The Relevance of the Source Language in Transfer Learning for ASR

Nils Hjortnæs
Indiana University
nhjortn@iu.edu

Niko Partanen
University of Helsinki
Helsinki, Finland
niko.partanen@helsinki.fi

Michael Rießler
University of Eastern Finland
Joensuu, Finland
michael.riessler@uef.fi

Francis M. Tyers
Department of Linguistics
Indiana University
Bloomington, IN
ftyers@iu.edu

Abstract

This study presents new experiments on Zyrian Komi speech recognition. We use DeepSpeech to train ASR models from a language documentation corpus that contains both contemporary and archival recordings. Earlier studies have shown that transfer learning from English and using a domain matching Komi language model both improve the CER and WER. In this study we experiment with transfer learning from a more relevant source language, Russian, and including Russian text in the language model construction. The motivation for this is that Russian and Komi are contemporary contact languages, and Russian is regularly present in the corpus. We found that despite the close contact of Russian and Komi, the size of the English speech corpus yielded greater performance when used as the source language. Additionally, we can report that already an update in DeepSpeech version improved the CER by 3.9% against the earlier studies, which is an important step in the development of Komi ASR.

1 Introduction

This study describes a Automatic Speech Recognition (ASR) experiment on Zyrian Komi, an endangered, low-resource Uralic language spoken in Russia. Komi has approximately 160,000 speakers and the writing system is well established using the Cyrillic script. Although Zyrian Komi is endangered, it is used widely in various media, and also in education system in the Komi Republic.

This paper continues our experiments on Zyrian Komi ASR using DeepSpeech, which started in Hjortnaes et al. (2020b). The scores reported there were very low, but in a later study we found that a language model built on more data increased the performance dramatically (Hjortnaes et al., 2020a). We are not yet at a level that would be immediately useful for our goals, but we continue to explore different ways to improve our result. This study uses the same dataset, but attempts to take the multilingual processes found in the corpus into account better.

ASR has progressed greatly for high resource languages and several advances have been made to extend that progress to low resource languages. There are still, however, numerous challenges in situations where the available data is limited. In many situations with the larger languages the training data for speech recognition is collected expressly for the purpose of ASR. These approaches, such as Common Voice platform, can be extended also the endangered languages (see i.e. Berkson et al., 2019), so there is no clear cut boundary between resources available for different languages. While having dedicated, purpose-built data is good for the performance of ASR, it also leaves a large quantity of more challenging but usable data untapped. At the same time these materials not explicitly produced for this purpose may be more realistic for the resources we intend to use the ASR for in the later stages.

The data collected in language documentation work customarily originates from recorded and transcribed conversations and/or elicitations in the target language. While this data does not have the desirable features of a custom speech recognition dataset such as a large variety of speakers and accents, and includes much fewer recorded hours, for many endangered languages the language documentation corpora are the only available source.

However, attention should also be paid to the differences in endangered language contexts globally. There is an enormous variation in how much one language is previously documented, and whether earlier resources exist. This also connects to the new materials collected, as some languages

Proceedings of the 4th Workshop on the Use of Computational Methods in the Study of Endangered Languages: Vol. 1 Papers, pages 63–69,
Online, March 2–3, 2021.

need a full linguistic description, and some already have established orthographies and variety of descriptions available. For example, in the case of Komi our transcription choice is the existing orthography which is also used in other resources (Gerstenberger et al., 2016, 32). Our spoken language corpus is connected with the entire NLP ecosystem of the Komi language, which includes Finite State Transducer (Rueter, 2000), well developed dictionaries both online[1] and in print (Rueter et al., 2020; Beznosikova et al., 2012; Alnajjar et al., 2019), several treebanks (Partanen et al., 2018) and also written language corpora (Fedina, 2019)[2]. We use this technical stack to annotate our corpus directly into the ELAN files (Gerstenberger et al., 2017), but also to create versions where identifiable information has been restricted (Partanen et al., 2020). Thereby our goal is not to describe the language from the scratch, but to create a spoken language corpus that is not separate from the rest of the work and infrastructure done on this language. From this point of view we need an ASR system that produces the contemporary orthography, and not purely the phonetic or phonemic levels. It can be expected that entirely undocumented languages and languages with a long tradition of documentation need very different ASR approaches, although still being under the umbrella of endangered language documentation.

In this work we expand on the use of transfer learning to improve the quality of a speech recognition system for dialectal Zyrian Komi. Our data consists of about 35 hours of transcribed speech data that will be available as an independent dataset in the Language Bank of Finland (Blokland et al., forthcoming). While this collection is under preparation, the original raw multimedia is available by request in The Language Archive in Nijmegen (Blokland et al., 2021). This is a relatively large dataset for a low resource language, but is still nowhere near high resource datasets such as Librispeech (Panayotov et al., 2015), which has about 1000 hours of English.

One of the largest challenges in our dataset is that there is significant code switching between Komi and Russian. This is a feature shared with other similar corpora (compare i.e. Shi et al., 2021). All speakers are bi- or multilingual, and use several languages regularly, so there are large segments where the language is in Russian, although the main language is Komi. There are also very fragmentary occurrences of Tundra Nenets, Kildin Saami and Northern Mansi languages, but these are so rare at the moment that we have not addressed them separately. In addition, none of the data is annotated for which language is being spoken, and we only have transcriptions in the contemporary Cyrillic orthographies of these languages, as explained above. We propose two possible methods to accommodate these properties of the data. First, we compare whether it is better to transfer from a high resource language, English, or the contact language, Russian. Second, we analyze the impact of constructing languages models from different combinations of Komi and Russian sources. The goal is to make the language model more representative of the data and thereby improve performance.

2 Prior Work

A majority of the work on Speech Recognition focuses on improving the performance of models for high resource languages. Very small improvements may be made through advances such as improving the network and the information available to it, as in Han et al. (2020) or Li et al. (2019), though as performance increases the gains of these new methods decrease. Another avenue is to try to make these systems more robust to noise through data augmentation (Braun et al., 2017; Park et al., 2019). As with improving the networks, however, these improvements become more and more marginal as performance increases.

As more models for ASR become available as open source (Hannun et al., 2014; Pratap et al., 2019), it becomes easier to develop these tools for low resource languages and to create best practice standards for doing so. This is the fundamental goal of Common Voice (Ardila et al., 2020). Others also work on individual languages, such as Fantaye et al. (2020) and Dalmia et al. (2018).

In the language documentation context we have seen a large number of experiments on endangered languages in the last few years, but often focusing on the datasets with a single speaker. Under this constraint a few hours of transcribed data has already shown to result in a relatively good accuracy, as shown by Adams et al. (2018). Also Partanen et al. (2020) report very good results on the extinct Samoyedic language Kamas, where

[1] https://dict.fu-lab.ru
[2] http://komicorpora.ru

the model was also trained with one speaker, for whom, however, a relatively large corpus exists. Under many circumstances it is realistic and important to record individual speakers in numerous recording sessions, and such collections appear to be numerous in the archives containing past field recordings, so there is no doubt that also single speaker systems can be useful, although not ideal.

Recently, Shi et al. (2021) also report very encouraging results on Yoloxóchitl Mixtec and Puebla Nahuatl, especially as the corpus contains multiple speakers. Our corpus is of a comparable size as is used in their experiments (Shi et al., 2021). Compared to their results our Komi accuracy, including the latest ones reported in this paper, are tens of percentages worse that what could be expected from the size of our corpus. This calls for wider experimentation at our dataset using different systems, which, we hope, will reveal more about how particularities of individual corpora impact the result.

Zahrer et al. (2020) also describe a language documentation project design where ASR tools are being integrated into actual workflows during the project. The end goal of our work is in line with this: we want Komi ASR to reach a level where it is useful for work on this language, and we see this happening through gradual steps where the system used is improved through different experiments.

What it comes to the usability and accessibility of ASR systems, Adams et al. (2020) describe their work on a user friendly interface for the language workers to train and use ASR tools. Cox (2019) has created an ELAN plugin, and the same approach was recently extended for DeepSpeech by Partanen (2021).

3 Methodology

3.1 Data

The Russian speech corpus we use is from Mozilla's Common Voice[3] project and contains about 105 hours of speech data (Ardila et al., 2020). The Komi data consists of about 35 hours of dialectal speech, and is described in Hjortnaes et al. (2020b).

3.2 Data Preprocessing

To prepare both our Komi and Russian data, we split it into 8/1/1 training, dev, and testing sets and cleaned any sections which were too long or too short as defined by DeepSpeech. The alphabet is based on the Komi data and not the text used to construct the language models as it is what determines the output of the network. We obtained our English model from DeepSpeech's publicly available release models[4].

3.3 DeepSpeech

We trained our models using Mozilla's open source DeepSpeech[5] (Hannun et al., 2014) version 0.8.2. We used this version because it was the latest release version at the time of these experiments. DeepSpeech is an end-to-end bidirectional LSTM neural network specifically designed for speech recognition. It consists of 5 hidden layers followed by a softmax layer where the 4th layer is the LSTM layer. The other hidden layers all use the ReLU activation function. The whole structure can be seen in figure 1, which shows an older version of the architecture with a unidirectional LSTM. For this experiment we used a dropout of 10% and a learning rate of 0.0001 with batch sizes of 128 for training, testing, and development sets. DeepSpeech automatically detects plateaus and reduces the learning rate by a factor of 10 when no further improvement is being made on the dev set.

We trained a Russian model using DeepSpeech from the standard random initialization using the hyperparameters defined above. DeepSpeech saves the best performing model, which we then use for transfer learning later.

3.4 Language Models

DeepSpeech outputs its best guess as to the transcription, but that is based entirely on the contents of the audio and does not account for spelling or punctuation. In order to address this, the output is put through a function which attempts to maximize the weighted probability of the model output and a probabilistic language model with two tuneable parameters. The first, α, determines how much the language model is allowed to edit the network output. The second, β, controls inserting spaces (Hannun et al., 2014). Our language models were constructed using Kenlm (Heafield, 2011)

[3] https://commonvoice.mozilla.org/

[4] https://github.com/mozilla/DeepSpeech/releases

[5] https://github.com/mozilla/DeepSpeech/

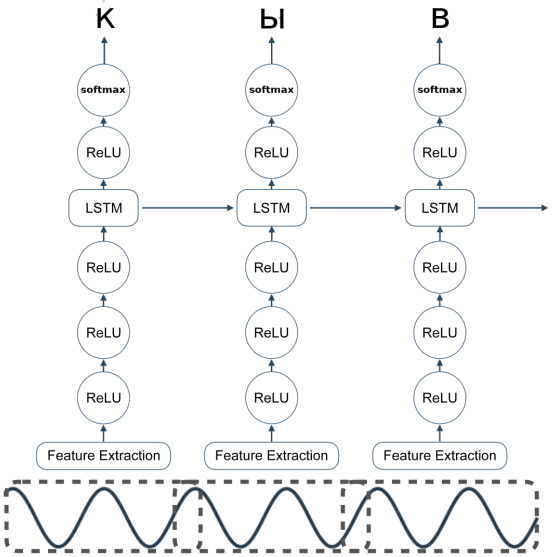

Figure 1: Mozilla's DeepSpeech architecture (Meyer, 2019)

on the 500000 most common words in the relevant text corpus.

We constructed three language models using various quantities of available Komi and Russian data. The first is exclusively Komi and is constructed using the *Komi-Zyrian corpora*[6] which consists of a main corpus of 1.39 million words in the literary domain and a 1.37 million word corpus from social media (see Arkhangelskiy, 2019). This serves as our baseline language model. The second includes all available text data from both the Komi corpora and the Russian Wikipedia dump from September 1st, 2020, which contains over 786 million tokens. We did not expect this largely Russian model to perform especially well, but include it anyway as an additional point of comparison. The last language model was constructed by cutting the Russian Wikipedia dump down to the same number of tokens as the combined Komi corpora such that the model is based on an equal amount of Komi and Russian data.

3.5 Transfer Learning

We trained our models using the transfer learning feature built into DeepSpeech which is based on Meyer (2019). This starts by training the model on a high resource language, re-initializing the last n layers, and switching to the target language. Both Meyer and Hjortnaes et al. (2020b) found that re-initializing the last 2 layers, the softmax and ReLU layer after the LSTM layer, yielded the best performance. The softmax function outputs a letter of the alphabet for each time stamp and re-initializing is necessary to accommodate the target language having a different alphabet than the source language.

To train the Komi model, we used transfer learning from English to Komi using each of the three language models defined above and used transfer learning from the Russian model we trained to Komi again using each of the three language models. We obtained the English model from DeepSpeech's released models for 0.8.2.

[6] http://komi-zyrian.web-corpora.net

Corpus	Size
Komi Speech	35 hours
Russian Speech	105 hours
Komi Literary	1.39M tokens
Komi Social	1.37M tokens
Komi Text Combined	2.76M tokens
Russian Wiki	786M tokens

Table 1: Token counts for the speech corpora and text corpora used to create the language models.

4 Results

The best Word Error Rate and best Character Error Rate were achieved by using English as the source language for the transfer and the language model constructed from equal parts Komi and Russian. This is, however, only a minor improvement as compared to using the language model constructed from Komi alone. When using Russian as the source language, performance drops regardless of the language model used. The worst performance was achieved when using Russian as the source language and the language model leveraging all available text.

5 Discussion

The biggest difference in performance between the models was between the English sourced models and the Russian sourced models. Though there is a significant amount of Russian data alongside the Komi due to code switching and borrowing, training from a Russian model did not yield any improvement. We interpret this as demonstrating that the amount of data in the source language is more important than the relevance of the source

	CER/WER	Language Model: Komi	Language Model: All Available Text	Language Model: Komi & Russian
Source Language	English	0.415/0.767	0.441/0.824	**0.414/0.765**
	Russian	0.478/0.830	0.499/0.870	0.477/0.830

Table 2: The best Character Error Rate (CER) and Word Error Rate (WER) for each combination of source language and language model (lower is better). Note the best WER and best CER may have come from different language model α and β parameters.

language to the dataset. Common Voice for English has over 1400 hours of validated data, as compared to the 105 hours of Russian data.

It is unsurprising that the language model constructed from all available text caused a reduction in performance, as the focus of the dataset is on Komi. The 786 million tokens in the Russian Wikipedia corpus dwarfed the 2.76 million tokens available for Komi, and because the language model was constructed using only the 500000 most common words in the text, there was probably very little Komi accounted for. This language model combined with the English source language still achieves a better performance than the Russian source language runs. We conclude from this that while both source language and language model are important, source language is more important.

Additionally, it can be discussed whether an ASR system that relies so strongly on the language model is the best architecture for endangered languages, especially with very agglutinative morphology. While a language model is capable of handling new words it has not encountered, it will presume them to be less likely regardless of their validity. In the case of Komi, new morphologically complex word forms are continuously encountered for the first time in the new recordings, and there is no way that a relatively small corpus would cover them perfectly, not to even mention the dialectal forms that are common. Still, the language model has proven to have an important role in our approach, and also other systems could possibly benefit from using it in one form or another.

6 Conclusions

We can report continuous improvements from the earlier studies by Hjortnaes et al. (2020b) and Hjortnaes et al. (2020a), and our CER improves by several percentages from the earlier best score. This appears to be, however, just due to a different DeepSpeech version, as otherwise the test setup was identical. Our results indicate that when using transfer learning to create ASR tools for minority languages, the size of the source language is more important than the similarity or contact. Having a larger quantity of training data in the source language allows the model to learn to interpret on a phonetic level. This improvement in phonetic understanding is more valuable and impactful on the performance of the model than having a more relevant but lower resource source language after transferring to the target language. We do note, however, that while this is true for this particular case, it does not necessarily hold true for any source/target language pairing.

Multilinguality is by no means the only challenge the used dataset offers. For example, the corpus has a large amount of overlapping speech, which is very frequent in the interviews. Most of the recordings have more than two participants, and participants were not discouraged to use interruptions and small verbal cues, as these are essential for normal communication, and the goal was to collect natural speech. Additionally the corpus has a large number of speakers, and many speakers are present in one recording only, so the untranscribed recordings are prone to contain speakers who are entirely unseen by the ASR model. Further work is needed to effectively leverage resources in closely related languages and contact languages, as the current choice of English in transfer learning is not motivated by anything other than the amount of available data.

Acknowledgments

This research was supported in part by Lilly Endowment, Inc., through its support for the Indiana University Pervasive Technology Institute. Niko Partanen and Michael Rießler collaborate within the project *Language Documentation meets Language Technology: The Next Step in the Description of Komi*, funded by Kone Foundation, Finland.

References

Oliver Adams, Trevor Cohn, Graham Neubig, Hilaria Cruz, Steven Bird, and Alexis Michaud. 2018. Evaluating phonemic transcription of low-resource tonal languages for language documentation. In *Proceedings of LREC 2018*.

Oliver Adams, Benjamin Galliot, Guillaume Wisniewski, Nicholas Lambourne, Ben Foley, Rahasya Sanders-Dwyer, Janet Wiles, Alexis Michaud, Séverine Guillaume, Laurent Besacier, Christopher Cox, Katya Aplonova, Guillaume Jacques, and Nathan Hill. 2020. User-friendly automatic transcription of low-resource languages: Plugging espnet into elpis.

Khalid Alnajjar, Mika Hämäläinen, Niko Partanen, Jack Rueter, et al. 2019. The open dictionary infrastructure for Uralic languages. In *II Meždunarodnaâ naučnaâ konferenciâ Èlektronnaâ pismennost narodov Rossijskoj Federacii: opyt, problemy i perspektivy*, pages 49–51. Baškirskaâ ènciklopediâ.

R. Ardila, M. Branson, K. Davis, M. Henretty, M. Kohler, J. Meyer, R. Morais, L. Saunders, F. M. Tyers, and G. Weber. 2020. Common Voice. In *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020)*.

Timofey Arkhangelskiy. 2019. Corpora of social media in minority Uralic languages. In *Proceedings of the Fifth International Workshop on Computational Linguistics for Uralic Languages*, pages 125–140.

Kelly Berkson, Samson Lotven, Peng Hlei Thang, Thomas Thawngza, Zai Sung, James C Wamsley, Francis Tyers, Kenneth Van Bik, Sandra Kübler, Donald Williamson, et al. 2019. Building a common voice corpus for laiholh (hakha chin). In *Proceedings of the Workshop on Computational Methods for Endangered Languages*, volume 2.

L. M. Beznosikova, E. A. Ajbabina, and R. I. Kosnyreva. 2012. *Slovar dialektov komi âzyka*. Kola.

Rogier Blokland, Vasily Chuprov, Maria Fedina, Marina Fedina, Dmitry Levchenko, Niko Partanen, and Michael Rießler. 2021. Spoken Komi Corpus. The Language Archive version.

Rogier Blokland, Vasily Chuprov, Maria Fedina, Marina Fedina, Dmitry Levchenko, Niko Partanen, and Michael Rießler. forthcoming. Spoken Komi Corpus. The Language Bank of Finland version.

Stefan Braun, Daniel Neil, and Shih-Chii Liu. 2017. A curriculum learning method for improved noise robustness in automatic speech recognition. In *2017 25th European Signal Processing Conference (EUSIPCO)*, pages 548–552. IEEE.

Christopher Cox. 2019. *Persephone-ELAN: Automatic phoneme recognition for ELAN users*. Version 0.1.2.

Siddharth Dalmia, Ramon Sanabria, Florian Metze, and Alan W Black. 2018. Sequence-based multilingual low resource speech recognition. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4909–4913. IEEE.

Tessfu Geteye Fantaye, Junqing Yu, and Tulu Tilahun Hailu. 2020. Investigation of automatic speech recognition systems via the multilingual deep neural network modeling methods for a very low-resource language, Chaha. *Journal of Signal and Information Processing*, 11(1):1–21.

Marina Serafimovna Fedina. 2019. Korpus komi âzyka kak baza dlâ naučnyh issledovanij. In *II Meždunarodnaâ naučnaâ konferenciâ Èlektronnaâ pismennost narodov Rossijskoj Federacii: opyt, problemy i perspektivy*, pages 45–48. Baškirskaâ ènciklopediâ.

Ciprian Gerstenberger, Niko Partanen, and Michael Rießler. 2017. Instant annotations in ELAN corpora of spoken and written Komi, an endangered language of the Barents Sea region. In Antti Arppe, Jeff Good, Mans Hulden, Jordan Lachler, Alexis Palmer, and Lane Schwartz, editors, *Workshop on the Use of Computational Methods in the Study of Endangered Languages (ComputEl-2)*, pages 57–66. Association for Computational Linguistics.

Ciprian Gerstenberger, Niko Partanen, Michael Rießler, and Joshua Wilbur. 2016. Utilizing language technology in the documentation of endangered Uralic languages. *Northern European Journal of Language Technology*, 4:29–47.

Wei Han, Zhengdong Zhang, Yu Zhang, Jiahui Yu, Chung-Cheng Chiu, James Qin, Anmol Gulati, Ruoming Pang, and Yonghui Wu. 2020. Contextnet: Improving convolutional neural networks for automatic speech recognition with global context. *arXiv preprint arXiv:2005.03191*.

Awni Hannun, Carl Case, Jared Casper, Bryan Catanzaro, Greg Diamos, Erich Elsen, Ryan Prenger, Sanjeev Satheesh, Shubho Sengupta, Adam Coates, and Andrew Y. Ng. 2014. Deep Speech.

Kenneth Heafield. 2011. Kenlm. In *Proceedings of the sixth workshop on statistical machine translation*, pages 187–197.

Nils Hjortnaes, Timofey Arkhangelskiy, Niko Partanen, Michael Rießler, and Francis M. Tyers. 2020a. Improving the language model for low-resource ASR with online text corpora. In *Proceedings of the 1st joint SLTU and CCURL workshop (SLTU-CCURL 2020)*, pages 336–341, Marseille. European Language Resources Association (ELRA).

Nils Hjortnaes, Niko Partanen, Michael Rießler, and Francis M. Tyers. 2020b. Towards a speech recognizer for Komi, an endangered and low-resource Uralic language. In *Proceedings of the Sixth International Workshop on Computational Linguistics*

of Uralic Languages, pages 31–37. Association for Computational Linguistics, Vienna.

Jinyu Li, Rui Zhao, Hu Hu, and Yifan Gong. 2019. Improving rnn transducer modeling for end-to-end speech recognition. In *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 114–121. IEEE.

Josh Meyer. 2019. *Multi-task and transfer learning in low-resource speech recognition*. Ph.D. thesis, University of Arizona.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. Librispeech: an asr corpus based on public domain audio books. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5206–5210. IEEE.

Daniel S Park, William Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Ekin D Cubuk, and Quoc V Le. 2019. Specaugment: A simple data augmentation method for automatic speech recognition. *arXiv preprint arXiv:1904.08779*.

Niko Partanen. 2021. *DeepSpeech-ELAN*. Version 0.1.1. 10.5281/zenodo.4486474.

Niko Partanen, Rogier Blokland, KyungTae Lim, Thierry Poibeau, and Michael Rießler. 2018. The first Komi-Zyrian Universal Dependencies treebanks. In *Proceedings of the Second Workshop on Universal Dependencies (UDW 2018)*, pages 126–132. Association for Computational Linguistics.

Niko Partanen, Rogier Blokland, and Michael Rießler. 2020. A pseudonymisation method for language documentation corpora. In Tommi A. Pirinen, Francis M. Tyers, and Michael Rießler, editors, *Proceedings of the Sixth International Workshop on Computational Linguistics of Uralic Languages*, pages 1–8. Association for Computational Linguistics.

Vineel Pratap, Awni Hannun, Qiantong Xu, Jeff Cai, Jacob Kahn, Gabriel Synnaeve, Vitaliy Liptchinsky, and Ronan Collobert. 2019. Wav2letter++: A fast open-source speech recognition system. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6460–6464. IEEE.

Jack Rueter, Paula Kokkonen, and Marina Fedina. 2020. *Komi-Zyrian to X lexica*. Version 0.5.1, December 7. 2020. 10.5281/zenodo.4309763.

Jack M. Rueter. 2000. Helsinkisa universitetyn kyv tuâlys Ižkaryn perymsa simpozium vylyn lyddömtor. In *Permistika 6 (Proceedings of Permistika 6 conference)*, pages 154–158.

Jiatong Shi, Jonathan D. Amith, Rey Castillo García, Esteban Guadalupe Sierra, Kevin Duh, and Shinji Watanabe. 2021. Leveraging end-to-end ASR for endangered language documentation: An empirical study on Yoloxóchitl Mixtec. ArXiv:2101.10877.

Alexander Zahrer, Andrej Zgank, and Barbara Schuppler. 2020. Towards building an automatic transcription system for language documentation: Experiences from muyu. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 2893–2900.