Introduction: Neural machine translation (NMT) has achieved excellent performance on many language pairs and translation tasks. However, two of the most pressing open problems in NMT are domain adaptation and *low-resource* scenarios [1] (i.e., language pairs for which large parallel corpora do not exist). This project will address both issues via cross-lingual domain adaptation in an extremely low-resource setting. The work is motivated by an urgent humanitarian crisis: refugees seeking asylum in the US who speak only Central American indigenous languages like K'iche', Mam, Kanjobal, and Mixtec [2]. The scarcity of interpreters in these languages makes it difficult for speakers to access legal services, and existing nonprofits and NGOs providing legal aid to refugees do not have adequate resources to provide interpreters. Additionally, existing methods for low-resource NMT do not scale down to the extremely low-resource situation for these Central American languages. Improving both extremely low-resource applicability and domain adaptation for NMT will increase the number of scenarios in which NMT is a viable solution. Additionally, domain adaptation techniques that work in the challenging low-resource setting will also improve NMT domain adaptation in higher-resource settings. Thus, this project addresses a challenging and widely applicable scientific problem with immediate humanitarian impacts.

Objectives and Hypothesis: The objective of my proposed research is a two-pronged approach to domain adaptation in low-resource NMT. I will address the domain issue by fine-tuning a pretrained NMT model on synthetic parallel data generated via backtranslation [3] of monolingual in-domain data in the target language. I will first validate my approach in a high-resource setting by fine-tuning a baseline Spanish-English model using backtranslated in-domain English data. Independent of the performance on low-resource languages, a Spanish-English translation system that is well-adapted for asylum testimonials will be of great use to organizations providing legal aid to refugees. Then, I will apply the fine-tuning via backtranslation approach to the low-resource case. In order to successfully backtranslate the English finetuning data, a decent English-LR model is needed. Thus, I plan to apply a combination of unsupervised NMT, transfer learning, multilingual translation models, and various data augmentation approaches to the problem of extremely low-resource NMT. Although various methods to augment parallel corpora have been proposed [4-7], they typically augment corpus size from a few hundred thousand sentences to a few million . They do not adequately address a context in which only thousands or tens of thousands of sentences are available, as is the case for K'iche' and other indigenous languages. This project will develop a linguistically-informed method to generate enough plausible, syntactically correct synthetic data that existing corpus augmentation techniques like backtranslation can be applied, bridging the chasm between researchers' assumptions about the amount of available monolingual data and the reality for many low-resource languages.

Method and Research Plan: In this research, I will use K'iche as a case study language to develop methods, then I will apply those methods to Kanjobal, Mam, and Mixtec. BLEU score, the standard evaluation metric for machine translation [8], will be the primary metric for evaluation of this proposed research. Possible BLEU scores range from 0 to 100, with higher scores indicating a closer match to the reference translation. In my exploratory work on the Magdalena Peñasco Mixtec dialect, achieving a maximum BLEU score of 7 with a model trained from scratch on 8000 Mixtec-English parallel sentences. After this preliminary work, I decided to use K'iche' as the test language in further study for two reasons: first, monolingual K'iche' speakers are more common among migrants to the US than monolingual Mixtec speakers;

second, there are more than fifty distinct dialects of Mixtec, each with varying degrees of mutual intelligibility, which makes collection and cleaning of data intractably difficult.

The first year of my work will focus on domain adaptation to asylum testimonials in the high-resource Spanish-English setting. High-resource NMT models are primarily trained on parallel data extracted from news articles, while LR models are generally trained on whatever portions of the Bible are available for a given language. Because the ultimate goal of my project is to translate first-person testimonials of asylees about the violence from which they are fleeing, I need to adapt the NMT models to fluently translate first-person narratives and to understand the specific vocabulary relevant to asylees. One way to accomplish this adaptation is to fine-tune a pretrained translation model on in-domain data. While parallel data is ideal, it is also possible to use monolingual target-language data via backtranslation. I plan to use the transcripts from the MALACH dataset [9], compiled from the USC Shoah Foundation's Video History Archive, which collects testimonials from Holocaust survivors. I hypothesize that these testimonials are similar enough to the stories of modern asylees to be useful as in-domain training data.

In the subsequent years of the project, I will apply the domain adaptation technique developed in the first year to the extremely low-resource setting. The first priority for Year 2 is to gather and clean a K'iche'-English parallel corpus (at least the New Testament and some Jehovah's Witness literature is readily available). Next, I will establish a baseline machine translation result for K'iche'-English using available parallel corpora with no data augmentation. Then, I will apply the same backtranslation for domain adaptation technique used for Spanish-English. However, I expect that the backtranslation results will be unsatisfactory because of the limitations of the English-K'iche' model used for backtranslation. Thus, I plan to develop a data augmentation scheme syntax-aware compositional data augmentation methods from [10], with some simplifications necessary due to data scarcity. Sentences produced with this method are guaranteed to be grammatically correct, but may not make sense semantically; however, they are much easier to produce than semantically correct sentences, especially in a low-resource setting. This research will investigate whether NMT systems can learn useful information from synthetic data that is syntactically, but not semantically, correct. In addition to data augmentation, I plan to explore whether the K'iche'-English and English-K'iche' models can be further improved by transfer learning and unsupervised NMT techniques.

Intellectual Merit: This work addresses a crucial gap in current methods for low-resource NMT: cases where there are fewer than 10,000 sentences of monolingual data available. My proposed data augmentation method would allow researchers to generate enough synthetic monolingual data to apply state-of-the-art methods in low-resource NMT to languages that would otherwise be intractable due to lack of data. If this data augmentation is successful, it will show that NMT systems can learn grammar and syntax from synthetic training data that is semantically very noisy, opening up future research on how exactly NMT systems learn grammar and how this learning differs from learning vocabulary.

Broader Impacts: The human rights impact of my proposed research is immediate and transformative. Interpreters for K'iche' and related languages are hard to find and often prohibitively expensive. Translation systems, even imperfect ones, would allow non-governmental organizations and pro bono immigration lawyers to help asylees they were previously unable to serve. Widespread access to legal assistance would speed up backlogged immigration courts and give thousands of asylum seekers per year a chance to enter the US legally. This work could save lives, because many of these asylees are children and teenagers fleeing from horrific gang violence.

References

[1] Koehn, Philipp, and Rebecca Knowles. "Six Challenges for Neural Machine Translation." *Proceedings of the First Workshop on Neural Machine Translation*. 2017.

[2] Medina, Jennifer. "Anyone Speak K'iche' or Mam? Immigration Courts Overwhelmed by Indigenous Languages" *New York Times* (2019).

[3] Sennrich, R., Haddow, B., & Birch, A. (2015). Improving neural machine translation models with monolingual data. *arXiv preprint arXiv:1511.06709*.

[4] Zhang, Jinyi, and Tadahiro Matsumoto. "Corpus Augmentation by Sentence Segmentation for Low-Resource Neural Machine Translation." *arXiv preprint arXiv:1905.08945* (2019).

[5] Li, Hongheng, & Heyan Huang. "Evaluating Low-Resource Machine Translation between Chinese and Vietnamese with Back-Translation". *arXiv preprint arXiv:2003.02197* (2020).

[6] Currey, Anna, Antonio Valerio Miceli-Barone, and Kenneth Heafield. "Copied Monolingual Data Improves Low-Resource Neural Machine Translation." *WMT 2017.*

[7] Fadaee, Marzieh et al. "Data Augmentation for Low-Resource Neural Machine Translation." *ACL 2017.*

[8] Papineni, Kishore, et al. "BLEU: a Method for Automatic Evaluation of Machine Translation." *ACL 2002.*

[9] Ramabhadran, Bhuvana, et al. USC-SFI MALACH Interviews and Transcripts English – Speech Recognition Edition LDC2019S11.

[10] Andreas, Jacob. "Good-enough compositional data augmentation." *arXiv preprint arXiv:1904.09545* (2019).