Saarland University
Department of Computational Linguistics

Master's thesis Proposal

Open Source Code and Low Resource Languages

Author:
Richard Littauer
Matriculation: 2539658

Supervisors:
Dr. Dietrich Klakow
Dr. Alexis Palmer

March 7, 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.

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

Montréal, March 7, 2018

Richard Littauer

Contents

1 Introduction **1**

2 Low Resource Languages: An Overview **4**
2.1 Definitions 4
2.2 Metrics 4
2.3 Digital presence 4
2.4 The current state of language diversity 4
2.5 Who makes resources for LRLs? 5
2.6 Language research funding 5

3 Open Source Code **6**
3.1 Defining *Open Source* 6
3.2 Where is open source code? 6
3.3 Digital Permanence and Storage 6
3.4 Data and privacy 6
3.5 Legal rights and liability 7
3.6 Military and enterprise solutions 7
3.7 Funding 7
3.8 Ethical reasons for using open source 7

4 Open Source Code for Low Resource Languages **8**
4.1 BLARK and beyond 8
4.2 NLTK and other open source libraries 8
4.3 A Database for Open Source Code 8
4.4 Linked Data 8

5 Case Studies **9**
5.1 Scottish Gaelic 9
5.2 Naskapi 9

6 Methods **10**
6.1 Choosing a license 10
6.2 Choosing repositories 10
6.3 Sharing code without a platform 10

7 Discussion **11**
7.1 Why isn't more code open? 11
7.2 How does open source demonstrably help? 11

8 Future Work **12**

8.1 Beyond Wikipedia and Ethnologue 12

9 Conclusion **13**

1 Introduction

At least half of the world's 7000 languages will be extinct this century (Grenoble, 2011, p. 27). Just over half of these languages have writing systems.[1] It is estimated that less than 5% of the world's languages will be used online or have significant digital presence (Kornai, 2013).

The majority of technological infrastructure used globally has been built with English, and serves English speakers. There are a few languages - perhaps thirty - with the combination of large populations with internet access, official governmental status, and Westernised, industrial economies which affords them a foothold on the web.[2]

English is the undisputed heavyweight as far as global written resources are concerned. Over half of the web's content is written in English (ibid.). 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.

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

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

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

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

instance, the most consistent translations cross-linguistically are of the Christian bible, which may not reflect the target language's culture.

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

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

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

Below, I will go into further depth about the state of endangered languages and computational resources in Section 2, and what different languages need in order to have digital presence. In Section 3, I'll define what open source is, and talk about issues relevant to open source code for under-resourced languages; specifically, data rights, liability, privacy, funding, military and industrial concerns, ethical reasons for using open source. I'll then in Section 4 talk about the state of open source code currently available online, in particular focusing on a database of open source code that I have built with the help of researchers around the world.

I'll touch on some specific examples of languages which could benefit from open source code in Section 5, focusing on Gaelic, an endangered language with tens of thousands of speakers but little online resources, and Naskapi, an endangered languages with only a thousand speakers which might be able to benefit from open

source code. The Naskapi case study will be largely informed by original research, as I engaged in field research at the town where most Naskapi live and talk to linguists working on literacy efforts for this language. In Section 6, I'll discuss how open source can help low resource languages, and in Section 7 I'll expound further at a high level on what open source enables for linguists and language communities. Finally, in Section 8 and Section 9 I'll discuss future work, and offer some concluding remarks.

2 Low Resource Languages: An Overview

In this section, I will outline the state of endangered languages.

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* languages mean. 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 define these terms: endangered, minority, low-resource, under-resourced, threatened, moribund, dormant, extinct, computer (such as JavaScript), revitalised, and constructed languages. This will help inform why I've chosen to focus on *low-resource* languages (LRLs), and specifically low-resource natural languages with living populations.

2.2 Metrics

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

2.3 Digital presence

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

Few of the metrics above take into account the level of digital literacy for a language. The possibility for a language to digitally ascend has been held up as a key component of judging a language's vitality by Kornai (2013).

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

2.4 The current state of language diversity

In this section, I am going to briefly go into detail about what diversity means for linguistics. This will be useful later for explaining how related languages can be used to bootstrap work in similar languages. For instance, Irish spell-checkers and

constitutional corpora from the EU can be used by Scottish Gaelic speakers with some tweaks in order to further improve their own systems.

2.5 Who makes resources for LRLs?

Here, I will explain briefly who makes language resources for these languages. I'll explain what I see as the main groups doing this work: professional translators, educators, missionaries (of multiple faiths, but mostly Christian), academics and native technologists. I'll explain each stakeholder and their canonical perspectives.

2.6 Language research funding

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

3 Open Source Code

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

3.1 Defining *Open Source*

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

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

3.2 Where is open source code?

Here, I will include a short section on how the open source world works. In particular, I'll answer the question of where code lives. I'll include a short overview and case study on GitHub, SourceForge, and some academic archival sites (UPenn, Max Planck, DFKI).

3.3 Digital Permanence and Storage

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

3.4 Data and privacy

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

3.5 Legal rights and liability

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

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

3.6 Military and enterprise solutions

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

3.7 Funding

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

3.8 Ethical reasons for using open source

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

4 Open Source Code for Low Resource Languages

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

4.1 BLARK and beyond

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

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

4.2 NLTK and other open source libraries

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

4.3 A Database for Open Source Code

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

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

4.4 Linked Data

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

5 Case Studies

5.1 Scottish Gaelic

Scottish Gaelic is a Celtic language spoken mainly in the United Kingdom, which UNESCO defines as *definitely endangered* [3].

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

5.2 Naskapi

Naskapi is a Cree language in the Algonquin family spoken in central Quebec MacKenzie and Jancewicz (1994), which UNESCO defines as *vulnerable* [4]. Virtually the entire population of around 800 Naskapi live within the reservation Kawawachikamach, around 10 miles from Schefferville, QC.

In October 2017 I travelled to Schefferville and interviewed linguists working on a Naskapi bible, visited the school and talked to teachers at length about language efforts there, and talked to individuals around the town about their thoughts on the language and how it is used. I'll include a summary of Naskapi here, outlining current efforts and future possibilities for the language, and how open source code can help.

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

6 Methods

6.1 Choosing a license

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

6.2 Choosing repositories

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

6.3 Sharing code without a platform

I'll outline a plan for peer-to-peer resource sharing, using IPFS (Benet, 2014) and other related tech. I'll mention a case study involving local indigenous communities in Guyana using peer-to-peer to track illegally logging on their land, and explain how this system could also be used for language development.[5]

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

7 Discussion

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

7.1 Why isn't more code open?

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

7.2 How does open source demonstrably help?

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

8 Future Work

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

8.1 Beyond Wikipedia and Ethnologue

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

9 Conclusion

Here I will conclude with some closing remarks.

References

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

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

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

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

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.

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

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.

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.

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

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.