Innu.

All told, computational work is exceedingly limited. There are some websites in Naskapi, which could be used to make a small corpus, but there are no currently active projects working on collecting corpora for the purpose of linguistic study, and neither is there an active academic community working on Naskapi outside of the SIL translators, who may occasionally publish a paper (or, of course, a dictionary or physical book).

While FLEx is open source, none of the linguists edit the code for it or use the codebase, depending on SIL International to keep the product up to date. Keymap is likewise not edited, although it is installed on local computers. There have been at least one Naskapi speaker who found and used a syllabic keyboard, but there has been no effort to standardise the syllabics in the schools or with other speakers, and the relevant code has not been shared in any official capacity by any party in the language community.

## 3. Open Source code

- 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

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

Not all research that is code based can be easily quantified as open source. For instance, [Afranaph](http://www.africananaphora.rutgers.edu/home-mainmenu-1) is a database of research on African languages. However, there is no code directly available to build your own database. Instead, you only have the option of searching their database. Other sites may use open source technology, but not be open source themselves. For instance, [TransNewGuinea](http://transnewguinea.org/about) has a colophon where they mention that they use Unicode, Django, Bootstrap, jQuery, Leaflet, PostgreSQL, and SQLite.

Keyboard layouts are another area where much i18n work has been focused. Link: https://github.com/HughP/MLKA

## 4. 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.

## 5. Funding

IARPA and DARPA both are involved with low resource languages and both of them may have their own institu-

[10] https://www.webonary.org/configuring-the-dictionary-inflex/

[11] https://keyman.com/

[12] https://github.com/keymanapp

[13] http://atlas-ling.ca/

tional 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.

## 6. 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?

## 7. 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.

## 8. Language Specific Needs

- Disambiguate low-resource language, minority 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.

### 8.1. Some Thoughts on NLTK

http://nltk.org/ is an free and open source library which uses the Python language, and enables users to interface with over fifty different corpora and lexical resources. A primer written by the main creators, (http://nltk.org/book), is used frequently in natural language processing classes written by the creators. It is licensed under the Apache 2.0 license, a common license [14]. On GitHub, there are currently 204 contributors listed , although the git history shows 234 (found by using the command 'git authors' ). Some of the resources within NLTK have to do with low resource languages. For instance, in 2015, NLTK added machine translation libraries, including popular ones such as IBM Models 1-3 and BLEU.

By open sourcing their code, the NLTK authors have allowed it to be adapted and re-used. Currently, there are several ports. One of these is the JavaScript language implementation, https://github.com/NaturalNode/natural. This has 6700 stars on GitHub, which is a good indicator of community vitality and use, and 88 contributors. The port is also open source, under an MIT license https://github.com/NaturalNode/natural#license.

## 9. 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

## 10. 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'

## 11. References


Comrie, B., (2013). *Writing Systems*. Max Planck Institute for Evolutionary Anthropology, Leipzig.

Grenoble, L. A., (2011). *Language ecology and endangerment*, pages 27–44. Cambridge Handbooks in Language and Linguistics. Cambridge University Press.

Kornai, A. (2013). Digital language death. *PloS one*, 8(10):e77056.

Llewellyn, K. and Ng-A-Fook, N. (2017). *Oral History and Education: Theories, Dilemmas, and Practices*. Palgrave Studies in Oral History. Palgrave Macmillan US.

MacKenzie, M. and Jancewicz, B. (1994). *Naskapi Lexicon*. Naskapi Development Corporation, Kawawachikamach, Quebec.

MacKenzie, M. (1980). *Towards a Dialectology of Cree-Montagnais-Naskapi*. University of Toronto.

Naskapi Development Corporation. (2007). *Naskapi New Testament*. Naskapi Development Corporation and Wycliffe Bible Translators.

Scannell, K. P. (2007). The crúbadán project: Corpus building for under-resourced languages.


[14] https://github.com/nltk/nltk/blob/develop/LICENSE.txt