Open Source Code Serving Endangered Languages

Richard Littauer, Hugh Paterson III
Saarland University, SIL International
Saarbrücken, Germany, Oregon, USA
richard.littauer@gmail.com, hugh_paterson@sil.org

Abstract

We present a database of open source code that can be used by low-resource language communities to build digital resources. Our database 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, database, endangered languages, code, computational resources

1. Introduction

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 (**?**). 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 a 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). Because the list is part of GitHub, and because it is not maintained, funded, or dependent upon a large institution or funding body (but rather run by independent contributors), the list itself is a novel way of crowd-sourcing data and resources from the linguists, computational linguists, and coders who would use the tools themselves.

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/, or other resources like GOLD (http://linguistics-ontology.org/), ELCAT (http://www.endangeredlanguages.com/), or OLAC (http://www.language-archives.org/) . 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 such a 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. Database structure

The list itself largely consists is a single Markdown file. This is useful for a couple of reasons; first, there's no need for a gateway or endpoint to access the content of the database, as there would be if it were coded in SQL, RDF, or some other relational database. Secondly, people can search the list using their browser and the standard search feature for any website. Finally, the list can be digested immediately instead of depending on searches to get complete coverage of the data.

2.0.1. Categories

Within the list, there are several main sections where we attempt to categorize the resources, based on user input and upon the best guesses of the maintainers about the functionality of the various tools. These sections are: Generic repositories (which includes massive dictionary and lexicography projects, single language lexicography projects, utilities, presentations of data, and software), i18n-related repositories, audio automation, text-automation, experimentation, flashcards, natural language generation, computing systems, android applications, Chrome extensions, FieldDB, FieldDB web-services and components and plug-

ins, academic research paper-specific repositories, example repositories, and language and code interfaces.
We also have two other lists: One of other open source linguistic organisations, on GitHub, and other OSS (Open Source software) organisations, and another for language-specific projects, which includes subsections for code which is relevant to: Amharic, Arabic, Bengali, Chichewa, Estonian, Georgian, Guarani, Hausa, Hindi, Høgnorsk, Inuktitut, Irish, Japanese, Kinyarwanda, Korean, Lingala, Malay, Malagasy, Migmaq, Minderico, Nishnaabe, Oromo, Quechua, Sami, Scottish Gaelic, Secwepemctsín, Somali, Tigrinya, Zulu.
Finally, we also include a short list of closed source resources which can still be utilized for free.

2.0.2. Example entry

Each entry is a single line, containing the name of the resource, a link to the resource, and a short description. If the resource is also a GitHub repository, we include a link to a badge that shows the amount of stars (similar to likes or favorites on other social media sites, and, generally, a good proxy for usage and developer uptake of the resource) for that repository.
Here is one such entry: "* [Chichewa ![GitHub stars](https://img.shields.io/github/stars/kscanne/chichewa.svg)](https://github.com/kscanne/chichewa) NLP resources for Chichewa"

3. Personas and Stakeholders

The list that we are maintaining is not aimed at any one group in isolation. Instead, there are several key groups whom we think may derive value from this list. These are as follows: Project Managers, Software Developers, Community Developers, and Linguists,

3.1. Project Managers

Project managers are generally linguists or community members who have been tasked with developing language resources, but generally don't have the skill to understand the technical aspects on their own (and thus are different from software developers). Many project managers do not have strong information technology project management backgrounds. However, they are often skilled linguists who have access to grant funding. The finer details of carrying out a project in a manner which benefits more than one language community is simply out of scope for many first time managers, and projects. We hope that project managers should be able to look at our list to be able to determine two things: 1. Has the project task, goals, or relevant deliverables already been accomplished for another language (or indeed, for the same language), and 2. where can they find the code base for that project so that they can integrate it into their own workflow. We hope, as well, that some project managers might be able to have a project used case and that they can use our list to answer the question of how to get funding for their idea.

3.2. Software Developers

Software developers are often looking to find pre-existing solutions, and to find pre-existing modules which can be applied to new use cases they are asked to solve. Every problem which has already been solved by someone else is time that does not have to be spent developing. Every major project today uses open source code in some capacity, partially for this reason. This is in some ways the opposite perspective from the project manager, who are looking to develop something new and may not be looking for extensible solutions to their problem. The developer is looking for use cases to which they can apply or reapply their code. We hope that our list makes this easy.
One developer, at least, has said that this was the case: "Thanks a lot for pointing me to the list. It is awesome! It has some really good tools and resources which would be very useful in a lot of things that I am doing. I shall definitely add some of my resources and tools to this list....I have also come across a very useful library - Poio-api - on your list which is a parser for most of these XML files that I work with." (Personal communication, 2015)

3.3. Community Developers doing Language Development

This may include language development organizations, or individuals who are "community members" looking to deploy their own solutions. They want to know ?does it work with my language? by which they often, but not always mean ?written language?. An additional complexity, is that there are different perspectives on what ?does it work? really mean. The degree of technologicalization in the majority culture affects the expectation of the kinds of real world tasks desired to be accomplished in the low-resourced language. Tasks might be sending SMS messages in a particular script. But sometimes, users of low-resource languages have a very different expectation and interaction with technology. For instance, members of deaf communities require video integration in their digital solutions more than many other kinds of Low-resource languages, but deaf communities may still need other tools commonly shared with written languages - like dictionary tools.

3.4. Linguists

We define linguists here as researchers who are not tech savy, but who are working in academia or directly with language communities from an academic perspective. Linguists, as such, are generally looking for patterns in language data. They want tools which are going to be easy to use to find the patterns they are looking for and present the data in ways which help others to understand the purpose and meaning of those patterns. As end users, they are more likely to be looking for tools which are useful out of the box, and so may not be able to appreciate all of the items in the list, but still may benefit from a quick search through it. As well, we provide many links to tools that can be used with ELAN (http://www.mpi.nl/corpus/html/elan/), Praat (http://www.fon.hum.uva.nl/praat/), and other audio software which are used on a day-to-day basis by many linguists themselves.

4. Commitments

Unlike most projects, which must depend upon institutions, private or public funding, or volunteers, this project has no

single point of failure. The list itself is currently hosted by Richard Littauer's GitHub account, but due to the nature of a Git repository and of collaborative work on GitHub, any replication of the list can be edited, stand alone, and be used if the original project goes down for any reason. The decentralized quality of git repositories makes this list a much less brittle solution than database or institutional repositories.

As well, there is a very low possibility of misuse; if any one person using the program decides to enforce their own viewpoint, perspective, or rules at the cost of any other user group or of the community, it is entirely possible for anyone else to make a copy of the list and to then use that as the source of truth going forward.

One possible issue with the list is that if a malicious user takes over the main list. Then, it would take some time for any other list to have the same clout in the community. This is, however, true of all existing databases. A good example, although not open data, is the Ethnologue, which recently added a paywall to their database about languages (but not to the ISO 639-3 standard which SIL International also stewards). The loss of a previously costlessly accessible resource became a major source of contention for many linguists, although SIL International had legitimate reasons for doing so (namely, the financial cost of hosting the database).

One of the future goals of the project is to develop a community where anyone can ask questions about the list and its resources, and other users can help out and give advice easily. While this is possible with GitHub issues, it depends upon a higher amount of usage of the list itself than currently exists. Marketing the list in tech conferences is one possible solution.

5. Conclusion

We have here outlined our reasons for developing a list of open source software for endangered languages. We hope that this resource is used by the community, and that this paper fosters discussion and awareness of open source resources.

6. Bibliographical References

7. Language Resource References