Open Source Resources for Low Resource Languages

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

Abstract

We present a database of open source code that can be used by low-resource language communities and developers to build digital resources. Our database is also useful to software developers working with those communities and to researchers looking to describe the state of the field when seeking funding for development projects.

Keywords: open source, under-resourced languages, database, endangered languages, code, computational resources

Contents

1. Introduction

1.0.1. The current state of language divesity

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 - perhaps thirty - with the combination of large populations, official governmental status, and Westernised, industrial economies which affords them a large foothold on the web. It is estimated that less than 5% of the world's languages will be used online or have significant digital presence (Kornai, 2013).

1.0.2. Digital presence

A language's digital presence is defined by: its ability to have tools which are easily accessible to users of the language, applications build on top of that tooling, large amounts of mono- and bilingual corpora, and an active amount of speakers (generally, in the millions). Generally, this also needs official backing in the form of an official status for a language, set at a national level. For instance, English is considered the *de facto* national language of United States, and the overwhelming majority of language users and computational software developed in the States is in English. In Europe, the EU mandates that citizens have the right to correspond with an EU institution (and receive a response back) in any of the 24 recognized official languages of the respective member states.[1] This includes relatively small languages, such as Maltese, which has a significant web presence and tooling considering there are only just under half a million speakers. [2] China is estimated to have more users of the internet than Japan, India, and the United States combined.[3] But while 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 (although the user interface will be in Chinese).

Neither official status nor population size necessarily means that a language will have resources or a digital presence, however. Romansh, although it is an official, provincial language of Switzerland, does not enjoy the same privileges as French or German, as there are only 40,000 speakers.[4] Min Dong, although it has over nine million speakers [5], does not afford the same privileges as Mandarin, as it is not the official language of China. It is also not true that a language needs to have a large government presence or access to military-grade technology to have digital resources - a notable example would be Haitain Creole, spoken by over ten million speakers in one of the world's poorest countries. After the 2010 earthquake, Google, Microsoft, and Carnegie Mellon collaborated to make a machine translation system, an overwhelmingly difficult achievement.(**?**)

1.0.3. Interested parties for smaller languages

There are, of course, thousands of languages which could be used as examples of little to no resources, without recognition, speakers, or high literacy or computational use. The majority of languages do not have access to large financial bodies (whether military, governmental, or private) willing to invest in technical development, nor in developing large corpora which could be used to develop, bootstrap, and train computational tools. In such cases, any development work is normally undertaken by a few different groups; higher education departments or labs, native speakers with some digital ability, and missionaries, or, more usually, a combination of all three.

These groups normally do not have access to large amounts of corpora, computational resources (such as servers, grids, or databanks), or, most importantly, manpower, time, and money. They are more often goal driven. For instance, Wycliffe translators are most interested in developing language literacy in order to translate the Bible into a language for first language speakers; they are often less interested in machine translation, Facebook groups, or developing literacy for schools unless it helps them towards this end. Native developers are normally more interested in developing apps or websites for locals, but they are often less interested in larger dictionaries, speech to text or text to speech resources, or use cases outside of their individual ken. Finally, academic staff are often forced to be myopic in scope in order to publish academically for their wider industry, and are less able to easily embark on decade-long timelines to developer singular resources. They are also often financially fragile, and depend on local investors or governmental budgets and stipends to fund their research, all of which often work on shorter timescales than missionaries or first language developers, who normally devote their entire lives to a particular language group. Finally, academics are often individual workers, working on a language family or small group at the expense of related languages, as they need are driven by the publish-or-perish academic market and do not usually have large amounts of funding for multilingual efforts.

1.0.4. Language research funding

1.0.5. Open source as a solution

1.0.6. Current Open Source resources

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

[1] https://europa.eu/european-union/topics/multilingualism_en

[2] https://www.ethnologue.com/language/mlt

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

[4] https://www.ethnologue.com/language/roh

[5] https://www.ethnologue.com/language/cdo

2. Language Case Studies

2.1. Metrics

There are various metrics would can be used to assess language health.

However, few of these metrics take into account the level of digital literacy for a language. The possibility for a language to digitally ascend has been held up as a key component for the life of that language by Kornai.

- 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.

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* [6]. A large corpus compiled by the An Crubádán project is available online [7] (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* [8]. 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[9]), 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.[10]

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 [11].

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

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

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

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

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

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

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

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) [12] 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 [13]. Users can export as a PDF (among other file formats), or export words to an online interface known as Webonary [14]. 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 [15], (free, open source software available on GitHub [16]) 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 atlas-ling.ca, which is a Canadian government-backed venture, originally cofounded by MacKenzie, who also worked on the Naskapi dictionary [17]. 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

3.1. Defining *Open Source*

Open Source is a complex term which refers to any code, not just code related to computational linguistics. At its core, *open source* refers to code which has a license which allows it to be available to freely inspect, use, or modify by anyone. It was introduced in 1998 by some programmers, in response to the Netscape browser's code being openly licensed and made available. *Open source* is one of

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

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

[14] https://www.webonary.org/configuring-the-dictionary-in-flex/

[15] https://keyman.com/

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

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

many terms which can be used to differentiate code which is either available or licensed permissively for re-use; other terms include *free software* and *libre software*, or the combination, *FLOSS* (free and and libre open source software). There is no standard definition of open source that is universally accepted.

Nor will universal acceptance be forthcoming. The issue regarding reconciliation between open source, free software, and the rest of the terms stems largely from a different of opinion between what constitutes open software. For some adherents, software itself ought to be free, as it is a result of human labour and because it is maximally helpful for others to never have to code that again. This idea contains within it the seed of the digital commons: like the commons in philosophical and economic literature, code can be viewed as a resource that belongs to humanity as a whole, and not the creators who initially fashioned it. In this sense, open source is a philosophy.

> Open source is a development methodology; free software is a social movement. For the free software movement, free software is an ethical imperative, essential respect for the users' freedom. By contrast, the philosophy of open source considers issues in terms of how to make software "better" - in a practical sense only. It says that nonfree software is an inferior solution to the practical problem at hand.
>
> *Richard Stallman* (Founder of GNU/Linux) [18]

Before continuing, a quick word on licenses. Licenses determine the legal rights to sharing code. A piece of code which is taken from a proprietary server and published on the internet is not necessarily open source. In this instance, the code may have been illegally copied and shared, but it is not licensed for free usage. Under no definitions is this considered open source. Indeed, this touches upon issues of digital copytheft and "piracy", which is a standard term used frequently in the media and in legal proceedings to attach a sense that copying code is the same as larceny or theft on the high seas. Avoiding the question of the validity of this viewpoint, it is important to focus on the license as the differentiating factor between code which has been released legally under an "open" definition or not.

There are many licenses which are considered to be open source, and there are several arbiters available which judge the validity of open source licensing. The Open Source Initiative maintains a list of approved licenses on their website: https://opensource.org/licenses.

Open source, on the other hand, under most definitions, does pertain to ethical concerns about the software's usage, but rather simply refers to whether or not it is permissively licensed and available for users.

The Open Source Institute, which originally coined the term *open source*, has several parameters by which open source software can be judge as being 'open' or 'closed' (that is, proprietary, non-permissively licensed, non-reusable, limited in usage to a set amount of people, and so on). It may be useful to list these terms below, as they are instructive about how open source can be a nuanced term. These terms are from the OSI's website [19].

1. Free Redistribution. The license shall not restrict any party from selling or giving away the software as a component of an aggregate software distribution containing programs from several different sources. The license shall not require a royalty or other fee for such sale.

2. Source Code. The program must include source code, and must allow distribution in source code as well as compiled form. Where some form of a product is not distributed with source code, there must be a well-publicized means of obtaining the source code for no more than a reasonable reproduction cost, preferably downloading via the Internet without charge. The source code must be the preferred form in which a programmer would modify the program. Deliberately obfuscated source code is not allowed. Intermediate forms such as the output of a preprocessor or translator are not allowed.

3. Derived Works. The license must allow modifications and derived works, and must allow them to be distributed under the same terms as the license of the original software.

4. Integrity of The Author's Source Code. The license may restrict source-code from being distributed in modified form only if the license allows the distribution of "patch files" with the source code for the purpose of modifying the program at build time. The license must explicitly permit distribution of software built from modified source code. The license may require derived works to carry a different name or version number from the original software.

5. No Discrimination Against Persons or Groups. The license must not discriminate against any person or group of persons.

6. No Discrimination Against Fields of Endeavor. The license must not restrict anyone from making use of the program in a specific field of endeavor. For example, it may not restrict the program from being used in a business, or from being used for genetic research.

7. Distribution of License. The rights attached to the program must apply to all to whom the program is redistributed without the need for execution of an additional license by those parties.

8. License Must Not Be Specific to a Product. The rights attached to the program must not depend on the program's being part of a particular software distribution. If the program is extracted from that distribution and used or distributed within the terms of the program's license, all parties to whom the program is

[18] https://www.gnu.org/philosophy/open-source-misses-the-point.html

[19] https://opensource.org/osd

redistributed should have the same rights as those that are granted in conjunction with the original software distribution.

9. License Must Not Restrict Other Software. The license must not place restrictions on other software that is distributed along with the licensed software. For example, the license must not insist that all other programs distributed on the same medium must be open-source software.

10. License Must Be Technology-Neutral. No provision of the license may be predicated on any individual technology or style of interface.

3.2. Open Source

This needs to be deleted.
- 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 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.

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 [20]. 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

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

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 ‘/.*(//githubċom/.*?/[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.