Saarland University
Department of Computational Linguistics

Master's Thesis

# Open Source Code and Low Resource Languages

*Author:*
Richard Littauer
Matriculation: 2539658

*Supervisors:*
Prof. Dr. Dietrich Klakow
Prof. Dr. Alexis Palmer

April 26, 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 selbstständig 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.

Richard Littauer

Montréal, April 26, 2018

## Acknowledgements

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. Jonathan Poitz provided formatting files for LaTeX. Fritz Van Deventer helped by suggesting nicer fonts. Many, many academics have helped advise me along the way towards this work: Bobbye Pernice, Stefan Thater, Ivana Kruijff-Korbayová and Hans Uszkoreit with their administrative assistance; Mike Rosner and Ray Fabri with a previous iteration on Maltese morphological parsing, which encouraged my interest in low resource languages; Christine Schreyer with discussions about constructed languages as low resource languages; Tyler Schnoebelen, Schuyler Erle, Robert Munro, and others from Idibon who advised on one iteration of this thesis; and finally, Alexis Palmer and Dietrich Klakow, who patiently advised me for years.
This work also draws heavily on an open source repository on GitHub, for which Gina Chiodo, Hugh Paterson III, Liling Tan, Ryan Txanson, Robert Forkel, Aidan Pine, Nick Heindl, Kevin Scannell, Sjur Moshagen, Waldir Pimenta, Joshua Olson, Edwin Ko, Arne Neumann, and Pablo Duboue are all contributors (in order of contributions) as of today, as well as the bots ReadmeCritic, greenkeeper[bot] (run by my friend, Gregor Martynus), and orthographic-pedant (run by Travis Hoppe).

Writing this paper involved using LaTeX[1] typeset with TeXShop[2]; Atom[3]; iTerm[4]; and Mac OSX 10.13[5], among a suite of other tools.

[1]https://www.latex-project.org/
[2]http://pages.uoregon.edu/koch/texshop/
[3]https://atom.io/
[4]https://iterm2.com/
[5]https://www.apple.com/macos/high-sierra/

# Contents

# List of Figures



# List of Tables

# 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.[6] It is estimated that less than 5% of the world's languages are 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.[7] 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 in any event, 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 spellcheckers, 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 (from here on, the 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.)

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

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

Another example would be 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.

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 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. In Section 4, I'll define what open source is, and talk about issues relevant to open source code for under-resourced languages. I'll then in Section 5 talk about the state of the open source ecosystem for low resource languages 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 6, 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 informed by original research, as I engaged in field research at the town where most Naskapi live and talked to linguists working on literacy efforts for this language. In Section 7, I'll discuss how open source can help low resource languages, and in Section 8 I'll expound further at a high level on what open source enables for linguists and language communities. Finally, in Section 9 and Section 10 I'll discuss future work, and offer some concluding remarks.

# 2 Low Resource Languages

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 describe these languages. 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 vitality. Finally, I'll mention the various different groups who work on and fund low resource development, and consider how they impact a language's development.

## 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.

All of these terms are controversial, and work within larger frameworks and ontologies. I'll cover some of these frameworks in Section 2.2 on metrics after giving this general overview of definitions.

### 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 biological species; 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 spoken examples of Guugu Yimithirr by first language speakers. The term is not completely 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, when used specifically within a terminological framework and not used generally to mean any language that is perceived to be in danger of extinction, 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, [8] although there are some struggling efforts to revive it (Hanks, 2017). On the other side of the northern Pacific, Haida has a similar amount of native speakers, but because of the amount of immersion programmes, government-funded schools, and new domains for the language, it is not considered moribund. An example of a new domain for Haida would be a recent motion picture filmed entirely in Haida with ethnically Haida actors who learned their lines from the elders. [9]

*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). Ethnologue defines dormant as a language which has no speakers, but there is still a community that attaches its ethnic identity to the language (Lewis and Simons, 2010).

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

[8] https://www.ethnologue.com/language/ain

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

of second language speakers until today, when there are now immersion programmes 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,[10] which was formerly a *literary* language (used mainly in relation to written texts) until revitalisation efforts began as a result of the creation of the Israeli 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 now submit that modern Hebrew is an imperfect descendant of historical Hebrew, which remains extinct, and that it reflects creolisation rather than language revitalisation (as Kornai (2013) does, citing Bickerton (2016); Izreel (2003)) 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 or lineage. These may be created to be logically resistant to ambiguity (such as Loglan or Lobjan (Okrent, 2009)); for a specific artistic purpose (such as Na'vi or Klingon, meant to be spoken by aliens in science fiction (Schreyer, 2015, 2011)); for scientific study, such as those used by evolutionary linguists for language games with participants to discern how language might have evolved (Scott-Phillips and Kirby, 2010); or such as used in the ubiquitous Wug test by scholars of language acquisition (Ratner and Menn, 2000)); or for political aims (such as Esperanto or Ido (Okrent, 2009)). Some of these may end up with thousands of speakers, including native speakers, and a huge sur-

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

plus of computational resources. For instance, Na'vi has a dictionary (Miller et al., 2018) that has been translated using computational tooling into over a dozen languages (including into Na'vi itself), and other dictionaries (Annis, 2018), grammars, spell checkers, and a morphological parser, Facebook translator,[11] and a Garmin audio file for navigation apps.[12] These languages are not normally considered as revitalised or dormant, but are instead mostly ignored or actively excluded (see Gibson (2016) for an example of this) 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.[13] 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).[14] 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[15] (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).[16] This has actually happened elsewhere, such as with Nisenan (Snyder, 2004). Another example might be Ishi, the last

[11] https://github.com/learnnavi
[12] https://learnnavi.org/media/
[13] https://gizmodo.com/5905595/the-atms-in-vatican-city-speak-latin
[14] https://spraakbanken.gu.se/swe/forskning/diabase
[15] https://wordhoardpress.com
[16] http://stories.schwa-fire.com/who_save_ayapaneco#chapter-113060

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, and is what Ishi consented to be called as a placeholder for his actual name (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 Official, *de facto*, *de jure*, majority, and minority languages

All of the former definitions were seen through the lens of language communities and vitality. However, there are other lenses through which languages as a whole can be viewed - for instance, politically and computationally.

Political definitions of language include *official* and *working* languages. Official languages are languages which are given a definitive status by a state, normally on the national level. On the supranational level (such as is the case with the EU or the ICC), they are generally termed working languages (which is different, in turn, from a *lingua franca*, which is a trade, bridge or link language used informally between groups who speak different languages themselves). These languages can be broken down into *de facto* an *de jure* languages - the latter are given legal status in the law, while the former do not have official legal status but are considered culturally and for most intents and purposes as the legal language. An example would be in the United States, where there is no *de jure* legal language, but the *de facto* language is English. This means that most resources are provided in English, and other languages are often ignored or not allocated resources by the law.

> These terms, as defined by Johnson (2013), distinguish policies from one another by virtue of their alignment between law and practice, respectively. Here, *de jure* policies are those disseminated in legal proclamations, typically being 'officially documented in writing' (p. 10). By contrast, *de facto* policy describes those policies that exist in *practice* [sic], crucially, without legal provenance or even *in spite* of existing *de jure* polices. (Hanks, 2017)

An example given by Hanks (2017) is the case of boarding schools in the United States and Canada for indigenous children, often forcibly removed from their home, where the *de jure* goal was to provide the children with a working knowledge of English, but the *de facto* result was that they were heavily discouraged (often through direct physical abuse to students who spoke in their language) from speaking their native tongues in the classroom or in the schools, with the result that many languages were directly endangered or lost. This has happened in many places, as well: for instance, Gaelic was forbidden in the classroom by English teachers, and children were beaten (for instance, slapped across the knuckles with a ruler) for using Gaelic.

Within a state, the proportion of population of speakers compared to the entire population generally determines wether a language is considered a *majority* or a *minority* language. Not all minority languages are endangered languages; for instance, Catalan, spoken by around nine million people in Catalonia and southern France, is not endangered, although it is a minority language and is not an official language of any country. There are arguments that it is the majority language for a stateless state. The same could be said of Tibetan, which is officially the minority language in a region of China, but is considered the be the majority language of the region of Tibet itself, which many view as its own state currently under illegal occupation (as with Hebrew and Israel, further political discussion is beyond the scope of this thesis.)

Some minority languages have legal status as minority languages. A good example would be in Canada, where minority languages in each province are given legal protection - for instance, English in Québec, where a majority of the speakers are Francophone, or French in Ontario, where the majority of the speakers are French. Sometimes languages with very small populations - such as indigenous languages spoken by First Nations communities in Canada - are given legal status, too, as is the case with Nunavut, a territory in Canada where two Inuit languages - Inuktitut and Inuinnaqtun - are granted legal status, although they are nationally minority languages, and although one of them, Inuinnaqtun, has less than 300 speakers and comprises only around 1% of the population of Nunavut. Another example would be Hawai'ian, which is the state language of Hawai'i since 1978, although it only has around 2000 native speakers, and is a minority language in Hawai'i.

### 2.1.3 Low resource, under resourced and incident languages

*Low resource* languages are languages which have fewer computational resources than any of the larger languages that dominate global discourse. There is no distinct cut-off for defining a low resource language versus a *high resource*, *resource-rich*, or just a *resourced* language. A *low resource* language can also be indiscriminately called a *under resourced* or *sparsely resourced* language, and occasionally can also be called a *non-central* language (Streiter et al., 2006). The disparity in resolved definitions reflects the focus of research, as generally researchers work with specific languages on computational models, and not on large databases where a precise definition is useful. Qualifiers are often included - for instance, Agić et al. (2015)'s paper, "If all you have is a bit of the Bible: Learning POS taggers for *truly* low-resource languages" (emphasis added). These qualifiers are generally not considered within a rigorous system of rank - for more on that, consider Section 2.2 on metrics below.

In the context of low resource languages, the majority of established work revolves around adapting existing systems from high resourced languages to low resource languages. In such a case, the *source* language is where the original system was originally trained or upon which it was built, while the *target* language is the language upon which the system is being used, tested, or adapted. These terms are largely context dependent. Similarly, *sparse* in particular is more often used to refer to a dataset, but can be used of a language when it is under resourced.

While hypothetically some languages could be defined as having no resources, there is no commonly used term such as 'resourceless'. In general, languages without writing systems fit in this category, and while it would potentially be interesting to train resources on audio-only vocabulary, this is generally not done computationally, but intensively by field linguists using specific tools such as dictionary applications or audio/video applications such as Praat (Boersma and Weenink, 2018), which allows you to view the waveforms for spoken corpora and annotate it. These resources - annotated (or not) corpora made by field linguists for a language - are, along with word lists and basic dictionaries, often the first resources for a given language, and are often not published but are accessibly only through corresponding with the linguist or team doing the work. A comparison with multimillion dollar projects such as Google Translate makes it clear that these languages would be considered under resourced.

Another couple of terms often used in this general context are *incident* or *surprise* languages. The latter is generally used for challenges, and was first used to describe the US Defense Advanced Research Projects Agency (DARPA) "Surprise Language Challenge", run by their Translingual Information Detection Extraction and Summarization (TIDES) programme in 2003. The challenge's goal was to see if a teams working on new languages they hadn't seen before (hence, 'surprise') could develop sufficiently useful resources and machine translation systems within a constrained period of time. (Oard, 2003) These sorts of challenges aren't limited to DARPA; for instance, there was a Workshop on statistical Machine Translation held at EMNLP 2011 (Callison-Burch et al., 2011). This workshop focused on a few tasks, one of which was based on the successful efforts by the Microsoft Translation team in 2010 to build a machine translation system for Haitian Creole that used SMS messages, after an earthquake there precipitated the need for a translation system between aid workers and spekaers of Haitian Creole, previously a low resource language (Lewis, 2010; Lewis et al., 2011). Haitian Creole, here, would be an *incident* language.

### 2.1.4 Computer languages

A *computer* or *computational* language is a formalised language used to communicate instructions to a machine. There are a large variety of names and variants, and the definition here may be construed as insufficient. For the purposes of this thesis, a computer language is for talking to a machine, and is demonstrably different than a human language, which is generally used for communicating with humans. This definition is important only in so much as it helps clarify that we are talking about human languages when we mean low resource or endangered languages, not computer languages. The relevancy, usage, or status of computer languages is largely irrelevant here, unless it touches on resources used on human languages. For instance, any grammar written in COBOL, a sixty year old language, may be less accessible to open source coders who write primarily in Python or JavaScript, two popular languages used on the web and in the FLOSS ecosystem today. This type of situation will be covered in more depth in Section 4.4.

Other terms used in exploring the theory of language, semiotics, or formal language theory - such as context-free or recursively-enumerable languages - are outside of the scope of this thesis unless they touch on human and low resource languages directly in some tangible way.

## 2.2 Metrics for language vitality

Language health or vitality is a topic of increasing scholarship and interest. Superficially, it makes sense to use a similar system to classify languages as one would classify biological species, using the metrics defined by the International Union for Conservation of Nature (IUCN).[17] They have nine levels of classification: Extinct, Extinct in the Wild, Critically Endangered, Endangered, Vulnerable, Near Threatened, Least Concern, Data Deficient and Not Evaluated. However, the system is not directly transferable - how would a dormant language be classified? One can see immediately that there is a need for a language-specific rating system.

It is perhaps unsurprising that there are various metrics which can be used to classify the health of a language and its community. In this section, I'll

[17] http://www.iucnredlist.org/

Summary of Fishman's GIDS

| GIDS | (adapted from Fishman 1991) |
|---|---|
| **LEVEL** | **DESCRIPTION** |
| 1 | The language is used in education, work, mass media, government at the nationwide level |
| 2 | The language is used for local and regional mass media and governmental services |
| 3 | The language is used for local and regional work by both insiders and outsiders |
| 4 | Literacy in the language is transmitted through education |
| 5 | The language is used orally by all generations and is effectively used in written form throughout the community |
| 6 | The language is used orally by all generations and is being learned by children as their first language |
| 7 | The child-bearing generation knows the language well enough to use it with their elders but is not transmitting it to their children |
| 8 | The only remaining speakers of the language are members of the grandparent generation |

Figure 1: A summary of GIDS from (Lewis and Simons, 2010, 105)

explain these metrics in detail, focusing on the GIDS, EGIDS, UNESCO, and LEI measurements, as suggested by Yang et al. (2017) as the main players in the field.

### 2.2.1 The Graded Intergenerational Disruption Scale (GIDS)

The Graded Intergenerational Disruption Scale, developed by Fishman (1991), is the earliest and most well known of the scales. It rates languages based on their domains of use, and on the amount of passing on and education which continues to the next generation through the parents. Figure 1 summarizes the different stages. Generally, as a language progresses and becomes more imperilled, it ends up further down the list. As a language ceases to be used in one domain, it becomes less likely that it will in the future, and more likely that parents will consider the language to be less useful than another. Over time, this causes the language to lose speakers (although the process is not inevitable; for examples, language policy in Quebec helped secure and revitalise the language over the past half century (Bourhis, 2001)).

### 2.2.2 The UNESCO measurement scale

Chronologically, the UNESCO rating was the next major scale in the field. The United Nations Educational, Scientific and Cultural Organization (UNESCO) is a specialised agency of the United Nations. In 2001, at the 31st Session of the UNESCO General Conference, they officially recognised that biodiversity, cultural diversity, and linguistic diversity are related. This viewpoint is relatively recent, and reflects increasing appreciation that culturally diverse regions tend to collocate with biodiverse regions, and that saving diversity implies saving both (Nettle and Romaine, 2000; Maffi, 2001; Anderson and Harrison, 2006; Krauss, 2007b; Gorenflo et al., 2012) (as discussed explicitly in Maffi et al. (2001), of which all of the authors were also members of the UNESCO Ad Hoc Expert Group on Endangered Languages). Encouragingly, UNESCO clarified at this event that sustaining and encouraging linguistic diversity lies within their charter.

In their publication from that conference, Brenzinger et al. (2003) lay out nine different metrics for measuring language vitality: six evaluate the vitality, two language attitudes, and one related to urgency of documentation. The UNESCO system is rigorous in its refusal to apply a single score to a language, as that would smooth over the complexities of language usage. The six factors for vitality are: intergenerational language transmission (as with GIDS), absolute number of speakers, proportion of speakers within the total population, trends in existing language domains, response to new domains and media, and materials for language education and literacy.

For each of these, they further break rating down into categories. For instance, when regarding intergenerational language transmission, they specify six different possible ratings - safe, unsafe, definitively endangered, severely endangered, critically endangered, and extinct - and equate each rating with a score from 5 to 0. Here one of the primary issues with the UNESCO rating can be seen (as pointed out by Lewis and Simons (2010)) - namely, that 'safe' is an incredibly large category that needs more fine-grained categories, as it would account for any GIDS-rated language above Level 6.

The three other factors they consider are: governmental and institutional language attitudes and policies including official status and use; community members' attitudes toward their own language; and the amount and quality of documentation. Each of these is also rated on a null to five scale. For documentation, only a superlative rating of five would be considered to be more than low-resourced, as a four rating would be given to a language where "There are one good grammar and a number of adequate grammars, dictionaries, texts, literature, and occasionally updated everyday media; adequate annotated high-quality audio and video recordings." Although useful for lin-

**Estimated Degree of Endangerment and Urgency for Documentation:**
**the case of three Venezuelan Indigenous Languages**

| Factors | Languages | | |
|---|---|---|---|
| | Mapoyo | Kari'ña | Sanɨma |
| Intergenerational Language Transmission | 0 | 2 | 5 |
| Absolute Number of Speakers | (7) | 650 | 2500 |
| Proportion of Speakers within the Total Population | 1 | 2 | 5 |
| Trends in Existing Language Domains | 0 | 2 | 5 |
| Response to New Domains and Media | 0 | 1 | --- |
| Materials for Language Education and Literacy | 1 | 3 | 0 |
| Governmental & Institutional Language Attitudes and Policies including Official Status & Use | 5 | 5 | 5 |
| Community Members' Attitudes toward Their Own Language | 2 | 3 | 5 |
| Amount and Quality of Documentation | 1 | 3 | 1 |

Figure 2: The UNESCO grading for three languages (Brenzinger et al., 2003, 23)

guists wishing to work in the language, this may not be enough for useful analysis and use by computational linguists.

In Figure 2, an example rating using this system, from the appendix of Brenzinger et al. (2003) itself, is included to get some grasp of how these grades work in parallel.

Importantly, UNESCO clarifies that it does not suggest using one metric over another, and that adding up the numbers in the scales - however easy that might seem, as all of the measurements except speaking population are scalar and hold the same number of levels - would be insufficient and not ideal. "**Languages cannot be assessed simply by adding the numbers**; we therefore suggest such simple addition *not be done* [sic]."

The UNESCO ratings for languages are listed in the *UNESCO Atlas of the World's Languages in Danger* (UNESCO, 2014).

### 2.2.3 The Extended GIDS (EGIDS)

Lewis and Simons, the authors of Ethnologue (Lewis et al., 2009)[18], pointed out some of the issues with GIDS which necessitate the creation of a new standard, and which could also eclipse or inform the UNESCO rating (Lewis and Simons, 2010). First, the levels are static, and don't account for directionality on the part of a language community up or down the strata. Second, there are language types which aren't included - for instance, there isn't a supranational level for extremely well-off languages, nor is there are level for extinct or dormant languages. Thirdly, GIDS focuses on intergenerational disruption in Level 5 and down, but in Level 4 and higher it focuses more on institutions, and this isn't accounted for well enough in the framework, which primarily focuses on parents as being the primary agents of language transmissions. Finally, the lower levels are not granular enough to cover the many complexities needed for language revitalisation groups.

EGIDS - the Expanded GIDS - serves these needs by providing more granular definitions. It also draws on the extensive knowledge of languages and their usage provided not only by Ethnologue, but also by the UNESCO Atlas and the community of linguists working with the Summer Institute of Linguistics (SIL), who fund and published Ethnologue. Figure 1 shows the main categories, taken from the Ethnologue website.[19] The table has been updated since Lewis and Simons (2010), in particular to also account for signed languages (Bickford et al., 2015). The addition of a Level 0 and two levels beneath the scale are evident, as well as more granularity in the GIDS scale, such as can be seen with Level 6, which now has two levels, Level 6a Vigorous and Level 6b Threatened.

Lewis and Simons (2010) also add another set of EGID levels which can be used to rate a language which is ascending in domains due to revitalisation efforts, which Figure 3 shows. This is generally useful, although it does suggest that a language uniformly descends or ascends, which may not be the case. The authors also spend time describing how to identify a language and decide which level best describes it.

They end with a quote from Fishman (2001), which explains further the purpose of EGIDS:

> Thus, any theory and practice of assistance to threatened languages-whether the threat be a threat to their very lives, on the one hand, or a much less serious functional threat, on the other hand-must begin with a model of the functional diversification of languages.

[18] Also a website available at https://www.ethnologue.com/

[19] https://www.ethnologue.com/about/language-status

| Level | Label | Description |
|---|---|---|
| 0 | International | The language is widely used between nations in trade, knowledge exchange, and international policy. |
| 1 | National | The language is used in education, work, mass media, and government at the national level. |
| 2 | Provincial | The language is used in education, work, mass media, and government within major administrative subdivisions of a nation. |
| 3 | Wider | Communication The language is used in work and mass media without official status to transcend language differences across a region. |
| 4 | Educational | The language is in vigorous use, with standardization and literature being sustained through a widespread system of institutionally supported education. |
| 5 | Developing | The language is in vigorous use, with literature in a standardized form being used by some though this is not yet widespread or sustainable. |
| 6a | Vigorous | The language is used for face-to-face communication by all generations and the situation is sustainable. |
| 6b | Threatened | The language is used for face-to-face communication within all generations, but it is losing users. |
| 7 | Shifting | The child-bearing generation can use the language among themselves, but it is not being transmitted to children. |
| 8a | Moribund | The only remaining active users of the language are members of the grandparent generation and older. |
| 8b | Nearly | Extinct The only remaining users of the language are members of the grandparent generation or older who have little opportunity to use the language. |
| 9 | Dormant | The language serves as a reminder of heritage identity for an ethnic community, but no one has more than symbolic proficiency. |
| 10 | Extinct | The language is no longer used and no one retains a sense of ethnic identity associated with the language. |

Table 1: Expanded Graded Intergenerational Disruption Scale (Lewis et al., 2018)

Revitalization EGIDS Levels

| | | |
|---|---|---|
| 6a | Vigorous | The language is used orally by all generations and is being learned at home by all children as their first language. |
| 6b | Re-established | Some members of a third generation of children are acquiring the language in the home with the result that an unbroken chain of intergenerational transmission has been re-established among all living generations. |
| 7 | Revitalized | A second generation of children are acquiring the language from their parents who also acquired the language in the home. Language transmission takes place in home and community. |
| 8a | Reawakened | Children are acquiring the language in community and some home settings and are increasingly able to use the language orally for some day-to-day communicative needs. |
| 8b | Reintroduced | Adults of the parent generation are reconstructing and reintroducing their language for everyday social interaction. |
| 9 | Rediscovered | Adults are rediscovering their language for symbolic and identificational purposes. |

Figure 3: A summary of EGIDS ascending levels (Lewis and Simons, 2010, 117)

> If analysts can appropriately identify the functions that are endangered as a result of the impact of stronger languages and cultures on weaker ones, then it may become easier to recommend which therapeutic steps must be undertaken in order to counteract any injurious impact that occurs. The purpose of our analyses must be to understand, limit and rectify the societal loss of functionality in the weaker language when two languages interact and compete for the same functions within the same ethnocultural community and to differentiate between life-threatening and non-life-threatening losses.

#### 2.2.4 The Language Endangerment Index (LEI)

Just as EGIDS expanded on GIDS, the Language Endangerment Index (LEI) was formed to resolve some of the issues with EGIDS, as well as to respond to GIDS, the UNESCO rating, and the rating in Krauss (2007a), another metric which focused almost exclusively on different ages of speakers and classified all languages with children speakers as 'stable', and all with over a million speakers as 'safe'. Lee and Van Way (2016) describe LEI for its use in The Catalogue of Endangered Languages (ELCat), part of the Google-powered Endangered Languages Project.[20] The project isn't only sponsored by Google, but

[20] endangeredlanguages.com

also by an American governmental National Science Foundation (NSF) grant[21], and is an ambitious project (like UNESCO and Ethnologue) to catalogue all languages and to provide specific metrics of language vitality.

The authors, in describing LEI, go into detail explaining how previous classifications, while they "highlight[s] the immensity of the problem at hand", can not easily apply to certain languages, and that these exceptions are critical to understanding whether the metrics are useful as opposed to being exceptions which prove the rule. Unlike the other papers, they explicitly mention some languages. For instance, they mention how Dwyer (2012) points out that Wutun, a Chinese-Tibetan-Mongolic language, is endangered due to a variety of factors, even if transgenerational transmission is not at risk - thus, GIDS or EGIDS may not satisfactorily categorise the language. A similar case could be made for Naskapi (see Section 6.2 for more on this).

The LEI uses four factors: intergenerational transmission, absolute number of speakers, speaker number trends (whether increasing or decreasing), and domains of use. Each of these is rated, like the UNESCO rating on a scale from null to five - however, unlike UNESCO, they add these numbers up to produce a single rating. The higher it is, the more likely the language is endangered. The scales are also somewhat different; for instance, number of speakers runs on orders of magnitude, with 100,000 being the top bound for a safe language (and not a million, like in Krauss (2007a)).

### 2.2.5 A response to qualitative metrics

Lee and Van Way (2016) point out further issues with some of the other assessments - most notably that "while the UNESCO framework is broad and its factors comprehensive, it does not give an overall vitality score to the language being assessed, making it difficult to compare accurately across different language" and that "while an assessment of the type and quality of documentation is doubtlessly important because it helps indicate the potential for revitalization and the urgency of further research, it is not clear that the type and quality of documentation directly affects the vitality of a language." These two points are interesting, because they reflect how the situation of Lee and Van Way (2016) influences their judgement and their decision in making LEI at all. The authors were aware that they were being overtly quantitative in their approach:

> Some may prefer a more nuanced examination of a language's vitality, with the view that the factors responsible for a language's endangerment are too complex to be compared across languages.

[21] https://www.nsf.gov/awardsearch/showAward?AWD_ID=1058096

> Researchers of this view would rally against quantitative measures, stating that quantitative measures can hardly be accurate. ... ELCat researchers, while sympathetic to these points of view, maintain that without understanding and investigating fundamental common factors responsible for language endangerment, very little progress will be made in assessing language vitality and, consequently, less can be done to help communities preserve their languages. ELCat strikes a balance between these different perspectives.

As Grenoble (2016) points out, this misses the point of qualitative rebuttals, by claiming that accuracy is the most salient argument. It doesn't have to be, as there are more pressing concerns. For instance, all of the metrics were built on the assumptions that quantifying language endangerment is useful, and that assessment directly leads to empowering communities to revitalise their language. Neither of these are directly backed up by empirical research (Grenoble, 2016). More pressingly, language itself is not indisputably something that is countable or measurable, and to think so is to reflect Western, modernist ideologies surrounding language, viewing a language as a distinct entity which is formalised in writing and education. Language could be viewed alternatively as inextricable from the speaker and the utterance, and this view is more likely to be taken by language groups which view themselves as separate from a nation-state or an ethnographic group (Bodó et al., 2017). To view language otherwise is to confine language to a countable, commodifiable entity in a post-colonial sense, which affects how the language is viewed and can have real effects on language communities. Even viewing linguistic biodiversity as something to be 'saved' raises ideological concerns, as Haspelmath (one of the main editors of Dryer and Haspelmath (2013)) notes.[22] Indeed, post-colonial attitudes towards language endangerment may be endemic; Newman (1998) certainly suggests that non-Western linguists cannot adequately document or revitalise their own languages without Western training, which presupposes that to be an informed researcher one must also conform to Western ideologies. Against this backdrop, Lee and Van Way (2016)'s claims that accuracy is something that can be attained seems to miss the mark; rather, the canonical approach to metrics is in itself a flawed approach that carries with it certain uncomfortable presumptions.

This thesis cannot hope to resolve these issues, nor is it meant to be an overview of the field of language vitality or endangerment as ideology. However, it is worth noting that metrics of language vitality do not exist in a vacuum, and that documentation and computational efforts are also a part of

[22] dlc.hypotheses.org/195

wider questions. Literacy is not a domain into which a language has to ascend to be seen as 'safe' or 'vital', and technological progress should not be viewed independently of an assessment of what exactly progress is.

Some things can be done, however. Terminologically, 'low-resource' is intentionally somewhat neutral, as compared to 'minority', 'endangered', or other terms that reflect Western viewpoints. Similarly, using the term *language vitality* as opposed to *language endangerment* "represents a significant shift in the representation of attitudes toward the rhetoric of indigenous languages to one away from dire predictions about endangerment to action-oriented attitudes about vitality and sustainability (Grenoble, 2016)." These terms will be used for the rest of this paper, and any statements about resource development should be viewed as part of a narrower question of digital development (in the sense of building resources) for a specific, almost naïvely countable view of language, unless otherwise specified (as in Section 6).

## 2.3 Digital presence

Digital presence, briefly alluded to previously, can be thought of as the amount of language data available through digital sources. A looser definition could be 'the amount of written text on the web', but this would miss out on several important considerations. First, linguistic data does not have to be written to be digitally encoded; videos and audio data are both examples of digital content which is often digitally encoded. In some cases, pictures are also relevant, especially for signed languages or for examples of written text, such as in the millions of scans of papyrus from the Egyptian city of Oxyrhynchus, which are being translated using a crowd-sourced system by thousands of volunteers (Williams et al., 2014), or for other language mediums, such as the khipu knot system used by the pre-Columbian Incan civilisation (Quilter and Urton, 2002). Secondly, the web (hereafter meant to refer to the World Wide Web) is not the only corpus of knowledge, nor is it the only network through which data can be accessed. Trivial example of other corpuses would be local files collected by individual field researchers that are backed up on hard drives; a similar example of another network would be a local area network in offline areas, or a university intranet.

However, the digital sphere can best be thought of schematically as a new domain for language use, and it is overwhelmingly today represented on the web. Ten years ago, it was fashionable to include references to the web "as a a corpus" (as Scannell (2007), for instance, cited Resnik (1999); Ghani et al. (2001); Kilgarriff and Grefenstette (2001), although the latter two were in reference to low-resource languages); today, it is more common to cite studies on digital natives such as the 20,000 citation-strong Prensky (2001) paper,[23] or to assume that the web, and occasionally phone networks, are the main locations for digital communication. The web is ubiquitous; not only are more than half of the global population connected to the internet,[24] but the internet, in developed countries, is used for all levels of communication, such as education, work, mass media, and in the home and local communities. Digital presence, then, is functionally the amount of usage on the web.

### 2.3.1 Finding resources

There are several resources which can be used to judge the amount of corpora for a language on the web, outside of papers defining metrics to judge these languages and to state whether they are endangered or thriving. The main resource for low resource languages is almost certainly the Crúbadán project, developed by Scannell (2007).[25] This is a massive crawler which looks for documents with trigram frequencies for particular languages by checking against a seed corpus for under-resourced languages developed from Wikipedia, the Jehovah's Witness translations, and translations of the UN Declaration of Human Rights (UNHR). It is often the only corpus for a low resource language on the web, as is the case with Naskapi (see Section 6.2).

Often, a translated Bible is the next best place to look for digital content. Biblical translations are so common as a first resource that there is a body of research that uses partial or full translations of the Bible for training NLP systems as a result (Chew et al., 2006; Agić et al., 2015). However, finding the bible or UNHR is often difficult. In these cases, it is best to look for aggregators of data. There are large projects which hold resources for linguists - for more, see Section 3.1. However, these resources aren't always directly reflective of a language's digital presence, but rather of the scope of resources available to computational linguists and natural language processing experts. They satisfy a different need, and tools such as Perseus[26] might show that there is work done on Latin, but it doesn't mean that there is a large Latin-speaking community that could be measured. Instead, organic corpora - such as collected from the web by Crúbadán - are most likely the best ways of measuring a language's foothold on the web.

[23]This number is from Google Scholar (https://scholar.google.com) accessed April 9, 2018.
[24]https://www.internetworldstats.com/stats.htm
[25]http://crubadan.org/
[26]http://www.perseus.tufts.edu/hopper/

### 2.3.2 Metrics for digital presence

Kornai (2013) outlined the first major metric for describing digital presence for a language. These metrics are needed because normal metrics aren't directly transferable to digital presence, as digital linguistic data is decoupled from speakers (it can survive beyond them), and because the digital domain is only one of a variety of domains for language usage. He divided languages into four possible categories: *Thriving*, *Vital*, *Heritage*, and *Still*. These can be thought of as a gradient, with digital ascent being the process of a language moving up the scale. Only 16 languages would be considered Thriving, all of which would be rated at 1 or higher on the EGIDs scale. Vital languages are those which may be in danger in the next hundred years, or show few signs of digital ascent - but they have a large population of speakers and at least some resources, such as a Bible or the UNHR; Heritage languages are dead or historic languages such as Latin which have large online presences that do not relate directly to a living language community; and Still languages show little to no presence on the web at all (although note that this does not mean that they endangered or moribund outside of the web.)

Kornai looks at five confluent factors; demographics, prestige, the identity function of the language, the level of software support, and Wikipedia presence for a language. Demographics and community size can be gathered by doing a quantitative analysis of all public data available in a language on the web, and by using this data size as a proxy for the amount of speakers of a language using the digital space. This has obvious limits, which Kornai points out, in that the data may not accurately reflect the amount of users, in that it is limited to public data accessible by researchers, and in that it doesn't give an accurate representation of passive consumption of multilingual data. It would be worth adding that this also doesn't give an accurate count of multilingual usage of a language. Prestige is an obvious factor for digital ascent; when a language community views one language more highly than another, it is more likely to create digital content in one than the other, regardless of social policies and to some extent speaker populations. Identity function relates largely to certain historical languages, like Latin and Classical Chinese, which have large corpora online but should not be considered in the same grouping as more vibrant, living languages.

Software support as a factor in digital presence could be identified with a variety of different metrics. Kornai lists various stages for a language on the road of digital ascent. First, localisation of internalisation (often expressed using the shorthand l10n or i18n, where the numbers refer to the length of the words) of the language script is the major milestone that separates languages which are ascending from still languages. While many scripts use the more common Roman, CJK, Cyrllic, or Arabic alphabets, there are hundreds which do not, and these languages have specific Unicode considerations which need to be met for the language to be used adequately. The next step would be word-level tools, such as dictionaries, stemmers, and spellcheckers - all of which depend, at some point, on standardisation (but not s13n) of the language. Finally, sentence level tools such as automatic translators can be used. Regarding support, the question of a language's status is straightforward: is there language support for an operating system provided by Apple or Microsoft? If so, then it is likely that the language is thriving or vital. If not, there is almost zero chance of it being so. Kornai also used the Crúbadán Project, UHDR and biblical presence, and presence on Omniglot and OLAC.

Kornai found that Wikipedia was a good example of a first port-of-call for language speakers on the web, and that it could be used to show that a language was "crossing the digital divide."

> The reason is that children, as soon as they start using computers for anything beyond gaming, become aware of Wikipedia, which offers a highly supportive environment of like-minded users, and lets everyone pursue a goal, summarizing human knowledge, that many find not just attractive, but in fact instrumental for establishing their language and culture in the digital realm. To summarize a key result of this study in advance: *No wikipedia, no ascent* [sic]. (Kornai, 2013)

Ultimately, Kornai found that the best indicator of a language's digital presence was their EGIDS rating. "The next best set of features indicated the quality of the wikipedia, followed by the number of L1 speakers, the size of the Crúbadán crawl, the existence of FLOSS spellcheckers, and the number of online texts listed in OLAC." (Kornai, 2013, 6) Overall, only 5% of the world's languages were seen as digitally ascending; like most results from this field, an increasingly dire statistic. As Kornai (2013, 10) writes:

> Unfortunately, at a practical level heritage projects (including wikipedia incubators) are haphazard, with no systematic programmes of documentation. Resources are often squandered, both in the EU and outside, on feel-good revitalization efforts that make no sense in light of the preëxisting functional loss and economic incentives that work against language diversity (Ginsburgh and Weber, 2011).

However, others have noted that Kornai may have been early in his predictions that most languages will not digitally ascend (Gibson, 2016).

In a follow-up paper, (Kornai, 2015) proposed adding a single number scale to assess digital ascent: "For the assessment we propose a simple log-linear formula that derives a single number *D* (digital vitality index) as a weighted sum of well-understood components such as the EGIDS ranking, (log) number of L1 speakers, (log) size of wikipedia, adjusted for quality, (log) crawl size, the existence of FLOSS spellcheckers, etc." The EGIDS ranking was considered objective, given that SIL linguists are generally interested in longer term work with communities as opposed to relatively short-lived or quantitative studies done by computational linguists. This log-linear formula was innovative for cleaning wikipedia, in particular, as it removes the likelihood of large wikipedias built by hobbyists with bots as being indicative of large language communities.

Gibson (2016) extends Kornai by adding two separate statuses for languages - that of *Emergent* and *Latent*. Emergent languages are those where there is data, but it is privately hidden in messaging applications or cellphone usage, and unlikely to be accessible by the crawlers and corpora agglomeration tools used in Kornai (2013). These would be identified by researchers in the field, and do not need to have locale or i18n setups before inception. Gibson cites Arabizi (as noted by Darwish (2013)), where numbers are used for sounds not present in standard Arabic, as an example; another might be the use of a forward slash to denote accents in early Irish Gaelic forums, as noted by Scannell (2007). Latent languages are languages which meet the following criteria: "stable intergenerational transmission of the language, an available model of writing the language, the availability of appropriate technology and infrastructure (internet, mobile phone coverage), fonts in which to write the language in the desired script, and communal desire to see the language used digitally." If all of these are met, then the language could ascend beyond still into vital. Such languages would be admittedly impossible to find by measurements, but this category would be helpful for linguists working in the field to determine how to best work with the language community to help bootstrap language development. Gibson also redefined *Still*, which Kornai (2013) had marked as languages which are 'unable' to ascend, while here they are merely 'unlikely'.

A more recent metric was also introduced in a draft by Soria et al. (2017), for the purposes of helping digital language planning for the EU, as part of the Digital Language Diversity Project.[27] Their scale has the following states: *Pre-digital*, *Dormant*, *Emergent*, *Developing*, *Vital*, and *Thriving*. Like Gibson, they exclude Kornai's *Heritage* status (oddly noting that Gibson also included it, which he hadn't for the same grounds), without sufficient explanation as to why dead languages are not relevant when there are communities based around them, some of which are communities with thousands of L2 speak-

[27] http://www.dldp.eu/content/reports-digital-language-diversity-europe

| Indicator | |
|---|---|
| 1. Evidence of connectivity<br>2. Digital literacy<br>3. Internet penetration or digital population size[5]<br>4. Character/script encoding<br>5. Availability of language resources | *digital capacity* |
| 6. Use for e-communication<br>7. Use on social media<br>8. Availability of Internet media<br>9. Wikipedia | *digital presence and use* |
| 10. Available Internet services<br>11. Localised social networks<br>12. Localised software<br>13. Machine translation tools/services<br>15. Dedicated Internet top-level domain | *digital performance* |

Figure 4: Indicators of digital vitality (Soria et al., 2017, 6)

ers. Dormant would be equitable to Latent, while Pre-digital would apply to languages without internet or cell connectivity for the speaking population. Emergent through Thriving are largely matters of scale. While Kornai used proxies for the five factors he mentioned, Soria et al. note that such factors are difficult to quantify; they remedy this by focusing on three indicators: "a group pertaining to a language digital *capacity* [sic], a group related to a language digital *presence and use*, and a group related to a language digital *performance*." (Soria et al., 2017, 5) An example of how these are used can be seen in Figure 4.

Soria et al. (2017) go into depth about each of these factors. As an example, for localised software, they propose the following scale in Table 2. They explain, for each scale, how to find information - for instance, they suggest asking local researchers and community members about the usage of "Windows, Mac OS X, Linux, Android, iOS, Microsoft Office, LibreOffice, Firefox, Chrome, Internet Explorer, Thunderbird, Adobe Creative Suite, Gimp" for judging localised software. However, they do not show metrics on any languages judged according to this scale, and they don't make it clear whether or not the different metrics ought to be summed to come up with a single number (an issue which Lee and Van Way (2016) raised with the UNESCO rating). In conclusion, while this is an interesting and in-depth metric, its wider applicability is not clear.

Each of these metrics suffers from growing pains. For instance, there is no metric as of yet which ranks English in its own category - something which

| Label | Grade | Localised software |
|---|---|---|
| none | 2 | Neither operating system nor general purpose software localised in the language |
| limited | 3 | At least one operating system (either desktop or mobile, either open or commercial) localised in the language |
| medium | 4 | At least one desktop and one mobile operating system (either open or commercial) + some general purpose software (a word processor and a browser) localised in the language |
| strong | 5 | Most used operating systems and general purpose software localised in the language; some specific purpose application software localised. |
| advanced | 6 | Main operating systems and application software localised in the language. |

Table 2: Scale for Localised Software (Soria et al., 2017, 21)

was seen as a large enough issue to cause the EGIDS authors to add another null ranking for supranational languages. As well, there hasn't been an integrated approach looking at quantitative and qualitative measurements together. The most substantial work on this has been Kornai's team, which has worked with funding from SIL International on a Digital Language Vitality database.[28] However, the future for this work as, at this moment, unclear (Kornai in personal communications, 2018).

[28]https://hlt.bme.hu/en/projects/lingvit

# 3 Resources

It makes sense at this point (if not earlier), to discuss what language resources are. There are two main types of resources: corpora and tools which act on corpora. They are inextricably linked, but the approaches towards building, archiving, and using either differ. This section seeks to answer one question: what resources are needed to take a language from no resources, to a thriving language with a large digital presence?

For digital vitalisation, Kornai (2015) proposes working on a pyramid approach: first build a corpus with active and engaged speakers, then l10n and i18n support; then word-level tooling such as spell checkers and morphological analysers; phrase and sentence level tooling such as parsers; and finally speech and character recognition and machine translation. This, in general, follows how most language development progresses. However, a finder grained understanding of the tools would be illuminating. While an exposition of all possible natural language processing tools is beyond the scope of this thesis, it is worth going into some depth about some of them.

It is worth noting here that there are different groups which work on each of the stages of language development. Abstractly, these could be defined as language communities and linguists, and the fields of computational linguistics and natural language processing (NLP). The first group are those - often not computational linguists by training or NLP researchers - who want their own language or the language they are studying to exist digitally and in some form. The initial step is generally to adopt any language script, whether preëxisting or ready-made for the language by linguists (for examples of this, see the Endangered Alphabets Project[29]) into Unicode, a standard for consistent character representation.[30] There are linguistic research groups that focus on this problem; for instance, the Script Encoding Initiative at Berkeley.[31]

Some of the people involved in this process may be computational linguists. Bender (2016) makes a distinction between the fields of computational linguistics and NLP: "computational linguistics is used to describe research interested in answering linguistic questions using computational methodology, while natural language processing describes research on automatic processing of human language for practical applications." It should be clear here that computational linguistics is a subfield of linguistics, and that the two are not always in sync, as for instance Kay (1997) points out when discussing improving machine translation (ML) by using informed linguists. Bender and Good (2010); Bender (2016)

[29]http://endangeredalphabets.com/
[30]https://unicode.org/
[31]http://linguistics.berkeley.edu/sei/index.html

goes further, suggesting that understanding language typology can drastically help with multilingual NLP. Many experts in NLP would not consider themselves computational linguists, but developers, just as many language developers would not consider themselves linguists. While navigating the field or looking at resources, it is important to keep these distinctions in mind, as they inform narratives concerning resource generation, scope, and efforts.

## 3.1 Resource Aggregators

I've already mentioned that Crúbadán (Scannell, 2007) is a good location to find monolingual texts from the web; however, this is but one of an almost infinite amount of corpora that might be of use to linguists, language activists, and to NLP practitioners. To find other resources can be an overwhelming task. To help solve this issue, there are a non-trivial number of large organisations and databases where it is possible to find resources - dictionaries, academic references, and occasionally software - on low resource languages. UNESCO (2011) for instance itemises hundreds of such resources. To give more of an idea of what these resources are like, here are some major examples:

- The Unicode Common Local Data Repository (CLDR) "provides key building blocks for software to support the world's languages, with the largest and most extensive standard repository of locale data available."[32] There are dozens of scripts available in Unicode.[33]

- The Endangered Languages Project (ELP), described above and in Lee and Van Way (2016) and online[34] has information on many under resourced languages.

- Ethnologue, which is both a book (Lewis et al., 2009) and an online resource,[35] is the most comprehensive resource describing the world's languages, such as population size and the general geographic locations of speakers. It is published by SIL International, an evangelical Christian non-profit organisation, and has proprietary paywalls for repeated access to content. Many SIL entries for specific languages include academic references.

- Glottolog[36] is an open source alternative to Ethnologue, developed at the Max Planck Institute for Evolutionary Anthropology. It has over 180,000

[32]http://cldr.unicode.org/

[33]https://www.unicode.org/standard/supported.html

[34]http://www.endangeredlanguages.com/

[35]https://www.ethnologue.com/

[36]http://glottolog.org/

references, with information on over eight thousand languages. (Hammarström et al., 2015)

- Omniglot, "the online encyclopaedia of writing systems and languages",[37] contains around writing information for around a thousand languages. (Ager, 2018)

- The Online Database of Interlinear Text (ODIN)[38] is a multilingual repository of annotated language data for 1274 languages.[39] The database is formed by crawling scholarly articles on the web and looking for interlinear glossed text (IGT), an industry standard for displaying corpora in academic linguistics by displaying the original datum, a morphosyntactic gloss, and a translation. These data are not massive, but they are useful in particular for training algorithms on structured data. As well, "ODIN was developed as part of the greater effort within the GOLD Community of Practice (Farrar and Lewis, 2007) and the Electronic Metastructure for Endangered Languages Data efforts (EMELD)[40], whose goals are to promote best practice standards and software, specifically those that facilitate interoperation over disparate sets of linguistic data." (Lewis and Xia, 2010)

- The Open Language Archives Community (OLAC), a worldwide virtual library of language resources (Simons and Bird, 2003).[41]

- Wikipedia,[42] "the largest and most popular general reference work on the Internet" (contributors, 2018) has a nontrivial amount of articles on low-resource languages, many of which have references themselves to Scholarly work. Kornai (2013), among others, notes that Wikipedia is one of the first ports-of-call for new language communities, and while it is not a precondition for having corpora on the web, it is a *sine qua non* for digital vitalisation. Thus Wikipedia has two purposes; documenting the language and its community (for instance, in the Naskapi Language article[43]), and providing a space for corpus development in the target language itself.

[37] http://omniglot.com
[38] http://odin.linguistlist.org
[39] Noted as of January 13, 2010; Accessed April 17, 2018. http://odin.linguistlist.org
[40] http://emeld.org/ and Farrar et al. (2002)
[41] http://www.language-archives.org/
[42] https://www.wikipedia.org/
[43] https://en.wikipedia.org/wiki/Naskapi_language

- The World Atlas of Language Structures (WALS) is a directory typological features which also includes academic references for many of the over two thousand languages presented. WALS is a curated resource, largely made by a team of 55 experts, and hosted by the Max Planck Institute for Evolutionary Anthropology (the same as Glottlog, and as other resources such as PHOIBLE[44] (Moran et al., 2014) and DOBES[45] (Wittenburg, 2003) related to taking an inventory of language structures). (Dryer and Haspelmath, 2013)

There are other resources: the CLARIN Virtual Language Observatory[46], the Linguistic Data Consortium at UPenn,[47] the ELRA,[48] META-SHARE,[49] the Association for Computational Linguistics' Wiki,[50] the NICT Universal Catalogue,[51] LT World[52] and so on. Providing an exhausting list would be exhausting - more pertinently, now that it is clear that there are resources, what ones are relevant to low resource languages?

## 3.2 BLARK and LRE maps

Soria et al. (2017) briefly mention "digital language survival kits" as one of the motivations for their paper - these are explicated more fully on the Digital Language Diversity Project's site.[53] This project is an EU initiative, through the Erasmus+ programme, and it aims to identify needs and provide "kits" for certain European low resource languages - specifically Basque, Breton, Karelian and Sardinian.

The use of the word "kit" is informative, as there is preëxisting literature on this topic in BLARK, or Basic Language Resource Kit. BLARK was developed by a joint initiative between the European Network of Excellence in Language and Speech (ELSNET), a European international umbrella for 145 different organisations in 29 countries, and the European Language Resources Association (ELRA), and first outlined in 1998 (Krauwer, 1998). The BLARK is defined as the "minimal set of language reosources that is necessary to do any precom-

[44] http://phoible.org/
[45] http://dobes.mpi.nl/
[46] https://vlo.clarin.eu
[47] https://www.ldc.upenn.edu/
[48] http://catalog.elra.info
[49] http://www.meta-share.eu/
[50] https://aclweb.org/aclwiki
[51] https://www.nict.go.jp/index.html
[52] http://www.lt-world.org/
[53] http://www.dldp.eu/en/content/digital-language-survival-kit

petitive research and education at all." (Krauwer, 2003, 4) In general, this comprises "written language corpora, spoken language corpora, mono- and bilingual dictionaries, terminology collections, grammars, modules (e.g. taggers, morphological analysers, parsers, speech recognisers, text-to-speech), annotation standards and tools, corpus exploration and exploitation tools, bilingual corpora, etc."

Krauwer (2003) has a comprehensive matrix in the appendix outlining technology that would be needed to provide a BLARK for Dutch, as outlined in a workshop documented in Binnenpoorte et al. (2002). In another paper, Maegaard et al. (2006) under NEMLAR (Network for Euro-Mediterranean LAnguage Resources) outlined the specific needs that BLARK specified which could be applied to Arabic, and actions which researchers took in order to develop resources to best fill in the grid. Both of the BLARK grids for Arabic provided in that paper are included here, in Figures 5 and 6, as they very usefully show not only the state of human language technology (HLT) resources for Arabic at the time, but also the categories thought sufficient. These categories - "prosody prediction", "alignment", "shallow parsing", and so on - are all terms which refer to a suite of resources that each reflect hundreds of papers from within the computational linguistics community.

The BLARK process - auditing a language, using a grid to identify what corpus and resource needs are necessary for language resources - has now been applied to Swedish (Elenius et al., 2008) and Bulgarian (Simov et al., 2004), and numerous South African languages (Grover et al., 2011), among others.

Unfortunately, BLARK (or ELARK, purportedly a more sophisticated version of BLARK for industry described in Mapelli and Choukri (2003), according to (Grover et al., 2011)) is a large grid, and may not work for languages without extensive funding models or support. For this, there is a smaller BLARK version, the BLARKette, which should work for low resource languages (although how a smaller version of a minimal set could be provided usefully is not clear).

> In order to accommodate this problem we have proposed the definition of a scaled down, entry-level version of the BLARK, targeting exclusively the research and (especially) the education community. It should be light and compact, not too demanding in terms of hard and software requirements, cheap, free from IPR issues, and ideally small enough to fit on a CD or DVD. We expect to release a first document, with tentative summary specifications, towards the end of 2006. Check the ELSNET site for news. (Krauwer, 2006)

The model of transportation for this - a CD, instead of a downloadable resource - shows that the concept has not aged well. There is a also surfeit

| | Monolingual Lexicon | Multi-/bilingual Lexicon | Thesauri, ontologies, wordnets | Unannotated Corpora | Annotated Corpora | Parallel Multi Ling Corpora | Multimodal corpora for (hand) OCR | Multimodal corpora for (typed) OCR |
|---|---|---|---|---|---|---|---|---|
| | | | | | | | | |
| Morphological comp.(infl, deriv., stemm., diacritic,...) | +++ | | | | ++ | | | |
| stat. | + | | | | +++ | | | |
| POS disambiguator/tagger | +++ | | | | | | | |
| stat. | + | | | | +++ | | | |
| | | | | | | | | |
| Diacritizer | +++ | | ++ | | | | | |
| stat. | | | | | +++ | | | |
| Sentence Boundary Detection (punctuation) | +++ | | | | ++ | | | |
| stat. | | | | | +++ | | | |
| Named Entity Recognition | +++ | | | | + | | | |
| stat. | | | | | +++ | | | |
| Word Sense Disambig. | +++ | | | ++ | ++ | | | |
| stat. | | | | | +++ | | | |
| Term extraction | +++ | | | +++ | | | | |
| stat. | | | | +++ | +++ | | | |
| Shallow parsing | +++ | | | | | | | |
| stat. | | | | | +++ | | | |
| Syntactic analysis comp. | +++ | | | | + | | | |
| stat. | | | | | +++ | | | |
| Semantic Analysis comp.(incl. Coreference res.) | +++ | | +++ | | | | | |
| Sentence synthesis and generation | +++ | | ++ | + | ++ | | | |
| Transfer tool (software) | | +++ | | | | | | |
| stat. | | | | | | +++ | | |
| Alignment | +++ | +++ | | | | + | | |
| stat. | | | | | | +++ | | |
| Grapheme recognition (for typewritten OCR), stat. | ++ | | | +++ | | | | +++ |
| Grapheme recognition (for handwritten OCR), stat. | ++ | | | +++ | | | +++ | |

Figure 5: A BLARK graph for Arabic, with written language applications and corresponding HLT modules, marked with importance (Maegaard et al., 2006, 775)

| | Dictation | Telephony speech applications | Embedded speech recognition | Transcription of broadcast News | Transcription of conversational speech | Speaker recognition | Dialect / language identification | “Emotion” Identification | Speaker Adaptation | Lips movement reading : | ‘topic’ detection, segmentation, topic boundaries | Speaker 2 speaker mapping | “Emotion/ Prosody” output | – Text to Speech (inc. formatted data e.g. databases) | – Synthesis by Concatenation : | – Customization to different voices | – Generation Lips Movement |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | | | | | | | | | | | | |
| Acoustic models | +++ | +++ | +++ | +++ | +++ | ++ | +++ | +++ | +++ | +++ | +++ | ++ | +++ | +++ | +++ | +++ | +++ |
| Language models | +++ | ++ | ++ | +++ | +++ | | ++ | | | | | | ++ | +++ | | | |
| Pronunciation lexicon | +++ | +++ | +++ | +++ | +++ | | | | | | | ++ | | +++ | | | |
| Lexicon Adaptation | + | + | + | + | + | | | | | | | ++ | | +++ | | | |
| Phoneme Alignment | + | + | + | + | + | + | ++ | | | | | ++ | | | | | |
| Prosody recognition | + | + | + | + | + | + | + | +++ | + | | | ++ | | | | | |
| Speech Units Selection | | | | | | | | | | | | | +++ | +++ | | | |
| Prosody prediction | | | | | | | | | | | | | +++ | +++ | | | |
| segmenter Speech / Silence: | ++ | + | ++ | ++ | ++ | + | ++ | ++ | + | + | + | | + | | | | |
| Sentence boundary detection: | + | + | + | + | + | + | + | ++ | + | + | + | | ++ | +++ | | | |
| Dialect / language identification | + | + | + | + | + | + | + | + | + | + | + | | | + | | | |
| (word) Boundary identification, | + | + | + | + | + | + | + | + | + | + | + | | ++ | | | | |
| Speech /Non-speech (music) detection: | + | + | + | + | + | + | + | ++ | + | + | + | | | | | | |
| Speaker recognition/identificatio n | + | + | + | + | + | + | + | + | + | + | + | ++ | | | | | |
| “Emotion” Identification | + | + | + | + | + | + | + | | + | + | + | ++ | ++ | | | | |
| Speaker Adaptation | ++ | + | ++ | + | ++ | + | + | + | + | + | + | ++ | | + | | | |
| Lips movement reading | | | | | | | | | | +++ | | | | | | | |
| Morphological comp.(infl, deriv., stemm., diacritic,...) | ++ | + | + | ++ | ++ | | | | | | | | | +++ | | | |
| POS disambiguator/tagger | ++ | + | + | ++ | ++ | | + | | | | | | | +++ | | | |
| Diacritizer | | | | | | | | | | | | | | +++ | | | |
| Named Entity Recognition | ++ | + | ++ | ++ | ++ | | | | | | | | | ++ | | | |
| Word Sense Disambig. | | | | | | | | | | | | | | ++ | | | |
| Shallow parsing | ++ | + | + | ++ | ++ | | | | | | | | | ++ | | | |
| Syntactic analysis comp. | ++ | + | + | ++ | ++ | | | | | | | | | ++ | | | |
| Sentence synthesis and generation | | | | | | | | | | | | | + | ++ | | | |
| Semantic Analysis | + | | | + | + | | | | | | + | | + | + | | | |

Figure 6: A BLARK graph for Arabic, with speech language applications and corresponding HLT modules, marked with importance (Maegaard et al., 2006, 776)

of references of BLARK or BLARKette in the past decade in the literature - Krauwer (1998) only has 31 references on Google Scholar (an imperfect but effective metric).[54] What happened? It is most likely (in my opinion) that building a BLARK for a language is too complex for language groups to perform, and lacks proper incentives. It requires an authoritative and intimate knowledge of a language's space by many researchers, all of whom must come together to identify gaps, often from proprietary institutions. This is a difficult task.

But this effort, in some sense, has expanded into LRE (Language Resources and Evaluation) maps within Europe. As described in Calzolari et al. (2010); Del Gratta et al. (2014); Mariani and Francopoulo (2015); Del Gratta et al. (2015), the Language Resources and Computation (LREC) conference organisers began asking conference participants who had submitted papers to fill out basic language resource grids when submitting papers. This effort was extended to ten different computational linguistics conferences, covering most large European languages and four regional Spanish languages. This data has been collected into matrices and a database that reflects language resources for a variety of languages. To date, this is the most comprehensive review of NLP per language that I'm aware of, with 4395 entries - however, it's worth noting that it is limited in scope. The 133 less-common languages represented in the LRE map represent only 414 entries. An example of the matrix for the high resource languages can be seen in Figure 7, which is a map of resources for various languages, cut off with a lower bound of 50 citations per resource type.

Several authors working on LRE maps are also authors of the Soria et al. (2017) paper; extending the LRE maps for low resource languages, and then intensifying efforts to develop low-hanging fruit for low resource languages is a logical next step for this research. The focus on European languages is expected; this may stem from the fact that LREC, the main conference series from which LRE data was drawn, is run by the European Language Research Association (ELRA). This fragmentation of the field is unsurprising, and happens in the reverse, as well: for example, Paricio Martín and Martínez Cortés (2010) cites a framework for upgrading low resource languages which is explained in a research paper written in Spanish, and, anecdotally, around half of the papers presented at the Ryukyuan Heritage Language Society's conference in Tokyo in 2012 (which I attended) were presented in Japanese. This is not to say that fragmentation and diversity of linguistics in academia is something to be avoided, but rather that it is a hurdle to be noted and worked with to avoid repeated work and splintered efforts.

[54]https://scholar.google.ca/scholar?cites=5069727220703395724

| | Bulgarian | Czech | Danish | Dutch | English | Estonian | Finnish | French | German | Greek | Hungarian | Irish | Italian | Latvian | Lithuanian | Maltese | Polish | Portuguese | Romanian | Slovak | Slovene | Spanish | Swedish | Other Europe | Asturian | Basque | Catalan | Galician | Arabic | Hindi | Japanese | Korean | Mandarin | Other | Multilingual | L.I. | N.A. | Total |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Corpus** | 22 | 33 | 32 | 56 | 702 | 13 | 15 | 135 | 133 | 30 | 26 | 1 | 74 | 12 | 10 | 2 | 25 | 48 | 23 | 5 | 15 | 104 | 44 | 157 | 1 | 16 | 18 | 7 | 84 | 32 | 65 | 10 | 94 | 216 | 21 | 37 | 47 | 2365 |
| **Lexicon** | 12 | 12 | 4 | 15 | 162 | 3 | 3 | 46 | 30 | 5 | 7 | 0 | 28 | 1 | 1 | 0 | 13 | 11 | 12 | 2 | 5 | 27 | 11 | 50 | 0 | 9 | 8 | 3 | 22 | 8 | 26 | 2 | 27 | 93 | 5 | 9 | 10 | 682 |
| **Tagger/Parser** | 1 | 3 | 0 | 2 | 68 | 1 | 0 | 24 | 14 | 5 | 3 | 1 | 11 | 0 | 1 | 0 | 6 | 5 | 2 | 0 | 0 | 12 | 4 | 17 | 1 | 4 | 2 | 1 | 11 | 1 | 7 | 0 | 15 | 18 | 4 | 51 | 20 | 315 |
| **Annotation Tool** | 2 | 2 | 1 | 4 | 44 | 1 | 1 | 9 | 5 | 1 | 1 | 0 | 5 | 1 | 1 | 1 | 1 | 2 | 2 | 1 | 2 | 7 | 1 | 4 | 0 | 1 | 2 | 0 | 10 | 0 | 2 | 0 | 5 | 13 | 3 | 93 | 23 | 251 |
| **Evaluation Data** | 2 | 5 | 2 | 5 | 101 | 0 | 1 | 13 | 14 | 1 | 1 | 1 | 5 | 0 | 0 | 0 | 0 | 4 | 3 | 0 | 1 | 15 | 2 | 11 | 0 | 2 | 1 | 1 | 8 | 4 | 9 | 0 | 14 | 9 | 2 | 4 | 8 | 249 |
| **Ontology** | 1 | 2 | 0 | 7 | 44 | 0 | 1 | 9 | 6 | 3 | 0 | 0 | 6 | 0 | 2 | 0 | 1 | 3 | 3 | 0 | 0 | 9 | 0 | 7 | 0 | 0 | 0 | 0 | 2 | 0 | 7 | 1 | 7 | 6 | 4 | 32 | 9 | 172 |
| **Grammar / Language Model** | 3 | 1 | 2 | 2 | 20 | 0 | 2 | 8 | 8 | 0 | 1 | 0 | 2 | 1 | 0 | 1 | 3 | 1 | 1 | 1 | 1 | 7 | 3 | 5 | 0 | 0 | 1 | 0 | 1 | 1 | 0 | 0 | 2 | 9 | 0 | 7 | 6 | 100 |
| **Terminology** | 1 | 1 | 0 | 3 | 22 | 2 | 0 | 11 | 7 | 2 | 1 | 0 | 5 | 1 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 8 | 0 | 5 | 0 | 2 | 2 | 0 | 0 | 0 | 2 | 0 | 1 | 0 | 1 | 3 | 2 | 84 |
| **Representation-Annotation Standards/Best Practices** | 1 | 2 | 0 | 4 | 22 | 1 | 0 | 3 | 3 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 2 | 1 | 4 | 0 | 0 | 0 | 0 | 2 | 0 | 1 | 0 | 4 | 3 | 1 | 20 | 5 | 83 |
| **Corpus Tool** | 0 | 0 | 0 | 3 | 13 | 1 | 0 | 1 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 1 | 2 | 0 | 0 | 1 | 3 | 1 | 5 | 0 | 1 | 1 | 0 | 3 | 0 | 0 | 0 | 1 | 0 | 0 | 36 | 3 | 78 |
| **Named Entity Recognizer** | 0 | 1 | 0 | 2 | 23 | 0 | 0 | 4 | 4 | 1 | 0 | 0 | 2 | 0 | 0 | 0 | 0 | 2 | 0 | 0 | 0 | 2 | 0 | 3 | 0 | 1 | 0 | 0 | 4 | 0 | 4 | 1 | 3 | 7 | 5 | 6 | 2 | 77 |
| **Language Resources/ Technologies Infrastructure** | 1 | 2 | 1 | 2 | 7 | 0 | 0 | 1 | 3 | 1 | 1 | 0 | 4 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 1 | 3 | 3 | 2 | 0 | 2 | 1 | 0 | 1 | 0 | 2 | 2 | 3 | 1 | 2 | 13 | 1 | 62 |
| **Machine Translation Tool** | 1 | 1 | 2 | 0 | 6 | 1 | 0 | 1 | 3 | 1 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 1 | 0 | 1 | 2 | 0 | 4 | 0 | 1 | 1 | 0 | 1 | 1 | 1 | 1 | 1 | 6 | 1 | 17 | 4 | 61 |
| **Evaluation Tool** | 0 | 0 | 0 | 1 | 13 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 30 | 11 | 59 |
| Total | 48 | 68 | 46 | 110 | 1364 | 23 | 24 | 286 | 265 | 56 | 44 | 5 | 158 | 18 | 18 | 4 | 55 | 87 | 51 | 9 | 28 | 222 | 73 | 293 | 2 | 48 | 41 | 12 | 157 | 50 | 135 | 17 | 190 | 414 | 58 | 532 | 207 | 5218 |

Figure 7: LRE maps for high resource languages (Mariani and Francopoulo, 2015, 460)

## 3.3 Who makes resources for languages?

Another hurdle which was briefly alluded to earlier was the plethora of large organisations, databases, or projects dedicated to cataloguing low resource languages. Each of these has differences in scope, funding, and incentives. However, large organisations are not the only groups working on language development, digital ascent, language revitalisation, or any other shared focus that relates to low resource languages.

As Hammarström (2015) points out, "language documentation and description is an extremely decentralized activity, carried out by missionaries, anthropologists, travellers, naturalists, amateurs, colonial officials, ethnographers and not least linguists over several hundred years." Language communities, amateur and professional linguists, educators, and language policy setters are most often involved in standardising a language and helping to document and revitalise low resource languages. Digitally, amateur computational linguists, and coders who are first language speakers of their own language are often the first to work on translating or migrating resources; this group is also often the first to set up Wikipedias in a local language (although this often leads to enthusiastic loners working outside of the main language communities) Soria et al. (2017). Beyond these groups, universities, local governments and businesses can also often develop language resources for low resource languages, as was the case with Rögnvaldsson et al. (2009). After these groups, large grant-driven institutions such as CLARIN or the NSF fund a large portion of language development, along with industry giants such as Google or Xerox, and large military research arms such as DARPA.

Unfortunately, the lion's share of the overall funding for language development goes to languages which are already resourced.

> Over the years the EU has invested massively in the development of language and speech technology, and many dedicated R&D programmes have had a significant impact on its advancement, including applications oriented towards solving the multilinguality problem... Unfortunately the strong industrial bias of recent EU programmes has led to a situation where the major part of the funding for language and speech technology goes to the major languages. This is not surprising, as industrial players will prefer to invest in the development and deployment of technologies for larger markets. As a consequence there has been only marginal support for the development of language and speech technology for the language communities that do not constitute profitable markets. As the development cost of such technologies is independent of the number

> of speakers of a language ("all languages are equally difficult") this has created a very unbalanced situation. (Krauwer, 2006)

Or:

> Were it not for the special attention DARPA, one of the main sponsors of machine translation, devoted to Haitian Creole, it is dubious we would have any MT aimed at this language. There is no reason whatsoever to suppose the Haitian government would have, or even could have, sponsored a similar effort (Spice, 2012). (Kornai, 2013, 9)

Another good example of where funding and incentives for language development can be controversial would be Ethnologue, which rate limits and has a paywall guarding usage of their database, even though they are widely recognised as one of the best informed databases for language data. SIL International also gatekeeps the standard ISO 639-3, which is the most widely used language code. By having a paywall on their data, they exclude the general public from having control of codes for their own languages. SIL has also come under criticism for their Christian missionary work, as it can be viewed as complicit in culture change, and by extrapolation, ethnocide (Dobrin, 2009; Dobrin and Good, 2009; Everett, 2009). This is just one example - and most likely one of the most extreme, not counting military work on languages used by insurgents in wars - of how organisations working on language resources may influence the work itself.

The funding of language resource development matters, because the way that the language community approaches language development affects the chance of survival for the language. This is one of the reasons that Grenoble (2016) pointed out that "language vitality" is a more politically correct term to use than "language endangerment", as it takes the focus away from loss and focuses attention on language ascent. Another reason that language funding matters is because the major players with funding will generally be able to out manoeuvre smaller groups with different resources. This can enforce language shift, and can render resources created by individual developers moot. For instance, the secwepemc-facebook[55] tool developed to automatically translate Facebook into low resource languages, created by the developer Neskie Manuel for his native Secwepemctsín, is no longer an active project and has not been updated, rendering it obsolete with Facebook UI changes, while automatic translation is provided for high resource languages natively by Facebook. Scannell, who helped port the secwepemc-facebook tool to Greasemonkey, was

[55]https://github.com/kscanne/secwepemc-facebook

one of the authors of Streiter et al. (2006), which suggested that developers for low resource languages use open source software pools in order to pool resources to enable them to overcome this - among other - issues facing low resource languages in particular.

As in Section 2.2.5, covering all of the potential issues with funding and the politics of language development is well beyond the scope of this paper. However, focusing on how open source can help low resource languages is not. But first; what do I mean by "open source"?

# 4 Open Source Code

## 4.1 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 low resource languages.

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 Linux programmers such as Eric Raymond, author of *The Cathedral and the Bazaar*[56](Raymond, 1999); Linus Torvalds, author of the Linux kernel[57] and Git[58]; Richard Stallman, founder of the GNU project[59] and the Free Software Foundation[60]; and others 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* and *libre* 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, and what free and open means. An oft-used expression is "free as in beer" as opposed to "free as in speech", where the first is used for gratis software which has no monetary price set on it, and the second is used to refer to software which is written without restriction. The term *libre* is most often used for this second definition, to differentiate the two meanings in English. Occasionally, the acronym FLOSS is used in open source parlance to refer to Free Libre Open Source Software, which is both gratis and libre software.

For some adherents, software ought to be free (gratis), as it is a result of human labour and because opening it up without cost maximises the utility function of that code, and minimises duplicated effort. This idea contains harks back to the idea of a 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.

[56]http://www.catb.org/ esr/writings/cathedral-bazaar
[57]https://www.kernel.org/
[58]https://git-scm.com/
[59]https://www.gnu.org/
[60]https://www.fsf.org/

> 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.[61] *Richard Stallman* (Founder of GNULinux)

However, for the most part, open source isn't disambiguated as a term, because authority for this task is delegated to the license put on a piece of software, which determines the legality and potential use. 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. The term open source under most definitions doesn't 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.

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 (OSI) maintains a list of approved licenses on their website.[62]

The OSI, whose founders were one of the original coiners of 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 and their definitions are from the OSI's website,[63] and are repeated below verbatim.

1. Free Redistribution. The license shall not restrict any party from selling or giving away the software as a component of an aggregate software dis-

[61] https://www.gnu.org/philosophy/open-source-misses-the-point.html
[62] https://opensource.org/licenses
[63] https://opensource.org/osd

tribution 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 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.

## 4.2 Open source licenses

These different terms and conditions are often conflated, and a legally-valid license which satisfies all of them is difficult to write on an *ad hoc* basis. For this reason most open source programming relies on using existing licenses, and copying them for specific projects. There are tools today to help make licensing more clear to naïve users, such as choosealicense.com, tldrlegal.com, and so on.

Some of the main licenses used in the wild are as follows:

- The MIT license, developed at MIT, is the most popular license on GitHub,[64] the world's largest repository of code, used in over 40% of the projects licensed there as of March 2015.[65] It is a a very permissive license, which allows commercial use, modification, distribution, sublicensing, and private use of any code so licensed. It also waives liability for the authors of the code, saving them from needing to worry about lawsuits in cases where their code would otherwise be liable - the code is granted as is, and what the user does with it is not the author's fault. The only restriction is that you need to include the license in any software which uses it.

- The Apache License 2.0, developed by the Apache Software Foundation,[66] is similar, but disallows users from trademarking code with the license, requires a few smaller modifications like stating code changes and adding a NOTICE file, if one exists, to derivational code, and also adds a patents clause for contributors.

[64] https://github.com
[65] https://blog.github.com/2015-03-09-open-source-license-usage-on-github-com/
[66] https://www.apache.org/licenses/

- The BSD licenses were developed for use with Berkeley Software Distribution, a Unix-like OS. There have been multiple iterations; the first, 4-clause license required every subsequent license to reference and acknowledge the original, ending with large lists of acknowledgements; a subsequent 3-clause license (often called the "New" BSD) removed this, but kept a clause which stated that usage does not imply endorsement by the original contributors; and this was removed in a 2-clause version, often called "Simplified" or the "FreeBSD" license.

- The GNU General Public License (GPL)[67] is the main example of copyleft licensing, where any derivative works that use GPL licensed code must also use a GPL license. This causes major issues when users want to combine code from multiple sources, some of whose licenses may conflict. For this reason, the GNU Library or "Lesser" General Public License (LGPL) was created, to allow only code under the LGPL to be accessible and modifiable openly, while all other code doesn't have to be. GPL also demands that users include installation instructions,

- Creative Commons licenses,[68] mostly used for sharing non-code material such as images and documents openly, was created by Lawrence Lessig, the founder of the Creative Commons organisation,[69] and may also be used for code projects. There are many licenses they offer, and some variants are copyleft licenses - in particular, "share-alike" clauses are an example of copyleft.

- The Unlicense,[70] created in 2010, is another option, which explicitly states that code is unlicensed, with no restrictions, and also with no liability for the authors (unlike code which is not licensed, which has stricter protections under US copyright law than code which specifically excludes a license). There is a Creative Commons Zero,[71] license which is similar, as well as the WTFPL license ("Do What The Fuck You Want Public License")[72] which, although intentionally comically profane, is non-trivial in that it is used in 11,714 different software projects on GitHub as of this writing.[73]

[67] https://www.gnu.org/licenses/
[68] https://creativecommons.org/licenses/
[69] https://creativecommons.org/
[70] https://unlicense.org/
[71] https://creativecommons.org/publicdomain/zero/1.0/
[72] http://www.wtfpl.net
[73] https://github.com/search?q=license%3AWTFPL

As is clear from these short descriptions, licenses are not easily interchangeable and they come with a range of suppositions about how the data ought to be used. Copyleft licenses (mostly GPL) require any derivative works to also be open source, which means that they cannot be used in proprietary codebases, leading to fragmentation of the code space and to legality issues in the long run. However, the effects of copyleft may be more perfidious, in that funders or developers may avoid projects altogether if they find a project has (or doesn't have) a copyleft license. The same could be said for liability waivers, or more especially the lack thereof. This is backed up in studies: for instance, two thirds of respondents for GitHub's open source survey in 2017 said that they value licensing as a major factor when contributing to a project.[74] Ultimately, licenses are complicated legal documents with various repercussions for how code is accessible.

## 4.3 Where is open source code?

For closed source or proprietary software, the code itself often isn't stored in the open or accessible to third parties. However, for open source software to be defined as open source according to OSI's definitions, it needs to be publicly accessible and well-publicised. This means that storing code on a server where it could technically be accessed via some protocol, or less ideally through a mail-order CD as Krauwer (2006) suggested, is not enough; instead, it ought to be linked to elsewhere and available for everyone to access. This raises the question: where is most open source code stored?

Unequivocally, GitHub[75] is the largest source of shared, open code on the internet, with 27 million users and 80 million repositories as of March 2018.[76] There have been several large-scale studies of its codebase by researchers (Gousios and Spinellis, 2012; Allamanis and Sutton, 2013; Gousios et al., 2014; Kalliamvakou et al., 2014; Beller et al., 2016) which confirm this. Other large repositories for code of a similar nature, include Sourceforge, with 430k projects and 3.7m users,[77] Bitbucket[78] with 5m users,[79] Launchpad[80] with

[74]http://opensourcesurvey.org/2017

[75]https://github.com

[76]https://github.com/about

[77]https://sourceforge.net, accessed April 18, 2018

[78]https://bitbucket.org/

[79]https://blog.bitbucket.org/2016/09/07/bitbucket-cloud-5-million-developers-900000-teams/

[80]https://launchpad.net/

4.2m users,[81] and Gitlab,[82] which holds the majority share of self-hosted Git platforms.[83] All of these platforms are based around Git, the versioning software developed by Linus Torvalds, used to store different versions of code for developers and teams, which lends itself particularly to shared code that can be updated easily by outside and community developers. It is worth mentioning that not all of these projects are public.

Self-hosted Git instances are a common way of storing proprietary code; one sets up a versioning system within a company, using the tools and set of social standards that developers are used to from working on open source code, but limit access to employees. This is what is meant by GitLab's statement that they host most self-hosted Git platforms. Git isn't the only possible versioning software for this; Google has their own versioning tool, Piper, which hosts the over two billion lines of code used by the majority of the company in a single repository.[84] Self-hosted Git instances are generally not open source. Generally, if someone wants to use a shared Git repository, they are limited to paying a fee for a hosting service, or using sites that have a freemium model where public repositories are free, but private or enterprises instances are not.

There are alternatives to cloud storage (the "cloud" here being a common metaphor for hosting on someone else's servers) with a hosting provider; one would be storing the code on your own website, and running your own server or building the user interface yourself. This is largely uncommon due to setup costs, but occasionally happens with academics and smaller teams who are not used to larger hosts or who are worried about the longevity of providers. This latter worry is founded; for instance, Google Code was closed after ten years of running in 2016, causing many projects to need to port to another service such as GitHub.[85] For academics, a common solution to offset setup and hosting costs is to use university websites and archives as a suitable place to store open source code. For instance, Giellatekno, a language-technology research group, and Divvun, a linked product development group, both work primarily on Sámi languages, and both use the same Subversion (another versioning system) database for storing their code (Moshagen et al., 2014), which is hosted by UiT The Arctic University of Norway.[86]

In a large part, the question of where to store information - especially academic information regarding languages - is one which the large archival sites mentioned in Section 2.3.1 were created to solve. In particular, this is true for

[81] https://launchpad.net/people
[82] http://gitlab.com/
[83] https://about.gitlab.com/is-it-any-good/
[84] https://www.wired.com/2015/09/google-2-billion-lines-codeand-one-place/
[85] https://code.google.com/
[86] http://giellatekno.uit.no/

non-code resources, such as audio and video corpora, which historically have been prioritised for storage over code due to the size and relative importance of the corpora, and due to the older industry standards of keeping all code related to research private, especially when that code was funding by enterprise. Many of these sites are repositories of metadata which pointed to individually hosted content, which made the links susceptible to link rot and offloaded the issue of storage altogether.

Today, however, there is a sea change towards putting computational work in the open. Occasionally, this means that academics point to the open source code for their papers on GitHub or elsewhere, or publish their software itself as a research object. For example, Mäkelä (2016) and Kleinberg and Mozes (2017) were published with the Journal of Open Source Software (JOSS)[87] (Smith et al., 2018), which peer-reviews, publishes, and assigns digital object identifiers (DOIs) to software as a way of recognising important academic work. The code for these papers is publicly available on GitHub. Incentivising academics to publish their code openly is difficult, as software is not weighted in job reviews the same way as research papers; however, there are other benefits such as reproducibility and transparency. There are efforts to align these incentives; for instance, The Austin Principles of Data Citation in Linguistics Berez-Kroeker et al. (2017) was created to emphasise the importance of citing, using, and storing linguistic data properly. Standardising open source paradigms in academia is an ongoing work.

## 4.4 Digital permanence and storage

Focusing a bit closer on the academic use case, we can easily imagine a case where a professor puts code related to research on a university server, only to see that server change hands, go offline, or become defunct if the professor leaves the university for a position elsewhere or if their focus changes. This is more true of grad students, who don't have the same locational longevity as staff. As mentioned briefly above, this can lead to link rot; links which formerly pointed to workable software may then point nowhere or to the wrong resource. For example, Bender (2016) points to http://resourcebook.eu/searchll.php as a location for the LRE maps mentioned in Calzolari et al. (2012); however, this link no longer resolves (as of this writing), and it is difficult to know what it ought to have pointed at.

This is an artefact of systemic defects; in a location-based protocol (such as the Hypertext Transfer Protocol (http) protocol used by most websites today), consistency of location is prioritised over consistency of content. If the content

[87] http://joss.theoj.org/

was pointed to using some more permanent reference, such as a DOI, than the object could be moved without issues, and the problem of link rot is largely solved.

Digital permanence is a larger issue than code placed in locations by individual actors, however. Large organisations may lose their funding, come to the end of their expected lifecycle, or decide to shutter or obfuscate projects upon which research or language communities may depend. A good example would be Google Code, mentioned above in Section 4.3. Another example might be listserv.heanet.ie,[88] which probably held the largest corpus of Irish data at one point, but which was unavailable to crawlers and depends upon the hosting of heanet.ie for continued service (Scannell, 2007). A final example might be the linguistic vitality database by Kornai's group mentioned earlier,[89] which is actively looking for more funding.

Aside from the problem of code actively being stored, there is another issue with code rot. Over time, the ecosystem around which code is built changes, and it becomes harder to reproduce the original environment where code was installed and executed, leading to the code itself becoming less useful (Eide, 2010). Some solutions to this problem involve using containers like Docker to emulate the original environment (Boettiger, 2015). While this research has largely been driven by a need to replicate scientific results (Schwab et al., 2000; Barnes, 2010; Ince et al., 2012), it is also relevant outside of academic research to enterprise and community solutions to difficult coding problems, such as natural language processing.

As computational languages naturally evolve, it is important to take into account that the code must also be maintained if it is going to find consistent usage. Maintenance is a difficult task that has few immediate incentives, and which generally involves long timelines. For more on this, see Section 4.5.

There are alternatives to using academic institutions as code providers; host your own, or use a larger institution that has space set aside for maintenance. Or, release the code publicly using whatever enterprise solution seems like it will last the longest. However, these do change - for instance, Sourceforge was very popular before GitHub rose to the top of the field, and now many projects are moving off of Sourceforge and onto GitHub (Finley, 2011), which takes time and effort (why is another matter, and may be related to network effects. For more work mining these networks, see Thung et al. (2013); Kalliamvakou et al. (2014)).

Another idea would be to store your data on peer-to-peer or decentralised networks, which lessen the risk of centralised storage facilities, but also require

[88]https://listserv.heanet.ie

[89]https://hlt.bme.hu/en/projects/lingvit

a peer to serve the files for longevity to be assured. Ultimately, the best bet is to build files and code which are actively used by the community; the long tail of disused projects are at the most risk, while more popular projects will find a way to survive. For more on this, see Section 7.3.

## 4.5 Funding

Open source code cannot by definition be sold directly for a profit; open source code must be freely available to all users. This raises an issue where funding for open source development is not direct in the sense of immediate fiscal returns. In this business environment, other funding models need to be pursued. The obvious, most common solution is to sell services on top of open source code, and give away the code itself for free. There are benefits to doing this. Giving away code can be seen as a marketing tactic, drawing other developers, or it may serve to develop a community of active developers who are interested in giving back to the original project without being employed by the core developer's company, or it may serve as a retention device keeping developers who prefer to work in the open happy, or it may serve as a way of verifying a level of security for the code itself, by allowing other participants to point out flaws in the system and fix them without needing to rely upon expensive and possibly ineffectual internal security audits.

For researchers, open sourcing code can be seen as a major time investment (FitzJohn et al., 2014; Lowndes et al., 2017), and although it can help reproducibility, it is not normally the primary source of sharing resource (which would be the scientific article). For researchers, funding needs to come from either salaries, from the researcher's free time, or from grants from larger institutions (not counting enterprise and interdisciplinary cross-overs). This is a serious barrier to open source work in the sciences.

For militaries and governments, there is little incentive to open source unless there is a direct mandate from their political constituents or legal process. Even when there are open challenges run by military branches - for example, the DARPA sponsored Low Resource Languages for Emergent Incidents (LORELEI) challenge[90] - there are often no demands that any resulting work be open sourced (although the initial challenge is open sourced as a way of inviting participation). Often, this is because the code itself has security concerns; for example, open sourcing speech recognition software for languages spoken by military targets in lossy situations (such as over cell networks) would only illuminate that such software exists. This example of security through closed

[90]https://www.nist.gov/itl/iad/mig/lorehlt-evaluations

source methodologies extends to enterprise; for Google to open all of their MT data would cause them to lose a competitive edge in the translation market.

For software developers outside of academia, militaries, governments, and large enterprises that have business advantages, however, open sourcing code can be a significant way to gain prestige, to improve and market developer relations, to market themselves to prospective clients and companies, and to contribute to their coding communities. There are a variety of ways of funding work within the open source model.

One direct way is to add payment schemes directly to source code or to a website, asking for donations. Another would be to use a collective community to allocate donations and funds; Open Collective[91] is an example of a company that helps do this for developers, some of whom are paid entirely through funds on the site.[92] Crowdfunding sites can also be useful for some developers. Patreon is a good example where makers can earn money directly through fan donations, while Kickstarter has been used many times to fund projects. For example, Dave Gandy, the developer for Font Awesome, an open source font resource, raised over a million dollars in a month from 35,550 backers for the next version of his product.[93] Code bounties, funds set by community members hoping to have other developers solve bugs, is another limited way of making money.[94] Cryptocurrencies may eventually present other ways of funding open source, either directly,[95][96][97] or through other avenues like initial coin offerings. Already, some companies are using initial coin offerings (similar to IPOs in the business world, but instead marking the launch of a new cryptocurrency) to fund development on open source, such as with Filecoin, which raised over 200 million for their coin development, of which many of the funds will go directly to open source projects run by the company Protocol Labs, such as IPFS (Benet, 2014) on GitHub.[98]

There are several guides online that outline other ways of funding open source.[99][100][101] In the end, the majority of open source developers are not remunerated for their work directly. Most open source work is unpaid, and

[91] https://opencollective.com/

[92] https://medium.com/open-collective/a-new-way-to-fund-open-source-projects-91a51b1b7aac

[93] https://www.kickstarter.com/projects/232193852/font-awesome-5

[94] https://www.bountysource.com/

[95] https://utopian.io/

[96] https://oscoin.io/

[97] https://gitcoin.co/

[98] https://coinlist.co/filecoin

[99] https://github.com/nayafia/lemonade-stand

[100] https://medium.com/open-source-life/money-and-open-source-d44a1953749c

[101] https://opensource.guide/getting-paid/

maintenance of open source software can be demanding and costly for developers who don't set expectations around levels of support for users. This is especially difficult for developers who do not have total control of their projects, such as is often the case with developers doing open source within a company.

More specifically, the problem of funds being directed to low resource languages is unlikely to be solved by any of the proposed solutions above. However, by banding together and sharing tools openly, computational linguists working on low resource languages can expedite their work. This methodology will be explored in Chapter 5.

## 4.6 Ethics and open source

The quote from Richard Stallman in Section 4.1 mentioned that "free software is an ethical imperative". This is, to put it mildly, a loaded statement, and comes from a philosophical viewpoint that not everyone agrees with. Open source, for all of its benefits, has serious drawbacks for developers involved in it.

For one, the overwhelming majority of open source coders on online communities are male, young, and white (Ghosh et al., 2002). A survey of 100k users from StackOverflow,[102] a large language-agnostic forum for support and technical questions, found that this has changed little since in the past fifteen years, with 92.9% of the users being male and 75% of them white.[103] Open source is disproportionally skewed towards already advantaged groups.

The incentives around open source contributions are also changeable, and while paid workers are more likely to contribute in the long run, users who contribute to code because of the value of the code to them are less likely to stay in the community for long periods of time (Roberts et al., 2006; Shah, 2006). Ultimately, it is hobbyists who end up working on code the longest, after the initial value to them has worn off (Shah, 2006). This has implications for low resource languages; is open source the best vehicle for developing language software, which may have long runways? On another note, is it ethical to implement a system where there is high burnout rate for developers who need it, when it may make more sense to find ways to fund direct work for a small core of dedicated developers?

These are a couple of small examples of where advocating open source is not a clearcut issue. This paper is not meant to provide a solid overview of all ethical issues; however, at least some of them are worth noting here as caveats. For low resource languages, open source coding presents a clear opportunity

[102] https://stackoverflow.com/

[103] https://insights.stackoverflow.com/survey/2018/

for allowing communities to work together, cross-linguistically and between stakeholders, with a minimum of friction caused by proprietary licensing. It is my opinion that any work which can be expedited or made redundant may be useful in a field where languages are dying at exorbitant rates.

# 5 Open Source Code for Low Resource Languages

Now that low resource languages (LRLs) have been described, and now that there has been a brief overview of open source as a software methodology, the reader will doubtless wonder - what is the state of open source code that can be used today by language communities?

Unfortunately, due to the decentralised nature of open source, this is an inherently difficult question to answer. In the ecosystem, there are a few strategies that can be used to inform an answer: use a specific task as a case study for what tools would be used, look at what resources are available from any of the main large data aggregators mentioned in Section 3.1, take a screenshot of the ecosystem based on some of the more-cited open source tool used for LRL NLP, examine linked open data, and sample relevant work on GitHub through a manually collected list of resources. Each of these strategies is employed in a subsection, below.

## 5.1 Case study: Mapping linguistic coördinates

The breadth of HLT is wide; choosing a specific task within it and then trying to perform that task as adequate as possible would be one way to figure out how much open source code exists, and what that looks like. For example, suppose we were interested in making dialect maps using language coördinates. This is an old research area in linguistics (Trudgill, 1983; Labov et al., 2005), and computational methods for mapping languages have been described in some research, including in the recently started *Journal of Linguistic Geography* (Labov et al., 2012).

For NLP, this is a nontrivial task, as l10n within a browser can depend upon geographical information from the user. For instance, if the client's browser does not send a `Accept-Language` header[104] in their requests to view a website, specifying languages the client understands by using ISO 639 tags[105], then the server may use the `NavigatorLanguage.language` object in JavaScript[106] to query for the language of the browser UI (normally set by the users depending on where they downloaded it), or they could ask the browser directly through the geolocation API (for instance, on Firefox[107][108]) to supply the geolocation of users and extrapolate plausible languages from this data. Knowing where the

[104]https://tools.ietf.org/html/rfc7231#section-5.3.5
[105]https://www.ietf.org/rfc/bcp/bcp47.txt
[106]https://www.w3.org/TR/html51/webappapis.html#language-preferences
[107]https://www.mozilla.org/en-US/firefox/
[108]https://developer.mozilla.org/en-US/docs/Web/API/Geolocation/Using_geolocation

user is likely to be, and what languages the user is likely to prefer using, could help with providing their native language automatically in the browser.

Gawne and Ring (2016) give a general overview of the mapping field currently, pointing out that the main resource for finding language geographical coördinates comes from the World Language Mapping System[109], a website owned and run by SIL, which are used for ISO 639-3 labelling, and by Glottolog and OLAC. The maps are under a closed license and must be purchased. Gawne and Ring (2016) also mention WALS, which uses its own geographical coördinates, and the ELP, which understandably uses Google Maps as its mapping program, and draws from multiple sources. They also mention Language Landscape,[110] a project which maps instances of language use on a map.

To use these geographic information systems (GIS), one needs to download licensed map data, which could be open or closed. Then, one has to have a mapping software to display that data. This software must also be appropriately licensed. Google Maps is not open source, although it is *open access*, in that it is free to use. An open source equivalent of Google Maps is Open Street Maps,[111] a community built tool that is permissively licensed as CC-BY-SA.[112] One could use data from Glottolog or the ELP and then map provide a map using Open Street Map while using entirely open source applications, but the end result could be reproduced on Google Maps with the same lack of restrictions - the only difference is that the engine making Google Maps would be a black box.

It is this mixed use case that is most common - researchers or NLP practitioners use a mix of open and closed resources, as needed. Gawne and Ring (2016) mention many programs: Google Earth[113] (closed source, free) for base maps; Geotag[114] (free, open source) and Photo KML[115] (free) for accessing GIS embedded in pictures taken on iPhones (closed); the KML and KMZ formats,[116] originally developed by Google for Google Earth but now standards implemented by the Open Geospatial Consortium[117] and licensed openly and freely; Koredoko[118] for viewing GIS data in photos (closed, free);

[109] http://www.worldgeodatasets.com/language/. Last accessed April 25, 2018.
[110] http://www.languagelandscape.org. Last accessed April 25, 2018.
[111] https://www.openstreetmap.org/
[112] https://www.openstreetmap.org/copyright
[113] https://www.google.com/earth/
[114] http://geotag.sourceforge.net/
[115] http://www.visualtravelguide.com/Photo-kml.html. This URL was provided in Gawne and Ring (2016), but may be down permanently.
[116] http://www.opengeospatial.org/standards/kml/
[117] http://www.opengeospatial.org
[118] https://itunes.apple.com/us/app/koredoko-exif-and-gps-viewer/id286765236

CartoDB[119] (proprietary) and CartoCSS[120] (free, open); TileMill[121] (free, open, but no longer maintained or updated) and MapBox[122] (open, freemium) ; QGIS[123] (free, open); the SQL[124] language (free, open - languages and formats also have licensing laws and can be copyrighted[125]); JPEG[126] and PNG[127] image formats (free, open); Adobe PhotoShop[128] (closed source, paid); and CartoHexa[129] (free, closed).

An example of a mixed workflow would be using a closed source application or website to shim open source data. For example:

> To give some more general locational context we downloaded some Open Access geopolitical boundaries for Nepal from the Global Administrative Areas website.[130] This data was downloaded as KMZ, which TileMill cannot read, so we opened the files in Google Earth (remember ... that KMZ is a compressed KML) and resaved them as KML, which TileMill can read. (Gawne and Ring, 2016, 228)

This particular use-case may have benefited from a specific tool which could convert KMZ to KML. A cursory look on GitHub shows 54 repositories that could be relevant[131], including one which does solely this task (albeit with Spanish documentation).[132] Using an entirely open source pipeline for working with language (or GIS data, as here) is rare, although it is hypothetically possible; however, one quickly runs into problems where open source is concerned, as each subsequent layer of computational processing must then depend upon open source - including the operating system (for instance, GNU/Linux as an open source alternative to the closed Mac OS), processor, silicon chips, and so on. (This is one of the reasons that copyleft

[119] http://cartodb.com
[120] https://github.com/mapbox/carto
[121] http://www.mapbox.com/tilemill
[122] https://www.mapbox.com/mapbox-studio
[123] http://www.qgis.org
[124] https://www.iso.org/committee/45342/x/catalogue/p/1/u/0/w/0/d/0
[125] Interestingly, constructed natural languages can also be licensed and copyrighted, leading to legal complications involving corporations suing fan communities for publishing documentation in a given language. Further discussion is out of scope here.
[126] https://www.iso.org/standard/54989.html
[127] https://www.iso.org/standard/29581.html
[128] https://www.adobe.com/products/photoshop.html
[129] http://www.colorhexa.com
[130] http://gadm.org/download.
[131] https://github.com/search?p=1&q=kmz+kml&type=Repositories
[132] https://github.com/fadamiao/kmz2kml

remains an issue in licensing.) Idiomatically put: there are turtles all the way down.

As Hu (2012); Hu et al. (2018) notes, the general trend in mapping software has been away from native (meaning on the OS level) applications and towards web applications, which may have a steeper learning curve, but which afford remote storage and access, and users over the Internet. WALS uses LeafletJS[133], an open source mapping software that uses Open Street Maps as an alternative to using an embedded Google Maps map using their API. Hu et al. (2018) suggests a workflow that uses Leaflet along with jQuery[134], an open source JavaScript utility library, to display GIS linguistic maps. Further study around using only FLOSS software for displaying GIS data for linguistics is necessary.

Cenerini et al. (2017) cite several open source software applications and libraries they used in their study mapping the Cree-Innu-Naskapi continuüm using data from the Algonquian Linguistic Atlas (Junker and Stewart, 2011),[135] but don't open source their own code. This would have been useful, specifically as replicating their study using R (Ihaka and Gentleman, 1996) would require researchers to write all of their own queries again. More on data privacy will be discussed in Section 5.6.

This was a small example, looking at only a couple of papers and showing how following open source methodology can be difficult, and how using mixed source applications is often necessary for research and linguistic information. This was a single use case, and every application involving NLP requires navigating software and licensing laws. My purpose in providing this study was to point out how describing the state of open source code that could be used for LRLs is not clear cut. One could argue that this case is reflective of linguistic software, as opposed to NLP or computational linguists. This arbitrary division is not useful, as all actors using language data that has been digitally encoded fall under the wider umbrella of users of human language technology. Languages do not exist within a vacuüm, and computational linguists using NLP to run deep learning artificial intelligence algorithms on spoken language corpora at scale depend upon previous work done by linguists, language communities, and researchers who spent time on the ground formalising orthographies, compiling dictionaries, and debating the finer points of linguistic minutiae.

That having been said, there are cases where using open source software is decidedly clear cut. For instance, if the goal is to build a part of speech tagger using two hours of annotation, you could use the low-resource-post-

[133] http://leafletjs.com/. Last accessed April 26, 2018.
[134] https://jquery.com/. Last accessed April 26, 2018.
[135] http://www.atlas-ling.ca/. Last accessed April 26, 2018.

tagging-2014 package developed as part of Garrette et al. (2013); Garrette and Baldridge (2013), and available on GitHub[136] without any other considerations than downloading Java and learning a bit of Scala, both free and open source languages. But this is a very limited use case, as this package was built as part of two scientific papers studying this narrowly scoped area.

## 5.2 LRL NLP available through data providers

Rather than exhaustively study each possible use case involving NLP, another strategy is to look at the databases where NLP practitioners, researchers, and language activists find code for their respective languages directly. Using the list of aggregators from Section 3.1, it is possible to give a general overview of what is available.

The first resource aggregator listed starts on the lower end of the language resource pyramid: Unicode's CLDR resources. Unicode is often the first port-of-call for a language team working on developing scripts for their language, unless the script is already using some preëxisting format (such as the Roman alphabet). CLDR has instructions on checking out their open source subversion repository online.[137] They also have a GitHub repository[138] and organisation with code for digesting the normally XML representation in JSON, the notation format used most often by JavaScript developers. However, CLDR isn't an aggregator - it's more of a suite of tools under one umbrella, as the scope is limited to working with the Unicode format.

Finding resources isn't easy. The Endangered Languages Project, for instance, contains information on over 3000 languages, and has 6830 languages.[139] None of these resources are code: the searchable formats are: Format, Image, Video, Document, Audio, Link, Guide. Glottolog only has academic references, and ODIN only has interlinear glossed text (IGT) corpora. Omniglot describes alphabets but doesn't index tooling for them. CLARIN has thousands of resources - but none of them are code, and you need to be an accredited researcher from a European institution to access them. The ELRA site provides hundreds of corpora resources - for purchase.

The Linguistic Data Consortium has a tool page,[140] where it notes five tools that may be useful for researchers using its data. These tools are Annotation Graph Kit (AGTK), The Champollion Toolkit, the LDC Word Aligner, Sphere

[136] https://github.com/dhgarrette/low-resource-pos-tagging-2014
[137] http://cldr.unicode.org/index/downloads. Last accessed on April 24, 2018.
[138] https://github.com/unicode-cldr/cldr-json. Last accessed on April 24, 2018.
[139] http://www.endangeredlanguages.com/resources. Last accessed on April 24, 2018.
[140] https://www.ldc.upenn.edu/language-resources/tools. Last accessed April 26, 2018.

Conversion tools, and XTrans, of which only Sphere has a non-standard license that allows use but may have more restrictions. This suite of tools is particularly useful for dealing with LDC data. This sort of tool and corpus bundling is common; when building a resource, the tools to manage that resource are included directly in a tools page. DoBes has the same type of page[141], where they mention tools developed at The Language Archive: ELAN,[142] a powerful tool for time aligned annotation of video or audio data; ARBIL, a metadata catalogue creation tool; LAMUS, a tool for uploading data and metadata into the DoBes archive and for managing existing collections; and LEXUS, a web-based lexicon tool. This scale is common, but there are some archives that have more tools listed. For instance, the Resource Network for Linguistic Diversity (RNLD), a largely Australian network, lists dozens of tools and applications that could be useful.[143] The list doesn't differentiate between bundled code that works as an application on the OS, and code which must be downloaded and run through a terminal. EMELD, the Electronic Metastructure for Endangered Languages Data (a short-term project run through LDC, the ELF, and the Universities of Arizona, Eastern Michigan, and Wayne State) has a similar list with hundreds of items.[144]

For field linguistics, this mixture of apps and small tooling is common, as is combining corpora with tools in some fashion. For instance, Caballero (2017) presents a fieldwork paper on Choguita Rarámuri (Tarahumara), an Uto-Aztecan language. In the paper, they mention using Microsoft Word[145] and Excel[146], SIL s Fieldwork Language Explorer (FLEx),[147] and show screenshots of Quicktime.[148] The corpus they present is stored on the Endangered Language Archive at SOAS, University of London[149] (Caballero, 2009). The majority of these tools are closed source, except for FLEx. However, they also mention using ELAN. Caballero (2017) made their own tools to work with ELAN, and they made this code available on GitHub.[150] In order to access the data, the reader is likely to have read the paper; thus the document, the code, and the corpus together form a unit of research, which are all used together.

[141]http://dobes.mpi.nl/archive_info/tools/. Last accessed April 26, 2018.
[142]https://tla.mpi.nl/tools/tla-tools/elan. Last accessed April 26, 2018.
[143]http://www.rnld.org/software. Last accessed April 26, 2018.
[144]http://emeld.org/school/toolroom/software/software-display.cfm. Last accessed April 26, 2018.
[145]https://products.office.com/en-CA/word. Last accessed April 26, 2018.
[146]https://products.office.com/en-CA/excel. Last accessed April 26, 2018.
[147]http://software.sil.org/fieldworks/download/. Last accessed April 26, 2018.
[148]https://support.apple.com/quicktime. Last accessed April 26, 2018.
[149]elar.soas.ac.uk/deposit/0056. Last accessed April 26, 2018.
[150]https://github.com/ucsd-field-lab/kwaras. Last accessed April 26, 2018.

OLAC, with hundreds of thousands of resources, has a tooling page,[151] which mainly helps with working with OLAC as opposed to pointing to resources which can be used with language data. Unfortunately, searching for software resources comes up short. A short look at a specific language, Naskapi, shows 23 resources[152], most of which are published papers - except for the Crúbadán archive, the Glottolog reference, typological references on WALS and on the Rosetta Project, and a pointer to resources noted on the LINGUIST List, which lists no resources when accessed.[153] Scottish Gaelic isn't much different (26 resources),[154] although it does point to some corpora.

META-SHARE, which aggressively pursues open access and open licensing for resources in its database, has 344 tools available, and allows easy searching for these tools,[155] although signing up as a user is necessary. The ACL Wiki has resources for 84 languages,[156]. From a random selection of these resources (including corpora), one could get an idea of how many resources are totally aggregated: Arabic (16), Navajo (1), Catalan (8), Faroese (4), Galician (20), Maltese (4), Irish (10). LT-World lists 523 separate tools (this number was reached by adding up all resource amounts listed on their language tools page[157] and assuming that there is no duplication of tools, which may be inaccurate). These tools are for all languages, and only a subset could be understood to apply to LRLs.

These collected resources are, to my knowledge, the main place to look for aggregated data around software resources on particular languages. This overview was brief; more fine-tuned exploration of the numbers of packages would likely not improve our understanding of the ecosystem. From this, it is clear that there are global open source software resources in the order of hundreds, not thousands. Considering that thousands of languages have not ascended digitally, this is neither unexpected nor ideal.

## 5.3 Multilingual NLP libraries

While searching for code that has been tagged with metadata noting the language it serves has some merits, there are also possibilities for using generic code on many languages. For instance, Bender (2016) explores the field of mul-

[151] http://www.language-archives.org/tools.html. Last accessed April 26, 2018.

[152] http://www.language-archives.org/language/nsk. Last accessed April 26, 2018.

[153] https://linguistlist.org/olac/search-olac.cfm?LANG=nsk. Last accessed April 26, 2018.

[154] http://www.language-archives.org/language/gla. Last accessed April 26, 2018.

[155] http://www.meta-share.org. Last accessed April 26, 2018.

[156] https://aclweb.org/aclwiki/List_of_resources_by_language. Last accessed April 26, 2018.

[157] http://www.lt-world.org/kb/resources-and-tools/language-tools. Last accessed April 26, 2018.

tilingual NLP (now decades old; for instance, Kay (1997) called for this in the 90s), pointing out that there is a growing body of research that uses language typology to abstract and identify language features which allow for applying NLP systems from one language to another.

> Businesses developing commercial products with NLP are interested in the markets represented by `low resource languages` (LRLs; i.e., those languages for which there are not many digitized data sets or basic NLP systems such as part-of-speech taggers or morphological or syntactic parsers), some of which represent very large populations in emerging economies. Finally, researchers looking to apply NLP techniques to assist in endangered language documentation are naturally interested in developing NLP systems that work across very diverse languages.(Bender, 2016, 646)

Bender (2016) goes on to mention the LinGo Matrix system (Bender et al., 2002; Drellishak and Bender, 2005) that can be used to create rule-based grammars for natural languages using linguistic typographical data. The LinGo Matrix, and all work within the DELPH-IN system (a collaboration looking mainly at the Head-Driven Phrase Structure Grammar and Minimal Recursion Semantics) is open source.[158]

They also mention projecting resources across languages, such as Yarowsky et al. (2001), who projected linguistic annotations like POS tags and noun phrase parsing from English to French and Chinese, by using bilingual texts that had been word-aligned. This was extended in the previously mentioned Agić et al. (2015), who similar POS tagger projection for one hundred LRLs. Their code is open source on Bitbucket.[159]

This avenue of research is fascinating and broad, because it allows for small tools to be applied to other LRLs at a minimal cost. A study involving looking at all of the available research, with an in-depth look in each scientific article that includes links to source code, would be warranted and welcomed. It is unfortunately largely out of scope for this thesis; it is enough, here, to know that open source code for LRLs is dependent upon academics working in this field sharing their code on large repositories, and that this code must also be adapted to each particular LRL, which, while an extensive task, is made easier through multilingual NLP and cross-linguistic projection.

At a lower level, there are NLP toolkits which are useful for working with LRL datasets, which are language agnostic. The most well known is arguably the Natural Language Toolkit (NTLK)[160] (Bird, 2006), a free and open source

[158] http://www.delph-in.net/wiki/index.php/Software. Last accessed April 26, 2018.
[159] https://bitbucket.org/lowlands/. Last accessed April 26, 2018.
[160] http://nltk.org. Last accessed April 26, 2018.

Python library that enables users to interface with over fifty different corpora and lexical resources, and which provides a suite of tools such as tokenizers and parsers which can be used in sparse data contexts. A primer written by the main creators (Bird et al., 2009)[161], is used frequently in natural language processing classes written by the creators. It is licensed under the Apache 2.0 license, an open source license [162]. On GitHub, there are currently 204 contributors listed[163], and the contribution history in Git shows 234 (found by using the command `git authors`). Some of the resources within NLTK work especially well with LRLs. 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, or reimplementations in another programming language which allows use in different coding language ecosystems. One of these is the JavaScript language implementation.[164] This has 6700 stars on GitHub, which, since they reflect favouritism from individual users, is a good indicator of community vitality and use, and 88 contributors. The port is also open source, under an MIT license.[165]

It is difficult to track usage of these open source software packages by LRL communities or researchers, as, once downloaded, there are no convenient metrics which lead back to the original source. Code, when run, generally leaves no trace. Again, the fundamental problem of tracking LRL open source software inhibits understanding the ecosystem, but it is clear from individual anecdotes and through scientific citations that work is being done in this area.

## 5.4 Linked open 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)

## 5.5 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

[161] Available online at http://nltk.org/book. Last accessed April 26, 2018.
[162] https://github.com/nltk/nltk/blob/develop/LICENSE.txt. Last accessed April 26, 2018.
[163] https://github.com/nltk/nltk/graphs/contributors. Last accessed April 10, 2018.
[164] https://github.com/NaturalNode/natural. Last accessed April 26, 2018.
[165] https://github.com/NaturalNode/natural#license. Last accessed April 26, 2018.

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.

## 5.6 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 LRLs, 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.

Sometimes, privacy revolves less around the users or the language communities, and more around researchers not wishing to open source their code until they are done developing their project, or until a grant ends, or until they are safe that they won't be scooped by other researchers. For instance, in a paper describing a tool for sharing interlinearized and lexical data in different formats, Kaufman and Finkel (2018) notes that "Kratylos will be made open-source and accessible to the public through a GitHub repository at the end of the current grant period. Kratylos is built entirely from open- source software itself and transcodes proprietary media formats into the open-source codecs Ogg Vorbis (for audio) and Ogg Theora (for video)." This is particularly insightful, as it shows that an understanding of open source can still be tied to initial closed-source development.

# 6 Case Studies

## 6.1 Scottish Gaelic

Scottish Gaelic is a Celtic language spoken mainly in the United Kingdom, which UNESCO defines as *definitely endangered*.[166] 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 [167] (Scannell, 2007).

ODIN has exactly 59 IGT entries for Scottish Gaelic.

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.

## 6.2 Naskapi

Naskapi is a Cree language in the Algonquin family spoken in central Quebec MacKenzie and Jancewicz (1994), which UNESCO defines as *vulnerable*.[168] 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[169]), 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

[166] http://www.unesco.org/languages-atlas/en/atlasmap/language-iso-gla.html
[167] http://crubadan.org/languages/gd
[168] http://www.unesco.org/culture/languages-atlas/en/atlasmap/language-id-2354.html
[169] https://en.wikipedia.org/wiki/Innu

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.[170]

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.

### 6.2.1 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.[171]

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 preëxisting 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

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

[171] 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) [172] 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.[173] Users can export as a PDF (among other file formats), or export words to an online interface known as Webonary.[174] 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,[175] (free, open source software available on GitHub [176]) 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.

### 6.2.2 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.[177] 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-

[172] https://software.sil.org/fieldworks/
[173] https://github.com/sillsdev/FieldWorks
[174] https://www.webonary.org/configuring-the-dictionary-in-flex/
[175] https://keyman.com/
[176] https://github.com/keymanapp
[177] 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.

### 6.2.3 Resources on the web

ODIN has exactly one IGT entry for Naskapi, from (Richards, 2004).

The Naskapi community website, run by the Naskapi Nation of Kawawachikamach, has a webpage on installing Naskapi syllabics.[178]

[178] http://www.naskapi.ca/en/Install-syllabics

# 7 Methods

## 7.1 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.

Work published without a license on a public site is not technically open source, either. When software is not licensed, it by default reverts (in the USA) to copyright where *all rights are reserved*, which is by definition not FLOSS. For this reason, it's important to add a license to code if it is in your purview to do so, and if you wish to follow the open source methodology.

## 7.2 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.

## 7.3 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.[179] I'll discuss 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.

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

# 8 Discussion

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

## 8.1 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.

## 8.2 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.

# 9 Future Work

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

## 9.1 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.

# 10 Conclusion

Here I will conclude with some closing remarks.

# References

Ager, S. (2018). Omniglot writing systems and languages of the world. http://omniglot.com. Last accessed April 17, 2018].

Agić, Ž., Hovy, D., and Søgaard, A. (2015). If all you have is a bit of the Bible: Learning POS taggers for truly low-resource languages. In *The 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference of the Asian Federation of Natural Language Processing (ACL-IJCNLP 2015)*.

Allamanis, M. and Sutton, C. (2013). Mining source code repositories at massive scale using language modeling. In *Proceedings of the 10th Working Conference on Mining Software Repositories*, pages 207–216. IEEE Press.

Anderson, G. D. and Harrison, K. D. (2006). Language hotspots: Linking language extinction, biodiversity, & human knowledge base. *Living Tongues Institute for Endangered Languages Occasional Papers*.

Annis, W. S. (2018). *Horen lì'fyayä leNa'vi: A Reference Grammar of Na'vi*. learnnavi.org.

Barnes, N. (2010). Publish your computer code: it is good enough. *Nature News*, 467(7317):753–753.

Beller, M., Bholanath, R., McIntosh, S., and Zaidman, A. (2016). Analyzing the state of static analysis: A large-scale evaluation in open source software. In *Software Analysis, Evolution, and Reengineering (SANER), 2016 IEEE 23rd International Conference on*, volume 1, pages 470–481. IEEE.

Bender, E. M. (2016). Linguistic typology in natural language processing. *Linguistic Typology*, 20(3):645–660.

Bender, E. M., Flickinger, D., and Oepen, S. (2002). The grammar matrix: An open-source starter-kit for the rapid development of cross-linguistically consistent broad-coverage precision grammars. In *Proceedings of the 2002 workshop on Grammar engineering and evaluation-Volume 15*, pages 1–7. Association for Computational Linguistics.

Bender, E. M. and Good, J. (2010). A grand challenge for linguistics: Scaling up and integrating models. *White paper contributed to NSF's SBE*, 2020:1–1.

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

Berez-Kroeker, A. L., Andreassen, H. N., Gawne, L., Holton, G., Kung, S. S., Pulsifer, P., Collister, L. B., the Data Citation and Attribution in Linguistics Group, and the Linguistics Data Interest Group (2017). The Austin principles of data citation in linguistics. Draft, Version 0.1.

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

Bickerton, D. (2016). *Roots of language*. Language Science Press.

Bickford, J. A., Lewis, M. P., and Simons, G. F. (2015). Rating the vitality of sign languages. *Journal of Multilingual and Multicultural Development*, 36(5):513–527.

Binnenpoorte, D., Cucchiarini, C., D'Halleweyn, E., Sturm, J., and De Vriend, F. (2002). Towards a roadmap for human language technologies: Dutch-flemish experience. In *Proceedings lrec 2002*.

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.

Bird, S., Klein, E., and Loper, E. (2009). *Natural language processing with Python: analyzing text with the natural language toolkit*. " O'Reilly Media, Inc.".

Bodó, C., Fazakas, N., and Heltai, J. I. (2017). Language revitalization, modernity, and the Csángó mode of speaking. *Open Linguistics*, 3(1):327–341.

Boersma, P. and Weenink, D. (2018). Praat: doing phonetics by computer [computer program].

Boettiger, C. (2015). An introduction to docker for reproducible research. *ACM SIGOPS Operating Systems Review*, 49(1):71–79.

Bourhis, R. Y. (2001). Reversing language shift in Quebec. *MULTILINGUAL MATTERS*, pages 101–141.

Brenzinger, M., Yamamoto, A., Aikawa, N., Koundiouba, D., Minasyan, A., Dwyer, A., Grinevald, C., Krauss, M., Miyaoka, O., Sakiyama, O., et al. (2003). Language vitality and endangerment. *Paris: UNESCO Intangible Cultural Unit, Safeguarding Endangered Languages*, 1:2010.

Caballero, G. (2009). Choguita raraÌĄmuri description and documentation. http://elar.soas.ac.uk/deposit/0056. Last accessed on April 26, 2018.

Caballero, G. (2017). Choguita rarámuri (tarahumara) language description and documentation: a guide to the deposited collection and associated materials. *Language Documentation and Conservation*, 11:224–255.

Callison-Burch, C., Koehn, P., Monz, C., and Zaidan, O. F. (2011). Findings of the 2011 workshop on statistical machine translation. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 22–64. Association for Computational Linguistics.

Calzolari, N., Del Gratta, R., Francopoulo, G., Mariani, J., Rubino, F., Russo, I., and Soria, C. (2012). The lre map. harmonising community descriptions of resources. In *LREC*, pages 1084–1089.

Calzolari, N., Soria, C., Del Gratta, R., Goggi, S., Quochi, V., Russo, I., Choukri, K., Mariani, J., and Piperidis, S. (2010). The lrec map of language resources and technologies. In *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*.

Cenerini, C., Junker, M.-O., and Rosen, N. (2017). Mapping dialectal variation using the algonquian linguistic atlas. *Language Documentation & Conservation*, 11:305–324.

Chew, P. A., Verzi, S. J., Bauer, T. L., and McClain, J. T. (2006). Evaluation of the Bible as a resource for cross-language information retrieval. In *Proceedings of the workshop on multilingual language resources and interoperability*, pages 68–74. 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*, chapter 141. Max Planck Institute for Evolutionary Anthropology.

contributors, W. (2018). Wikipedia — Wikipedia, the free encyclopedia. https://en.wikipedia.org/wiki/Wikipedia. Last accessed April 17, 2018.

Darwish, K. (2013). Arabizi detection and conversion to arabic. *arXiv preprint arXiv:1306.6755*.

Del Gratta, R., Frontini, F., Khan, A. F., Mariani, J., and Soria, C. (2014). The LREMap for under-resourced languages. *CCURL 2014: Collaboration and Computing for Under-Resourced Languages in the Linked Open Data Era*, page 78.

Del Gratta, R., Frontini, F., Monachini, M., Pardelli, G., Russo, I., Bartolini, R., Goggi, S., Khan, F., Quochi, V., Soria, C., et al. (2015). Visualising italian language resources: a snapshot. *CLiC it*, page 100.

Dobrin, L. M. (2009). Sil international and the disciplinary culture of linguistics: Introduction. *Language*, 85(3):618–619.

Dobrin, L. M. and Good, J. (2009). Practical language development: Whose mission? *Language*, 85(3):619–629.

Drellishak, S. and Bender, E. M. (2005). Coordination modules for a crosslinguistic grammar resource. *D epartam ento de Inform ática Facu ld ad ede C iências da U niversid ad ede L isb oa C am po G rande, 1749–016 L isboa Portu gal*, page 29.

Dryer, M. S. and Haspelmath, M., editors (2013). *WALS Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig.

Dwyer, A. M. (2012). Tools and techniques for endangered-language assessment and revitalization. *Minor. Lang. Today's Glob. Soc.*

Eide, E. (2010). Toward replayable research in networking and systems. *Position paper presented at Archive*, page 5.

Elenius, K., Forsbom, E., and Megyesi, B. (2008). Language resources and tools for swedish: A survey. In *LREC*.

Everett, D. L. (2009). *Don't sleep, there are snakes: Life and language in the Amazonian jungle*. Profile Books.

Farrar, S., Lewis, W., and Langendoen, T. (2002). A common ontology for linguistic concepts. In *Proceedings of the Knowledge Technologies Conference*, pages 10–13.

Farrar, S. and Lewis, W. D. (2007). The gold community of practice: An infrastructure for linguistic data on the web. *Language Resources and Evaluation*, 41(1):45–60.

Finley, K. (2011). Github has surpassed sourceforge and google code in popularity. *ReadWrite, 02-Jun-2011.*

Fishman, J. A. (1991). *Reversing language shift: Theoretical and empirical foundations of assistance to threatened languages*, volume 76. Multilingual matters.

Fishman, J. A. (2001). *Can threatened languages be saved?: Reversing language shift, revisited: A 21st century perspective*, volume 116. Multilingual Matters.

FitzJohn, R., Pennell, M., Zanne, A., and Cornwell, W. (2014). Reproducible research is still a challenge. *http://ropensci.org/blog/2014/06/09/reproducibility*.

Garrette, D. and Baldridge, J. (2013). Learning a part-of-speech tagger from two hours of annotation. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 138–147.

Garrette, D., Mielens, J., and Baldridge, J. (2013). Real-world semi-supervised learning of pos-taggers for low-resource languages. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 583–592.

Gawne, L. and Ring, H. (2016). Mapmaking for language documentation and description. *Language Documentation & Conservation*, 10:188–242.

Ghani, R., Jones, R., and Mladenić, D. (2001). Mining the web to create minority language corpora. In *Proceedings of the tenth international conference on Information and knowledge management*, pages 279–286. ACM.

Ghosh, R. A., Glott, R., Krieger, B., and Robles, G. (2002). Free/libre and open source software: Survey and study.

Gibson, M. (2016). Assessing digital vitality: analytical and activist approaches. *CCURL 2016 Collaboration and Computing for Under-Resourced Languages: Towards an Alliance for Digital Language Diversity*, page 46.

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

Ginsburgh, V. and Weber, S. (2011). *How many languages do we need?: The economics of linguistic diversity*. Princeton University Press.

Gorenflo, L. J., Romaine, S., Mittermeier, R. A., and Walker-Painemilla, K. (2012). Co-occurrence of linguistic and biological diversity in biodiversity hotspots and high biodiversity wilderness areas. *Proceedings of the National Academy of Sciences*, 109(21):8032–8037.

Gousios, G. and Spinellis, D. (2012). Ghtorrent: Github's data from a firehose. In *Mining software repositories (msr), 2012 9th ieee working conference on*, pages 12–21. IEEE.

Gousios, G., Vasilescu, B., Serebrenik, A., and Zaidman, A. (2014). Lean ghtorrent: Github data on demand. In *Proceedings of the 11th working conference on mining software repositories*, pages 384–387. ACM.

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.

Grenoble, L. A. (2016). A response to 'Assessing levels of endangerment in the Catalogue of Endangered Languages (ELCat) using the Language Endangerment Index (LEI)', by Nala Huiying Lee & John Van Way. *Language in Society*, 45(2):293–300.

Grover, A. S., Van Huyssteen, G. B., and Pretorius, M. W. (2011). The south african human language technology audit. *Language resources and evaluation*, 45(3):271–288.

Hammarström, H. (2015). Glottolog: A free, online, comprehensive bibliography of the world's languages. In *3rd International Conference on Linguistic and Cultural Diversity in Cyberspace*, pages 183–188. UNESCO.

Hammarström, H., Forkel, R., Haspelmath, M., and Bank, S. (2015). Glottolog 2.3. *Leipzig: Max Planck Institute for Evolutionary Anthropology. Available online at http://glottolog.org*.

Hanks, D. H. (2017). Policy barriers to Ainu language revitalization in Japan: When globalization means English. *Working Papers in Educational Linguistics (WPEL)*, 32(1):5.

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

Hu, S. (2012). Multimedia mapping on the internet using commercial apis. In *Online Maps with APIs and WebServices*, pages 61–71. Springer.

Hu, S., Karna, B., and Hildebrandt, K. (2018). Web-based multimedia mapping for spatial analysis and visualization in the digital humanities: a case study of language documentation in nepal. *Journal of Geovisualization and Spatial Analysis*, 2(1):3.

Ihaka, R. and Gentleman, R. (1996). R: a language for data analysis and graphics. *Journal of computational and graphical statistics*, 5(3):299–314.

Ince, D. C., Hatton, L., and Graham-Cumming, J. (2012). The case for open computer programs. *Nature*, 482(7386):485.

Izreel, S. (2003). The emergence of spoken israeli hebrew. *Corpus Linguistics and Modern Hebrew: Towards the Compilation of the Corpus of Spoken Israeli Hebrew*, pages 85–104.

Johnson, D. (2013). *Language Policy*. Springer.

Junker, M.-O. and Stewart, T. (2011). A linguistic atlas for endangered languages: www. atlas-ling. ca. In *Proceedings of EDULEARN 11: International Conference on Education and New Learning Technologies*, pages 4–6.

Kalliamvakou, E., Gousios, G., Blincoe, K., Singer, L., German, D. M., and Damian, D. (2014). The promises and perils of mining github. In *Proceedings of the 11th working conference on mining software repositories*, pages 92–101. ACM.

Kaufman, D. and Finkel, R. (2018). Kratylos: A tool for sharing interlinearized and lexical data in diverse formats. *Language Documentation & Conservation*, 12:124–146.

Kay, M. (1997). The proper place of men and machines in language translation. *machine translation*, 12(1-2):3–23.

Kilgarriff, A. and Grefenstette, G. (2001). Web as corpus. In *Proceedings of Corpus Linguistics*, volume 2001, pages 342–344.

Kleinberg, B. and Mozes, M. (2017). Web-based text anonymization with node. js: Introducing netanos (named entity-based text anonymization for open science). *The Journal of Open Source Software*, 2017.

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

Kornai, A. (2015). A new method of language vitality assessment. *Linguistic and Cultural Diversity in Cyberspace*, page 132.

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

Krauss, M. (2007a). Classification and terminology for degrees of language endangerment. *Language diversity endangered*, 181:1.

Krauss, M. E. (2007b). Keynote-mass language extinction and documentation: The race against time. *The vanishing languages of the Pacific rim*, pages 3–24.

Krauwer, S. (1998). Elsnet and elra: A common past and a common future. *ELRA Newsletter*, 3(2):4–5.

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.

Krauwer, S. (2006). Strengthening the smaller languages in europe. In *Proc. of 5th Slovenian and 1st International Language Technologies Conference*, pages 9–10.

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

Labov, W., Ash, S., and Boberg, C. (2005). *The atlas of North American English: Phonetics, phonology and sound change*. Walter de Gruyter.

Labov, W., Preston, D., Britain, D., et al. (2012). Journal of linguistic geography. *Journal of Linguistic Geography*, 1(1).

Lee, N. H. and Van Way, J. (2016). Assessing levels of endangerment in the Catalogue of Endangered Languages (ELCat) using the Language Endangerment Index (LEI). *Language in Society*, 45(2):271–292.

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

Lewis, M. P. and Simons, G. F. (2010). Assessing endangerment: expanding Fishman's GIDS. *Revue roumaine de linguistique*, 55(2):103–120.

Lewis, M. P., Simons, G. F., Fennig, C. D., et al. (2009). *Ethnologue: Languages of the world*, volume 16. SIL international Dallas, TX.

Lewis, M. P., Simons, G. F., Fennig, C. D., et al. (2018). Language status | Ethnologue. https://www.ethnologue.com/about/language-status. Last accessed on April 24, 2018.

Lewis, W. (2010). Haitian Creole: How to build and ship an MT engine from scratch in 4 days, 17 hours, & 30 minutes. In *14th Annual conference of the European Association for machine translation*. Citeseer.

Lewis, W. D., Munro, R., and Vogel, S. (2011). Crisis MT: Developing a cookbook for MT in crisis situations. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 501–511. Association for Computational Linguistics.

Lewis, W. D. and Xia, F. (2010). Developing odin: A multilingual repository of annotated language data for hundreds of the world's languages. *Literary and Linguistic Computing*, 25(3):303–319.

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.

Lowndes, J. S. S., Best, B. D., Scarborough, C., Afflerbach, J. C., Frazier, M. R., O'Hara, C. C., Jiang, N., and Halpern, B. S. (2017). Our path to better science in less time using open data science tools. *Nature ecology & evolution*, 1(6):0160.

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.

Maegaard, B., Krauwer, S., Choukri, K., and Jørgensen, L. (2006). The blark concept and blark for arabic. In *Proc. LREC*, pages 773–778.

Maffi, L. (2001). *On biocultural diversity: Linking language, knowledge, and the environment*. Smithsonian Inst Pr.

Maffi, L., Krauss, M., and Yamamoto, A. (2001). The world languages in crisis: Questions, challenges, and a call for action. In *Conference Handbook on Endangered languages of the Pacific Rim*, pages 75–78. Endangered Languages of the Pacific Rim Project.

Mäkelä, E. (2016). Las: an integrated language analysis tool for multiple languages. *The Journal of Open Source Software*, 2016.

Mapelli, V. and Choukri, K. (2003). Report on a (minimal) set of lrs to be made available for as many languages as possible, and map of the actual gaps. *ENABLER Deliverable D*, 5:22.

Mariani, J. and Francopoulo, G. (2015). Language matrices and a language resource impact factor. In *Language Production, Cognition, and the Lexicon*, pages 441–471. Springer.

Miller, M., Aean, T., Tuiq, and Littauer, R. (2018). *Na'vi-English Dictionary, version 13.63*. learnnavi.org.

Moran, S., McCloy, D., and Wright, R., editors (2014). *PHOIBLE Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig.

Moshagen, S., Rueter, J., Pirinen, T., Trosterud, T., and Tyers, F. (2014). Open-source infrastructures for collaborative work on under-resourced languages. *CCURL 2014: Collaboration and Computing for Under-Resourced Languages in the Linked Open Data Era*, pages 71–77.

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

Nettle, D. and Romaine, S. (2000). *Vanishing voices: The extinction of the world's languages*. Oxford University Press on Demand.

Newman, P. (1998). We has seen the enemy and it is us: the endangered languages issue as a hopeless cause. *Studies in the Linguistic Sciences*, 28(2):11–20.

Oard, D. W. (2003). The surprise language exercises. *ACM Transactions on Asian Language Information Processing (TALIP)*, 2(2):79–84.

Okrent, A. (2009). *In the land of invented languages: Esperanto rock stars, Klingon poets, Loglan lovers, and the mad dreamers who tried to build a perfect language*. Spiegel & Grau.

Paricio Martín, S. and Martínez Cortés, J. (2010). New ways to revitalise minority languages: the repercussions of the internet in the case of aragonese. *Digithum. The humanities in the digital age*, 0(12).

Prensky, M. (2001). Digital natives, digital immigrants part 1. *On the horizon*, 9(5):1–6.

Quilter, J. and Urton, G. (2002). *Narrative threads: Accounting and recounting in Andean khipu*. University of Texas Press.

Ratner, N. B. and Menn, L. (2000). In the beginning was the wug: Forty years of language elicitation studies. *Methods for studying language production*, pages 11–26.

Raymond, E. (1999). The cathedral and the bazaar. *Knowledge, Technology & Policy*, 12(3):23–49.

Resnik, P. (1999). Mining the web for bilingual text. In *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics*, pages 527–534. Association for Computational Linguistics.

Richards, N. W. (2004). The syntax of the conjunct and independent orders in wampanoag. *International Journal of American Linguistics*, 70(4):327–368.

Roberts, J. A., Hann, I.-H., and Slaughter, S. A. (2006). Understanding the motivations, participation, and performance of open source software developers: A longitudinal study of the apache projects. *Management science*, 52(7):984–999.

Rögnvaldsson, E., Loftsson, H., Bjarnadóttir, K., Helgadóttir, S., Whelpton, M., Nikulásdóttir, A. B., and Ingason, A. K. (2009). Icelandic language resources and technology: Status and prospects. In *Proceedings of the NODALIDA 2009 workshop Nordic Perspectives on the CLARIN Infrastructure of Language Resources*, volume 5 of *NEALT Proceedings Series*, pages 27–32. Northern European Association for Language Technology (NEALT).

Scannell, K. P. (2007). The crúbadán project: Corpus building for under-resourced languages. In *Building and Exploring Web Corpora: Proceedings of the 3rd Web as Corpus Workshop*, volume 4, pages 5–15.

Schreyer, C. (2011). Media, information technology, and language planning: what can endangered language communities learn from created language communities? *Current Issues in Language Planning*, 12(3):403–425.

Schreyer, C. (2015). The digital fandom of Na'vi speakers. *Transformative Works & Cultures*, 18.

Schwab, M., Karrenbach, M., and Claerbout, J. (2000). Making scientific computations reproducible. *Computing in Science & Engineering*, 2(6):61–67.

Scott-Phillips, T. C. and Kirby, S. (2010). Language evolution in the laboratory. *Trends in cognitive sciences*, 14(9):411–417.

Shah, S. K. (2006). Motivation, governance, and the viability of hybrid forms in open source software development. *Management science*, 52(7):1000–1014.

Simons, G. and Bird, S. (2003). The open language archives community: An infrastructure for distributed archiving of language resources. *Literary and Linguistic Computing*, 18(2):117–128.

Simov, K. I., Osenova, P., Kolkovska, S., Balabanova, E., and Doikoff, D. (2004). A language resources infrastructure for bulgarian. In *LREC*.

Smith, A. M., Niemeyer, K. E., Katz, D. S., Barba, L. A., Githinji, G., Gymrek, M., Huff, K. D., Madan, C. R., Mayes, A. C., Moerman, K. M., et al. (2018). Journal of open source software (joss): design and first-year review. *PeerJ Computer Science*, 4:e147.

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

Soria, C., Quochi, V., Russo, I., Gurrutxaga, A., and Ceberio, K. (2017). A digital language vitality scale and indicators. Draft.

Spice, B. (2012). Carnegie Mellon releases data on Haitian Creole to hasten development of translation tools. http://www.eurekalert.org/pub_releases/2010-01/cmu-cmr012710.php. Last accessed August 27, 2012.

Streiter, O., Scannell, K. P., and Stuflesser, M. (2006). Implementing nlp projects for noncentral languages: instructions for funding bodies, strategies for developers. *Machine Translation*, 20(4):267–289.

Thung, F., Bissyande, T. F., Lo, D., and Jiang, L. (2013). Network structure of social coding in github. In *Software maintenance and reengineering (csmr), 2013 17th european conference on*, pages 323–326. IEEE.

Trudgill, P. (1983). *On dialect: Social and geographical perspectives*. Blackwell, Oxford.

UNESCO (2011). Directory international cooperation programs for the protection and promotion of languages and linguistic diversity. Accessible at http://www.unesco.org/new/fileadmin/MULTIMEDIA/HQ/CLT/pdf/International%20co

UNESCO, A. (2014). UNESCO atlas of the world's languages in danger. http://www.unesco.org/culture/languages-atlas/. Last accessed October 19, 2014.

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

Williams, A. C., Wallin, J. F., Yu, H., Perale, M., Carroll, H. D., Lamblin, A.-F., Fortson, L., Obbink, D., Lintott, C. J., and Brusuelas, J. H. (2014). A computational pipeline for crowdsourced transcriptions of Ancient Greek papyrus fragments. In *Big Data (Big Data), 2014 IEEE International Conference on*, pages 100–105. IEEE.

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.

Wittenburg, P. (2003). The dobes model of language documentation. *Language Documentation and Description*, 1:122–139.

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.

Yarowsky, D., Ngai, G., and Wicentowski, R. (2001). Inducing multilingual text analysis tools via robust projection across aligned corpora. In *Proceedings of the first international conference on Human language technology research*, pages 1–8. Association for Computational Linguistics.