Open Source Code Serving Endangered Languages

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

Abstract

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

Keywords: open source, under resourced languages, ontology, database, endangered languages, code

1. Paper

Almost half of the approximately 7,000 currently spoken languages are expected to become extinct this century; it is estimated that less than 5% of these will be used online or have significant digital presence (Kornai 2013). Languages which do not have significant digital resources (often called low-resource, under-resourced, or minority languages) risk extinction from loss of prestige, specific domain usage (such as online), and ultimately loss of speakers. Many language communities and academics working with them try to prevent this by developing tools, websites, and resources for their languages. However, these approaches are fragmented, often incurring large developmental and funding costs for single, non-extensible use cases. To make this process easier for all stakeholders, we have built the first database (to our knowledge) of all open source code projects related to low-resource languages.

Our database is structured as an simple list in Markdown format, hosted in a GitHub repository. GitHub is the largest online network of open source code and allows for parallel collaborative development, while providing critical collaborating support via a built in wiki, issue tracking and comments on new suggestions. The features provided by Github are increasingly important to academics and the field of education (Zagalsky 2015). These features were not available via previous code and text sharing solutions such as SourceForge and are often not available via institutional repositories. Our list is structurally simple, shareable, easily updated, and part of a wider cultural movement on GitHub of using Markdown files as simple databases (Sorhus 2013). Our list monopolizes on the low barrier of entry on GitHub (Storey et al. 2014). Our list currently describes over 241 open source projects, which includes specific sections for extensible code for over 26 different languages.

Our solution is open source, not just in the sense of code and data availability (or disclosure) but in the sense of collaborative involvement and inclusive discussion. By using a solution like GitHub we were able to reach out to both developers and researchers. Our list is updatable more rapidly than previous solutions like http://lingtransoft.info/. We mitigate the risk of a single point of failure (a recent issue with LinguistList) by working in a distributed fashion. The data in our list is open and can be copied and reused by anyone, something not necessarily true of previous solutions. By choosing a commercial solution, we sidestep many of the institution issues associated with archives and repositories currently servicing academics.

Our ultimate goal is a collaboratively built and maintained resource for highlighting useful, extensible code for low-resource languages. We would like to share our current efforts and welcome communication with the wider academic linguistics community.

2. Bibliographical References

3. Language Resource References