Language Family Data Scheduling for Low-Resource Speech Translation

Jennifer Huang

A thesis submitted
in partial fulfillment of the
requirements for the degree of

Master of Science

University of Washington

2026

Reading Committee:
Gina-Anne Levow, Chair
Scott Lundberg

Program Authorized to Offer Degree:
Department of Linguistics

University of Washington

Abstract

Language Family Data Scheduling for Low-Resource Speech Translation

Jennifer Huang

Chair of the Supervisory Committee:
Professor Gina-Anne Levow
Department of Linguistics

Low-resource languages often underperform on natural language processing tasks when compared to high-resource languages due to their limited amount of data. In particular, supervised speech translation tasks require parallel language data in the form of source language audio paired with target translation language text, which is very difficult to obtain for languages which already lack abundant resources. Given a low-resource language, we study the effects of incorporating high-resource language data with the low-resource language data in the training of low-resource speech translation models. We choose the high-resource language by selecting one which is the closest in the phylogenetic language family tree to the low-resource language, in this case focusing on the high-resource language of Welsh and low-resource language of Irish. By exploring the effects of incorporating the Welsh data with the Irish data through various data scheduling configurations, we find that data schedules which utilize both Welsh and Irish data bolster performance on the naturalistic Irish dataset. However, when the Irish dataset is augmented with synthetic data, it is more beneficial to solely use the Irish data.

TABLE OF CONTENTS

Page

List of Figures iii

List of Tables v

Glossary vi

Chapter 1: Introduction 1

Chapter 2: Literature Survey 3
2.1 Speech Translation Models 3
2.2 Curriculum Learning 5
2.3 Language Vector Representation 6

Chapter 3: Methodology 7
3.1 Language Selection 7
3.2 Datasets 9
3.3 Model 10
3.4 Experiments 13
3.5 Evaluation 16

Chapter 4: Results 18
4.1 Quantitative Results 18
4.2 Qualitative Results 26

Chapter 5: Discussion 28
5.1 Decoder Freezing Leads to Higher Performance for Mixed Language Fine-Tuning 28
5.2 Decoder Freezing Performance on Pure Low-Resource Language Fine-Tuning Varies Based on Presence of Synthetic Data 29

5.3 Synthetic Data Augmentation Facilitates Better Overall Semantic Learning . 30
5.4 Mixed Language Data Scheduling is Useful Only in the Naturalistic-Only Low-Resource Language Dataset Scenario 31

Chapter 6: Conclusion and Future Work 33
6.1 Summary and Conclusions 33
6.2 Limitations 34
6.3 Future Work 35

Bibliography 37

Appendix A: Model Weights and Code 42

LIST OF FIGURES

Figure Number Page

3.1 High-level Whisper model architecture. Components shown in blue are frozen during fine-tuning, excluding layer normalization. 10

4.1 chrF++ metric on high-resource evaluation dataset over high-resource training portion only of 100% HR and then 100% LR data schedule. 21

4.2 COMET metric on high-resource evaluation dataset over high-resource training portion only of 100% HR and then 100% LR data schedule. 21

4.3 chrF++ metric on low-resource evaluation dataset over training for 100% LR naturalistic-only experiment. 22

4.4 COMET metric on low-resource evaluation dataset over training for 100% LR naturalistic-only experiment. 22

4.5 chrF++ metric on low-resource evaluation dataset over training for mixed language naturalistic-only experiments. Start and end of linear phase of curriculum schedule A indicated by the dashed vertical lines. Only the LR portion of the 100% HR and then 100% LR data schedule is depicted. 23

4.6 COMET metric on low-resource evaluation dataset over training for mixed language naturalistic-only experiments. Start and end of linear phase of curriculum schedule A indicated by the dashed vertical lines. Only the LR portion of the 100% HR and then 100% LR data schedule is depicted. 23

4.7 chrF++ metric on low-resource evaluation dataset over training for synthetic data augmentation experiments. Start and end of linear phase of curriculum schedule A indicated by the dashed vertical lines. Only the LR portion of the 100% HR and then 100% LR data schedule is depicted. 24

4.8 COMET metric on low-resource evaluation dataset over training for synthetic data augmentation experiments. Start and end of linear phase of curriculum schedule A indicated by the dashed vertical lines. Only the LR portion of the 100% HR and then 100% LR data schedule is depicted. 24

4.9 Sample validation sentence prediction outputs and their references from experiment with no fine-tuning. 25

4.10 Sample validation sentence prediction outputs and their references from naturalistic-only fine-tuning on curriculum learning schedule B experiment. 25
4.11 Sample validation sentence prediction outputs and their references from synthetic data augmentation fine-tuning on 100% LR experiment. 26

LIST OF TABLES

Table Number Page

3.1 Common experiment model hyperparameters. 12
3.2 Naturalistic experiment model step hyperparameters. 13
3.3 Synthetic data augmentation experiment model step hyperparameters. . . . 13

4.1 Baseline experiment results. 95% confidence interval for BLEU and chrF++ shown in the square brackets next to their respective values. 19
4.2 Naturalistic-only experiment results. 95% confidence interval for BLEU and chrF++ shown in the square brackets next to their respective values. 20
4.3 Synthetic data augmentation experiment results. 95% confidence interval for BLEU and chrF++ shown in the square brackets next to their respective values. 20
4.4 Test results on low-resource language compared to official IWSLT 2026 submissions. 20

A.1 Mapping of experiment configurations to model weights. 42

GLOSSARY

AUTOMATIC SPEECH RECOGNITION: A task which converts from a speech audio signal into written text.

CURRICULUM LEARNING: A machine learning technique in which the model is trained on examples which get progressively more difficult as training progresses, inspired by human learning progression in education.

LOW-RESOURCE LANGUAGE: A language which lacks the parallel data at scale needed for standard supervised learning.

PYHLOGENETIC DISTANCE: A measure of how closely related a given two languages are based on their ancestral family tree.

SPEECH TRANSLATION: A task which converts from speech in one language in one language to text in another language.

ACKNOWLEDGMENTS

I would like to thank my advisor, Professor Gina-Anne Levow, for her time and investment in me. She was a regular sounding board who provided me with wisdom and advice throughout my thesis writing process. I would also like to thank my reader, Scott Lundberg, for his time and valuable feedback on this work.

Thank you to Atul Kr. Ojha for helping run the IWSLT 2026 test evaluations on my outputs to present in this thesis, as well as Antonios Anastasopoulos for connecting us.

I would also like to thank my dearest family and friends for their support throughout this process: Mom, Dad, Joseph, and Anda.

Finally, thanks be to God, my Lord and Savior. Through His abounding and unceasing grace, He has given me the time, abilities, and opportunity to do this work.

DEDICATION

This thesis is dedicated to my late grandpa, Professor Hsu Loke Soo, the inaugural head of the Department of Computer Science at the National University of Singapore. More than his numerous academic achievements, he was a tangible example to me of what it looked like to live each day with curiosity and joy.

Chapter 1

INTRODUCTION

Speech translation aims to convert speech from a source language into the text of a target language. In doing so, it not only converts from audio signal to written text but also converts from one language to another. This is highly useful in cross-lingual contexts, improving communication and accessibility across linguistic barriers. It is used in many applications today where the source and target languages are different, including meeting subtitling, clinical healthcare documentation, and voice assistants. As such, low latency and high accuracy are paramount in this task.

In cases where both the source and target languages have large amounts of data, speech translation has high accuracy. However, performance is significantly lower in cases where at least one of the languages is low-resource due to data scarcity. In low-resource settings, there is not enough parallel language data to properly train the traditional supervised models used for speech translation. Using the conventional models and methods results in low accuracy, leading to poor usability and adoptability by downstream applications. Thus, alternative approaches apart from the conventional must be taken in order to produce a viable speech translation model for low-resource languages.

In this thesis, we investigate if introducing data from a high-resource language phylogenetically similar to a low-resource language in the training process of end-to-end speech translation models can improve performance on low-resource languages. Given the vector representation of a low-resource language, we find its close neighbors based on phylogenetic distance and choose the closest neighbor which is high-resource. We go over past related work in this domain which motivates this approach in Chapter 2.

In Chapter 3, we detail our experiment methodology. At a high level, we aim to ex-

plore the incorporation of high-resource language data in the model training process within several different configurations: transfer learning where the model is first trained on the high-resource language and then trained on the low-resource language, a static 50%/50% split between the high-resource language data and the low-resource language data, and data-driven curriculum learning. For data-driven curriculum learning, we start training with the data from the high-resource language. As the training process progresses, we introduce the data from the low-resource language through a data scheduler.

Overall, through our conducted experiments and analysis presented in Chapter 4 and Chapter 5, we find that utilizing data from high-resource phylogenetic neighbors can improve low-resource language performance, but in limited contexts. In Chapter 6, we summarize our findings into conclusions and present directions for future work. This includes expanding our study to more language pairs and performing targeted experimentation to determine dataset influence over phylogenetic mixed language data scheduling effectiveness.

Chapter 2

LITERATURE SURVEY

2.1 Speech Translation Models

Traditionally, speech translation is done using a cascaded, modular approach where a first component transcribes from source language audio to source language text and a second component takes the source language text output of the first component and translates it to target language text [28]. However, this means that any errors from the transcription step will get propagated to the translation step, as the translation step does not have the direct context of the source audio to help with error correction.

More recently, there has been an emphasis on the direct end-to-end approach to speech translation, focusing on a single model which directly takes the source language audio as input and outputs the target language text [6]. Avoiding the distinct separate steps of transcription followed by translation allows the model to have access to the source audio when deciding the final target text, which helps the output more accurately capture features present in the audio such as prosody [33]. Furthermore, this avoids the cascaded transcription model to translation model error propagation issue as there are no longer separate distinct modules for transcription and translation. Not only does this end-to-end approach improve accuracy, but also latency as we no longer require a cascading system where one module must complete before the next one can start [26].

There are multiple different model architectures for end-to-end speech translation today. The traditional transformer-based encoder-decoder architecture used for automatic speech recognition can be extended to bridge across languages to perform speech translation, where the source language audio is encoded and the target language text is decoded. An example of this is Whisper [18], a weakly supervised model trained on 680,000 hours of multi-task

audio to perform automatic speech recognition and speech translation tasks. The source audio's log-Mel spectrogram is processed by Whisper's speech encoder and is decoded into text with the help of task and source language decoder prompt tokens. Another example is SeamlessM4T-v2 [4], a massively multilingual and multimodal end-to-end speech translation model for multiple language pairs of the shared task. It supports speech-to-speech, speech-to-text, text-to-speech, and text-to-text translation. For speech-to-text translation, the model encodes the audio waveform using w2v-BERT 2.0 [4] and a length adapter, aligning acoustic features with text tokens. The No Language Left Behind (NLLB) transformer decoder [15] then generates the translated text.

Besides encoder-decoder models, recent research has invested in speech large language models (LLMs), which are pre-trained LLMs extended to process speech. One such model is AudioPaLM [22], a fusion of text-based and speech-based language models. The model uses self-supervised speech encoders and neural audio codecs to take the input audio as a continuous representation and transform it to a discrete representation which can then be used as input into the pre-trained decoder-only LLM. As LLMs have already been proven to have very good text-to-text translation capabilities, they are able to learn and produce reasonably accurate mappings from the source audio tokens to the target language text output.

2.1.1 Low-Resource Speech Translation Models

In high-resource language settings, the traditional cascaded approach for speech translation typically works well out-of-the-box, as the automatic speech recognition step is already generally highly accurate and leaves little error to propagate to the translation step. However, when the source language is low-resource, upstream transcription errors are more frequent and thus more likely to impact the downstream translation. Ambiguity from the first transcription step is lost, as its output must be passed sequentially directly to the second translation step. For high-resource languages where the original language transcription does not have much ambiguity, this is less important, but for low-resource languages, ambiguity in its

original transcription is significant and thus this is more important.

Sonawane [25] investigated creating robust cascaded systems for the 2026 International Workshop on Spoken Language Translation (IWSLT) low-resource speech translation task, focusing on the Bhojpuri to Hindi and Irish to English language pairs. Through evaluation of multiple automatic speech recognition models and translation routing strategies, they found that cascaded pipelines for those language pairs require careful design, as the accuracy of the transcription step is the decisive factor in end-to-end performance.

In previous iterations of IWSLT on the same shared task, Robinson [21] and Meng [13] both fine-tuned an end-to-end model, Seamless4MT-v2 [4], for multiple low-resource languages. In both cases, they found that fine-tuning improves performance only on languages which the model was exposed to during training. As such, they concluded that the best-performing approach varies based on the specific language pair and the model being used.

2.2 Curriculum Learning

Curriculum learning is a model training technique used to train neural networks which is inspired by human education [5]. The model is trained starting with simple examples and difficulty is gradually introduced, which can be done either at the task or the data level. Curriculum learning is especially useful for low-resource settings, as it makes the best use of the limited data available and helps avoid overfitting. In contrast with transfer learning [16], where a model already trained for one task is used as the starting point for a new related task, curriculum learning focuses on the order and difficulty of the examples that are introduced to the model during the training process.

Curriculum learning has been applied to improve the performance of end-to-end speech translation models, typically at the task level. For example, in Du [11] the model is first trained on the "easiest" task of transcribing from the source audio to the source language text, then is trained on the "intermediate" task of translating from the source audio and the source language text to the target language text, and finally is trained on the "hardest" task of producing the source language text and target language text from the source audio.

They found that with this technique, less than 10 hours of speech audio for each language was needed to achieve state-of-the-art performance for low-resource settings. In another case, Wang [30] found that a combination of both transfer learning and curriculum learning is ideal for training speech translation models, as they employed a task-based curriculum pre-training method which led to significant speech translation performance improvement.

2.3 Language Vector Representation

While the data for low-resource languages is scarce, each language has additional characteristics besides its audio and written data that can be leveraged. These include features such as language syntax, geographical location, and phonology, which can be represented as language-specific vectors such as in lang2vec [12]. This language vector representation allows us to calculate the distance between vectors, revealing other languages which are linguistically similar in various aspects to a given low-resource language. This gives us additional context that could help improve low-resource language task accuracy.

Chapter 3

METHODOLOGY

3.1 Language Selection

3.1.1 Translation Target Language

For the speech translation task, we focus on instances where the source language is low-resource and the target language is English. There are several reasons for doing so, one of which is that the Whisper model, which we use in our experiments and is detailed in Subchapter 3.3, only supports speech translation where the target language is English. Furthermore, having the final output as English text also enables the evaluation to be done using the standardized metrics (detailed more in Subchapter 3.5) which are used to measure performance against the baseline in shared tasks such as those from IWSLT [24].

3.1.2 Source Low-Resource and High-Resource Languages

We leverage the low-resource language datasets from the 2026 International Workshop on Spoken Language Translation (IWSLT) [24], which has a shared task for low-resource speech translation. For training and validation, these datasets take the form of low-resource language speech audio paired with their corresponding English text translations. For testing, the datasets consist of the low-resource language speech alone. The IWSLT 2026 shared task datasets which have the target language of English include the low-resource source languages of Bemba, Central Kurdish, Irish, Igbo, Hausa, Catalan, and Yoruba.

Of these low-resource languages, we select one which can be paired with a high-resource language using the process detailed as follows. Using lang2vec [12] where each language is represented by a vector of a combination of geographical, phylogenetic, and typological

features, we select the high-resource language which is closest in similarity to our low-resource language. This is done through the URIEL knowledge base [12], which includes distance measures between the lang2vec vectors for 4005 languages. These distance measures include geographical, phylogenetic, and typological distances.

We focus on phylogenetic distance, which measures how close two given languages are in the language family tree, looking at historical lineage. We choose this distance metric with the hypothesis that knowing a high-resource language, which shares common historical derivations with a low-resource language, gives us an advantage in learning the low-resource language. Given a low-resource language vector, we use URIEL to find its closest neighbors based on phylogenetic distance and choosing the closest neighbor which is high-resource. We define high-resource languages by their presence in standard large-scale multilingual speech datasets for speech translation, such as CoVoST 2 [12] which is detailed in Subchapter 3.2.1. This high-resource language should have sufficient data for the model to perform reasonably well on it.

Following the above process, we select the low-resource language of Irish and the high-resource language of Welsh for our experiments, which have a phylogenetic distance of 0.375 (where distances range from 0 to 1, with 1 being the most distant) between them. The 2022 Ireland census revealed that 1.8 million people were able to speak Irish; however, only about 72,000 people spoke Irish daily outside of the education system [7]. On the other hand, in the Wales census in 2021 about 538,000 residents spoke Welsh, but over half of them spoke it daily [31, 32], meaning that there were noticeably more daily active community speakers for Welsh than Irish. Irish and Welsh both belong to the Insular Celtic branch of the Indo-European language family [23]. However, Irish is on the Goidelic sub-branch while Welsh is on the Brittonic sub-branch. These sub-branches are primarily differentiated by the evolution of the *k^{w} sound, as Goidelic shifted it to /k/ while Brittonic shifted it to /p/.

3.2 Datasets

3.2.1 High-Resource Dataset

We use CoVoST 2 [12] as the high-resource language dataset. This dataset is an industry standard today as the current largest open-source dataset of its type, and contains paired data for 21 languages' audio and their corresponding English text translations. The audio originates from Mozilla Common Voice 4.0 [2] and the translations were determined through the work of human professional translators passed through automated sanity checks. Of this dataset, we only use the data from the selected high-resource language of Welsh. The Welsh portion of this dataset consists of 2,624 audio clips and their corresponding English text translations with a 1,242/691/691 train/validation/test split, totaling 3 hours worth of data.

3.2.2 Low-Resource Dataset

As mentioned in Chapter 3.1, we use the Irish to English dataset from the IWSLT 2026 shared task for low-resource speech translation [24]. The original dataset consists of around 13 audio hours from 9,320 audio clips translated into English text, with a 7,478/1,120/722 train/validation/test split. Each audio clip corresponds to one spoken sentence. The audio was derived from the news domain, books, Mozilla Common Voice 4.0 [2], and the Living-Audio-Dataset[1]. Details of how the English translations were determined were not provided.

In addition to the original dataset, IWSLT also provides access to 196 hours of synthetic data for training, which is given as 86,324 Irish audio clips and their corresponding English translations. This dataset was created by taking text datasets from OPUS [27]. These OPUS datasets are EUbookshop, which is a body of publications from from various European institutions, Tatoeba, which is a crowdsourced database of sentences and translations, and Wikimedia, which is a collection of sentences from Wikipedia. Moslem [14] then synthesized the audio from these text datasets using Azure Speech.

---

[1]https://github.com/Idlak/Living-Audio-Dataset

Putting both the naturalistic and synthetic datasets together, the Irish to English dataset consists of approximately 209 hours of audio speech data and the corresponding English text translations. We add the synthetic data to the training partition of the combined dataset, yielding a combined 93,802/1,120/722 train/validation/test split. We run two sets of experiments, each with different low-resource datasets: one set purely using the naturalistic dataset and one set utilizing the combined naturalistic and synthetic dataset.

3.3 Model

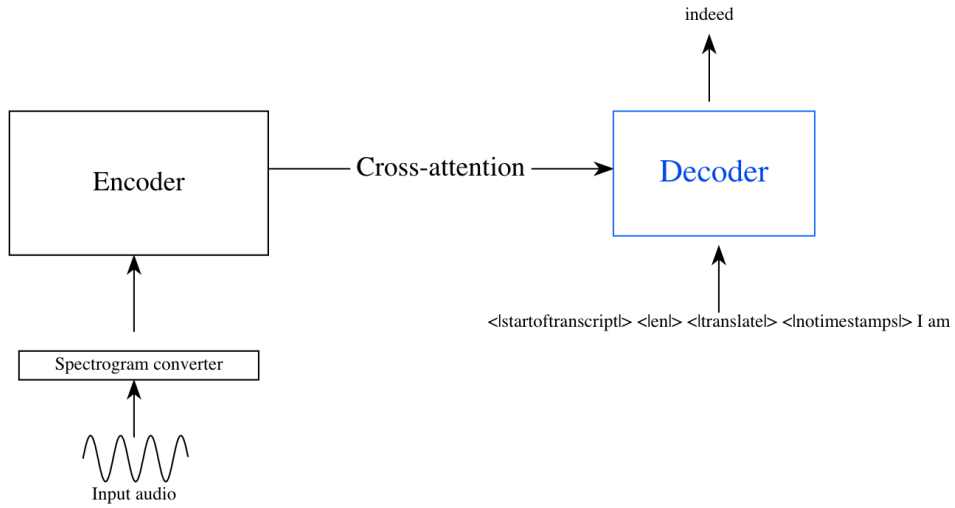

Figure 3.1: High-level Whisper model architecture. Components shown in blue are frozen during fine-tuning, excluding layer normalization.

We use the standard pre-trained transformer-based encoder-decoder multilingual Whisper model [18], as it is an industry standard baseline today. Whisper has been trained on 680,000 hours of multilingual audio from the internet and can perform transcription, translation, language identification, and voice activity detection. Out of these 680,000 hours, 125,739 (18.5%) are data for translation from various source language audios to English, with the top five source languages being Korean, Chinese, Japanese, Welsh, and Russian. Specifically, the

Welsh data accounts for 8,263 of these hours, making up approximately 6.6% of the speech translation dataset. We opt to use the Whisper-small model which includes 244 million parameters, a size which is large enough to learn meaningful acoustic features but yet small enough to fine-tune on limited compute resources.

To perform speech translation to the English language with Whisper, the task token must be set to "translate", which is shown in Figure 3.1. Because Whisper does not natively support the Irish language, the source language token is set as "en" as a placeholder across all experiment runs. Setting this token as "en" means that the decoder is conditioned on English language priors, which allows us to evaluate how well the model is able to map Irish acoustic features through the decoder in a zero-shot manner. We choose to use English as the language token, as English dominates the pre-training dataset and thus its decoder embeddings are the richest and most robustly trained.

As this model is already pre-trained, we utilize transfer learning across all of our experiments by using the model out-of-the-box and fine-tuning it. We fine-tune using one GPU per experiment. The GPUs used vary among NVIDIA GeForce RTX 2080 Ti, L40, L40S, A100, and T4 depending on availability. The fine-tuning is done through various data scheduling configurations, which are detailed in Chapter 3.4.2.

3.3.1 Frozen Decoder

As Whisper is a pre-trained model, it has already learned a general representation of the patterns in acoustic and textual language data that it has been trained on, as represented by its learned model parameter values. Naively fine-tuning the pre-trained model will cause the parameter values to be overwritten and potentially lead to catastrophic forgetting, undermining the benefits of using a pre-trained model in the first place.

Stickland [9] experimented with several configurations of freezing various parts of an encoder-decoder model for machine translation and found that specifically for language pairs that have an English target language, the highest performance is achieved when encoder-decoder attention and layer normalization are unfrozen and the rest of the decoder is frozen.

We apply these findings to the model configuration in our experiments, namely keeping the entire encoder unfrozen, the encoder-decoder attention unfrozen, and the decoder layer normalization unfrozen while freezing the rest of the decoder parameters. This is shown visually in Figure 3.1, where components in blue are frozen during fine-tuning, excluding layer normalization.

3.3.2 Hyperparameters

Parameter	Value
Learning rate	$5 * 10^{-5}$
Maximum generated length	225
Label smoothing factor	0.1
Weight decay	0.0
Number of beams	1
Number of warmup steps	1000
Number of high-resource language fine-tuning steps	5000

Table 3.1: Common experiment model hyperparameters.

We use the hyperparameters as defined in Table 3.1 for all experiments, many of which are standard across similar Whisper fine-tuning experiments.

For experiments purely using the naturalistic dataset as the low-resource dataset, we use the step hyperparameters as defined in Table 3.2. The meanings of the step hyperparameters specific to curriculum schedules A and B are detailed in Chapter 3.4.2. As the introduction of the synthetic data increases the amount of low-resource training data by more than 16 times, we increase the value of the step hyperparameters accordingly, which is reflected in Table 3.3.

Step type	Value
Maximum steps	5000
Curriculum schedule A warmup steps	1500
Curriculum schedule A annealing steps	2000
Curriculum schedule B target steps	500

Table 3.2: Naturalistic experiment model step hyperparameters.

Step type	Value
Maximum steps	15000
Curriculum schedule A warmup steps	2000
Curriculum schedule A annealing steps	5500
Curriculum schedule B target steps	1500

Table 3.3: Synthetic data augmentation experiment model step hyperparameters.

3.4 Experiments

3.4.1 Baseline Experiments

To have an accurate baseline for performance comparison of the effects of the introduction of the high-resource and low-resource data in the fine-tuning of the model, we conduct a series of baseline experiments. We first keep all of the Whisper model unfrozen, allowing all model parameters to be changed and report its results on all data scheduling configurations as detailed in Chapter 3.4.2. Having this baseline allows us to empirically determine the usefulness of freezing these model parameters in the context of this particular low-resource speech translation task.

In addition, we also generate the predictions of the Whisper model straight out-of-the-box with no fine-tuning, and report the evaluation metrics on these predictions. This allows us to clearly see if any of the fine-tuning configurations are degrading or improving performance compared to what Whisper already natively provides.

3.4.2 Data Scheduling Experiments

Following curriculum learning, we consider data from the high-resource language of Welsh to be the more "simple" data, as the pre-trained model has already learned the language's representations and thus should not have much difficulty with data from that language. As the training process progresses, we introduce the more "complex" data from the low-resource language and remove the more "simple" data from the high-resource language through a data scheduler $\lambda(t)$, where t is the training time step and $\lambda(t)$ is the proportion of low-resource language data in the training batch at t. It follows that $1 - \lambda(t)$ is the proportion of high-resource language data in the training batch at t. In these experiments, most of the decoder parameters are frozen throughout the fine-tuning, as detailed in Chapter 3.3.1.

We fine-tune several versions of the model with various definitions of $\lambda(t)$ in order to properly determine the effect of curriculum learning on the model in these settings. Note that because our low-resource language dataset contains more samples than our high-resource dataset, we implement a wrap-around strategy, as necessary, when constructing the batches including the smaller high-resource dataset so that all of the larger low-resource dataset is still seen while preserving the desired ratio as determined by the value of $\lambda(t)$.

First, we fine-tune the model purely on the low-resource language data, and thus the value of t does not affect the value of $\lambda(t)$:

$$\lambda(t) = 1$$

Second, we emulate transfer learning during the fine-tuning stage by training the model first on the high-resource language data and waiting until convergence, then training on the low-resource language data, where T_{switch} is the time step at which the model converges on the high-resource language data. The value of T_{switch} is empirically determined to be 5000, as shown in Table 3.1. The performance trajectory over training for the purely high-resource language training portion is captured in Figures 4.1 and 4.2. This data schedule is

represented by the following:

$$\lambda(t) = \begin{cases} 0, & \text{if } t < T_{switch} \\ \\ 1, & \text{if } t \geq T_{switch} \end{cases}$$

Third, we fine-tune the model on a static, constant 50%/50% split between the low-resource language data and the high-resource language data. In this case, the value of t still does not affect the value of $\lambda(t)$:

$$\lambda(t) = 0.5$$

Fourth, we begin using curriculum learning by increasing $\lambda(t)$ as a linear function with respect to t within three phases, where T_{warmup} is the number of curriculum warmup steps and $T_{annealing}$ is the number of curriculum annealing steps. We refer to this as curriculum schedule A:

$$\lambda(t) = \begin{cases} 0.5, & \text{if } t < T_{warmup} \\ \\ 0.5 + 0.5(\frac{t - T_{warmup}}{T_{annealing}}), & \text{if } T_{warmup} \leq t < T_{warmup} + T_{annealing} \\ \\ 1, & \text{if } t \geq T_{warmup} + T_{annealing} \end{cases}$$

Finally, we utilize curriculum learning with $\lambda(t)$ as an exponential function which starts heavily biased towards the high-resource language data and exponentially decays its high-resource language data proportion in the training batch in favor of low-resource language data. In the equation below, α is a dynamically-defined hyperparameter and T_{target} is the step at which only 5% of the data in a training batch should be from the high-resource language. We utilize a high-resource language proportion threshold of 0.02, meaning that once the

calculated proportion of high-resource language is below 2%, we override that proportion to 0% to free the model's capacity. We refer to this as curriculum schedule B:

$$\alpha = \frac{ln(10)}{T_{target}}$$

$$\lambda(t) = \begin{cases} 1 - 0.5e^{-\alpha t}, & \text{if } 0.5e^{-\alpha t} \geq 0.02 \\ 1, & \text{if } 0.5e^{-\alpha t} < 0.02 \end{cases}$$

We perform experiments using each of the above data scheduling configurations, once with a completely unfrozen model using the purely naturalistic low-resource language data (the baseline experiments), once with a frozen decoder model using the purely naturalistic low-resource language data (the naturalistic-only experiments), and once with a frozen decoder model using the combined naturalistic and synthetic low-resource language data (the synthetic data augmentation experiments).

3.5 Evaluation

We use the COMET [20], BLEU [29], and chrF++ [17] metrics to perform evaluation on the output English text, which are the same metrics used to evaluate accuracy of the submission results for the IWSLT 2026 shared task. COMET is a score between 0 and 1 which measures semantic accuracy, where closeness to 1 is correlated with high quality output aligned with human judgment. To compute the COMET score, we run inference with our prediction and reference outputs using the default wmt22-comet-da model [19]. This is a reference-based regression model built on top of XLM-RoBERTa [8] and trained on human evaluation direct assessments from the 2017-2020 instances of the Conference on Machine Translation (WMT). The chrF++ metric ranges from 0 to 100 and captures precision by calculated character-level n-gram overlaps and adding word-level unigrams and bigrams. Lastly, BLEU ranges from 0 and 100 and performs quality evaluation by counting the number of matching n-grams between the output and reference texts. Of these metrics, COMET is the highest correlated

with human judgment, as it accounts for semantic similarity cases such as synonyms and paraphrasing, while BLEU and chrF++ measure more surface-level exact match. All three metrics are standard to report in this field today, yielding a comprehensive picture of both deep and surface-level translation quality.

We report the final values post-finetuning of these evaluation metrics on the low-resource Irish language validation set for all model versions trained. For BLEU and chrF++ values for the low-resource language, we also compute the non-parametric bootstrap 95% confidence interval with 1000 bootstrap resamples over the 1120 validation samples. We do not do the same for COMET due to its high latency, as its values are computed from model inference. Furthermore, we present the evaluation results on the official test set for the IWSLT 2026 shared task and compare it to the Irish-English results that were officially submitted to IWSLT 2026. Note that the official IWSLT 2026 scoring process for Irish did not include the COMET metric, and thus we only report BLEU and chrF++ values on the test dataset for comparison.

For reproducibility and reference purposes, the model fine-tuning and evaluation code is available on GitHub and the trained model weights are available on HuggingFace, as referenced in Appendix A.

Chapter 4

RESULTS

We follow the methodology described in Chapter 3 and obtain both quantitative and qualitative results from our experiments. We present them here for comparison and analysis.

4.1 Quantitative Results

We report the evaluation results on the validation dataset from the baseline experiments in Table 4.1, frozen decoder naturalistic-only experiments in Table 4.2, and frozen decoder synthetic data augmentation experiments in Table 4.3. In these tables, the best-performing BLEU, chrF++, and COMET scores across all data schedules for each set of experiments are marked in bold.

The trajectories of chrF++ and COMET values on the validation dataset over the training process for each respective experiment set are shown in Figures 4.1, 4.2, 4.3, 4.4, 4.5, 4.6, 4.7, and 4.8. Along with the evaluation metric trajectories, we also show the start and end of the linear phase of curriculum schedule A through the dashed vertical lines, giving us an idea of what the performance looks like with respect to the data schedule. We do not indicate the target 5% high-resource language step boundary corresponding to curriculum schedule B in the graphs, as the first step interval checkpoint at which we evaluate the model (step 1250 for the naturalistic-only experiments and step 2500 for the synthetic data augmentation experiments) is already past this target step. In the tables and graphs, "LR" refers to low-resource language data and "HR" refers to high-resource language data.

It is worthwhile to note that compared to the overall range of the evaluation metrics, the scores across all configurations are relatively low compared to high-resource language speech translation benchmarks today. This highlights the difficulty of this particular low-

resource language task. We report the 95% confidence interval computed through bootstrap resampling for the BLEU and chrF++ metrics in order to derive a measure of statistical significance despite low overall scores. This is shown in the square brackets next to their respective values.

Our evaluation results on the IWSLT 2026 test dataset are in the same range as the official IWSLT 2026 submissions for Irish to English speech translation, as shown in Table 4.4 where the values are pulled from the official IWSLT 2026 campaign findings [1]. These submissions include MTU Primary [25], a cascaded system comprised of a Wav2Vec2 CTC model [3] trained specifically for Irish ASR and a NLLB-200 model for direct translation [10], and MTU Contrastive 1 [25], which uses the same setup as the primary but employs a pivot-based translation technique for NLLB-200 where the transcribed Irish is first translated to Hindi, then to English. The details of the SLC submission were not made available.

Notably, all of the official IWSLT submissions were under the unconstrained condition, meaning that the systems could be trained with any resource and were not just constrained to the datasets provided by the conference organizers. This matches with our system, as we made use of the Whisper model which was already pre-trained with other datasets and also utilized the Welsh dataset in addition to the conference's shared task Irish dataset. As can be seen from Table 4.4, the best performing configurations we present have an increase of 0.8 points in BLEU and 9.0 points in chrF++ over the best performing official submission to IWSLT 2026.

Data schedule	BLEU	chrF++	COMET
Not fine-tuned	0.595 [0.179, 0.692]	12.189 [11.823, 12.814]	0.374
Fine-tuned on 100% LR	0.715 [0.426, 0.940]	12.630 [12.330, 12.940]	**0.419**
Fine-tuned on 100% HR and then 100% LR	0.674 [0.404, 0.913]	12.567 [12.256, 12.871]	0.417
Fine-tuned on static 50% LR and 50% HR	**1.157** [0.702, 1.516]	**13.629** [13.154, 13.935]	**0.419**
Fine-tuned on curriculum learning schedule A	0.846 [0.530, 1.095]	13.068 [12.713, 13.390]	**0.419**
Fine-tuned on curriculum learning schedule B	1.131 [0.771, 1.434]	13.464 [13.131, 13.795]	0.416

Table 4.1: Baseline experiment results. 95% confidence interval for BLEU and chrF++ shown in the square brackets next to their respective values.

Data schedule	BLEU	chrF++	COMET
Fine-tuned on 100% LR	0.515 [0.225, 0.875]	12.666 [12.034, 13.020]	0.374
Fine-tuned on 100% HR and then 100% LR	2.625 [1.931, 3.008]	17.613 [16.671, 18.104]	0.442
Fine-tuned on static 50% LR and 50% HR	2.899 [2.115, 3.271]	18.603 [17.493, 19.037]	0.443
Fine-tuned on curriculum learning schedule A	2.980 [2.351, 3.403]	18.496 [17.828, 18.982]	**0.444**
Fine-tuned on curriculum learning schedule B	**3.134** [2.339, 3.630]	**18.653** [17.674, 19.124]	0.443

Table 4.2: Naturalistic-only experiment results. 95% confidence interval for BLEU and chrF++ shown in the square brackets next to their respective values.

Data schedule	BLEU	chrF++	COMET
Fine-tuned on 100% LR	**9.120** [7.250, 9.971]	**29.553** [27.774, 30.403]	**0.538**
Fine-tuned on 100% HR and then 100% LR	8.631 [6.837, 9.421]	28.986 [27.166, 29.707]	0.537
Fine-tuned on static 50% LR and 50% HR	8.716 [6.966, 9.475]	29.522 [28.059, 30.308]	**0.538**
Fine-tuned on curriculum learning schedule A	7.432 [6.320, 8.351]	27.426 [26.513, 28.309]	0.518
Fine-tuned on curriculum learning schedule B	6.833 [5.144, 7.522]	27.419 [26.182, 28.140]	0.518

Table 4.3: Synthetic data augmentation experiment results. 95% confidence interval for BLEU and chrF++ shown in the square brackets next to their respective values.

Data schedule	Dataset	BLEU	chrF++
Not fine-tuned	-	0.11	12.0
Fine-tuned on 100% LR	Naturalistic	0.20	12.0
Fine-tuned on 100% LR	Combined	**3.20**	**25.0**
Fine-tuned on 100% HR and then 100% LR	Naturalistic	1.00	18.0
Fine-tuned on 100% HR and then 100% LR	Combined	3.10	**25.0**
Fine-tuned on static 50% LR and 50% HR	Naturalistic	0.90	19.0
Fine-tuned on static 50% LR and 50% HR	Combined	**3.20**	**25.0**
Fine-tuned on curriculum learning schedule A	Naturalistic	0.80	18.0
Fine-tuned on curriculum learning schedule A	Combined	2.60	24.0
Fine-tuned on curriculum learning schedule B	Naturalistic	1.20	19.0
Fine-tuned on curriculum learning schedule B	Combined	2.50	24.0
IWSLT 2026: MTU Contrastive 1	-	2.3	15.0
IWSLT 2026: SLC Primary	-	2.4	10.0
IWSLT 2026: MTU Primary	-	2.4	16.0

Table 4.4: Test results on low-resource language compared to official IWSLT 2026 submissions.

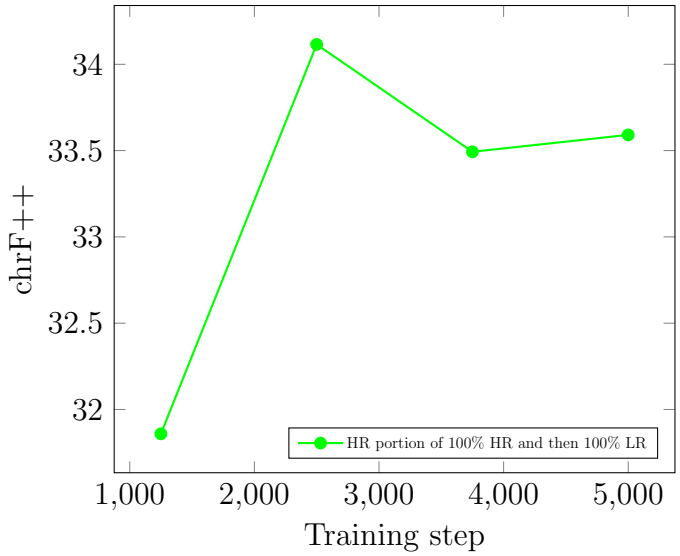

Figure 4.1: chrF++ metric on high-resource evaluation dataset over high-resource training portion only of 100% HR and then 100% LR data schedule.

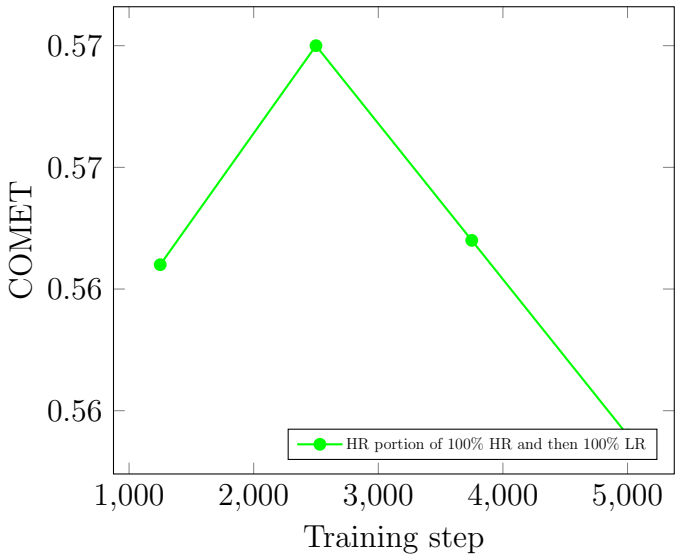

Figure 4.2: COMET metric on high-resource evaluation dataset over high-resource training portion only of 100% HR and then 100% LR data schedule.

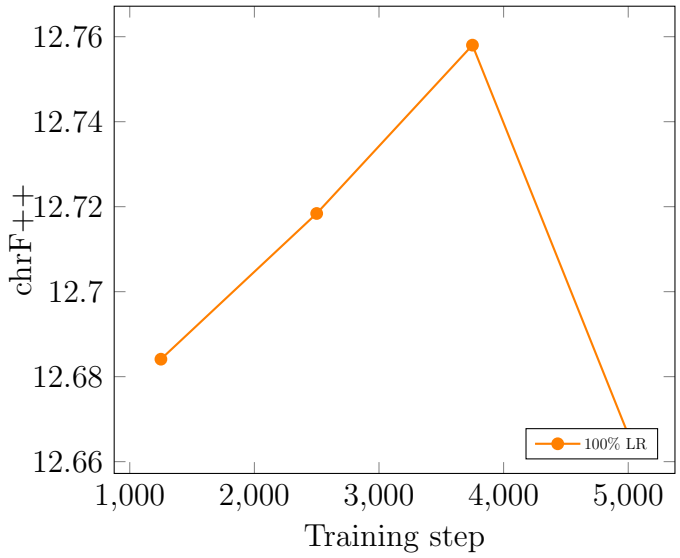

Figure 4.3: chrF++ metric on low-resource evaluation dataset over training for 100% LR naturalistic-only experiment.

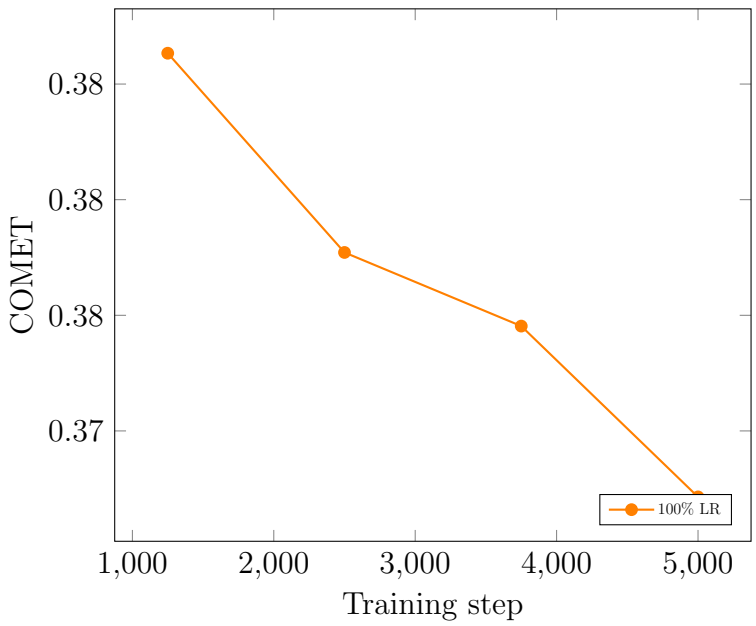

Figure 4.4: COMET metric on low-resource evaluation dataset over training for 100% LR naturalistic-only experiment.

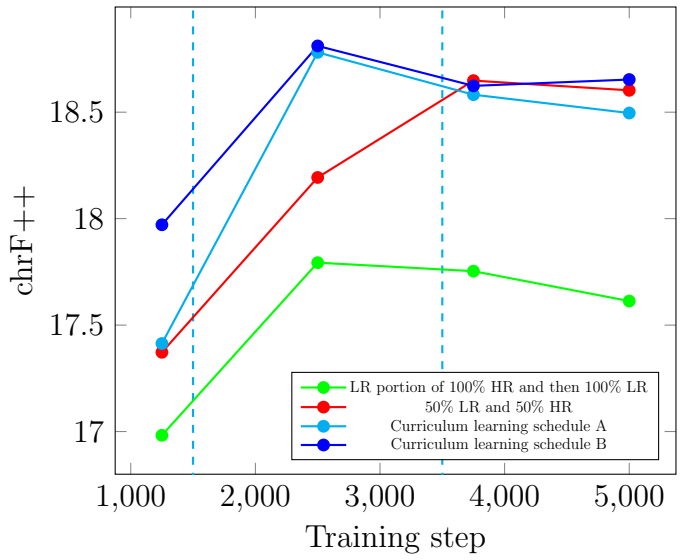

Figure 4.5: chrF++ metric on low-resource evaluation dataset over training for mixed language naturalistic-only experiments. Start and end of linear phase of curriculum schedule A indicated by the dashed vertical lines. Only the LR portion of the 100% HR and then 100% LR data schedule is depicted.

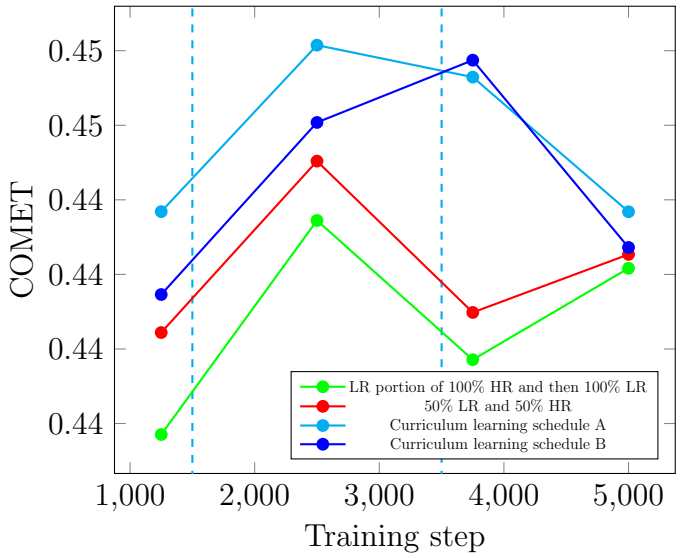

Figure 4.6: COMET metric on low-resource evaluation dataset over training for mixed language naturalistic-only experiments. Start and end of linear phase of curriculum schedule A indicated by the dashed vertical lines. Only the LR portion of the 100% HR and then 100% LR data schedule is depicted.

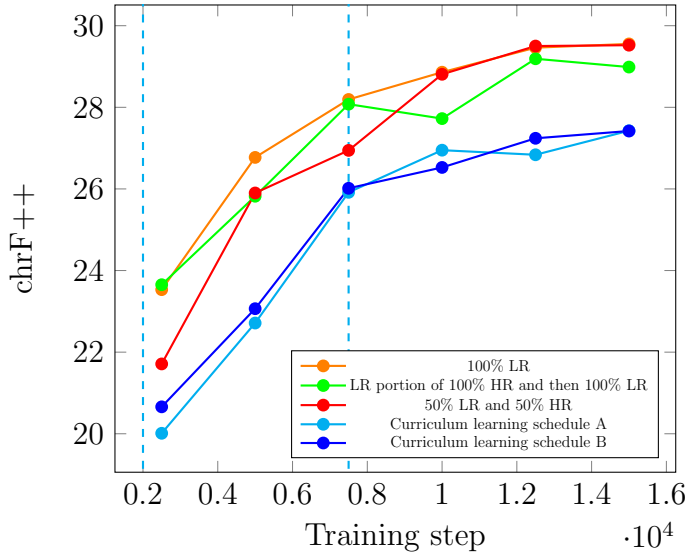

Figure 4.7: chrF++ metric on low-resource evaluation dataset over training for synthetic data augmentation experiments. Start and end of linear phase of curriculum schedule A indicated by the dashed vertical lines. Only the LR portion of the 100% HR and then 100% LR data schedule is depicted.

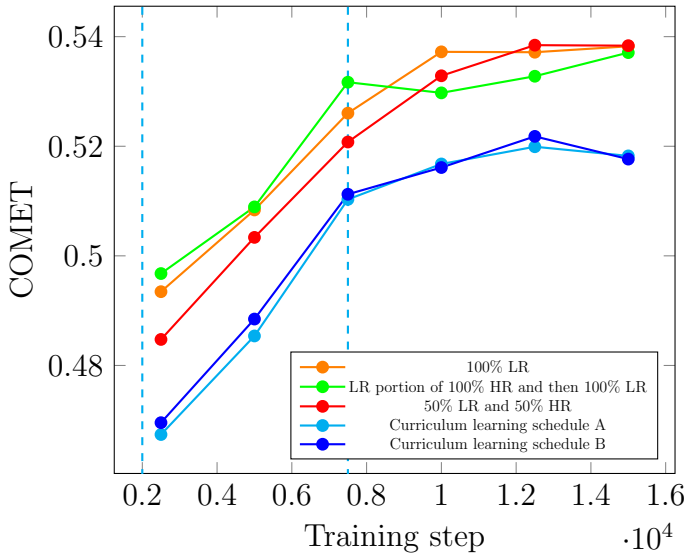

Figure 4.8: COMET metric on low-resource evaluation dataset over training for synthetic data augmentation experiments. Start and end of linear phase of curriculum schedule A indicated by the dashed vertical lines. Only the LR portion of the 100% HR and then 100% LR data schedule is depicted.

REF: I don't want to settle down.

PRED: Nimin nam sa cruxies.

REF: In a report, written by inspectors from the Department

PRED: Idoris, a screve kickery, own Rhine.

REF: Michael D. Higgins has always been independent.

PRED: The Michael D Higgins Reeve Navs Block

Figure 4.9: Sample validation sentence prediction outputs and their references from experiment with no fine-tuning.

REF: I don't want to settle down.

PRED: I don't think she is.

REF: In a report, written by inspectors from the Department

PRED: The door is for writing.

REF: Michael D. Higgins has always been independent.

PRED: Máire had a pig on the nose block.

Figure 4.10: Sample validation sentence prediction outputs and their references from naturalistic-only fine-tuning on curriculum learning schedule B experiment.

REF: I don't want to settle down.

PRED: I don't want to sit down

REF: In a report, written by inspectors from the Department

PRED: In a document written by inspectors from the Department.

REF: Michael D. Higgins has always been independent.

PRED: Michael D'Oy Higgins was never dependent.

Figure 4.11: Sample validation sentence prediction outputs and their references from synthetic data augmentation fine-tuning on 100% LR experiment.

4.2 Qualitative Results

We show sample low-resource language outputs and their corresponding references from the validation dataset for the non-fine-tuned model in Figure 4.9, the best-performing naturalistic-only data schedule configuration in Figure 4.10, and the best-performing synthetic data augmentation data schedule configuration in Figure 4.11. Together, these examples represent the general range of worst to best performing outputs over all experiments run.

Listening to the source audio and comparing it to the reference output, we find that the results from the non-fine-tuned model in Figure 4.9 are simply direct transcriptions of the audio which have not been translated. Given that this is the first time that the model is being exposed to Irish audio, these results follow as the model has no reference of how to perform this particular speech translation task.

Conversely, Figure 4.10 and Figure 4.11 show cohesive English phrases and sentences. Some naturalistic-only results in Figure 4.10 show poor semantic knowledge, for example, "The door is for writing", but the prediction outputs do show some syntactic and word-level

similarities with that of the references. The synthetic data-augmented prediction results are much stronger semantically and overall closer to the references; however, details like proper noun translation and specific word choice (for example, “has always been independent” versus “was never dependent”) make it extremely rare for an exact match between prediction and reference to occur.

Chapter 5

DISCUSSION

From the baseline experiment results in Table 4.1, we find that baseline performance, where the model is completely unfrozen, is extremely low overall. Fine-tuning does appear to relatively improve the baseline performance across the BLEU, chrF++, and COMET metrics, with the fine-tuned static 50%/50% split between the low-resource and high-resource languages achieving the highest scores in all three metrics amongst the different baseline configurations.

5.1 Decoder Freezing Leads to Higher Performance for Mixed Language Fine-Tuning

After implementing decoder freezing, low-resource language performance on BLEU, chrF++, and COMET improves noticeably for the models fine-tuned on 100% high-resource and then 100% low-resource, 50%/50% static fine-tuning, and curriculum learning, as seen by comparing Table 4.2 with Table 4.1 ($p < 10^{-3}$ with $N = 1000$ bootstrap resamples for both BLEU and chrF++; +0.025 in COMET between the unfrozen 100% HR and then 100% LR schedule and its frozen decoder counterpart). These configurations all have data schedules which mix low-resource with high-resource language data. They achieve similar evaluation metric values among each other, with the highest BLEU and chrF++ performance coming from curriculum learning schedule B and the highest COMET performance coming from curriculum schedule A.

This points to the need for strategic use of model parameter freezing. When the encoder is fine-tuned on both Irish and Welsh data, it learns a general, non-overfitted representation of the input audio which then allows it to utilize the decoder representations which were

already acquired during the extensive pre-training phase. On the other hand, when the decoder is not frozen, it risks catastrophic forgetting.

5.2 Decoder Freezing Performance on Pure Low-Resource Language Fine-Tuning Varies Based on Presence of Synthetic Data

Interestingly, looking at Table 4.2, the frozen decoder model fine-tuned only on 100% naturalistic low-resource language data does not follow the same trend of improved performance as its mixed data counterparts, instead showing slightly decreased performance in BLEU and COMET scores compared to the baseline unfrozen model fine-tuned on only 100% naturalistic low-resource. Figures 4.3 and 4.4 show that for this particular experiment configuration, validation performance is not only relatively low but actually ends up worse after fine-tuning. We see that chrF++ increases slightly during the first three quarters of the fine-tuning process but then decreases at step 5,000, while COMET consistently decreases over fine-tuning.

From this, we learn that in the naturalistic-only low-resource language data context, decoder freezing helps improve performance only when the model is fine-tuned on a mixture of high-resource and low-resource data. On the other hand, referencing Table 4.3, when the synthetic low-resource language data is included, the frozen decoder model fine-tuned on 100% low-resource language data performs significantly better than the unfrozen 100% low-resource language data baseline in Table 4.1 ($p < 10^{-3}$ with $N = 1000$ bootstrap resamples for both BLEU and chrF++; +0.119 in COMET), even outperforming the other fine-tuning configurations in Table 4.3 where the high-resource language data is included.

This divergence in results between the naturalistic-only and synthetic data augmentation experiments could mean that in scenarios where low-resource data is extremely limited, such as in the 13 hour naturalistic dataset case, the unfrozen encoder overfits to the very specific distribution of the 13 hours of data, which is then bottlenecked by the frozen decoder which expects an encoder output in its pre-trained representation space. When the 209 hour low-resource language dataset is used instead, the encoder learns a more general distribution as it is exposed to more variety, and thus is more compatible with the frozen decoder pre-trained

target language representation. The model is then able to leverage the priors that it has learned during pre-training to a far greater extent.

5.3 Synthetic Data Augmentation Facilitates Better Overall Semantic Learning

Excluding the outlier case mentioned above for the frozen decoder model fine-tuned only on 100% naturalistic low-resource language data, looking at the chrF++ metric on the validation data over the fine-tuning process in Figures 4.5 and 4.7, we see a general steady increase and plateau for both the naturalistic-only and synthetic data augmented experiments. This shows that the model is learning the surface-level exact match well over the fine-tuning process.

With regards to the COMET metric, across all data schedules for naturalistic-only experiments we see in Figure 4.6 that the COMET score starts low at training step 1250, peaks in the middle between training steps 2500 and 3750, and arrives at a value below the peak at training step 5000. This is directly in contrast with the co-occurring chrF++ training trajectory in Figure 4.5, which shows a steady increase and plateau at the highest point as the training finishes. As the Whisper model utilizes cross-entropy loss, which emphasizes exact token match, to facilitate training, it concurs that as loss is minimized throughout training, the chrF++ score will improve and thus increase. However, cross-entropy loss does not include semantic similarity and thus does not have a direct correlation with the COMET score.

On the other hand, as seen in Figure 4.8, all data schedules for the synthetic data augmentation experiments show a steady increase in COMET score followed by a plateau as training converges. Thus, the general trajectory of the COMET metric over fine-tuning for naturalistic-only experiments in Figure 4.6 differs from that for synthetic data augmented experiments in Figure 4.8.

While the model appears to be improving on chrF++ throughout fine-tuning in both the naturalistic-only and synthetic data augmented experiments, the model only improves

on COMET throughout fine-tuning when the training dataset is augmented with synthetic data. This shows that synthetic data augmentation aids with overall comprehensive learning, contributing to performance improvement over training at both the surface level and at the semantic level.

5.4 Mixed Language Data Scheduling is Useful Only in the Naturalistic-Only Low-Resource Language Dataset Scenario

The addition of synthetic data to the low-resource language data improves fine-tuned model performance significantly. Namely, all fine-tuned models show a noticeable increase of at least 3 in BLEU, 8.5 in chrF++, and 0.07 in COMET in Table 4.3 compared to Table 4.2 ($p < 10^{-3}$ with $N = 1000$ bootstrap resamples for both BLEU and chrF++ between the highest scoring naturalistic-only experiment and worst scoring synthetic data augmentation experiment). Fine-tuning on only 100% combined naturalistic and synthetic low-resource language data results in the highest performance in BLEU, chrF++, and COMET out of all fine-tuning data scheduling configurations.

Looking at curriculum schedule A's performance trajectory with respect to its linear phase represented by the dashed vertical lines, we see in Figures 4.7 and 4.8 that with synthetic data augmentation, chrF++ and COMET both increase steadily over the linear phase. However, after the linear phase only the low-resource language data remains, and the metric no longer increases at the same rate but rather plateaus while other the performance from other data schedules continue to climb. Interestingly, curriculum schedule B appears to follow a similar performance trajectory pattern as curriculum schedule A despite not following the same data schedule.

On the other hand, in the naturalistic-only case, the mixed language data schedules achieve the best performance, as seen in Table 4.2 ($p < 10^{-3}$ with $N = 1000$ bootstrap resamples for both BLEU and chrF++; +0.069 in COMET between the 100% low-resource schedule and curriculum learning schedule B). These findings indicate that phylogenetic mixed language data scheduling is beneficial primarily when the low-resource language data

is limited to the 13 hour naturalistic Irish dataset case. This suggests that phylogenetic mixed language data schedules can lead to performance improvements, but only when the primary dataset itself lacks the scale and diversity to form a generalizable, accurate representation of the task.

Chapter 6

CONCLUSION AND FUTURE WORK

6.1 Summary and Conclusions

In this study, we explored methods to improve low-resource language speech translation. Specifically, given a low-resource language, we found the most phylogentically similar high-resource language to it and incorporated its data with that of the low-resource language. Using the encoder-decoder based Whisper model, data was incorporated during model fine-tuning in various different ways: fine-tuning transfer learning where a pre-trained model was first fine-tuned on all of the high-resource language data and then fine-tuned on all of the low-resource language data, a static 50%/50% split between high-resource and low-resource language data throughout the fine-tuning process, a two curriculum learning schedules. These curriculum learning schedules consisted of one where high-resource data was linearly decreased over time and replaced with low-resource language data, and another where high-resource language data was exponentially decreased over time and replaced with low-resource language data. In addition, we investigated the effects of freezing the model decoder during fine-tuning under these different data scheduling configurations. We ran further experiments where the low-resource language data utilized varied from a purely naturalistic dataset to a combined naturalistic and synthetic dataset.

We find that decoder freezing, namely, freezing all decoder parameters besides encoder-decoder attention and layer normalization, leads to higher performance than the baseline when both low-resource and high-resource language data is utilized during model fine-tuning. When only low-resource language data is used, we see that decoder freezing only leads to improvements when the low-resource language dataset is augmented with synthetic data. We conclude that decoder freezing is a highly beneficial strategy when the input data passed

to the encoder has a generalizable representation, whether through a mixture of different languages or through an extensive dataset of one language, which in turn allows the learned priors of the decoder to be leveraged most effectively. Conversely, decoder freezing does not prove to be effective when the input data is limited and specific, as this leads to encoder overfitting that misaligns representations with the pre-trained priors of the frozen decoder.

We also find that synthetic data augmentation is valuable in facilitating proper semantic learning during the fine-tuning process. Without synthetic data, fine-tuning does not lead to steady improvement in semantic accuracy, contrasting with the improvement of surface-level exact match. From this, it is deduced that augmenting limited datasets with synthetic data can introduce the depth and variety needed for a model to form a generalizable, comprehensive representation at both the semantic and surface level.

Lastly, we see that language family data scheduling leads to overall accuracy improvement only in the naturalistic low-resource language dataset scenario. When the low-resource language dataset is augmented with synthetic data, the best performance comes from purely using the low-resource language data. This shows that the primary dataset being utilized has an impact on the effectiveness of data scheduling. We conclude that phylogenetic mixed language data scheduling can be beneficial in contexts where the primary dataset lacks the depth and comprehensiveness needed for the model to learn a generalizable representation. If the primary dataset is already representative, it is better to train purely on itself, as introducing other inputs unecessarily exposes the model to contexts which it will not need to perform in.

6.2 Limitations

We only explore the encoder-decoder model architecture for speech translation in these experiments, as all experiments were performed using a pre-trained Whisper model. Furthermore, due to compute and time constraints we are limited to the Whisper-small model, which only has 244 million parameters, and reported results are from only one fine-tuning training run for each experiment configuration.

In our experiment decoding pipeline, we use set the source language token to "en" as a placeholder due to the lack of native support for Irish. This conditions the decoder to use English acoustic priors, which may have interference or bias towards English phonetic mappings while processing Irish speech.

Our experiments are limited to the Irish/Welsh low-resource/high-resource pair, and translation to the target language of English. We also do not use any speech translation datasets outside of the IWSLT 2026 shared task dataset for Irish and the CoVoST 2 dataset for Welsh. Notably, the CoVoST 2 Welsh dataset only includes 3 hours of data while the IWSLT 2026 Irish dataset includes 209 total hours of data.

6.3 Future Work

Recognizing the limitations of this study with regards to the usage of the English source language token, future work could involve instead using the Welsh source language token "cy" in the decoding process. As Welsh and Irish share greater Celtic phonetic and structural overlap than English and Irish, using Welsh acoustic priors may lead to improved performance when Irish is the source language. Investigating this would directly reveal the tradeoffs in using a structurally and phonetically similar language versus using the robust decoder embeddings of a high-resource anchor language.

It would also be beneficial to extend the experiments to more low-resource/high-resource language family pairs to examine if the same conclusions regarding phylogenetic mixed language data scheduling hold. As can be seen from the drastic change in results after adding the synthetic Irish data, the qualities of a given low-resource language dataset have a big impact on performance. Thus, we recognize that other low-resource languages may perform differently based on the specifics of their available data. Additionally, given that the model we used only performs speech translation to the English language, it would also be useful to extend this work to include other speech translation models which will allow for translating to a variety of other languages.

Furthermore, future work could focus on what specific low-resource language dataset

qualities lead to differences in the impact of phylogenetic mixed language data scheduling. Our experiments surfaced two of these qualities: the size of the dataset and the nature of the data (naturalistic vs synthetic, or a combination of both).

This could be investigated by performing the experiments utilizing mixed language data with naturalistic-only data and then with synthetic-only data, where both dataset sizes are equivalent. Analyzing the results would allow us to determine whether the nature of the data alone has an impact on curriculum learning performance.

To determine dataset size influence, it would be useful to perform size benchmarking experiments to determine any size thresholds at which this style of mixed language data scheduling is relevant. By slowly increasing the amount of low-resource language data present in multiple iterations of mixed language data scheduling experiments, we would be able to verify if there are any distinct patterns or correlations between data amount and performance improvement. Especially as each low-resource language has varying amounts of data usable for the speech translation task, this will help inform future work in the area where decisions of which model training technique to use can be informed by the amount of data present.

BIBLIOGRAPHY

[1] David Ifeoluwa Adelani, Victor Agostinelli, Antonios Anastasopoulos, Luisa Bentivogli, Ondřej Bojar, Sébastien Bratières, Marine Carpuat, Fabrício Carraro, Roldano Cattoni, Mauro Cettolo, Lizhong Chen, Marcello Federico, Marco Gaido, Mahendra Gupta, HyoJung Han, Ali Hatami, Lewis C. Howe, Dávid Javorský, Yejin Jeon, Marek Kasztelnik, Antoine Laurent, Danni Liu, Nam Luu, Min Ma, Dominik Macháček, Marie Maltais, Evgeny Matusov, John McCrae, Chutong Meng, Chandresh Kumar Maurya, Mohammad Mohammadamini, Yasmin Moslem, Kenton Murray, Satoshi Nakamura, Matteo Negri, Jan Niehues, Atul Kr. Ojha, John E. Ortega, Siqi Ouyang, Sara Papi, Peter Polák, Fabian Retkowski, Stephanny Sánchez, Beatrice Savoldi, Claytone Sikasote, Matthias Sperber, Sebastian Stüker, Katsuhito Sudoh, Marie Tahon, Marco Turchi, Alexander Waibel, Patrick Wilken, Rodolfo Joel Zevallos, Vilem Zouhar, and Maike Züfle. Speech translation and metrics in 2026: Findings of the IWSLT campaign. In Elizabeth Salesky, Antonios Anastasopoulos, Matteo Negri, and Marcello Federico, editors, *Proceedings of the 23rd International Conference on Spoken Language Translation (IWSLT 2026)*, pages 336–422, San Diego, USA (in-person and online), July 2026. Association for Computational Linguistics.

[2] Rosana Ardila, Megan Branson, Kelly Davis, Michael Kohler, Josh Meyer, Michael Henretty, Reuben Morais, Lindsay Saunders, Francis Tyers, and Gregor Weber. Common voice: A massively-multilingual speech corpus. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4218–4222, Marseille, France, May 2020. European Language Resources Association.

[3] Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. *CoRR*, abs/2006.11477, 2020.

[4] Loïc Barrault, Yu-An Chung, Mariano Coria Meglioli, David Dale, Ning Dong, Mark Duppenthaler, Paul-Ambroise Duquenne, Brian Ellis, Hady Elsahar, Justin Haaheim, et al. Seamless: Multilingual expressive and streaming speech translation. *arXiv preprint arXiv:2312.05187*, 2023.

[5] Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. Curriculum learning. In *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML '09, page 41–48, New York, NY, USA, 2009. Association for Computing Machinery.

[6] Alexandre Berard, Olivier Pietquin, Christophe Servan, and Laurent Besacier. Listen and translate: A proof of concept for end-to-end speech-to-text translation. *CoRR*, abs/1612.01744, 2016.

[7] Central Statistics Office. Census of population 2022: Profile 8 – The Irish language and education. Technical report, Central Statistics Office, Cork, Ireland, December 2023.

[8] Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised cross-lingual representation learning at scale. *CoRR*, abs/1911.02116, 2019.

[9] Asa Cooper Stickland, Xian Li, and Marjan Ghazvininejad. Recipes for adapting pre-trained monolingual and multilingual models to machine translation. In Paola Merlo, Jorg Tiedemann, and Reut Tsarfaty, editors, *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3440–3453, Online, April 2021. Association for Computational Linguistics.

[10] Marta R. Costa-jussà, Jeff Wang, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, and et al. Scaling neural machine translation to 200 languages. *Nature*, 630(8018):841–846, 2024.

[11] Yexing Du, Youcheng Pan, Ziyang Ma, Bo Yang, Yifan Yang, Keqi Deng, Xie Chen, Yang Xiang, Ming Liu, and Bing Qin. Making LLMs better many-to-many speech-to-text translators with curriculum learning. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar, editors, *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12466–12478, Vienna, Austria, July 2025. Association for Computational Linguistics.

[12] Patrick Littell, David R. Mortensen, Ke Lin, Katherine Kairis, Carlisle Turner, and Lori Levin. URIEL and lang2vec: Representing languages as typological, geographical, and phylogenetic vectors. In Mirella Lapata, Phil Blunsom, and Alexander Koller, editors, *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 8–14, Valencia, Spain, April 2017. Association for Computational Linguistics.

[13] Chutong Meng and Antonios Anastasopoulos. GMU systems for the IWSLT 2025 low-resource speech translation shared task. In Elizabeth Salesky, Marcello Federico, and Antonis Anastasopoulos, editors, *Proceedings of the 22nd International Conference on Spoken Language Translation (IWSLT 2025)*, pages 289–300, Vienna, Austria (in-person and online), July 2025. Association for Computational Linguistics.

[14] Yasmin Moslem. Leveraging synthetic audio data for end-to-end low-resource speech translation. In Elizabeth Salesky, Marcello Federico, and Marine Carpuat, editors, *Proceedings of the 21st International Conference on Spoken Language Translation (IWSLT 2024)*, pages 265–273, Bangkok, Thailand (in-person and online), August 2024. Association for Computational Linguistics.

[15] NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Holger Schwenk, et al. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*, 2022.

[16] Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10):1345–1359, 2010.

[17] Maja Popović. chrF++: words helping character n-grams. In Ondřej Bojar, Christian Buck, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, and Julia Kreutzer, editors, *Proceedings of the Second Conference on Machine Translation*, pages 612–618, Copenhagen, Denmark, September 2017. Association for Computational Linguistics.

[18] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine Mcleavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 28492–28518. PMLR, 23–29 Jul 2023.

[19] Ricardo Rei, José G. C. de Souza, Duarte Alves, Chrysoula Zerva, Ana C Farinha, Taisiya Glushkova, Alon Lavie, Luisa Coheur, and André F. T. Martins. COMET-22: Unbabel-IST 2022 submission for the metrics shared task. In Philipp Koehn, Loïc Barrault, Ondřej Bojar, Fethi Bougares, Rajen Chatterjee, Marta R. Costa-jussà, Christian Federmann, Mark Fishel, Alexander Fraser, Markus Freitag, Yvette Graham, Roman Grundkiewicz, Paco Guzman, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Tom Kocmi, André Martins, Makoto Morishita, Christof Monz, Masaaki Nagata, Toshiaki Nakazawa, Matteo Negri, Aurélie Névéol, Mariana Neves, Martin Popel, Marco Turchi, and Marcos Zampieri, editors, *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 578–585, Abu Dhabi, United Arab Emirates (Hybrid), December 2022. Association for Computational Linguistics.

[20] Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. COMET: A neural framework for MT evaluation. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language*

Processing (EMNLP), pages 2685–2702, Online, November 2020. Association for Computational Linguistics.

[21] Nathaniel Romney Robinson, Kaiser Sun, Cihan Xiao, Niyati Bafna, Weiting Tan, Haoran Xu, Henry Li Xinyuan, Ankur Kejriwal, Sanjeev Khudanpur, Kenton Murray, and Paul McNamee. JHU IWSLT 2024 dialectal and low-resource system description. In Elizabeth Salesky, Marcello Federico, and Marine Carpuat, editors, *Proceedings of the 21st International Conference on Spoken Language Translation (IWSLT 2024)*, pages 140–153, Bangkok, Thailand (in-person and online), August 2024. Association for Computational Linguistics.

[22] Paul K Rubenstein, Chulayuth Asawaroengchai, Duc Dung Nguyen, Ankur Bapna, Zalán Borsos, Félix de Chaumont Quitry, Peter Chen, Dalia El Badawy, Wei Han, Eugene Kharitonov, et al. Audiopalm: A large language model that can speak and listen. *arXiv preprint arXiv:2306.12925*, 2023.

[23] Paul Russell. *An introduction to the Celtic languages*. Routledge, 2014.

[24] Elizabeth Salesky, Antonios Anastasopoulos, Matteo Negri, and Marcello Federico, editors. *Proceedings of the 23rd International Conference on Spoken Language Translation (IWSLT 2026)*, San Diego, USA (in-person and online), July 2026. Association for Computational Linguistics.

[25] Pournima Sonawane and Haithem Afli. ADAPT–MTU HAI at IWSLT2026: Robust cascaded speech translation for Bhojpuri–Hindi and Irish–English. In Elizabeth Salesky, Antonios Anastasopoulos, Matteo Negri, and Marcello Federico, editors, *Proceedings of the 23rd International Conference on Spoken Language Translation (IWSLT 2026)*, pages 58–67, San Diego, USA (in-person and online), July 2026. Association for Computational Linguistics.

[26] Matthias Sperber and Matthias Paulik. Speech translation and the end-to-end promise: Taking stock of where we are. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7409–7421, Online, July 2020. Association for Computational Linguistics.

[27] Jörg Tiedemann. Parallel data, tools and interfaces in OPUS. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Mehmet Uğur Doğan, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2214–2218, Istanbul, Turkey, May 2012. European Language Resources Association (ELRA).

[28] Alex Waibel, Ajay Jain, Arthur McNair, Joe Tebelskis, Louise Osterholtz, Hiroaki Saito, Otto Schmidbauer, Tilo Sloboda, and Monika Woszczyna. Janus: Speech-to-speech translation using connectionist and non-connectionist techniques. *Advances in neural information processing systems*, 4, 1991.

[29] Changhan Wang, Anne Wu, and Juan Miguel Pino. Covost 2: A massively multilingual speech-to-text translation corpus. *CoRR*, abs/2007.10310, 2020.

[30] Chengyi Wang, Yu Wu, Shujie Liu, Ming Zhou, and Zhenglu Yang. Curriculum pre-training for end-to-end speech translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3728–3738, 2020.

[31] Welsh Government and Office for National Statistics. Welsh language in Wales (Census 2021). Statistical Bulletin SB 42/2022, Welsh Government, Cardiff, Wales, December 2022.

[32] Welsh Language Commissioner. Position of the Welsh language: Census 2021 analysis. Technical report, Office of the Welsh Language Commissioner, Cardiff, Wales, December 2022.

[33] Giulio Zhou, Tsz Kin Lam, Alexandra Birch, and Barry Haddow. Prosody in cascade and direct speech-to-text translation: a case study on korean wh-phrases. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 674–683, 2024.

Appendix A
MODEL WEIGHTS AND CODE

All model weights are available at https://huggingface.co/jjnhuang, with the mapping from experiment configuration to HuggingFace model name shown in Table A.1. The "Combined" dataset in the table refers to the dataset consisting of both naturalistic and synthetic data. Code for model fine-tuning and evaluation is available at https://github.com/jhuangjennifer/msthesis-curriculumlearning-lowresource-s2tt (contact the author for access).

Data schedule	Dataset	Frozen decoder?	HuggingFace model name
Fine-tuned on 100% LR	Naturalistic	No	jjnhuang/whisper-small-finetuned-iwslt-irish-no-frozen-decoder
Fine-tuned on 100% LR	Naturalistic	Yes	jjnhuang/whisper-small-finetuned-iwslt-irish
Fine-tuned on 100% LR	Combined	Yes	jjnhuang/whisper-small-finetuned-iwslt-irish-synthetic
Fine-tuned on 100% HR and then 100% LR	Naturalistic	No	jjnhuang/whisper-small-finetuned-iwslt-irish-transfer-welsh-no-frozen-decoder
Fine-tuned on 100% HR and then 100% LR	Naturalistic	Yes	jjnhuang/whisper-small-finetuned-iwslt-irish-transfer-welsh
Fine-tuned on 100% HR and then 100% LR	Combined	Yes	jjnhuang/whisper-small-finetuned-iwslt-irish-transfer-welsh-synthetic
Fine-tuned on static 50% LR and 50% HR	Naturalistic	No	jjnhuang/whisper-small-finetuned-iwslt-irish-welsh-50-50-no-frozen-decoder
Fine-tuned on static 50% LR and 50% HR	Naturalistic	Yes	jjnhuang/whisper-small-finetuned-iwslt-irish-welsh-50-50
Fine-tuned on static 50% LR and 50% HR	Combined	Yes	jjnhuang/whisper-small-finetuned-iwslt-irish-welsh-50-50-synthetic
Fine-tuned on curriculum learning schedule A	Naturalistic	No	jjnhuang/whisper-small-finetuned-iwslt-irish-welsh-curriculum-no-frozen-decoder
Fine-tuned on curriculum learning schedule A	Naturalistic	Yes	jjnhuang/whisper-small-finetuned-iwslt-irish-welsh-curriculum
Fine-tuned on curriculum learning schedule A	Combined	Yes	jjnhuang/whisper-small-finetuned-iwslt-irish-welsh-curriculum-synthetic
Fine-tuned on curriculum learning schedule B	Naturalistic	No	jjnhuang/whisper-small-finetuned-iwslt-irish-welsh-curriculum-b-no-frozen-decoder
Fine-tuned on curriculum learning schedule B	Naturalistic	Yes	jjnhuang/whisper-small-finetuned-iwslt-irish-welsh-curriculum-b
Fine-tuned on curriculum learning schedule B	Combined	Yes	jjnhuang/whisper-small-finetuned-iwslt-irish-welsh-curriculum-synthetic-b

Table A.1: Mapping of experiment configurations to model weights.