Endangered Language Preservation: A Model for Automatic Speech Recognition Based on Khroskyabs Data

Ruiyao Li, Yunfan Lai
Trinity College Dublin, Trinity College Dublin
College Green, Dublin 2, Ireland, College Green, Dublin 2, Ireland
lir4@tcd.ie, yunfan.lai@tcd.ie

Abstract

This is a report on an Automatic Speech Recognition (ASR) experiment conducted using our Khroskyabs data. With the impact of information technology development and globalization challenges on linguistic diversity, this study focuses on the preservation crisis of the endangered Khroskyabs language, a language falling under the Gyalrongic language group (Glottocode: guan1266). We used Automatic Speech Recognition technology and the Wav2Vec2 model to transcribe the Khroskyabs language. Despite challenges such as data scarcity and the language's complex morphology, preliminary results show promising character accuracy from the model. Additionally, the linguist also has given relatively high evaluations to the transcription results of our model. Therefore, the experimental and evaluation results demonstrate the high practicality of our model. At the same time, the results also reveal issues with high word error rates, so we plan to augment our existing dataset with additional Khroskyabs data in our further studies. This study provides insights and methodologies for using Automatic Speech Recognition to transcribe and protect Khroskyabs, and we hope that this can contribute to the preservation efforts of other endangered languages.

Keywords: ASR, Gyalrong, Khroskyabs

1. Introduction

According to Moseley (2010), in recent decades, alongside the development of information technology, there has been a gradual reduction in the diversity of human languages. Particularly with the challenges of globalization, the preservation of many Asian languages, such as the Khroskyabs language, is facing a crisis. Therefore, we hope to apply automatic speech recognition tools to transcribe some traditional stories in Khroskyabs into IPA, thereby protecting the language and culture by preserving these traditional stories in Khroskyabs.

1.1. Endangered Language Preservation

Khroskyabs is a language in Gyalrongic language group spoken in western Sichuan, China. Currently, there are about 9000 native speakers of Khroskyabs. The transmission of Khroskyabs relies entirely on speech, as it lacks a writing system. It is Gong (2017) indicates that the Gyalrongic language group is classified as endangered, gradually heading towards extinction under the pressure of Sichuanese Mandarin and the Amdo Tibetan. Our fieldwork on the Khroskyabs language also observed that the local people, due to pursuing education and work opportunities outside, have become less proficient in speaking the Khroskyabs language compared to earlier generations. Additionally, there are no specialized schools teaching the Khroskyabs language. Furthermore, the lack of a written system for Khroskyabs exacerbates its preservation challenges.

The preservation of the Khroskyabs language is important. Due to its long-standing use in secluded mountainous regions, minimally affected by external linguistic influences, Khroskyabs, just like many other Gyalrongic languages, has retained a substantial amount of ancient Sino-Tibetan features (Gong, 2017). It holds significant importance in Sino-Tibetan historical linguistics, as it preserves the complex consonant clusters and verb morphology in proto-Sino-Tibetan. Additionally, the Khroskyabs language is a highly morphologically rich language, characterized by numerous verb affixes and root alternations. These features of it are beneficial for the study of Sino-Tibetan historical linguistics, underscoring the urgent need for attention to its endangered status. Beyond its scholarly value, preserving this language also supports the cultural identity and heritage of its speakers, promoting inclusion and underscoring the importance of linguistic diversity. These considerations drive our pursuit of new preservation methods, including the application of automatic speech recognition tools, to protect the Khroskyabs language for the benefit of both academia and its native speaker communities.

1.2. Method

This section outlines the methodology employed in our study, focusing on the selection of the

2nd *EURALI Workshop*, pages 36–40
25 May, 2024. © 2024 ELRA Language Resource Association: CC BY-NC 4.0

Khroskyabs language as our subject and the implementation of the Wav2Vec2 model for automatic speech recognition.

1.2.1. The Source of the Data

The language we have chosen is Khroskyabs, which belongs to the western branch of the Gyalrongic language group (Sun, 2000a,b; Huang, 2001; Lai, 2017). Khroskyabs is among the less spoken languages within this group.

Protecting endangered languages faces a significant challenge: transcription. Linguists may spend up to half an hour transcribing just one minute of audio. Therefore, using automatic speech recognition can expedite and streamline the transcription process for endangered languages, enabling us to efficiently document and preserve them. However, in the process of automatic speech recognition, a large amount of input data is required to train the model. Compared to the data for many endangered languages (Guillaume et al., 2022), the dataset for Khroskyabs is notably larger (Lai, Yunfan, unpublished). These data include recordings of local elders telling traditional stories in Khroskyabs and transcriptions by the linguist, ensuring transcription accuracy. Because the model cannot recognize the punctuation, we removed all punctuation marks. In this experimental training of our model, we only used one hour of Khroskyabs data to assess the model's utility when faced with languages lacking ample annotated data. The previous data format was .txt, but in our training, we required the data format to be .eaf, which necessitated re-splitting the audio and inputting transcriptions in ELAN. Therefore, moving forward, we plan to augment the amount of Khroskyabs data to enhance the model's accuracy after putting more date into ELAN. Afterwards, the dataset will be uploaded to Pangloss to make it publicly available.

1.2.2. The Model Selection

Currently, there are several automatic speech recognition tools available, and for low-resource languages, there are some data augmentation techniques that can help improve ASR systems (Bartelds et al., 2023). For our project, we have selected the XLS-R-Wav2Vec2 model fine-tuned for low-resource languages (O'Neill et al., 2023). This model has shown promising results in the context of Newar and Dzardzongke languages spoken in Nepal (O'Neill et al., 2023).

The Wav2Vec2 model employs multitask learning to optimize both its audio feature extractor and language model components, thereby enhancing its performance on low-resource languages. Importantly, the model supports transfer learning, allowing knowledge transfer from a related high-resource language model to improve the training process and performance of the low-resource language model.

In this study, we will demonstrate the development of an automatic speech recognition model for Khroskyabs using the model. For fine-tuning the model, several hyperparameters were configured to optimize the training process. The training used a per-device train batch size of 8, combined with gradient accumulation steps set to 2. The model was set to train for a total of 50 epochs. Additionally, a learning rate of 3e-4 was chosen.

In Section 2, we will present the experimental results concerning Khroskyabs transcription. In Section 3.1, we will discuss the challenges faced and potential future improvements. Lastly, we will have a conclusion in Section 4.

2. Evaluation of the Results Using Khroskyabs

In this section, we will showcase the model trained using Khroskyabs data as the foundation and discuss the outcomes of our training.

2.1. Experimental Results

In our experiment, we used one hour of Khroskyabs data. The Khroskyabs dataset comprises six audio recordings, each featuring different speakers, thereby adding a challenge to the model training process.

The quality of our automatic speech recognition system is evaluated using two metrics: character error rate (CER) and word error rate (WER). Both metrics quantify the disparity between the recognized text and the original text, with character error rate focusing on character-level errors and word error rate on word-level errors. These are two classic metrics used to evaluate automatic speech recognition systems.

The Figure 1 illustrates the average word error rate at each step of the training process across iterations, ranging from 100 to 1400, when training with one hour of Khroskyabs data.

From here, it can be observed that after 100 to 600 iterations of training, the results were far from satisfactory, with the word error rate approaching nearly one hundred percent. However, after further training, particularly at 1200 iterations, the word error rate decreased to eighty-seven percent.

Although the results above may not be entirely satisfactory, we can also observe the median character and word error rates for each checkpoint, as depicted in Figure 2.

From this table, it can be observed that at the first checkpoint, the median character error rate

Step	Training Loss	Validation Loss	Wer
100	6.402600	3.848435	1.000000
200	3.320600	3.311335	1.000000
300	3.281700	3.279070	1.000000
400	3.212000	3.196068	1.000000
500	2.755600	2.177797	1.010610
600	1.357800	1.550854	1.031830
700	0.870800	1.424223	0.893899
800	0.617200	1.517462	0.920424
900	0.508900	1.593333	0.888594
1000	0.400400	1.622587	0.875332
1100	0.296900	1.762700	0.877984
1200	0.282100	1.816226	0.875332
1300	0.225100	1.911786	0.899204
1400	0.212300	1.914231	0.899204

Figure 1: Word error rate across iterations

```
01/400 Median CER:      1.0
01/400 Median WER:      1.0

01/800 Median CER:      0.213
01/800 Median WER:      0.75

01/1200 Median CER:     0.192
01/1200 Median WER:     0.667
```

Figure 2: The median character and word error rates at each checkpoint

is 1.0, and the median word error rate is also 1.0. However, by the third checkpoint, the median character error rate further decreases to 0.192, while the median WER decreases to 0.667. These results indicate that as training progresses, both the character-level and word-level error rates of the model gradually decrease. This outcome is more satisfactory, particularly considering that the model is trained on only one hour of language data, with a character error rate of 0.19 already being low.

In evaluating the performance of our model, we also have box plots to visually analyze the character error rate and word error rate of the Wav2Vec2 model, as shown in Figure 4.

We can see from the box plots that the character error rate results demonstrated satisfying performance, with a median close to 0 and a very compact interquartile range. It suggests that the majority of characters were accurately recognized. Although there were a few minor outliers, they had minimal impact on the overall performance. In contrast, the word error rate median was relatively higher, indicating that recognition errors at the word level were more common and dispersed, and the distribution of word error rate included a

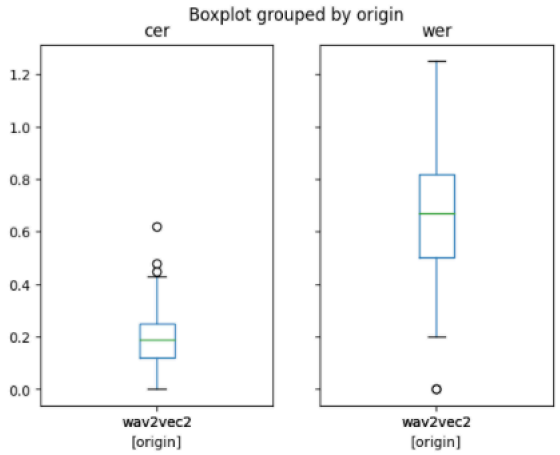

Figure 3: Boxplots of character and word error rates

significant outlier. Overall, these findings suggest that our model exhibits greater stability and accuracy in character recognition.

2.2. Evaluation by the Linguist

In addition to using classical character and word error rates as our model evaluation metrics, we also sought the opinion of the linguist regarding the transcription quality. In Figure 2.2, we compare the transcription results produced by our model with those transcribed by the linguist.

Figure 2.2 displays the transcription results for three randomly selected sentences. The top row shows the transcriptions provided by the linguist for the recordings, while the bottom row presents the transcriptions generated by our trained automatic speech recognition model. Portions highlighted in red indicate errors in our model's transcription, while parts within parentheses denote omissions in our model's transcription. It can be observed the model demonstrates good accuracy in character recognition, with occasional errors in discerning vowels, distinguishing between voiced and voiceless consonants, and occasionally omitting some consonants and tones.

From the results shown in this figure, the number of corrections required to achieve transcription quality appears to be lower than the quantity indicated by the character error rate we obtained earlier. Discrepancies between assessments of classic evaluation methods and assessments by linguists are also mentioned in Guillaume et al. (2022). The linguist, who is also the annotator of the training data and specializes in Khroskyabs, has also provided a positive evaluation of the model's accuracy. This suggests that the practical utility of the model we trained may be higher than what is evaluated by character and word er-

1. ɐ̂ɛə ŋəɲɟə̂jə tædpɑ̂fstænpɑtəɣə ætə̂jə jəʁûpʰræsce zjɐ̂ŋ vî nærôdpɑ rəŋó təɣî
ɐ̂ɛə ŋəɲɟə̂jə tædpɑ̂fstænbɑtəɣə ætə̂(jə) jóɣpʰráéscеə zjɐ̂ŋ vî nærû̂dpɑ rəŋóə
2. ɲəmɑ̂ nɑ̂riɛəvæ mdɐ̂ʁzə gərgôɣ ndzɐ̂jtə lodzê pʰádtəmɲədtəɣu vîskə
ɲəmɑ̂ nɑ̂riɛəvæ mdɐ̂ʁzəgərgô (ɣ)o(n)dzɐ̂jtə lodzê pʰátəɣmɲəddtə(ɣu) vîskə
3. jóŋ ŋəɲɟə̂jə jóŋ χtsʰətsʰə̂ jóŋ mêr nærjé cə̂ tʰjɐ̂ ŋókɑ næsɑ̂ɲ næpʰrǽɲrætə χtsʰətsʰə̂ ɛəntɛʰê nɑví rəŋó
joáɣ ŋəɲɟə̂jə jáŋ χtsətsʰə̂ntə jóŋmé(r)nærjé cə tʰjɐ̂ ŋókɑrætəɛə næsɑ̂(ɲ)næpʰrǽɲnrætə̂ χtsʰətsʰə̂ ɛəntɛʰê nɑví rəŋó

Figure 4: Comparison between the transcription result of the linguist and the model

ror rates.

3. Reflections And Further Studies

Now, we can observe that the model we trained has demonstrated a satisfactory level of accuracy in transcribing Khroskyabs. In this section, we will critically reflect on our approach and propose some possible further studies.

3.1. Reflections on the Model

Although our model has demonstrated a low character error rate, our results also reveal a higher word error rate, which is likely associated with the complex morphology of Khroskyabs. This indicates that our model currently lacks the capability to accurately capture word boundaries and has not fully adapted to the unique phonological and morphological characteristics of Khroskyabs. It shows the complexities of transcribing low-resource languages, where limited data availability and linguistic diversity pose significant challenges.

Furthermore, the model we developed struggles with accurately transcribing Chinese loanwords. In our data, Khroskyabs is transcribed using the International Phonetic Alphabet, while Chinese loanwords are transcribed using the Pinyin system. This has led to a higher error rate in processing Chinese loanwords. Additionally, the limited occurrence of Chinese loanwords in speech exacerbates the model's challenges in handling them.

3.2. Further Studies

To address the issue of a high word error rate, we plan to augment our existing dataset with additional Khroskyabs data. During this round of training, we used one hour of Khroskyabs data, and we aim to double this amount by incorporating an additional hour of data. This expansion is expected to enrich our dataset, providing a broader linguistic base that could enhance the model's understanding of the complex morphology.

To address the challenge of low transcription accuracy for Chinese loanwords, we plan to revise the original data, retranscribing all the Chinese loans and replacing Pinyin with IPA. Also, we plan to increase the presence of Chinese loanwords in our training dataset, which could potentially improve the model's proficiency in accurately processing these loanwords.

4. Conclusion

In the experiment, we demonstrated the transcription of endangered languages such as Khroskyabs using automatic speech recognition technology and the Wav2Vec2 model. Our results, after many training iterations, showed a median word error rate of 0.67 and a character error rate of 0.19. These results indicate an optimistic outcome in character accuracy and have been highly rated by linguists. However, we still face challenges, notably the high word error rate, likely due to the model's insufficient morphological understanding of the language. In the future, we plan to incorporate more data to enhance the model's transcription accuracy.

5. Bibliographical References

Martijn Bartelds, Nay San, Bradley McDonnell, Dan Jurafsky, and Martijn Wieling. 2023. Making more of little data: Improving low-resource automatic speech recognition using data augmentation. *arXiv preprint arXiv:2305.10951*.

Xun Gong. 2017. The Morphology of the Gyalrongic language group and Old Chinese. *Ancient Scripts and Historical Phonology of Chinese: Fudan Journal of Chinese Civilization Studies*, pages 134–156.

Séverine Guillaume, Guillaume Wisniewski, Cécile Macaire, Guillaume Jacques, Alexis Michaud, Benjamin Galliot, Maximin Coavoux,

Solange Rossato, Minh-Châu Nguyên, and Maxime Fily. 2022. Fine-tuning pre-trained models for Automatic Speech Recognition, experiments on a fieldwork corpus of Japhug (Trans-Himalayan family). In *Proceedings of the Fifth Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pages 170–178.

Bufan Huang. 2001. Research on the Language Belonging of Guanyinqiao. *Language and Linguistics*, 2(1):69–92.

Yunfan Lai. 2017. *Grammaire du Khroskyabs de Wobzi*. Ph.D. thesis, Université Sorbonne Paris Cité.

Christopher Moseley. 2010. *Atlas of the World's Languages in Danger*. Unesco.

Alexander O'Neill, Marieke Meelen, Rolando Coto-Solano, Sonam Phuntsog, and Charles Ramble. 2023. Language Preservation through ASR.

Jackson T-S Sun. 2000a. Parallelisms in the verb morphology of Sidaba rGyalrong and Lavrung in rGyalrongic. *Language and linguistics*, 1(1):161–190.

Jackson T-S Sun. 2000b. Stem alternations in Puxi verb inflection: Toward validating the rGyalrongic subgroup in Qiangic. *Language and linguistics*, 1(2):211–232.

6. Language Resource References

Lai, Yunfan. unpublished. *Siyuewu Khroskyabs texts*. Unpublished field notes.