MULTIMODAL CORPUS USING MULTIMODAL DICTIONARY IN LOHORUNG

Prof. Jens Allwood[1], Sagun Dhakhwa[2], Bhim Narayan Regmi[2], Prasanna Shrestha[2]

[1]University of Gothenburg, Sweden
[2]Centre for Communication and Development Studies, Nepal

jens.allwood@ling.gu.se, rite2sagun@gmail.com, bhimregmi@gmail.com, prasannashrestha.64@gmail.com

ABSTRACT

Lohorung is one of the minority Tibeto-Burman languages of Kirati group spoken in the Eastern part of Nepal. It lacks a written tradition. Crowd sourcing can be a good way to document a language in terms of resources, especially in the context of a linguistically rich country like Nepal. We have built a multimodal dictionary, browsing and authoring tool with which Lohorung community members, even if they have limited computer experience, can collect multimodal data, (see below) author the dictionary and browse the information in it. The collected information can be used for building a multimodal corpus for further analysis of Lohorung. This paper gives a brief overview of the multimodal dictionary application, its uses and its potentials.

Index Terms— multimodal corpus, multimodal dictionary, Lohorung

1. INTRODUCTION

The Lohorung multimodal dictionary is an online dictionary and data collection tool designed to support a language community in collecting, authoring, and browsing a multimodal corpus and an online encyclopedic dictionary through self-sustainable community information centers.

Multimodal corpora often include video, audio, and textual records. Some multimodal corpora contain records of interaction extracted from recordings of naturally occurring conversations. These may consist of a variety of different records of everyday communication from a range of different discursive environments; from face-to-face to digital media [1]. Thus, they may include a variety of different types of media. The multimodal dictionary allows users to add audio, video and textual description of an object, event or word, collected from various parts of the Lohorung community. The audio and video data can be collected using many types of recording devices like, audio recorders, mobile phones, and digital cameras, following a guideline which allows the data to be acceptable. The guideline mainly focuses on the quality of recording which can be related to noise, visibility, focus and distance of the speaker(s).

We intend to use crowd sourcing to collect data from the Lohorung community. If we are successful, this will speed up the work on data collection, however quality of data could be a big concern. A study of the quality of linguistic annotation, using crowd sourcing in Amazon's Mechanical Turk system, showed that the difference between using expert and non-expert annotators was not significant, when clear instructions were given on each task succinctly [3]. Later work done by Tilly et al., related to the use of crowd sourcing in linguistic studies showed that the quality of the result is also comparable to results obtained from labs [2]. This earlier work serves as a basis for our work, using crowd sourcing for data collection and annotation of Lohorung.

Though projects like the "Open Mind Initiative" has already attempted to make tasks of annotating word sense and commonsense word relations sufficiently "easy and fun" to entice users into freely labeling data starting from 1999, an entire language has not been documented (collection and annotation) using crow sourcing so far [4]. We intend to move in this direction, by following the concept of *Games with a Purpose* where the main motivator for the crowd participation is fun [5]. Data collection and entry tasks are designed to provide entertainment of producing multimedia content in Lohorung and to get social recognition as an additional bonus.

The Lohorung Multimodal Dictionary was built as a part of the project *Work on Lohorung and Nepali spoken language*, which is being funded by Swedish Research Council (VR), in collaboration with the development agency SIDA. The project has been jointly undertaken by the SCCIIL Interdisciplinary Center at the University of Gothenburg (Sweden), the Centre for Communication and Development Studies (Nepal) in collaboration with 'Lohorung Yakhaba Samaj (Nepal) a social organization of the Lororung community.

This project has two parts with slightly differing objectives based on the status of the involved two languages. The Nepali part of the project, is a continuation of an already begun multimodal project, which is a part of the Nepali National Corpus. Besides having about 240K of annotated corpus material, this part has multimodal recordings of different social activities of about 260K which are still not annotated. The whole corpus is planned to include about 500K corpus annotated with basic spoken language features like pause, silence, and overlap. The Lohorung part also aims at a corpus of similar size and nature with additional works on making a sketch grammar, a basic dictionary and a cultural profile. The additional work on Lohorung has been envisaged for documentation and vitalization since it is one of the endangered languages in Nepal.

The multimodal corpora of Nepali and Lohorung are supposed to be the base for the study of the communicative features of these languages which can be compared to spoken language features in Swedish.

Below, we will begin with a general overview of Lohorung. Then we will give a brief description of the dictionary application building process, the dictionary building process and the use of the data.

Figure 1: Lohorung Villagers

2. THE LOHORUNG LANGUAGE

The Lohorung language is a Tibeto-Burman language grouped under the Kirati languages spoken by the Lohorung people residing mainly in the Sankhuwasabha district (Eastern part of Nepal). Lohorung speakers are also distributed in city areas of other districts like Dharan, Dhankuta, Kathmandu Valley, and Darjeeling. The main occupation of the Lohorung people is agriculture while some are in government services, education, military service (Nepal, India and British army) and other professions. The Lohorung language is rich in oral literature. Their oral narrative called Mundum is the pillar of their folklore and folklife. There has been some preliminary study of the Lohorung language which includes a basic dictionary in Lohorung with a grammatical sketch [6].

According to the 2001 Population Census of Nepal, there are 1207 Lohorung speakers which is about 0.01% of the total population of Nepal [1] [7]. Most of the Lohorung speakers were found in the Ilam district. Being small, Lohorung is on the verge of extinction because of the dominance of Nepali and other languages. The aim of the project is to study, preserve and vitalize the Lohorung Language in Nepal.

The introduction of a writing system to the Lohorung community is another purpose of the project. The Lohorung community has decided to use an extended Devanagari based writing system for Lohorung. The current, Devanagari system cannot be directly used for Lohorung because there are many sounds that cannot be represented by the Devanagari or Nepali writing system. As extended Devanagari has not yet been developed for Lohorung, we are currently using basic Devanagari and Roman script for English in the multimodal dictionary.

3. THE LOHORUNG MULTIMODAL DICTIONARY

Figure 2: Navigation

The Lohorung Multimodal Dictionary is a web based application where different words in Lohorung along with their meaning, pronunciation, description, examples and related audio, picture and video information can be added by authors and retrieved. Each word also contains its Nepali & English meanings and pronunciations. The system supports collation of Unicode encoding which allows for data written in Devanagari or any other script. Hence, operations like sorting and searching can easily be performed.

Currently, the user interface supports three languages namely: Lohorung, Nepali and English but it can be adapted to other languages according to need by creating new language packs. After localization, the multimodal dictionary system could also be used by user of the new

[1] No Census data about Nepal is available after 2002.

language to learn Lohorung by using navigation in the new language or by just using pictorial navigation. Also the system can be used to document other languages than Lohorung.

Words of different parts of speech can be added and categorized in the system. Different categories and sub-categories can be created by authors and words can be tagged into multiple categories. In the figure above, we can sample categories of nouns namely, relation, ornament, medicine, crops, human body, and nature. The crop category has two sub-categories: cash and food crops. Similarly, they can still be sub-categorized. For example, an apple can fall in to the local fruit sub-category of the fruit category and culture category at the same time. The system can be broadly divided into a dictionary browser and an administrator panel whose description is given below:

3.1. Dictionary Browser

In the front end (user side), the UI design is simple to understand and easy to use. There is a navigation menu with categories with their respective sub-categories of Lohorung words. Figure 2 shows an example of categories and sub-categories of nouns; however, it can contain other types of words too. Figure 2 also shows how navigation has been developed so that people with limited literacy can navigate through the dictionary.

Figure 3: Multimodal Dictionary Browser

Users can easily find different words in specific categories. There is search option which allows users to find required words by typing a keyword. It also supports scripts or alphabets from initial to final symbol, to enable searching for the word start with a specific script. Figure 3 shows the detailed information of a word (Rice in this case) and its related pictures. The word can also contain video and audio data. The left pictures are the search results with similar tags.

3.2. Administration Panel

In the Admin (administration) panel, there are two types of users, main admin and normal user(s). The Admin panel can manage the words entries by, for instance, adding words in their respective category or editing/deleting words. Normal users can only add and view entries. The Admin panel has a search component to search for words from a large database. It also has a media component to manage the audio, video and pictures of what the words denote. The administration control is fully managed by the Lohorung community while the research team gives support when needed.

4. USE OF THE MULTIMODAL DICTIONARY

The Multimodal dictionary application was built with multiple purposes, which are as follows:

4.1. To enable Multimodal Corpus building through crowd sourcing

The primary purpose of the project is to build a multimodal corpus of the Lohorung Language, to make possible a multimodal analysis of the language. We intend to use crowd sourcing of audio and video data from the Lohorung community with the help of this web application. Crowd sourcing is particularly useful because it can help us to collect a large amount of data in a comparatively short amount of time. It can also provide greater variety and representativity of the collected data.

Using the system, any interested native speaker will be provided user ID and password and he/she can add contents online but the data will be made public only after it has been checked and approved by an administrator. A trusted administrator is a person from the community who is nominated by the community, and who has been trained for this purpose. Hence the administrator is authorized by the community to check the data in terms of use in the community and the technical quality of the data.

4.2. Teaching and Learning Lohorung

One of the main aims of the Lohorung project is to vitalize the Lohorung Language. Vitalization of a language implies making it more widely used. A Lohorung multimodal dictionary will be a good resource for both teaching and learning in Lohorung in the future.

4.3. Documentation of Indigenous knowledge

Much indigenous knowledge in an oral culture like Lohorung is dying with senior members. The newer generation is less interested in learning that knowledge. The dictionary will help the community to preserve the knowledge digitally.

5. THE WEB APPLICATION BUILDING PROCESS

The Lohorung Community members are the main users of the multimodal dictionary. Most of the community members live in the countryside and they are the most important resources of the language because Nepali is the dominating language in the urban areas. Recent development in telecommunication in Nepal makes access to the Internet and to Mobile phones easier in the countryside of Nepal and hence availability of technology is not as large a barrier anymore. However, experience with computers, Internet and data collection is still a challenge while working with rural communities.

We have focused on the ease of use through interactivity of the web application. We followed an agile methodology while developing this application in order to involve the community members directly in the application building process 0. Our main focus was collaboration between the research team and the Lohorung community for the development of the application.

5.1. Requirement Analysis

We wanted to use a User Centered Design (UCD) process in order to know what the users want and need from the beginning 0. The first step of conceptualizing the requirements of the multimodal dictionary was to collect specifications of the dictionary. This was finalized in a workshop program where more than 30 Lohorung Community members participated along with our team in April 2009.

During the workshop we introduced the concept of multimodal dictionary and asked the community members to form groups in order to discuss and suggest their requirements. One of their requirements was related to ownership of their indigenous knowledge. They were scared that non-Lohorung people would capitalize on their indigenous knowledge and technologies, since the multimodal dictionary will also contain multimodal information on their culture and indigenous knowledge. This led us to a requirement of administrative privileges in the dictionary for the Community leaders who can decide whether or not to allow sensitive data to be published for the public.

5.2. User Centered Interaction Design

Since, most of the community members living in the Lohorung villages have no prior experience with computers, designing interaction supported by a good interface was quite challenging. We involved community members in the design of the user interfaces and user interaction of the application right from the beginning. Incremental informal meetings with community representatives to show the recent development in the web application and to get quick feedback from them were part of the software building process.

5.3. Usability Studies & Acceptance Tests

One of the goals of Multimodal dictionary design was to make it usable for users with limited literacy and computer experience. In order to ensure this, we performed usability a study and refined the user interface using its results.

Figure 4: Usability Study with a Lohorung Speaker

We used hallway testing with Lohorung community members because it is an easy way to find most of the problems in the design with 5 test users [9]00. We invited 5 users, some skilled and some unskilled users of computer. They were from the age group of 30 to 67, but only one of them was female. One of the persons was semi-literate.

The testing was performed in the usability test setup at Centre for Communication and Development Studies, Office (see figure above). We had given a set of predefined tasks to each user and recorded their activities with their consent. We also observed their performance and took brief interviews with them about their experiences. The interviews basically consisted of open questions related to their experience, problems, and comments.

During the test, we found that the users were quite excited to use a multimodal dictionary in their own language. They were happy to see and listen to media in Lohorung. Users didn't find it difficult to navigate through the system. It was really interesting to see that a user, who is semi-literate and was using a computer for the first time, could easily start using the system after a minute of instruction. However, some users suggested adding some functionality like to add audio for word pronunciation in all three languages (Lohorung, Nepali, and English) and to add categories with sub categories to the words. They also requested the system to be modified to allow the addition of verbs, adjectives, and other parts of speech. Initially the dictionary allowed only nouns.

This feedback was really important in order to improve the system so that the system can be accepted by the community. After completing the improvements according their suggestions, the users from the Lohorung community were again invited for the final acceptance test. This time there were two users; one was more experienced with computers than the other one. The two persons agreed to be our trainers in the Lohorung village. We allowed them to use the system and observe if there were any problems. During the observation, they found no problem in using the system and were already starting to add data to the system. The system was accepted by them. After this acceptance test, we gave them a brief training on how to collect data and on how to enter the data in the multimodal dictionary admin panel. We based our training on the data collection guidelines discussed in next section.

6. DICTIONARY BUILDING PROCESS

The dictionary building process is quite different from the software development process. In the dictionary building phase we focused primarily on data collection and entering the collected data in the online multimodal dictionary administration panel.

6.1. Making a plan and guidelines for data collection

The initial phase of the data collection was the planning of the data collection process. Field work was a real test and challenge for the collaboration between the research team and the community. We prepared a plan for data collection, where initially our research team will train the community trainers. These community trainers were selected from the community members who had been actively participating in the software development process.

The sources of data are the activities of Lohorung people. Thus, a questionnaire has been developed to make a survey of daily, weekly, monthly, yearly and any other periodically performed activities related to livelihood from recreational to spiritual needs. The questionnaire has been taken as freely modifiable basic guidelines which provide information on the activities to be recorded.

The activities known through the questionnaire will be listed and an appropriate schedule to record them will be prepared. These schedules include the date, time, place, and people and objects of recording.

The plan also includes the modes of recording, i.e., visual, audio, and audio-visual. We have been collecting and will collect data of archive quality. This, among other things, means that it should have a life length of a certain period and be exportable to a variety of formats. Spoken corpora will be built by transcribing the audio/visual data collected with a target of at least 500 k words in running text. Most of the recordings will be made in natural settings or minimally controlled settings. This is done to record authentic multimodal communication in the Lohorung Community. This also makes the data collection easier and more natural for native Lohorung speakers.

6.2. Training the trainers

We organized a week long training program to two community trainers of which one member is from the city and other member resides in the Lohorung village of Shankhuwasabha. The training basically consisted of introduction to media collection guidelines, basic operation of audio recorder and video camera, and operation of administration panel of multi-modal dictionary. They were also trained in administering the questionnaires, and received explanations of the objectives, methodologies and envisaged outcome and outputs of the project.

6.3. Training and awareness to the community members

After the training, we sent the trainers to do fieldwork in the Lohorung village. The trainers are currently training the Lohorung community members by orienting them about the project, and running awareness programs in the village. The trainers also collect data from the field and upload the collected data in the online multi-modal dictionary from their Internet station at Khadbari, Sankhuwasabha.

6.4. Data Collection

We have focused on user experience and we plan to use a social game model to investigate user's involvement in order to improve the crowd participation in line with Kazai et al's findings about relation between the affordances of the system, the incentives of the social game, and the behavior of the assessors [8]. Though, in our case it is not really a game but social fame that will be an incentive to the contributors. Data Collection is a continuous process and we intend to involve as many community members as possible in the data collection and in contributing to the multimodal dictionary system. Currently, only a handful number of people are actively participating in the process and we hope to see this handful of people grow in to a crowd.

7. CHALLENGES

We have faced many challenges during the web application development process of which requirement collection and interaction design, so far, were the most important. They were important because, we need to make the community members feel the ownership of the multimodal dictionary. Apart from these technical challenges, sustaining the crowd involvement for collecting our 1 million word spoken corpus is a managerial challenge.

Loss of interest in the system can be one challenge while conflict between different contributors could be another.

Such conflicts can arise because of different claims concerning the meaning or pronunciation of the words. Currently, there is no big conflict as only few people are contributing to the dictionary. We have introduced identification of dialect and village name in order to partly manage such conflicts, technically. But it will be interesting to see how the crowd will react if a bigger conflict develops.

8. USE OF COLLECTED DATA AS CORPUS

There will be photographs, audio, audio-visual and written texts collected through the media center, using crowd sourcing methodology. The individual words, phrases and some larger units of expression like idioms and proverbs will be entered into the online multimodal encyclopedic dictionary. The entries can be followed by a description in either of or all the modalities, i.e., writing, audio, visual or audio-visual. Among the collected audio or audio-visual recordings, multimodal interactive spoken language data will be used in creating the corpus. This is in the line of objectives of the project that aims to study spoken language features with special focus on multimodal communication.

The data will be transcribed using a Devanagari based phonemic scheme as well as a simple Roman based phonemic scheme so that it could be understood both by native Lohorung, and other people who have access to Devanagari or Roman. The transcribed texts will be annotated at least for three basic spoken language features - pause, silence and overlap. The metadata also will be maintained. The corpus will be analyzed for grammar, lexicon and communicative features. The analysis of communicative features will be concentrated on interpersonal communication specifically feedback and own communication management which will be compared to Nepali and Swedish to investigate the similarities and differences between geographically close but genealogically different (Nepali and Lohorung), genealogically close but geographically different (Nepali and Swedish), and both genealogically and geographically different (Swedish and Lohorung) languages.

9. REFERENCES

[1] Adolphs, S., Knigh, D., “Building a spoken corpus: What are the basics?”, In: In O’Keeffe, A. , McCarthy, M., ed. The Routledge Handbook of Corpus Linguistics, Routledge , 2010

[2] Tilly, H. et al, Crowdsourcing and language studies: the new generation of linguistic data, Proceedings NAACL Workshop on Creating Speech and Language Data With Amazon's Mechanical Turk, 2010

[3] Snow, R., Brendan O., Daniel J., and Andrew T. Ng. 2008. Cheap and fast—but is it good?: evaluating non-expert annotations for natural language tasks, Proceedings of the Conference on Empirical Methods in Natural Language Processing, pages 254–263., 2008

[4] Timothy, C., and Rada M., Building a sense tagged corpus with Open Mind Word Expert. In Proc. of the Workshop on ”Word Sense Disambiguation: Recent Successes and Future Directions”, 2002.

[5] Von, A. L., Dabbish, L., Designing games with a purpose. Commun, ACM 51(8):58–67, 2008

[6] Yadava, Y. P., Rai N. K., Regmi B. N., Rai T. and Rai T., "Lohorung-Nepali-English: A Basic Dictionary" A report submitted to The National Foundation for Development of Indigenous Nationalities, Kathmandu, 2004.

[7] 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.

[8] Kazai, G., Milic-Frayling, N., Costello, J., “Towards methods for the collective gathering and quality control of relevance assessments.” In: Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval (SIGIR 2009), ACM, pp 452–459, 2009.

[9] Borysowich, C., http://it.toolbox.com/blogs/enterprise-solutions/sample-usability-test-plan-17826, 2011.

[10] Amber, S., “Examining the Agile Manifesto“ http://www.ambysoft.com/essays/agile Manifesto.html, 2011

[11] Nielsen, J., Usability Engineering, Academic Press Inc, ISBN: 0125184069, p 165, 1994

[12] Nielsen, J., “Why You Only Need to Test with 5 Users”, http://www.useit.com/alertbox/20000319.html, 2000

[13] Travis, D., An introduction to UCD principles through narrative, User Focus, p 11-14, 2009