# Exploring Transfer Learning Pathways for Neural Machine Back Translation of Eskimo-Aleut, Chicham, and Classical Languages

**Aaron Serianni**
SIL International
7500 W. Camp Wisdom Road
Dallas, TX 75236-5629 USA
serianni@princeton.edu

**Daniel Whitenack**
SIL International
7500 W. Camp Wisdom Road
Dallas, TX 75236-5629 USA
dan_whitenack@sil.org

## Abstract

Back translations are an important resource for those reviewing the quality of candidate translations. We explore various transfer learning techniques to create automated back translations in low resource scenarios with neural machine translation models. Results from Eskimo-Aleut, Chicham, and classical languages suggest that transfer learning using related language data improves back translation quality, even when the domain of the related language data does not match the target domain.

## 1 Introduction

Of the 7000+ currently spoken languages (David M. Eberhard et al., 2021), only a small fraction have sufficient corpora to train state-of-the-art neural machine translation (NMT) models. Translation of content into local, low resource languages remains a largely manual process, and, for important documents, translations need to be checked to ensure quality. Creation of back translations (from the target language to the source language) is one widely-used tool utilized by human reviewers to ensure translation quality (Brislin, 1970). Back translation also provides a mechanism by which all of human society can better understand and benefit from the culture, values, thoughts, and world views expressed in local languages.

Transfer learning for NMT is a technique where a parent model is trained to perform a task closely related to the target translation task. A child model is then initialized on the parent model parameters and further trained on the target task (Zoph et al., 2016). We explore transfer learning pathways for NMT back translation of Saint Lawrence Island (or Central Siberian) Yupik, Latin, and Chicham language texts. Transfer learning allows us to utilize data from related languages to improve NMT back translation results for low resource languages. In some experiments, parent and child models are trained on data from the same domain (e.g., literature, news, or software documentation), and in other experiments the domain of parent and child models differ. Transfer learning improves back translation results in both cases, which suggests that this technique is generally useful in low resource NMT back translation scenarios.

## 2 Languages and Data Sources

Ethnologue reference data (David M. Eberhard et al., 2021) was used to determine parent-child language pairs based on proximity within the Ethnologue language taxonomy and availability of corresponding data. For all NMT experiments the back translation language is taken to be English.

Our first pair of parent and child languages is Inuktitut [iku] and Saint Lawrence Island Yupik [ess] (“Yupik”), which are both Eskimo-Aleut languages. This pairing allows us to train a parent model (Inuktitut to English) on a large related language corpus in a first domain, and then fine-tune a child model (Yupik to English) on a relatively small corpus in a second domain. We use the Nunavut Hansard Inuktitut-English Parallel Corpus (Joanis et al., 2020) for Inuktitut-English, which includes transcriptions of legislative proceedings. For Yupik-English, Bible texts were sourced from the Digital Bible Platform[1].

[1] https://www.digitalbibleplatform.com/

Our second parent-child pair isolates the scenario of domain adaptation (Chu et al., 2017) within a single language (Latin [lat]). In this case, the parent model is trained on text from the OPUS parallel corpus (Tiedemann and Nygaard, 2004), made up of translations from various domains. The child model is trained on data from a single domain, Bible data sourced from the Digital Bible Platform.

To isolate the scenario of transfer learning using in-domain data from related languages, we used three Chicham languages from northern Peru and eastern Ecuador: Shuar [jiv], Shiwiar [acu], and Awajun [agr]. Their datasets were also gathered from the Digital Bible Platform. The parent model is trained on the combination of all data from the three languages (Shuar, Shiwiar, and Awajun to English), and then each child model was individually fine-tuned on data from the respective language.

The sizes of the datasets described above are in Table 1. A randomized split of 70-15-15% for train-validation-test with a maximum of 10,000 samples for each validation and test set was used across all datasets.

| **iku-eng** | **ess-eng** | **OPUS lat-eng** | **DBP lat-eng** | **acu-eng** | **agr-eng** | **jiv-eng** |
|---|---|---|---|---|---|---|
| 1,145,965 | 4,938 | 39,480 | 5,004 | 4,780 | 5,004 | 4,915 |

Table 1: The total number of sentence pairs in each language after preprocessing, inclusive of splits

## 3 Methods

To pre-process the data, we romanize the scripts, which reduces the load on our models (Schwartz et al., 2020). Since Yupik, all Chicham languages, and Latin already use the Latin alphabet as their writing system, romanization was not necessary, and only diacritics were removed. We used the rule-based system provided by Joanis et al. (2020) to romanize Inuktitut syllabics.

After romanization, data was normalized to remove extraneous characters and segment sentences. To filter our data, the `fast_align` library (Dyer et al., 2013) was used to create source-target word pairs within corresponding sentences. Using the distances from these word pairs, we calculated a mean absolute error for each sentence pair, removing the worst 10% quantile from each dataset. We use the `subword-nmt` library to generate byte-pair encodings of the source and target languages, forming vocabularies of subwords (Sennrich et al., 2016). For transfer learning, byte-pair encodings must be generated from the union of the parent and child datasets (Nguyen and Chiang, 2017).

We chose to utilize transformer-based models for all experiments as these models are widely used throughout NMT community (Vaswani et al., 2017; Popel and Bojar, 2018). Specifically, our models used a transformer-based encoder and decoder with an embedding size of 512. All training was implemented using the JoeyNMT framework (Kreutzer et al., 2019) with the Adam optimizer, cross entropy loss, and an initial learning rate of 0.0002. We used early stopping to prevent overfitting during fine-tuning, and the parent models' layers were not frozen. BLEU scores were calculated using `sacreBLEU` (Post, 2018) as the evaluation metric.

## 4 Results and Discussion

| | **ess-eng** | **lat-eng** | **acu-eng** | **agr-eng** | **jiv-eng** |
|---|---|---|---|---|---|
| **Child Data Only** | 11.07 | 16.02 | 9.46 | 13.81 | 7.65 |
| **With Transfer Learning** | 16.41 | 25.93 | 17.85 | 23.66 | 13.14 |
| **Percent Improvement** | 48.24% | 61.86% | 88.69% | 71.33% | 71.76% |

Table 2: Improvements in BLEU scores with transfer learning.

BLEU score results from all experiments are included in Table 2. Across all investigated language families and parent-child combinations, transfer learning substantially improves BLEU scores for child-to-English translation. These improvements cannot be compared directly across combinations, as each language family have drastically different morphologies, and the size of the datasets differ as well.

However, we can conclude that for back translation of texts into English, transfer learning is generally beneficial to improve accuracy of NMT transformer models. Although the benefit is maximized when the domains of the parent and child models match, transfer learning still boosts performance when parent and child model domains do not match.

## References

Richard W. Brislin. 1970. Back-Translation for Cross-Cultural Research. *Journal of Cross-Cultural Psychology*, 1(3):185–216. _eprint: https://doi.org/10.1177/135910457000100301.

Chenhui Chu, Raj Dabre, and Sadao Kurohashi. 2017. An Empirical Comparison of Domain Adaptation Methods for Neural Machine Translation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 385–391, Vancouver, Canada. Association for Computational Linguistics.

David M. Eberhard, Gary F. Simons, and Charles D. Fennig. 2021. *Ethnologue: Languages of the World*. SIL International, Dallas, Texas, twenty-fourth edition.

Chris Dyer, Victor Chahuneau, and Noah A Smith. 2013. A Simple, Fast, and Effective Reparameterization of IBM Model 2. *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 644–648.

Eric Joanis, Rebecca Knowles, Roland Kuhn, Samuel Larkin, Patrick Littell, Chi-kiu Lo, Darlene Stewart, and Jeffrey Micher. 2020. Le corpus parallèle inuktitut – anglais du Hansard du Nunavut 3.0The Nunavut Hansard Inuktitut–English Parallel Corpus 3.0, January. Type: dataset.

Julia Kreutzer, Joost Bastings, and Stefan Riezler. 2019. Joey NMT: A Minimalist NMT Toolkit for Novices. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP): System Demonstrations*, pages 109–114, Hong Kong, China. Association for Computational Linguistics.

Toan Q Nguyen and David Chiang. 2017. Transfer Learning across Low-Resource, Related Languages for Neural Machine Translation. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 296–301.

Martin Popel and Ondřej Bojar. 2018. Training Tips for the Transformer Model. *The Prague Bulletin of Mathematical Linguistics*, 110(1):43–70, April. arXiv: 1804.00247.

Matt Post. 2018. A Call for Clarity in Reporting BLEU Scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Belgium, Brussels. Association for Computational Linguistics.

Lane Schwartz, Francis Tyers, Lori Levin, Christo Kirov, Patrick Littell, Chi-kiu Lo, Emily Prud'hommeaux, Hyunji Hayley Park, Kenneth Steimel, Rebecca Knowles, Jeffrey Micher, Lonny Strunk, Han Liu, Coleman Haley, Katherine J. Zhang, Robbie Jimmerson, Vasilisa Andriyanets, Aldrian Obaja Muis, Naoki Otani, Jong Hyuk Park, and Zhisong Zhang. 2020. Neural Polysynthetic Language Modelling. *arXiv:2005.05477 [cs]*, May. arXiv: 2005.05477.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural Machine Translation of Rare Words with Subword Units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

Jorg Tiedemann and Lars Nygaard. 2004. The OPUS corpus - parallel and free http://logos.uio.no/opus/. *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, page 4, May.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, \Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Barret Zoph, Deniz Yuret, Jonathan May, and Kevin Knight. 2016. Transfer Learning for Low-Resource Neural Machine Translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1568–1575, Austin, Texas. Association for Computational Linguistics.