NoLoR: An ASR-Based Framework for Expedited Endangered Language Documentation with Neo-Aramaic as a Case Study

Matthew Nazari [1,2]

[1] John A. Paulson School of Engineering and Applied Sciences, Harvard University
[2] The North-Eastern Neo-Aramaic Database Project, University of Cambridge
matthewnazari@college.harvard.edu

Abstract

The documentation of the Neo-Aramaic dialects before their extinction has been described as the most urgent task in all of Semitology today. The death of this language will be an unfathomable loss to the descendents of the indigenous speakers of Aramaic, now predominantly diasporic after forced displacement due to violence. This paper develops an ASR model to expedite the documentation of this endangered language and generalizes the strategy in a new framework we call NoLoR.

Introduction

Aramaic is the oldest continuously spoken Semitic language in the world. Today, the language is highly endangered and is expected to be within its final generation or two of fluent speakers. Speakers of the Neo-Aramaic dialects, which are comprised of Christian and Jewish communities across the Middle East (see Fig. 1), have been the victims of much violence over the past century, displacing them from their homelands. Such violence includes the Assyrian genocide of World War I (Gaunt 2020) and the systematic ethnic cleansing by Islamic State terrorist groups (Omer 2022). The death of this language will be an unfathomable loss for the descendents of these communities, to whom language is an essential part of their culture and identity.

This problem is not unique to Aramaic. Over 40% of spoken languages are endangered, and in the next 100 years, about 90% of spoken languages are expected to become extinct (Moseley and Nicolas 2010). Unfortunately, technologies like the internet and smartphones only further marginalize endangered language by disproportionally promoting the usage of high-resource languages like English (NLLB Team et al. 2022). Although the field of low-resource NLP appears to be addressing this disparity, it is predominantly focused on developing luxury technologoies, such as translators for already documented languages. However, such work overlooks languages that are underdocumented, without a written tradition, or with very few speakers remaining. Attention must be paid to such situations where the priority is to document the language before it is extinct.

Figure 1: Area where North-Eastern Neo-Aramaic dialects are spoken by indigenous communities (in red). The city of Urmi is the center for one of the most common dialects.

Endangered Language Documentation

In order to preserve endangered languages, they must be documented. The field of endangered language documentation emerged as its own field of linguistics in the 1990s. In this field, researchers document the grammar and oral history of a language by collecting samples of speech accompanied with a written transcription. However, there is a "transcription bottleneck" limiting the scale of documentation efforts (Shi et al. 2021). Annotating speech data is an extremely costly, lengthy, and esoteric process only performable by the few experts of that language. This severely limits the speed and capacity at which documentation efforts can operate. Hence, it is important that tools and frameworks exist to address this problem.

This paper constructs a framework for documenting endangered languages, employing ASR to overcome the transcription bottleneck. The effectiveness of this strategy is demonstrated by the application of NoLoR in the documentation of the Neo-Aramaic dialect of the Assyrian Christians of Urmi (C. Urmi).

The NoLoR Framework

The No- to Low-Resource (NoLoR) framework outlines the strategy of developing an ASR model to expedite the doc-
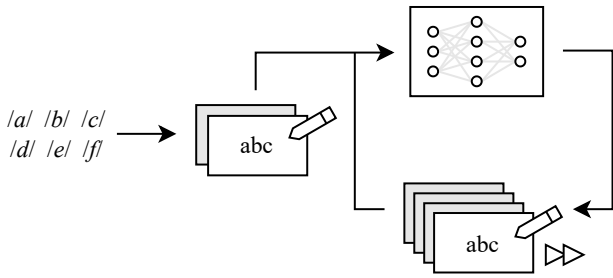
Figure 2: The NoLoR framework describes a positive feedback loop following two preliminary stages. In this loop, language documentation teams collect more data and their efficiency to transcribe increases.

umentation of C. Urmi in a manner generalizable to other language documentation efforts. The NoLoR framework is comprised of four parts:

Step 1: Defining a Phonemic Orthography. Many endangered languages are unwritten or may not have any pre-existing documentation efforts. Defining a phonemic orthography is hence necessary to begin the documentation process or align with pre-existing ones.

Step 2: Building an Initial Dataset. After collecting and transcribing interviews, folktales, and other narrations, we can construct an initial dataset ready for a computational pipeline.

Step 3: Training an ASR Model. An ASR model can be trained from the initial dataset by fine-tuning a pretrained ASR model using careful, task-specific optimizations.

Step 4: Expanding the Dataset. The transcription of subsequently collected data will be expedited and more data will then be created.

Steps 1 and 2 allows us to enter the positive feedback loop created by steps 3 and 4. Each iteration of this loop produces an ASR model that can become accurate enough to transcribe samples with only minimal expert intervention.

Contributions

This paper proves the NoLoR framework to have been an impactful contribution to both the Assyrian community and to the field of endangered language documentation through its application to the documentation of C. Urmi. Furthermore, this paper contributes a speech dataset, an ASR model, and AssyrianVoices, an online application to crowdsource speech data specifically for future machine learning tasks.

Related Work

Within Language Documentation

Researchers regularly use software to assist in documenting languages. Such software includes FieldWorks Language Explorer, which allows users to define a lexicon of the target language and to morphologically tag text (SIL International 2007). The software SayMore provides a means of file

	Language	Labelled	Unlabelled	CER (%)
Xu et al. 2020	Lithuanian	1.35hr	11.8hr	**10.3**
Zahrer et al. 2020	Muyu	2.05hr	0.0hr	48.3
This paper	Neo-Aramaic	**0.58hr**	**0.0hr**	12.5

Table 1: Comparison of previous examples of low-resource ASR and ours

organization and supplies an interface for audio transcription (SIL International 2010). Softwares like these are quite outdated, and they mainly aim to improve transcribing efficiency with convinient interfaces. Meanwhile, NoLoR is itself a framework that aims to automate the transcription process entirely.

Within Low-Resource NLP

The low-resource NLP community has frequently used endangered languages as a case study in various tasks, but not often ASR (Hedderich et al. 2020; Ranathunga et al. 2021; Adams et al. 2017). A main goal of this field is to eliminate communication boundaries between communities. Almost no work has explicitly sought to uplift the status of endangered languages by improving the workflow of fieldworkers documenting them.

The exception to this is one work that offers a general-purpose ASR model to assist in the documentation of the endangered Muyu language of New Guinea (Zahrer, Zgank, and Schuppler 2020). This ASR model was trained to learn a universal phonetic representation of human speech by training on speech data from 3 languages (American English, Austrian German, and Slovenian). Such model can then be used on other lanuages, including endangered ones. To overcome the low-resource barrier, we also exploit a general representation of human speech by fine-tuning a wav2vec 2.0 pretrained on not 3 languages, but 53 languages then a phonetically adjacent one. This difference, alongside many others, results in the ASR model produced by NoLoR to vastly outperform with much less data (see Table 1). Additionally, NoLoR rewards researchers by improving the accuracy of the ASR as more documentation work is collected through periodic fine-tuning the ASR. Finally, this paper emperically demonstrates that the ASR model was able to meaningfully improve transcription times.

Examining the literature more generally, there have been many attempts to train an ASR model with low training data. However, even works that fall under the category of "extremely low-resource" assume dozens of hours of unlabelled data (Xu et al. 2020). Our approach, using a phonemic orthography performing careful optimizations, proves to be on par with such work with far less data.

We also develop AssyrianVoices, an online application for crowdsourcing both speech samples of existing text examples and new text samples. This format is directly inspired by the Common Voice project started by Mozilla (Ardila et al. 2019). A seperate platform was developed for Neo-Aramaic in order to allow for the display of text in multiple transliteration schemes to maximize accessibility, something not possible with Common Voice.

Natural	Phonemic	Phonetic
'strewn'	/strun/	[stɹuːn]
'tenth'	/tɛnθ/	[tʰɛ̃nθ̥]
'zine'	/zin/	[ziːn̩]
'clean'	/klin/	[kɬiːn]

Table 2: A phonemic orthography is a sweet-spot representation of human language from both a language documentation perspective and a machine learning one.

Specifying a Phonemic Orthography

Specifying a phonemic orthography allows us to:

a) work with existing language documentation efforts or pave the way for future ones

b) minimize the amount of initial data we need to collect to train a minimally viable ASR model

In the case of C. Urmi, we align ourselves with existing work and use the orthographic standard set by Geoffrey Khan in his description of the dialect (Khan 2016).

The Importance of Orthography

In language documentation, the language of interest may not have a written tradition, or it may use a seperate esoteric literary language that is greatly misaligned with the vernacular. This was especially true for langauges of indigenous communities, such as the Navajo language of North America and the Neo-Aramaic dialects. In these cases, a very first step in documentation is to design an effective orthography (Cahill and Karan 2008). A phonemic orthography is one that assigns each unique unit sound in a language (phone) to a unique character. This scheme has been classically accepted by linguists as most ideal (Pike 1971). However, most contemperary opinions acknowledge the importance of other features of a language beyond phonemic analysis that ought to be represented in the orthography (Venezky 2004). In English, whose orthography is not phonemic, apostraphes to represent possessive constructs are an example of this. In the orthography we selected for C. Urmi, the double hyphen '=' indicates an enclitic boundary.
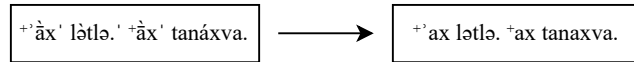
Figure 3: Example of refining original transcriptions for machine learning tasks.

Unsurprisingly, a phonemic orthography is a great written representation of spoken language in machine learning tasks since the written and the waveform representations of human speech are in principle being maximally aligned (Feng et al. 2021a; Khare et al. 2021). This is different from "natural" orthographies, like that of standard French or English, which have frequent irregularities and letters with multiple possible sounds, making it more difficult for speech recognition models to learn. One might then consider a phonetic
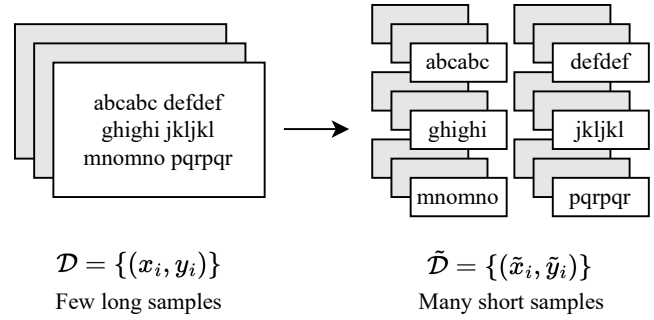

Figure 4: Data from the language documentation effort will not be formatted for machine learning tasks and must be processed accordingly.

orthography, like the International Phonetic Alphabet (IPA), but this is demonstrated as being too nuanced for ASR models to learn without massive amounts of examples.

The orthography of C. Urmi is phonemic with special symbols for certain features like enlitic boundaries and suprasegmental emphasis. Most of these are left in, with the exception of markers for features notorious for being difficult to pick up on such as vowel stress and intonation boundaries (Bosch 2000; Vicsi and Szaszák 2006).

Building an Initial Dataset

Building an initial dataset from preliminary documentation efforts allows us to train an ASR model that can expedite the transcription of future speech samples. In the case of C. Urmi, a labelled speech dataset of 35 minutes was built. This dataset has been made publically available[1] and is released under a Creative Commons CC0 license to encourage follow-up work.

Optimizing Data for Speech Recognition

Preliminary documentation efforts will produce a dataset $\mathcal{D} = \{(x_i, y_i)\}$ of audio-transcript pairs. Since a priority in preserving oral history is eliciting speech in the form of interviews and folktales, each pair will be too long to train a speech recognition model. Not only that, but when completely done by hand, the transcripts $y_i \in \mathcal{Y}$ may essentialize the utterance by not transcribing mishaps like repeated words. Hence, a new dataset $\tilde{\mathcal{D}} = \{(\tilde{x}_j, \tilde{y}_j)\}$ must be processed where $\tilde{y}_j$ is the exact transcript of $\tilde{x}_j$ and all $\tilde{x}_j$ are singificantly shorter (see Fig. 3).

The preliminary data for C. Urmi consisted of 8 examples, some as long as 11 minutes in length. This was carefully segmented into roughly 600 samples each no longer than 15 seconds.

Training an ASR Model

Training an ASR model allows us to transcribe future collected samples more efficiently. Intuitively, the ASR model,
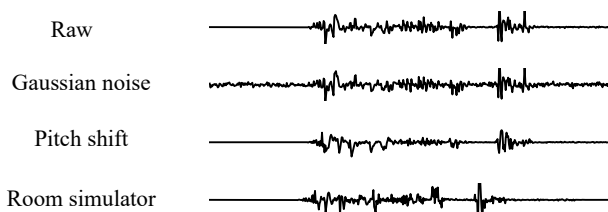

Figure 5: Data Augmentation

| Length | Without ASR | | With ASR | | Speedup |
	Time	CER (%)	Time	CER (%)	
15sec	3min	0.0	89sec	0.0	2.0×
1min	25min	0.0	7min	0.0	3.6×
3min	91min	0.5	17min	1.1	5.3×
5min	132min	0.8	21min	2.1	6.3×

Table 3: Transcribing 8 unique speech samples with and without ASR at varying lengths.

which might not be very accurate trained on only on the initial dataset, will output an initial draft of the transcription which human oversight can correct. In the case of C. Urmi, a wav2vec2.0 model fine-tuned on Persian and C. Urmi data achieved a CER of 12.5% and expidited transcriptions up to $6.3\times$. This model has been made publically available[2] to again encourage follow-up work.

Deep Learning Over Classical Models

With the immense quantity of speech data available today, deep learning has outperformed all traditional approaches to ASR (Malik et al. 2021). However, training these models can require up to 10,000 hours of labelled data, whereas traditional approaches generally require orders of magnitude less data. Training these traditional models is an extremely intensive process that would be difficult to perform repeatedly in NoLoR. This training data must also be phoneme-aligned, which does not work if simply collecting the unlabelled speech data is challenging enough.

Selecting a wav2vec 2.0 Checkpoint

Thankfully, end-to-end models like wav2vec 2.0 (Baevski et al. 2020) leverage pre-training to frontload the learning of human speech representations. This means that with very little training data, a wav2vec 2.0 model pretrained on 100,000+ hours of 53 multilingual speech data can be fine-tuned for other tasks (Chen and Rudnicky 2021), including ASR for other languages.

The wav2vec 2.0 architecture takes in as input raw waveform $\mathbf{x}_i \in \mathcal{X}$. An encoder network $f : \mathcal{X}^u \to \mathcal{Z}$ comprised of convolutional layers outputs a feature representation $\mathbf{z}_i \in \mathcal{Z}$ of the original waveform. Context is encoded into the feature representation using the context network $g : \mathcal{Z}^v \to \mathcal{C}$. This network mixes multiple latent representations $\mathbf{z}_i, \mathbf{z}_{i-1}, \ldots, \mathbf{z}_{i-v}$ into a single contextualized representation $\mathbf{c}_i$ using more convolutional layers. These contextualized representations can then be used as inputs into a language model and decoded into text.

The encoder and context networks are trained by self-learning from an enourmous amount of multilingual data. The initial dataset will not be large enough to fine-tune these layers to the specific phonology of the language of interest. If we don't fine-tune these layers, we run risk of being handicapped by generic contextualized representations. Hence,

---

in the case of C. Urmi, we use a wav2vec 2.0 checkpoint fine-tuned on a phonologically adjacent language, Persian.

Language Model

A character-level tokenizer was used as a language model. The alternate to this would be using a subword tokenizer with a relatively small vocabulary size of 256 or 512. This was attempted but proved to have poor results. Training a subword tokenizer requires sample text to learn the the frequencies of each subword. However, with such a limited amount of text data in existance, such a tokenizer proved to be too specific and unable to identify novel words or verb inflections. This behavior is expected according to Zipf's law.

The loss used to decode the contextualized representations was a connectionist temporal classification (CTC) loss (Graves 2012). CTC loss is necessary for training on data that is not phoneme-aligned.

Fine-tuning

In order to fine-tune from the wav2vec 2.0 checkpoint, the encoder and context networks were frozen during training and the CTC language model was concatenated to the end of the context network for fine-tuning. By feeding the contextualized representations into a dense layer, a probability distribution over possible characters is generated which can be fed into a CTC loss.

Overfitting on the training split of the dataset was a huge issue, as soon discussed. A very careful sweep of hyperparameters had to be done to discover which combination had the best regularizing effects.

Data Augmentation

As will be the case in many language documentaiton efforts, the number of speakers comprising the initial dataset may be very low. In the case of our dataset, all speech came from a single elderly woman. Machine learning models are interpolators, meaning it is often difficult to generalize their performance on out-of-distribution data (Neyshabur et al. 2017).

In our situation, it means the ASR model will overtrain on this specific speaker's voice by learning that her unique pronunciation habits are instrinsic to C. Urmi as a whole. We use data augmentation to combat this issue, which is a proven technique to improve model robustness and accuracy in low-resource NLP settings (Feng et al. 2021b). For each example, we apply augmentations (see Fig. 5) each at varying degrees. The augmentations were chosen specifically to counteract the speed of the speaker's utterance, the pitch of the speaker's voice, and the clarity of the microphone.
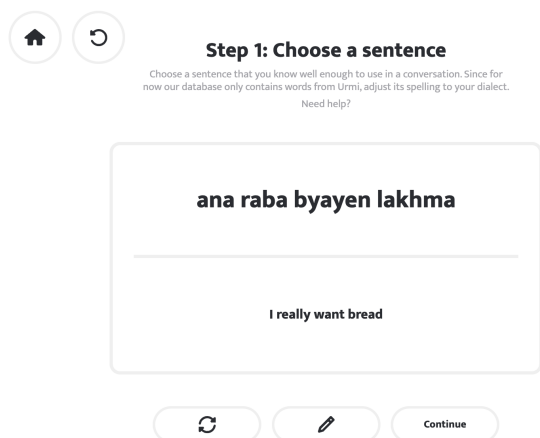
Figure 6: The interface for speakers to find a sentence they would be comfortable reading, which in this case in a simplified spelling.

Evaluation and Impact

To evaluate the ASR model, we looked to the diasporic Assyrian population of Armenia and travelled to the southern Assyrian villages near Verin Dvin. Here, we collected several samples of people speaking about the history of the village as well as various jokes and anecdotes. One striking story, in particular, comes from a grandmother who describes in grief how she is discouraged to speak to her children in her language since her daughter had married an Armenian outside of the village. With our work, her voice and story will be documented.

Regarding the efficacy of the ASR model, we clearly observe dramatic improvements to the speed of documentation when assisted by the ASR model (see Table 3). Although the ASR model achieved a test accuracy of 12.5% CER, the accuracy for certain samples reached as low as 27.5%. In these examples, the speakers exhibited alternate pronunciations of certain phones (like [ʤ] for /ʝ/ rather than [ɣ]). This articulation, less common in the C. Urmi dialect (Khan 2016), had become ubiquitous in this diasporic community. Additionally, when inputting several minutes into the ASR at once, entire sentences were misconstrued which decreased the accuracy significantly. However, given these challenges, the ASR still provided adequate drafts that proved to be great aids in the transcription process, up to $6.3\times$ faster for longer, more difficult transcriptions.

We noticed that with the ASR model, we tended to assume we were done transcribing early, which meant we allowed mistakes to make it through which we corrected later. We observed this behavior even without the ASR model but to a smaller extent.

Expanding the Dataset

To expand this dataset, we collect and transcribe more speech samples as we did in Armenia. Beyond this, however, we developed AssyrianVoices, an online crowdsourcing application for speakers of Neo-Aramaic around the world to donate speech samples.

The application is designed to attentively consider the backgrounds of the users. For example, users are able to specify which dialect they identify with. While the phonemic orthography is useful for transcription and modelling purposes, some characters are not readily legible to non-linguists. In the real world, speakers will use a standard English keyboard and phonetically spell words in a fairly non-standardized way. Some trends appear, such as representing the voiceless postalveolar fricative [š] with the sequence /sh/ as done in English. A simplification of the phonemic script is hence provided.

Conclusions

Language documentation is vital to the preservation of endangered languages, but transcribing speech samples is a costly and time-consuming process. Within the low-resource NLP community, there is virtually no discussion about how to facilitate the transcription process.

This paper provides the NoLoR framework for expedited endangered language documentation and demonstrates its efficacy using Neo-Aramaic as a case. By doing so, we produced an ASR model that helped document the oral history of the diasporic Assyrian community of Armenia. We then released our dataset of annotated speech samples and launched a crowdsourcing campaign with the goal of collecting data to further explore the role of AI in language documentation.

Next Steps

There are many future steps that will further improve the efficacy of NoLoR as well as develop it into a more out-of-the-box approach. One such improvement would involve building a model to segment the long-form audio samples automatically in order to shorten the step of building an initial dataset. Another direction is to implement a system where fieldworkers can simply dump their new transcribed samples and have the ASR model periodically fine-tune itself. This way, an engineer does not have to be on staff in order to continuously improve the ASR model. The most immediate work, however, is to continue language documentation efforts with C. Urmi and see how many iterations of NoLoR is necessary to eliminate the need for human intervention entirely.

An area of interest specific to Neo-Aramaic is creating programs to educate members of the community of the importance of documentation efforts like AssyrianVoices and teach them how to use it.

Acknowledgements

References

Adams, O.; Makarucha, A.; Neubig, G.; Bird, S.; and Cohn, T. 2017. Cross-Lingual Word Embeddings for Low-Resource Language Modeling. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, 937–947. Valencia, Spain: Association for Computational Linguistics.

Ardila, R.; Branson, M.; Davis, K.; Henretty, M.; Kohler, M.; Meyer, J.; Morais, R.; Saunders, L.; Tyers, F. M.; and Weber, G. 2019. *Common Voice: A Massively-Multilingual Speech Corpus*. arXiv.

Baevski, A.; Zhou, H.; Mohamed, A.; and Auli, M. 2020. wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations.

Bosch, L. t. 2000. Emotions: what is possible in the ASR framework. In *ISCA Tutorial and Research Workshop (ITRW) on Speech and Emotion*.

Cahill, M.; and Karan, E. 2008. Factors in designing effective orthographies for unwritten languages. *SIL International*.

Chen, L.-W.; and Rudnicky, A. 2021. Exploring Wav2vec 2.0 fine-tuning for improved speech emotion recognition. *arXiv preprint arXiv:2110.06309*.

Feng, S.; Żelasko, P.; Moro-Velázquez, L.; Abavisani, A.; Hasegawa-Johnson, M.; Scharenborg, O.; and Dehak, N. 2021a. How Phonotactics Affect Multilingual and Zero-Shot ASR Performance. In *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 7238–7242.

Feng, S. Y.; Gangal, V.; Wei, J.; Chandar, S.; Vosoughi, S.; Mitamura, T.; and Hovy, E. 2021b. A Survey of Data Augmentation Approaches for NLP.

Gaunt, D. 2020. The Long Assyrian Genocide. *Collective and State Violence in Turkey*, 56–96.

Graves, A. 2012. *Connectionist Temporal Classification*. Berlin, Heidelberg: Springer Berlin Heidelberg. ISBN 978-3-642-24797-2.

Hedderich, M. A.; Lange, L.; Adel, H.; Strötgen, J.; and Klakow, D. 2020. A Survey on Recent Approaches for Natural Language Processing in Low-Resource Scenarios.

Khan, G. 2016. *The Neo-Aramaic Dialect of the Assyrian Christians of Urmi (4 vols)*. Brill.

Khare, S.; Mittal, A. R.; Diwan, A.; Sarawagi, S.; Jyothi, P.; and Bharadwaj, S. 2021. Low Resource ASR: The Surprising Effectiveness of High Resource Transliteration. In *Interspeech*, 1529–1533.

Malik, M.; Malik, M. K.; Mehmood, K.; and Makhdoom, I. 2021. Automatic speech recognition: a survey. *Multimedia Tools and Applications*, 80(6): 9411–9457.

Moseley, C.; and Nicolas, A. 2010. *Atlas of the world's languages in Danger*. UNESCO Pub.

Neyshabur, B.; Bhojanapalli, S.; McAllester, D.; and Srebro, N. 2017. Exploring generalization in deep learning. *Advances in neural information processing systems*, 30.

NLLB Team; Costa-jussà, M. R.; Cross, J.; Çelebi, O.; Elbayad, M.; Heafield, K.; Heffernan, K.; Kalbassi, E.; Lam, J.; Licht, D.; Maillard, J.; Sun, A.; Wang, S.; Wenzek, G.; Youngblood, A.; Akula, B.; Barrault, L.; Gonzalez, G. M.; Hansanti, P.; Hoffman, J.; Jarrett, S.; Sadagopan, K. R.; Rowe, D.; Spruit, S.; Tran, C.; Andrews, P.; Ayan, N. F.; Bhosale, S.; Edunov, S.; Fan, A.; Gao, C.; Goswami, V.; Guzmán, F.; Koehn, P.; Mourachko, A.; Ropers, C.; Saleem, S.; Schwenk, H.; and Wang, J. 2022. No Language Left Behind: Scaling Human-Centered Machine Translation.

Omer, M. 2022. Religious Freedom in Iraq in 2021.

Pike, K. L. 1971. Phonemics: A technique for reducing languages to writing.

Ranathunga, S.; Lee, E.-S. A.; Skenduli, M. P.; Shekhar, R.; Alam, M.; and Kaur, R. 2021. Neural Machine Translation for Low-Resource Languages: A Survey.

Shi, J.; Amith, J. D.; García, R. C.; Sierra, E. G.; Duh, K.; and Watanabe, S. 2021. Leveraging End-to-End ASR for Endangered Language Documentation: An Empirical Study on Yoloxóchitl Mixtec.

SIL International. 2007. FLEx.

SIL International. 2010. SayMore.

Venezky, R. L. 2004. In search of the perfect orthography. *Written Language & Literacy*, 7(2): 139–163.

Vicsi, K.; and Szaszák, G. 2006. Prosodic cues for automatic phrase boundary detection in ASR. In *International Conference on Text, Speech and Dialogue*, 547–554. Springer.

Xu, J.; Tan, X.; Ren, Y.; Qin, T.; Li, J.; Zhao, S.; and Liu, T.-Y. 2020. LRSpeech: Extremely Low-Resource Speech Synthesis and Recognition.

Zahrer, A.; Zgank, A.; and Schuppler, B. 2020. Towards Building an Automatic Transcription System for Language Documentation: Experiences from Muyu. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, 2893–2900. Marseille, France: European Language Resources Association. ISBN 979-10-95546-34-4.