Open Source Code and Low Resource Languages: Thesis Proposal

Richard Littauer *on* 24.01.2018

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, it needs to have 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 to bootstrap 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 (with help from others) of open source code serving these languages. Looking to the future, I discuss steps for helping save languages from virtual extinction.

Sections

- Intro
- An overview of low resource languages
- Open source code
- Open source code for low resource languages
- Case studies
- Discussion
- Future work
- Conclusion

Introduction

- Briefly cover the topics: languages are dying, what computational linguistics is, how to help languages digitally ascend, how open source can help.

- Outline the paper in general.

2. LRL: An Overview

- Define the terms being used
- Talk about how to gauge a language's endangerment
- What digital presence for a language means
- Why language diversity matters for computational linguistics
- Who makes resources?
- And who funds them?

3. Open Source Code

- Define “open source”
- Where does open source code happen
- What does a long term goal for permanence and storage for code look like?
- How do we deal with privacy?
- Legal rights?
- What about the military? Or industry? How do they influence open source?
- What about funding for open source?
- Are there ethical questions to cover?

4. Open Source Code for LRLS

- The Basic Language Resource Kit (BLARK)
- NLTK, other large libraries
- The endangered-languages database I set up
- Linked data (what happened to the semantic web?)

5. Case Studies

- Scottish Gaelic

- Naskapi

6. Methodologies

- Specific recommendations for how to use Open Source to help endangered languages

- Recommendations for a license, and for choosing repositories

- How to share code without a platform using p2p

6. Discussion

- Why isn’t more code open?

- How does open source *actually* help?

7. Future Work

- What’s after Wikipedia and Ethnologue?

8. Conclusion

- Wrap it up with some nice words.