An Online Platform for Community-Based Language Description and Documentation

Rebecca Everson, Wolf Honoré, and Scott Grimm

February 26, 2019

Introduction

The talk presents work in progress to develop a technological platform designed to accelerate and make more inclusive the process of documenting and describing languages

Two pieces of interlocking front-end technology:

- a mobile app
  - speakers submit written text, pictures, voice recordings and/or videos
  - functionality and simplicity is reminiscent of WhatsApp
- an online platform
  - serves as a Community Language Portal for each language
  - organizes submitted data for presentation to the community

Introduction

One source of inspiration: developing a dictionary through rather standard methodologies that were available at the inception of the project

- 2011-2018 Grimm developed Dagaare-English Dictionary together with Mark Ali, College of Education, Winneba, Ghana
- developed 7000 word dictionary in Toolbox
- collaborated both in Ghana and via email
- converted it to LaTeX for Language Science Press (Everson & Grimm 2017)

Introduction

While the final dictionary was a good outcome, process was challenging:

- extremely time-consuming
- limited dialectical variation recorded
- no audio or visual information
- without further modification, information is not directly usable for other purposes

Technological advances provide tremendous potential to accelerate and improve this and similar workflows

Introduction

Beyond accelerating certain workflows for field linguistics, we are trying to address some challenges broadly facing researchers engaged in language documentation and description:

I. community benefit

II. time with and access to community members

III. trade-off between recording breadth of sociolinguistic diversity and depth of grammatical and lexical content

IV. making documentation products available to the community and to diverse academic communities

Community benefit

A core goal of many documentation projects is to develop resources that benefit the speaker communities (Cameron et al., 1992; Dwyer, 2006; Yamada, 2007; Czaykowska-Higgins, 2009)

Obstacles:

- direct exposure to the linguist or documentation team often highly limited for general population
- often the goal of documentation products (e.g., archived materials) remains abstract, especially as no tangible products are typically developed until final stages of the project

Time with and access to community members

A fieldworker documenting a language has finite time and access to directly engage with the community

Obstacles:

- geopolitical issues of access (conflict; unsafe conditions)
- less dramatically, community members often have their own busy lives
- these difficulties tend to increase exponentially the greater number of speakers the fieldworker tries to engage

Analytic depth vs. sociolinguistic diversity

Researchers are often faced with a choice between depth or breadth:

- Collect a large amount of materials from a small group of speakers to develop in-depth treatments of particular language phenomena
- Or collect fewer materials from a larger more demographically diverse set of speakers to better address social and linguistic variation

Accessibility of documentation products

The creation of digital resources that are “multi-purpose” is far from trivial (Bird and Simons, 2003)

- Often a language (documentation) corpus is most useful for that researcher, and less so for other linguists or the community

Despite a broader conversation in the language documentation community to move towards more “participatory” archive models (Shilton and Srinivasan, 2007; Good, 2011; Dobrin and Holton, 2013), community involvement with language archives or other products often remains infrequent, as there are often serious barriers to access.

- Access to the Endangered Language Archive (ELAR), for instance, requires moderate computer literacy and literacy in English.

Accessibility of documentation products

An outstanding challenge is to develop data collection methods which:

I. are participatory and develop significant presentation products

II. deliver high-quality archivable materials

III. use data formats which support language research as well as the development of computational systems.

Addressing Challenges

While there is no one-size-fits-all solution, the mobile app and online platform we are developing aims to make progress on meeting these challenges:

- permit field linguists, community educators and other stakeholders to serve in the capacity of “language community coordinator”, assigning tasks that are collaboratively achieved with the community

The back-end of this platform provides a structured database, which serves multiple purposes:

- construction of data sets for traditional and computational linguistics research
- development of language resources for the community
- and preparation for archival preservation

Addressing Challenges

Using mobile app technology permits distributing fieldwork tasks to a large number of consultants who can complete the tasks at their convenience.

- circumvents the bottleneck of time and access since the researcher and speaker need not be in the same location
- can address some aspects of trade-off between depth of description and diversity of speakers:
  - potential access to a large number of speakers
  - ongoing ability to contact them and gather data through a mobile app long after the end of a field trip

Addressing Challenges

We seek to build community engagement through a front-end Community Language Portal, where community members can contribute and comment on the contributions of others.

- This provides immediate relevance of speakers' efforts

The success of any community engagement will vary based on community attitudes towards their language, culture and preservation, and one of the central questions we explore is what serves as motivation for contributing language-based resources across different communities.

Extended Prototype Example

We will walk through a prototype we are currently testing through which language researchers and communities can collaboratively develop lexicographic resources

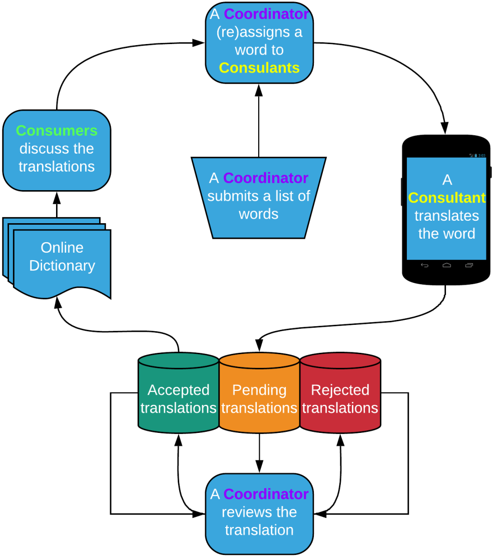
- the platform is much broader and can be used for a range of documentation/description activities
- emphasis on very simple design to increase user accessibility

User Types

Three primary user types:

- **coordinator** assigns tasks (typically linguist and/or community member)
- **consultants** respond via the mobile app (written text, pictures, audio or video)
- **consumers** are users/community members using the Community Language Portal (CLP)
  - web interface that organizes and displays the accepted submissions
  - provides community members with an ongoing view of their community's contributions along with their own

Lifecycle of a word

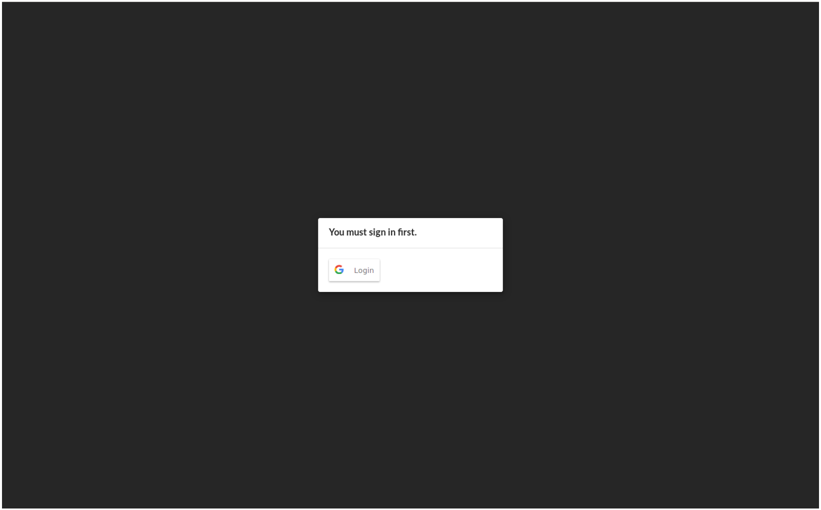

A Coordinator (re)assigns a word to Consulants
Consumers discuss the translations
A Coordinator submits a list of words
A Consultant translates the word
Online Dictionary
Accepted translations
Pending translations
Rejected translations
A Coordinator reviews the translation

Login to mobile app

Translate words (offline)

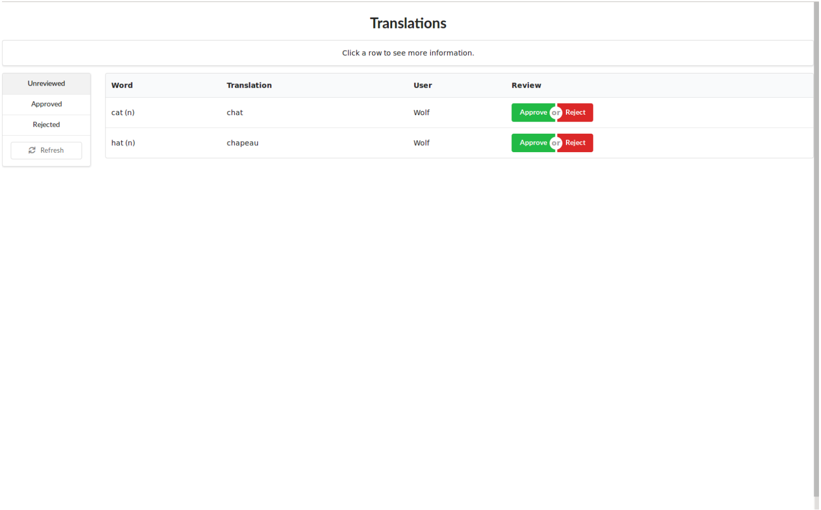
Submit translations

Coordinator console login



You must sign in first.

View all new submissions

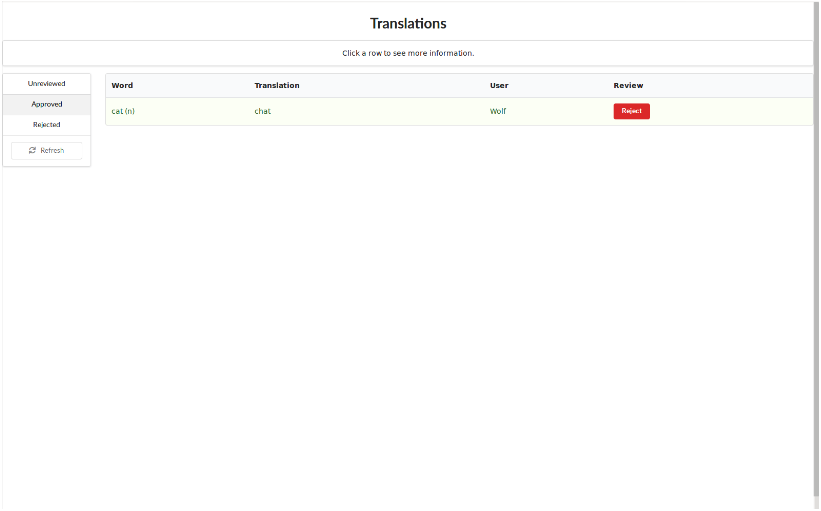


Translations

Click a row to see more information.

Word	Translation	User	Review
cat (n)	chat	Wolf	Approve or Reject
hat (n)	chapeau	Wolf	Approve or Reject

See all approved submissions

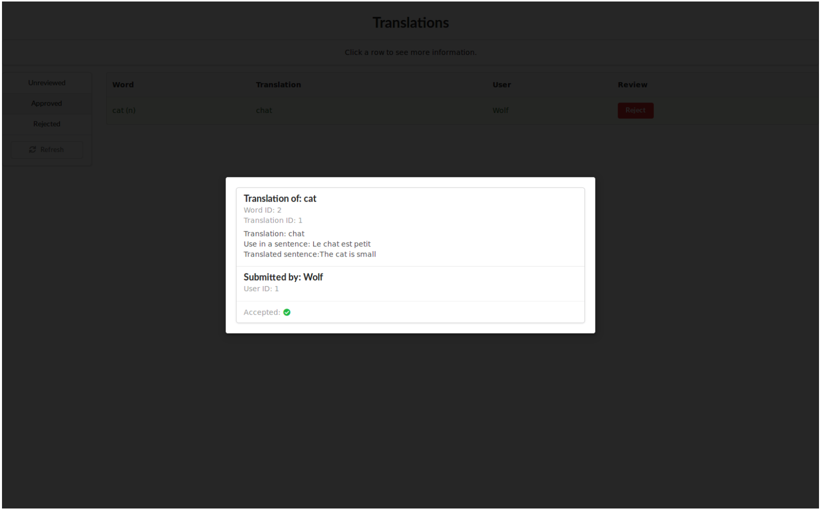


Translations

Click a row to see more information.

Word	Translation	User	Review
cat (n)	chat	Wolf	Reject

View in detail



Translations

Click a row to see more information.

Word	Translation	User	Review
cat (n)	chat	Wolf	Reject

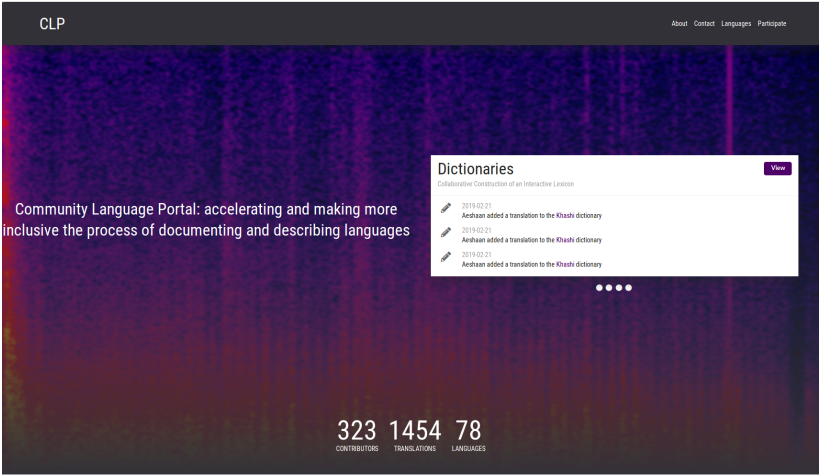
Translation of: cat
Word ID: 2
Translation ID: 1

Translation: chat
Use in a sentence: Le chat est petit
Translated sentence:The cat is small

Submitted by: Wolf
User ID: 1

Accepted: ✓

Community Language Portal



Community Language Portal: accelerating and making more inclusive the process of documenting and describing languages

Dictionaries

Collaborative Construction of an Interactive Lexicon

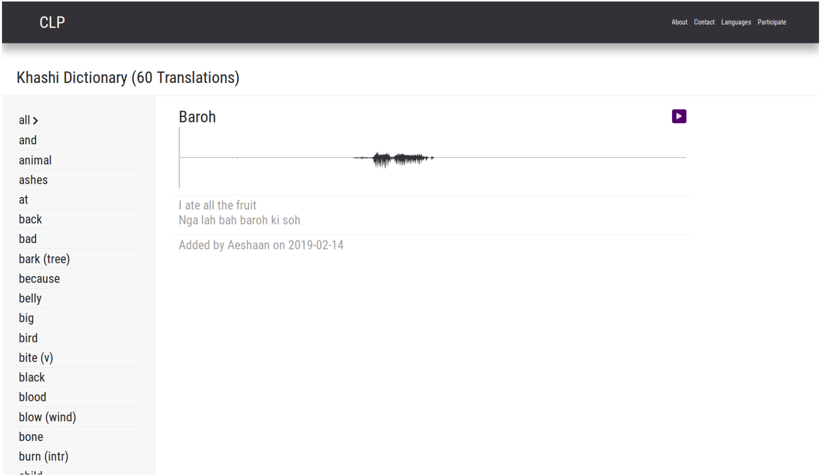
2019-02-21
Aeshaan added a translation to the Khashi dictionary

2019-02-21
Aeshaan added a translation to the Khashi dictionary

2019-02-21
Aeshaan added a translation to the Khashi dictionary

323	1454	78
CONTRIBUTORS	TRANSLATIONS	LANGUAGES

View Dictionary



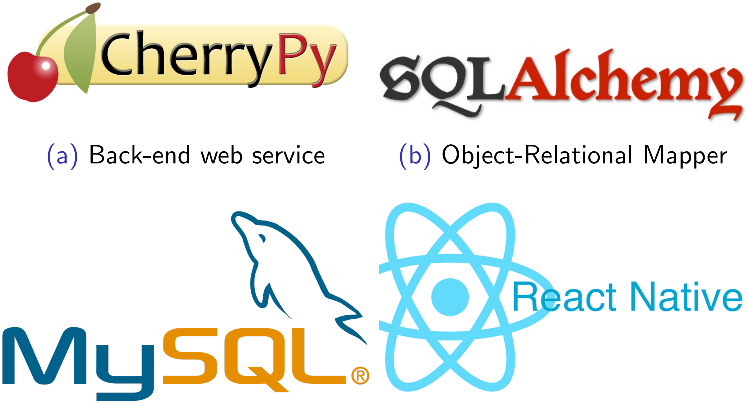
Khashi Dictionary (60 Translations)

Baroh

I ate all the fruit
Nga lah bah baroh ki soh

Added by Aeshaan on 2019-02-14

Technical overview



(a) Back-end web service

(b) Object-Relational Mapper

(c) Database

(d) Mobile JavaScript framework

- RESTful API
- Handles all communication between database and client layers
- Lightweight HTTP server

SQLAlchemy

- Object-Relational Mapper
- Utilize abstraction in Python while maintaining performance
- Transparent, readable Python adaptations of SQL commands

React Native

- Mostly platform agnostic
- Develop mobile app in JavaScript, compiles to native OS language
- Widely supported by industry giants

Future platform enhancement

Community discussion platform

- Inspired by StackOverflow
- Invite discussion from community members about translations, sample sentences, pronunciation, dialect variation
- Utilize upvoting/downvoting system

Consultant enhancements

- Skip or mark words for reassignment
- Mark words as irrelevant/not translatable in target language
- Receive feedback from coordinator
- Train *expert consultants* to complete more challenging assignments (transcription, annotation)

Coordinator enhancements

- Reassign words to specific consultants
  - If previous consultant did not provide an adequate sample sentence
  - If another consultant has a better device for recording audio
- Assign consultants and words to subgroups
  - Solicit information from populations with specialized vocabularies (farmers, weavers, etc.)

Future use

Building in interoperability and multi-purpose compatibility into our back-end design:

- archival preservation
- data sets for linguistics research
- development of language resources (dictionaries, literacy materials)
- conversion into standard data serialization schemas common in natural language processing to feed the use of NLP tools, e.g. multilingual parsing from raw text to universal dependencies

Outlook

Immediate next step is broad user testing:

- different communities
- different devices
- different degrees of technology and internet access

Thanks

Thanks to the Digital Scholarship Lab at the University of Rochester (Emily Sherwood, Joshua Romphf) for their work on the Community Language Portal, to Dan Weiner for early work on mobile interface, and to Maya Abtahian, Nadine Grimm and Aaron White for discussion.

Bird, S. and Simons, G. (2003). Seven dimensions of portability for language documentation and description. *Language*, 79:557–582.

Cameron, D., Frazer, E., Harvey, P., Rampton, M., and Richardson, K. (1992). *Researching language: Issues of power and method*. Routledge.

Czaykowska-Higgins, E. (2009). Research models, community engagement, and linguistic fieldwork: Reflections on working within Canadian indigenous communities. *Language Documentation and Conservation*, 3(1):1550.

Dobrin, L. M. and Holton, G. (2013). The documentation lives a life of its own: The temporal transformation of two endangered language archive projects. *Museum Anthropology Review*, 7(1-2):140–154.

Dwyer, A. M. (2006). Ethics and practicalities of cooperative fieldwork and analysis. *Essentials of language documentation*, pages 31–66.

Good, J. (2011). Data and language documentation. In Austin, P. K. and Sallabank, J., editors, *The Cambridge Handbook of Endangered Languages*, pages 212–234. Cambridge University Press, Cambridge.

Shilton, K. and Srinivasan, R. (2007). Participatory appraisal and arrangement for multicultural archival collections. *Archivaria*, 63:87–101.

Yamada, R.-M. (2007). Collaborative linguistic fieldwork: Practical application of the empowerment model. *Language Documentation and Conservation*, 1(2):257–282.