Saarland University
Department of Computational Linguistics

Master's thesis Proposal

Open Source Code and Low Resource Languages

Author:
Richard Littauer
Matriculation: 2539658

Supervisors:
Dr. Dietrich Klakow
Dr. Alexis Palmer

March 15, 2018

Abstract

Of the roughly seven thousand languages currently spoken, less than fifty have a significant digital presence. In order for a language to be used digitally and to survive in the long term, it's speakers may need to develop computational resources: orthographies, dictionaries, grammars, spell checkers, parsers, and more. Instead of depending on closed source code from large providers, researchers and communities can leverage open source code as a means of bootstrapping digital language development. In this thesis, I discuss the state of the field for low resource languages, what open source code is and how this methodology can help languages. I provide two cases studies, looking in detail at Gaelic and Naskapi, and I describe a database I have developed for open source code serving these languages. Looking to the future, I suggest steps for helping save languages from being lost.

My specific contributions in this thesis include not only the first published analysis of open source code specifically regarding endangered languages, and an exposition of the only database of open source resources, but also the first independent fieldwork with Naskapi that pertains to its digital presence. I also outline how researchers and developers can change their processes to help make their work more effectual in the long term.

Eidesstattliche Erklärung
Hiermit erkläre ich, dass ich die vorliegende Arbeit selbststÃd'ndig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel verwendet habe.

Declaration
I hereby confirm that the thesis presented here is my own work, with all assistance acknowledged.

Acknowledgements
Jonathan Poitz provided formatting files, but did not advise on the content. This thesis is based loosely on a paper presented at the LREC CCURL Workshop in July 2016 in Slovenia (Littauer and Paterson III, 2016). Hugh Paterson III was a coauthor on that paper.

Montréal, March 15, 2018

Richard Littauer

Contents

1 Introduction

At least half of the world's 6000-odd languages will be extinct this century (Krauss, 1992; Grenoble, 2011). Just over half of these languages have writing systems.[1] It is estimated that less than 5% of the world's languages will be used online or have significant digital presence (Kornai, 2013).

The majority of the world's computational technology has been built by English, with English manuals, English interfaces, and by English speakers. The most prevalent language spoken by users of this technology is also English. There are a few languages - around thirty - with the combination of large populations with internet access, official governmental status, and industrial economies which affords them some native computational technology, in particular on the World Wide Web, the largest global network for sharing code and written material.

English is the undisputed heavyweight as far as global written resources are concerned.[2] Over half of the web's content is written in English. The next largest languages are Russian, German, Spanish, Japanese, and French - with a combined population of well over a billion speakers. Portuguese, Italian, and Chinese have the next largest amount of content - but each of them only covers between 2 and 3% of the web's content - followed by Polish, Turkish, Dutch, and Korean with over 1%. Suffice to say, the graph of global written content is not skewed towards language diversity as a norm. This is not surprising, as around 90% of the world's languages are spoken by less than 10% of its people (Bernard, 1992).

In part, these high-resource languages depend upon shared code. Put simply (and therefore ungracefully), a literacy system affords written corpora, and written corpora can be used by researchers to either build tools for that language or to adapt tools from other languages. These tools might be spell-checkers, parsers, input systems, or later on speech recognition and generation software, semantic analysers, or machine learning and translation systems, among others.

This culturally shared body of code is most often developed in closed environments with consumer endpoints, by the military or large businesses. For instance, the World Wide Web, the largest shared corpus of written language, started with support from the Massachusetts Institute of Technology (MIT) and the Defense Advanced Research Projects Agency (DARPA). (This helps to explain why most of the web is written in English.) Another example would be

[1] https://www.ethnologue.com/enterprise-faq/how-many-languages-world-are-unwritten-0

[2] https://w3techs.com/technologies/history_overview/content_language

Google Translate, which uses massive bilingual corpora to provide automatic translation services for free online, but whose code is proprietary and owned by Google.

While the enterprise pathway works well for large languages where populations of speakers can be leveraged to provide funding, the majority of the world's languages are not able to develop their own computational resources - either grammars, corpora, or code. Instead, they must rely on small groups of researchers, limited funding, and a grab-bag of written resources when they have them. For instance, the most consistent translations cross-linguistically are of the Christian bible, which may not reflect the target language's culture.

Incidentally, there is something to be said for spoken language corpora, which may be more prevalent in some cases than written resources (especially in a region with a history of radio transmissions in the local language, for instance). However, the direct use of spoken language corpora for building language resources is limited and generally requires more processing and development time (not to mention storage), compared to cheap, written data.

In this thesis, I will examine methodology that can be used by linguists, researchers, and language developers to help their languages "digitally ascend" (as Kornai (2013) puts it) - to bootstrap their corpora creation, write grammars, transform other language's tools and research to their own languages, and to ultimately enable their communities to speak and share their knowledge computationally. This methodology goes under the broad label of *open source* software. Open source software is code which has been developed and made available for free, without concessions about how it is to be used or who uses it. This allows coders to use code which they personally haven't built without allocating funds for it, thus freeing up significant portions of research and development costs for making tools. At present, the majority of the world's code depends on some level on open source software - for instance, Linux, and much of the World Wide Web, depends on open source code.

In the field of computational linguistics, however, there are a deficit of resources which are licensed and available as open source. This largely stems from the need to financially recoup expenses for development, on licenses mandated by research groups or military funders, and on a lack of awareness of how open source code works by developers. Another consideration is that an open source label does not ensure that the code is worth using, maintained, relevant, or in scope for a given domain.

Below, I will go into further depth about the state of endangered languages and computational resources in Section 2, and what different languages need in order to have digital presence. In Section 3, I'll define what open source is, and talk about issues relevant to open source code for under-resourced languages;

specifically, data rights, liability, privacy, funding, military and industrial concerns, ethical reasons for using open source. I'll then in Section 4 talk about the state of open source code currently available online, in particular focusing on a database of open source code that I have built with the help of researchers around the world.

I'll touch on some specific examples of languages which could benefit from open source code in Section 5, focusing on Gaelic, an endangered language with tens of thousands of speakers but little online resources, and Naskapi, an endangered languages with only a thousand speakers which might be able to benefit from open source code. The Naskapi case study will be largely informed by original research, as I engaged in field research at the town where most Naskapi live and talk to linguists working on literacy efforts for this language. In Section 6, I'll discuss how open source can help low resource languages, and in Section 7 I'll expound further at a high level on what open source enables for linguists and language communities. Finally, in Section 8 and Section 9 I'll discuss future work, and offer some concluding remarks.

2 Low Resource Languages: An Overview

In this section, I will outline the state of low resource languages. First I will define contrasting and distinct terms which are often used to these languages, which inform how one can approach a language. Then, I will talk about language demographics and metrics used to categorise languages as having low resources, before moving on to discuss digital presence as a term for understanding language endangerment today. Finally, I'll go into depth further about the current state of language diversity (both in research and demographically), and mention the various different groups who work on and fund low resource development, and how considering their impact influences a language's digital presence.

2.1 Definitions

Before going further, it makes sense to define what the terms *endangered*, *minority*, *low* and *under-resourced*, and other terms like *threatened* mean when they refer to a language. Ultimately, they refer as a whole to languages which are in peril in some way. However, there have slightly different meanings in different contexts, and according to the scale and metric applied.

In this section, I will generally define these terms: *endangered*, *moribund*, *extinct*, *dormant*, *revitalised*, *historic* and *constructed* languages; *minority*, *low-resource*, *under-resourced*, *incident* and *surprise* languages; and finally *computer* or *computational* languages. This will help inform why I've chosen to focus on low resource languages, and specifically low resource natural languages with living populations.

2.1.1 Endangered, revitalised, and extinct languages

Endangered languages are human languages that are in danger of extinction. The term is borrowed from the scientific literature describing animals; just as there exists as very real possibility that one day there will be no more Australasian Bittern specimens in the wilds of Australia, it is also possible that one day there may be no living speakers of Guugu Yimithirr. The term is not complete analogous; we can still read Tocharian texts, but Tocharian is not considered to be a living language, but *extinct*, as there are no speakers who use it regularly (and who are not scholars of obscure languages).

Endangered languages are normally languages which have a high amount of speakers, and crucially are still teaching children the language. Children ensure that the language will live on to the next generation, and when this chain breaks, it is almost impossible to resurrect a language. A language would

be endangered when it can be assumed that children will stop learning the language in the next hundred years (according to Krauss (1992)). This can be difficult to judge, as the rate of deterioration can be high. For instance, Breton had over a million speakers in 1950, but today the numbers may be as low as 200,000. Its future is uncertain.

Moribund languages are languages which are critically endangered, in that there are no children currently learning the language and using it frequently, although there are speakers. Ainu is a good example, with roughly ten native speakers still living, all of whom are over 80 years old. Haida has a similar amount of native speakers, but because of the amount of immersion programs, government-funded schools, and new venues for the language such as a motion picture filmed entirely in Haida with ethnically Haida actors who learned their lines from the elders, [3] it is not considered moribund.

Dormant or *sleeping* languages are a stage beyond moribund languages. They have no living fluent speakers. This does not mean that the language is extinct. An example would be Mutsun, an Ohlone or Costanoan language formerly spoken near San Juan Bautista, California, whose last known fluent speaker Ascensión Solórsano passed away in 1930. However, in the late 90s, the Mutsun people (recognised formally as the Amah Mutsun Tribal Band) began a revitalisation project using the extensive documentation left behind by linguists, anthropologists, and a Catholic mission priest, and now there are several conversational (albeit no fluent) speakers (Warner et al., 2007).

Often, dormant languages only come to attention when they are considered a *revitalised* language. As Warner et al. (2007) notes, "Daryl Baldwin did indeed teach himself his then-dormant ancestral language, Myaamia, and is now raising his children largely in the language (Hinton, 2001; Leonard, 2004)." Before Baldwin's work, Myaamia would have been considered a dormant language. Another example would be Manx, which lost all of its native speakers (the last being Ned Maddrell, who died in 1974 (Wilson, 2008)), but retained a score of second language speakers until today, when there are now immersion programs for children and over a thousand speakers of the language (Clague et al., 2009). Between 1974 and a vague point somewhere in the past couple of decades where a child could consider Manx as their first language, the language was dormant; now, however, it is revitalised.

The most famous example of a revitalised language is Hebrew, with a speaking population of over eight million,[4] which was formerly a literary language until revitalisation efforts began as a result of the creation of the Israeli

[3] https://www.nytimes.com/2017/06/11/world/americas/reviving-a-lost-language-of-canada-through-film.html

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

state in the early 20th century, where it is now an official language and not in a state of endangerment. Hebrew is a good example of why the often synonymous terms such as 'endangered' and 'revitalised' should be considered as differentiable.

While on the subject of Hebrew, it is worth mentioning that the initial efforts to revitalise it were often maligned by both Jewish communities and linguists, for a variety of reasons. First, the Jewish faith had traditionally viewed Hebrew as a holy tongue, and many religiously conservative Jews objected to the sacrilegious use of it for day-to-day matters, preferring Aramaic or Yiddish. Many also objected on the grounds that its use was connected to Zionism (why is well beyond the scope of this thesis). But most pertinently, linguists objected because they viewed revitalisation as an impossibility. If the language was dead, than it would be impossible to accurately bring it back, as literary texts are not sufficient at adequately capturing all of the intricacies of a language and how it is used. Clearly, with millions of first language speakers, this is no longer a valid point; these critics can now claim that modern Hebrew is an imperfect descendant of historical Hebrew, which remains extinct, and they are likely right to do so. Revitalisation is not always an ethically or logistically clear process.

This is especially true for *constructed* languages, which are *a priori* languages invented by a linguist or a community without a historical speaking community. These may be created to be logically resistant to ambiguity (such as Loglan or Lobjan), for a specific artistic purpose (such as Na'vi or Klingon, meant to be spoken by aliens in science fiction), for scientific study (such as those used by evolutionary linguists for language games with participants to discern how language might have evolved, or such as used in the ubiquitous Wug test by scholars of language acquisition) or for political aims (such as Esperanto or Ido). Some of these may end up with thousands of speakers, including native speakers, and a huge surplus of computational resources. Na'vi has a dictionary that has been translated using computational tooling into over a dozen languages, for instance, and morphological parser, spell checkers, and a Facebook translator. These languages are not normally considered as revitalised or dormant, but are instead mostly ignored by the scientific community altogether.

Heading back to natural languages, Latin would largely not be considered a revitalised language either, although there are immersion schools and some daily usage by the Catholic liturgy. These domains are specific and do not extend into normal life, on the whole. This doesn't mean it doesn't have some computational resources, however - the ATMs in the Vatican use Latin as a user

interface language.[5] Old Swedish, likewise, has some computational resources (admittedly, from a single research group that is humorously aware of the lack of general global interest in the field).[6] Latin would normally be considered a *historic* language, like Ancient Greek or Old English. All of these languages, while extinct themselves, have direct descendants (the Romance languages, modern Greek, and English, respectively), but this is not always the case.

Gothic is considered *extinct* today, as it has no direct descendants, although it is still studied, and although there is a small community of writers who continue to use the language, and at least one publishing company which publishes modern work in Gothic[7] (incidentally run by, of all people, me). Not all languages have sufficient texts to be revitalised or used today: Etruscan, Minoan, and Pictish are good examples.

One could argue that some languages may be considered dormant even if there are native speakers alive, if they do not speak the language. For instance, there are a few cases where a couple of speakers are left of a language, but they don't speak it to each other due to interpersonal differences. Most famously, there is the apocryphal story of Ayapeneco, where a global même ensued from an imagined feud between the last two speakers, to the point where Vodafone released a video claiming that they helped bring the men together to save the language (to the chagrin of actual linguists and anthropologists who had worked on the language for decades).[8] This has actually happened elsewhere, such as with Nisenan (Snyder, 2004). Another example might be Ishi, the last Yahi and a speaker of Yana, who explained that he had no name, because there was no other Yahi man to formally introduce him. Ishi means 'man' in Yana (Kroeber and Robbins, 1973).

Such cases are extreme, and there will be exceptions to almost any of these categories. Even for living languages, questions of identification can be difficult. For instance, Gil (2009) points to at least a dozen different interpretations of what Riau Indonesian might technically be. Defining language is beyond the scope of this thesis - however, I would be amiss not to mention this problem here.

2.1.2 Minority, low and under resourced, and incident languages

Minority languages are are spoken by a stable, but small, population (for example, Maltese or Hawai'ian); and low or under resourced languages, which are spoken by a significant population but underrepresented on the web (for

[5] https://gizmodo.com/5905595/the-atms-in-vatican-city-speak-latin
[6] https://spraakbanken.gu.se/swe/forskning/diabase
[7] https://wordhoardpress.com
[8] http://stories.schwa-fire.com/who_save_ayapaneco#chapter-113060

instance, Quechua). These languages share certain characteristics in common; the most pertinent is sparse data and a lack of resources, ranging from spell-checkers to grammars to machine translation corpora. Other under-resourced languages that do not fall under this list include constructed languages (for instance, Klingon or Na'vi), computer languages (for instance, Javascript or Lua), and extinct languages that are so sparse as to be rendered computationally irrelevant for most purposes (for instance, Tocharian).

2.2 Metrics

There are various metrics would can be used to assess language health. In this section, I'll explain these metrics in detail, focusing on the UNESCO, GIDS, EGIDS, and LEI measurements, as suggested by Yang et al. (2017).

2.3 Digital presence

In this section, I am going to explain what digital presence is. This is more than just defining language endangerment - instead, this is about how do we quantify a language's existence digitally, either on the web or offline in archives.

Few of the metrics above 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 of judging a language's vitality by Kornai (2013).

I'll describe his assessment here, and explain why an alternative assessment would also be good. For instance, Wikipedia is, in my opinion, not a good judge of a language's health, as it is a closed ecosystem with diminishing returns for users who are bilingual.

2.4 The current state of language diversity

In this section, I am going to briefly go into detail about what diversity means for linguistics. This will be useful later for explaining how related languages can be used to bootstrap work in similar languages. For instance, Irish spell-checkers and constitutional corpora from the EU can be used by Scottish Gaelic speakers with some tweaks in order to further improve their own systems.

2.5 Who makes resources for LRLs?

Here, I will explain briefly who makes language resources for these languages. I'll explain what I see as the main groups doing this work: professional translators, educators, missionaries (of multiple faiths, but mostly Christian), aca-

demics and native technologists. I'll explain each stakeholder and their canonical perspectives.

2.6 Language research funding

Here, I'll go into more depth about funding, as we've outlined who works on LRLs and who would fund research, and why. This will further inform the basis for the work of the previous section. I'll talk about DARPA MT funding in the 20th century, as well as other efforts such as CLARIN.

3 Open Source Code

Changing tack, here I will talk about what *open source* means. This is important - otherwise, this thesis is just a rehash of current existing computational work on LRLs.

3.1 Defining *Open Source*

Open Source is a complex term which refers to any code, not just code related to computational linguistics.

Here, I'll define what I mean by Open Source. This will largely inform the next section where I talk about its use for LRLs.

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 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 difference 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 more of a philosophical theme than a technical term.

> 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.[9] *Richard Stallman* (Founder of GNULinux)

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

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 directly below, as they are instructive about how open source can be a nuanced term. These terms are from the OSI's website [10].

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.

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

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 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 Where is open source code?

Here, I will include a short section on how the open source world works. In particular, I'll answer the question of where code lives. I'll include a short

overview and case study on GitHub, SourceForge, and some academic archival sites (UPenn, Max Planck, DFKI).

3.3 Digital Permanence and Storage

Universities and institutions have short timelines and are largely dependent on specific, allocated, and thus finite funding. Here, I'll answer the question: What other models are there for data storage? What concerns are there?

3.4 Data and privacy

Here, I'll talk specifically about data rights and privacy, in regards to 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.

3.5 Legal rights and liability

Here, I'll talk about specific licenses used in Open Source, and how they apply to code. I'll try to keep this brief.

I'll also talk about liability wavers - a separate issue from licenses. I'll talk about the standard liability wavers used with the MIT license, and other issues that might arise for language code specifically.

3.6 Military and enterprise solutions

In this section, I will talk about how open source meshes with military and enterprise development.

3.7 Funding

Here, I'll talk about funding again - but in terms of open source code. This will be a short section.

3.8 Ethical reasons for using open source

Finally, I want to close with a discussion of the moral and ethical reasons for using open source, and whether or not these concerns are relevant to computational linguists.

4 Open Source Code for Low Resource Languages

In this section, I'll move on to the real meat of this thesis; how is open source code used for computational linguistics, and specifically for LRLs.

4.1 BLARK and beyond

First, I am going to talk about BLARK - the Basic LAnguage Resource Kit proposed by Krauwer (2003) - and what a language needs digitally as a base layer to digitally ascend. I haven't talked specifically about how computational linguistics addresses low resource languages yet - the preceding sections have largely been showing the state of the field and what open source is. We'll get to open source eventually, but here I want to cover the tools needed for a language.

I'll then mention tools here that can be used after a language has some digital presence - basically, what makes an LRL a resourced language.

4.2 NLTK and other open source libraries

Here, I'll explain some open source resources that can be used to bootstrap development; for instance, http://nltk.org/, a free and open source library which uses the Python language by Bird (2006), and enables users to interface with over fifty different corpora and lexical resources.

4.3 A Database for Open Source Code

Here, I'll talk about a database of open source code. Specifically, I'll mention my own work building https://github.com/RichardLitt/endangered-languages, described first in Littauer and Paterson III (2016), and what it contains and who has worked on it with me. I'll cover the main tools, what kind of tools were included, and why I built the database on GitHub in this way.

I'll also include diagnostics on how it has been used and how the tools it mentions have been used - what percentage have been downloaded, and so on.

4.4 Linked Data

Here, I'll briefly talk about related efforts with the Open Linguistics Working Group's (Chiarcos et al., 2012) work on open source data reflected on the semantic web.(Chiarcos et al., 2013)

5 Case Studies

5.1 Scottish Gaelic

Scottish Gaelic is a Celtic language spoken mainly in the United Kingdom, which UNESCO defines as *definitely endangered* [11]. Gaelic - sometimes called Scots Gaelic, simply Gaelic, or the Gaelic - is a Goidelic or Q-Celtic langauge, along with Manx and Irish (also sometimes called Irish Gaelic, but here always referred to as Irish). This means that, while related to the Brythonic languages of Welsh, Cornish and Breton, it is different enough to not be able to benefit from the many resources available in Welsh, which, while endangered, has a much stronger academic interest and presence in the United Kingdom, with roughly half a million speakers.

A large corpus compiled by the An Crubádán project is available online [12] (Scannell, 2007).

As it is similar to Irish, it is a good example of how code from related languages can be used to bootstrap efforts to build code for its own language. I'll talk in depth about the language, its structure and grammar as related to code, its users and their use cases, and efforts to use code to make Scottish Gaelic digitally ascend.

5.2 Naskapi

Naskapi is a Cree language in the Algonquin family spoken in central Quebec MacKenzie and Jancewicz (1994), which UNESCO defines as *vulnerable* [13]. 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[14]), 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

[11] http://www.unesco.org/languages-atlas/en/atlasmap/language-iso-gla.html
[12] http://crubadan.org/languages/gd
[13] http://www.unesco.org/culture/languages-atlas/en/atlasmap/language-id-2354.html
[14] https://en.wikipedia.org/wiki/Innu

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.[15]

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.

5.2.1 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 [16].

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

[15]http://www12.statcan.gc.ca/census-recensement/2016/dp-pd/index-eng.cfm
[16]https://www.wycliffe.org/

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) [17] 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 [18]. Users can export as a PDF (among other file formats), or export words to an online interface known as Webonary [19]. This allows language workers to automatically create a useable, free dictionary for members of the community.

Naskapi uses the Inuit 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 [20], (free, open source software available on GitHub [21]) 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.

5.2.2 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 [22]. 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 indi-

[17] https://software.sil.org/fieldworks/
[18] https://github.com/sillsdev/FieldWorks
[19] https://www.webonary.org/configuring-the-dictionary-in-flex/
[20] https://keyman.com/
[21] https://github.com/keymanapp
[22] http://atlas-ling.ca/

cated 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 utilised to jump-start 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. Keyman 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.

6 Methods

6.1 Choosing a license

I'll give some recommendations on a license, both for individuals and for larger companies. I am not a lawyer, so this will be short and tempered.

6.2 Choosing repositories

I'll talk about my actual recommendations for storing code. I'll talk about how GitHub is a business, and its aims may not be aligned with researchers interested in long term archival, and similar concerns.

6.3 Sharing code without a platform

I'll outline a plan for peer-to-peer resource sharing, using IPFS (Benet, 2014) and other related tech. I'll mention a case study involving local indigenous communities in Guyana using peer-to-peer to track illegally logging on their land, and explain how this system could also be used for language development.[23]

[23] https://www.digital-democracy.org/

7 Discussion

Here, I want to drive home the point; how open source can help languages. Specifically, I will cover:

7.1 Why isn't more code open?

Finally, I'll go into a little detail on the question of why more hasn't been open sourced, and how to find open source resources.

7.2 How does open source demonstrably help?

I'll talk about use cases where open source has actually helped languages. This will include, for instance, NLTK case studies.

8 Future Work

Here, I'll talk about where to go next.

8.1 Beyond Wikipedia and Ethnologue

I'll talk about the shortcomings of both Wikipedia as a service, and Ethnologue as a provider of language data. Specifically, I want to draw attention to how Wikipedia treats its long-term contributors, and how Ethnologue charges exorbitant fees for using its data, and what we can do to improve this.

9 Conclusion

Here I will conclude with some closing remarks.

References

Benet, J. (2014). IPFS-content addressed, versioned, p2p file system. *arXiv preprint arXiv:1407.3561*.

Bernard, H. (1992). Preserving language diversity. *Human organization*, 51(1):82–89.

Bird, S. (2006). NLTK: the natural language toolkit. In *Proceedings of the COLING/ACL on Interactive presentation sessions*, pages 69–72. Association for Computational Linguistics.

Chiarcos, C., Hellmann, S., Nordhoff, S., Moran, S., Littauer, R., Eckle-Kohler, J., Gurevych, I., Hartmann, S., Matuschek, M., and Meyer, C. M. (2012). The open linguistics working group. In *LREC*, pages 3603–3610.

Chiarcos, C., Moran, S., Mendes, P., Nordhoff, S., and Littauer, R. (2013). Building a linked open data cloud of linguistic resources: Motivations and developments. In *The People's Web Meets NLP*, pages 315–348. Springer.

Clague, M. et al. (2009). Manx language revitalization and immersion education. *E-Keltoi: Journal of Interdisciplinary Celtic Studies*, 2:165–198.

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

Gil, D. (2009). What is Riau Indonesian? *Studies in Malay and Indonesian Linguistics*.

Grenoble, L. A. (2011). Language ecology and endangerment. In Austin, P. K. and Sallabank, J., editors, *The Cambridge Handbook of Endangered Languages*, pages 27–45. Cambridge University Press, Cambridge.

Hinton, L. (2001). Sleeping languages: Can they be awakened. *The green book of language revitalization in practice*, pages 413–417.

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

Krauss, M. (1992). The world's languages in crisis. *Language*, 68(1):4–10.

Krauwer, S. (2003). The basic language resource kit (BLARK) as the first milestone for the language resources roadmap. *Proceedings of SPECOM 2003*, pages 8–15.

Kroeber, T. and Robbins, R. (1973). *Ishi: Last of His Tribe*. Turtleback Books.

Leonard, W. (2004). The acquisition of miami: Findings from a field study. In *36th Annual Algonquian Conference, Madison, WI*.

Littauer, R. and Paterson III, H. (2016). Open source code serving endangered languages. In Soria, C., Pretorius, L., Declerck, T., Mariani, J., Scannell, K., and Wandl-Vogt, E., editors, *Proceedings of LREC 2016 Collaboration and Computing for Under-Resourced Languages: Towards an Alliance for Digital Language Diversity (CCURL) Workshop*, Portorož, Slovenia.

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

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

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.

Snyder, G. (2004). *The Practice of the Wild: Essays*. Shoemaker & Hoard.

Warner, N., Luna, Q., and Butler, L. (2007). Ethics and revitalization of dormant languages: The mutsun language. *Language Documentation & Conservation*, 1(1):1–1.

Wilson, G. N. (2008). The revitalization of the manx language and culture in an era of global change. In *Proceedings of the Third International Small Island Cultures Conference*, pages 74–81. Small Islands Cultures Research Initiative Sydney.

Yang, C., O'Grady, W., and Yang, S. (2017). Toward a linguistically realistic assessment of language vitality: The case of Jejueo. *Language Documentation and Conservation*, pages 103–113.