SELF DOCUMENTATION OF ENDANGERED LANGUAGES

Sagun Dhakhwa

Centre for Communication & Development Studies, Nepal

Jens Allwood

SSKKII Interdisciplinary Center, University of Gothenburg, Sweden

ABSTRACT

Several minority languages are on the verge of extinction in Nepal, especially when they don't have a generally accepted writing system and occur in an area where Nepali (the official language) is predominantly used. Lohorung is an example, which is spoken among the Lohroung Rai communities of Sankhuwasabha, a hilly district of eastern Nepal. Older generations of Lohorung are the only experts in Lohorung but they have limitations in reading and writing English or Nepali. The documentation of Lohorung and other similar endangered languages is important. If the right tools and techniques are used, we believe that self documentation is one of the best ways, to document a language. We have developed an online platform using which community members can collaboratively self document their language. The platform is multimodal dictionary authoring and browsing tool and it has been developed with the focus on usability, ease of use and productivity.

Index Terms— Self documentation, multimodal dictionary, crowd-sourcing, Lohorung,

1. INTRODUCTION

Language documentation is the process of recording of the linguistic practices of a speech community, such as a collection of recorded and transcribed texts [1]. Traditionally language documentation has been done by trained professionals, linguists. However, self documentation is a method in which native speakers, often non-linguists (but with some basic training), document their own language. We have developed an online encyclopedic dictionary platform using which native speakers can collaboratively document their language. In this paper we share our experiences of our endeavors to start an initiative where we have empowered Lohorung community members to self document their language. In section 2, we will discuss our approach of self documentation, in section 3 we will introduce the multimodal dictionary tool in brief, in section 4 we will discuss how this tool tries to incorporate best practices of language documentation, in section 5 we will discuss about challenges and finally conclude in section 6 by reporting the progress that has been made in Lohorung community.

2. CROWD-SOURCING: THE APPROACH

Crowd-sourcing is a type of participative online activity in which an individual, an institution, a non-profit organization, or company proposes to a group of individuals of varying knowledge, heterogeneity, and number, via a flexible open call, the voluntary undertaking of a task. [2]. Crowd-sourcing is not a new approach in the Internet world, where tasks, both creative and mundane are commonly outsourced to hundreds of contributors. Very successful large projects like Wikipedia are contributed by thousands of contributors. However, crowd-sourcing may not be limited to the Internet; it can also be done offline, where tasks are not only related to information. The potential of crowd-sourcing can be very significant in documentation of small and endangered languages like Lohorung because crowd-sourcing is not just cost effective but also less time consuming. In contrast, traditional approach of language documentation is a time consuming and expensive. Few well trained linguists work both on the field for language documentation. Languages like Lohorung are being encroached at the rate which cannot be preserved by a handful of linguists. In countries like Nepal there are hundreds of other similar languages.

In our approach, the data collection is done by the crowd after a very basic training while the linguists use the data to produce language description and other materials.

Practically, crowd-sourcing might sound difficult to achieve because out of 1207 Lohorung speakers (according to 2001 Population Census of Nepal, new data is not available) most are old generations [3]. They are often illiterate and have little or no experience of digital equipments. But, they are the only but source of the language. Most of the younger generations have stopped using their native language because of the dominance of Nepali. However, many of them are literate in English and Nepali and may have some acquaintances with computers and digital equipments. This might often be the case in other endangered languages of Nepal and the World.

We encourage families to work together as a crowd where parents and grandparents contribute linguistically and younger generations contribute in collecting data and uploading it to the online platform. Many Lohorung families have already been working in teams in order to self document Lohorung based on their knowledge and technical skills.

Loss of interest is a big challenge in crowd-sourcing and it is inevitable in our case too. We have used social prestige through attention as the main motivator for participation by acknowledging the contribution and rating the highest and best contributors on the online platform [4]. Acknowledgement is given in the form of positive comment by other users. Word of the day and latest added word features on the homepage also gives acknowledgement to the best entries. A good ranking increases social prestige to the contributors and in the case of Lohorung it is even more valuable because the contributors know each other personally as Lohorung is a small language. Ranking are done on the basis of approved complete entries. Contributors compete for social prestige and more and more entries are added to the system.

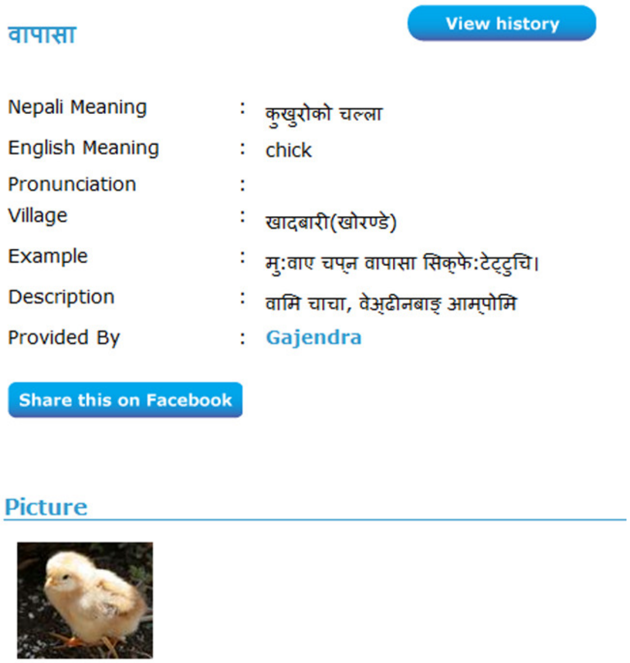
Since we are using an online application platform for crowd-sourcing, usability of the application will be another key factor for the success of the crowd-sourcing. We have followed user centered designs methods for interaction design of the application, working closely with the Lohorung community [5]. We have been doing various usability tests taking continuous feedbacks and improving the user experience of the application based on results. We envisage that a well planned interaction design and user experience will increase the chances of contributions and hence help in self documentation of a language.

3. MULTIMODAL DICTIONARY PLATFORM

The multimodal dictionary is an online web application developed in PHP, Code Ingiter framework, with MySQL database backend. It uses MVC (Model View Controller) architectural pattern making it a robust and easy to maintain system [6]. The system can be broadly divided into a dictionary browser and an administration panel:

3.1. Dictionary browser

The dictionary browser is the public browsing interface, through which a user can search a word textually or graphically. The users can view the word's description in text, audio, pictures or video. Even if the dictionary is based on the words of a target language (here Lohorung) a user who is new to the target language or illiterate can also navigate through the system using pictures to find a word. Apart from pictorial representations, words can be browsed using textual cues or directly searched using a conventional search box. The browser can also be used to write comments on a dictionary entry in order to encourage the contributors to correct any errors in their entries.



वापासा

Nepali Meaning	:	कुखुरोको चल्ला
English Meaning	:	chick
Pronunciation	:	
Village	:	खादबारी(खोरण्डे)
Example	:	मु:वाए चप्न वापासा सिक्फे:टेट्टुचि।
Description	:	वामि चाचा, वेअ्दीनबाङ् आम्पोमि
Provided By	:	Gajendra

Figure 1: Multimodal Dictionary Entry

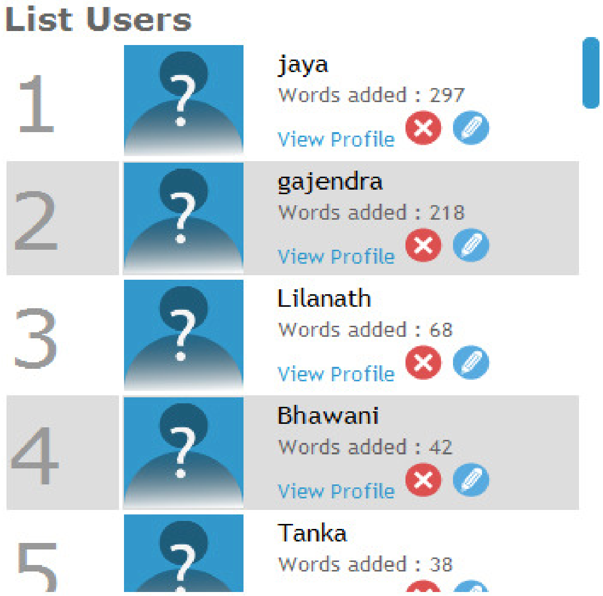
Also the users can add the entries into their watch list so that any activity related to that entry can be kept in track. There is a version control system in the dictionary using which different versions of an entry can be tracked. Also an entry can be switched back to an old version in case of a mistake or vandalism.

On the front end (user side), the UI (user interface) design is simple to understand and easy to use. The application is easy to learn, intuitive and a fun experience. For your reference, a dictionary entry looks like the picture shown in Figure 1.

Social prestige has been used as a reward for contributors. We have presented a contributor rating system in the home page which will list the top contributors as shown in Figure 2, latest word added to the dictionary and word of the day which is elected by the administrator. Another way of attention is a feature to share a contributed entry in Facebook, which can also be useful to bring more traffic to the multimodal dictionary platform. These are the ways to give attention to the contributors so that they continue to contribute to the dictionary and contributors are encouraged to compete with each other yielding high quality contributions.

3.2. Administration panel

In the administration panel, there are two types of users, main administration and contributors. The main administration can manage the word entries like adding words in their respective category and editing/deleting them.


List Users

1 jaya
Words added : 297
View Profile

2 gajendra
Words added : 218
View Profile

3 Lilanath
Words added : 68
View Profile

4 Bhawani
Words added : 42
View Profile

5 Tanka
Words added : 38

Figure 2: User listing based on their contribution

Contributors can only add and view entries while administrative users can monitor other user's activities and moderate new entries made by users. The administration panel has a search component to search words from a large database. It also has a media management component to manage the audio, video and pictures related to a word.

The administration panel can also be used to edit entries contributed by other administrators (or contributors). The version control mechanism is a good way to switch between versions and compare different versions in order to correct the entries. The version mechanism can also be used to study the trends of crowd-sourcing and to study how meanings and usage of a word evolves over the time.

4. ISSUES RELATED TO BEST PRACTICES

The multimodal dictionary is a tool to collect multimodal data (audio, video and text) from crowd-sourcing. The application is available in the Internet making it accessible to virtually everyone in the web. Currently, we are extending the system to allow transcription features in the system so that crowd can also transcribe and annotate the collected media collaboratively and online.

Web is an open standard and is open to all. The contents of the dictionary are stored in remote web servers which can use cloud technology. This ensures availability and continuity of the content because web hosting companies use state of the art technology to store data and take the responsibility of data migration when they are upgrading their servers.

We are also aware of the issues related to content and format of the data that is collected through crowd-sourcing [1].

4.1. Content

In the content, coverage and accountability are two important things that we have considered. The multimodal dictionary is open to the categories of entries and it gives the liberty to the contributors to decide upon the categories of entries. This ensures the contributions to be on the categories of topics which are culturally important to the language community and languages are rich in those areas. The comment feature also allows a non contributor to suggest a new area or entry to be contributed on. An entry in the dictionary stores the details of the contributors, participant, genre of recordings, transcription of the contents in the form of description and examples. In near future, transcription and annotation features will also be available online, which can be used to transcribe the data. Availability of media, its metadata and transcriptions will ensure accountability of data.

4.2. Format

For the data to be available to all the users, the format of data is also an important aspect. We have used open standards like HTML and Unicode encoding schemes. Lohorung was not a written language until recently, when introduced Devanagari based writing system for the Lohorung. Devanagari is the most appropriate system because, Nepali is widely used and Lohorung speakers are well acquainted with Nepali Devanagari. However, there are certain sounds that are not represented in Nepali Devanagari for which we have introduced new symbols available in Devanagari Unicode set. This makes sentences written in Lohorung readily available to all the users in Unicode. For media file storage we have used the formats that are pervasively used in Windows, Mac or Linux.

5. CHALLENGES

There are several challenges that we have encountered and dealt, during the process of developing the Multimodal Dictionary. This section will briefly describe the challenges and our approach to tackle them.

5.1. Literacy and IT experience

For a country like Nepal, literacy can be an issue in using a dictionary, because nearly 41% of the population is non-literate [7]. However, there is a growing trend of sending children to school so the younger generation now has better literacy rate and many of this generation use mobile phones too. Even if this generation is, perhaps, not the best resource for a language to be documented using a multimodal dictionary, their involvement will certainly encourage participation of older family members and can hence aid in the vitalization of the language.

5.2. Use of Crowd-sourcing for data collection

As the productivity in crowd-sourcing exhibits a strong positive dependence on attention, it is a challenge to keep the contributors to a crowd-sourcing platform [4]. Our idea of fun and social prestige therefore has to be well communicated in the dictionary user experience, as social prestige is a way of showing attention to one's contribution. Contributors are motivated by the comments and suggestions by other contributors and/or audiences.

5.3. Availability and sustainability

Availability of the dictionary application is an issue because the dictionary building process is expected to remain an ongoing process. The contents have to be continuously updated if we are to vitalize the language. Communities in remote villages have limited resources to keep the system alive because even if people voluntarily contribute data, there is a cost in running computers, internet and the web application. To mitigate this problem, we have tried to develop a concept of a "self-sustainable telecenter" in a village which will be run by the village community and where the cost of the multimodal dictionary project will be maintained by the income from such telecenter. Telecenter give basic facilities of Internet, email, telephone, photocopiers and so on to the community, when not used for multimodal dictionary project. In communities other than Lohorung, this model of sustainability could be useful in the case of developing countries like Nepal because cost of maintaining a website can be pretty high in such economies.

5.4. Quality versus availability

Since the digital contents of the multimodal dictionary are accessed over the Internet, their sizes should be optimal at the available bandwidth. Though Internet bandwidth has been increasing comparatively over the years, available bandwidth in rural villages and cities of Nepal are still limited. This can also be true if the system is to be used for another endangered language elsewhere in the world. To mitigate this, compressed formats of media files (video, audio and images) has to be used. We are currently using flv for video, mp3 for audio and jpg and png for images. We have used these formats because they are pervasive file formats in Windows, Mac or Linux machines. For further processing of data, quality is an important feature. In order to transcribe a video, picture and sound quality has to be acceptable. There is a tradeoff between quality of video and level of compression of the files when working online. Though, the quality of video has to be compromised at times, if proper guidelines are used while recording the media, low quality video and audio can still be useful for transcription. One should always be ready for low quality data, but thanks to the crowd-sourcing, low quality media can always be replaced by high quality contributions.

6. CONCLUSION

Despite of the challenges in the self documentation of Languages by community members can be less expensive and efficient way to document endangered languages. There might be several questions regarding quality of data to one's mind. But from our experience, we found that the issue of quality of data can be managed by proper training of the leaders of the community who will train other members and they will in turn train others. Training on use of electronic devices like digital cameras, sound recorders, computers, multimodal dictionary and guidelines on collecting data are adequate and necessary to collect acceptable data.

Till date, we have more than 900 entries of word in the online Lohorung multimodal dictionary. In order to increase participations, we have been organizing several training programs and meetings with the Lohorung contributors. We have a group of more than 15 regular contributors who have been active since the beginning of this project. Though availability of Internet, power and electronic devices might be a difficulty sometimes, contributors are enthusiastic to contribute and social prestige has been the encouraging factor for most of them. Hence, we conclude that multimodal dictionary is the useful and only tool for self documenting endangered languages.

6. REFERENCES

[1] S. Bird and G. Simons, "Seven Dimensions of Portability for Language Documentation and Description," *Language*, pp. 557-582, 2002

[2] E. Estellés-Arolas and F. González-Ladrón-de-Guevara, "Towards an integrated crowd-sourcing definition," *Journal of Information Science*, Sage Journals, Vol 38, pp. 189-200, 2012

[3] Central Bureau of Statistics National Planning, Commission Secretariat His Majesty's Government of Nepal (CBS) and UNFPA, "Population Census 2001: National Report", CBS and UNFPA, Kathmandu, 2002.

[4] B.A. Huberman, D. M. Romero and F. Wu, "Crowdsourcing, attention and productivity," *Journal of Information Science*, Sage Journals, Vol 35, pp. 758-765, 2009

[5] C. Abras, D. Maloney-Krichmar and J. Preece, "User-Centered Design," *In Bainbridge, W. Encyclopedia of Human-Computer Interaction*, Thousand Oaks: Sage Publications, 2004

[6] E. Gamma, R. Helm, R. Johnson and J. Vlissides, *Design Patterns: Elements of Reusable Object-Oriented Software*, Pearson, pp 4-6, 2000.

[7] http://www.unicef.org/infobycountry/nepal_nepal_statistics .html, *Nepal Statistics*, Unicef, accessed 2012.