WAV2GLOSS: Generating Interlinear Glossed Text from Speech

Taiqi He[1], Kwanghee Choi[1], Lindia Tjuatja[1], Nathaniel R. Robinson[2], Jiatong Shi[1], Shinji Watanabe[1], Graham Neubig[1], David R. Mortensen[1], Lori Levin[1]
[1]Language Technologies Institute, Carnegie Mellon University
[2]Center for Language and Speech Processing, Johns Hopkins University

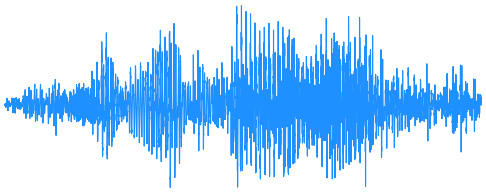
Abstract

Thousands of the world's languages are in danger of extinction—a tremendous threat to cultural identities and human language diversity. Interlinear Glossed Text (IGT) is a form of linguistic annotation that can support documentation and resource creation for these languages' communities. IGT typically consists of (1) transcriptions, (2) morphological segmentation, (3) glosses, and (4) free translations to a majority language. We propose WAV2GLOSS: a task in which these four annotation components are extracted automatically from speech, and introduce the first dataset to this end, FIELDWORK:[1] a corpus of speech with all these annotations, derived from the work of field linguists, covering 37 languages, with standard formatting, and train/dev/test splits. We provide various baselines to lay the groundwork for future research on IGT generation from speech, such as end-to-end versus cascaded, monolingual versus multilingual, and single-task versus multi-task approaches.

1 Introduction

Working against the overwhelming tide of social and historical forces, linguists and community activists from around the world have set out to record endangered languages while they are still actively spoken. As a first step, these documentary efforts involve—quite literally—recordings, then transcriptions, translations, and other annotations. The ultimate goal of such efforts is often to take a large volume of recorded speech and annotate it with Interlinear Glossed Text (IGT).

IGT is the *lingua franca* of documentary linguistics. Most IGT now follows a set of conventions called the Leipzig glossing rules (Comrie et al., 2008) but other formats are in use. An example of IGT from Vydrina (2022) is shown in Figure 1.



wd:	n	sìginde		yan	de
sr:	n	sigi	-nde	yan	de
ur:	n	sigi	-len	yan	le
gl:	1.SG	sit	-PC.RES	that	FOC
tr:	"I live here."				

Figure 1: A representation of a single Kakabe utterance in the FIELDWORK corpus: speech paired with annotations. wd is the unsegmented transcription; ur and sr are the underlying and surface representations; gl is a morpheme-by-morpheme gloss; and tr is a free translation into the metalanguage.

It consists of an unsegmented transcription (wd), underlying (ur) and surface forms (sr) segmented into morphemes, morpheme labels (glosses; gl), and free translation (tr) aligned with one another and the source audio recording. The lines labeled sr, ur, and gl are most important to linguists and language teachers. Together, they tell the user how linguistic form maps into linguistic function, enabling the creation of many valuable resources and a variety of useful analyses.

Most linguistic field recordings, though, never make it to IGT (Seifart et al., 2018). Simply transcribing field data (without other annotations) can take up to one hour per minute of recorded speech (Do et al., 2014). Adding additional annotations is even more expensive. This bottleneck keeps vast collections of field recordings from achieving their full documentary potential.

Producing IGT from audio is a tractable problem. While linguists can do little or nothing to address the underlying causes of language endangerment—which form an intersecting lattice of political, cul-

[1]https://huggingface.co/datasets/wav2gloss/fieldwork

Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 568–582

tural, and economic factors—speech and natural language processing researchers can do even less. Technologists can, however, facilitate the efforts of field linguists and language workers to *document* endangered languages by developing technologies that make the mammoth tasks of annotating field data surmountable (Shi et al., 2021a,b). For example, they can develop models that automatically generate first-pass transcriptions from raw speech. Research has demonstrated that such models can speed up transcription dramatically (Amith et al., 2021).

In direct support of the goal of language documentation, we propose a new speech and language processing task: WAV2GLOSS. This task assumes recorded speech as the only input. The output consists of aligned annotations for transcription (with and without segmentation), glossing, and translation. In order to allow the research community to participate in this task, we introduce the following:

1. The FIELDWORK Corpus, a speech+IGT dataset for 37 languages—drawn from five archives of linguistic field data—with a standard format and train/dev/test splits.
2. Four subtasks crucial for language documentation: prediction of transcription, underlying representation, gloss, and translation. To our knowledge, this is the first attempt to extract these annotations directly from speech.
3. Benchmarks based on well known speech and NLP models, including end-to-end and cascaded approaches to predicting IGT from speech.

2 Dataset

We present the FIELDWORK dataset, a collection of linguistic field recordings with audio that has been transcribed and glossed in IGT. We build upon DoReCo (Seifart et al., 2022) and Multi-CAST, which are curated collections of field data, as well as data released through the COCOON repository,[2] and data produced by the INEL project[3] and NINJAL (Nakagawa et al., 2021). Our main contributions are selecting data for which there is both audio and gloss; compiling this data into a single structured, computer accessible dataset; and providing transcription and glossing benchmarks for each language in the dataset. Our work would not have been possible without the dedicated work of expert field linguists and speaker communities.

Our first step in adapting the FIELDWORK corpora for our four subtasks is to select languages where both audio and IGT are available. There are some overlaps between DoReCo, Multi-CAST, and INEL. For each language that appears more than once in those three sources, we only use the data from the source that contains the highest number of utterances. We do not attempt to merge corpora of the same language from multiple sources. We also require all languages we select to have a permissive CC license without a No Derivatives (ND) restriction.[4] The list of language corpora selected along with their license information, and the number of hours of data available in training and combined dev and test splits is shown in Table 1. See Figure 2 for an overview of the data processing pipelines.

Annotated data come in a variety of formats such as JSON or XML. Most of our source data come in XML-based formats—with ELAN (Brugman and Russel, 2004) being the most popular—as well as EXMARaLDA (Schmidt and Wörner, 2014) and Pangloss DTD (Michailovsky and Jacobson, 2001).[5] In ELAN and EXMARaLDA, annotations are text strings with beginning and end time stamps, organized into tiers including underlying form, surface form, transcription, gloss, and unique ID.

We extract annotations associated with utterances along with their corresponding audio spans, convert all audio files to WAV format with a single channel and 16 kHz sampling rate, and store annotations in an intermediate YAML based format (Mortensen et al., 2023) that is easier to process, read, and edit. We manually inspect the annotations for conversion errors and non-speech markers.

Finally, we partition the corpus into train/dev/test splits. We create partitions that contain full documents in order to preserve the contextual information. Assuming each document covers a different topic and has slightly different recording conditions, using full documents will make modeling more challenging and realistic, since the dev and test splits will be out-of-distribution, with minimal overlap in content between the splits. For each language, we look at the number of utterances in total to determine the splits. If there are fewer than 200

[2] https://cocoon.huma-num.fr/exist/crdo?lang=en

[3] https://www.slm.uni-hamburg.de/en/inel.html

[4] We base our decision on a layperson's reading of the Creative Commons licenses, which counts cleaned, reformatted, and standardized versions of datasets as derivative works.

[5] See von Prince and Nordhoff (2020) for a more detailed overview of ELAN and common annotator practices.

Glottocode	Name	CC Type	Train (h)	Dev+Test (h)
DoReCo				
beja1238	Beja (Vanhove, 2022)	BY-NC	1.55	0.29
ruul1235	Ruuli (Witzlack-Makarevich et al., 2022)	BY	0.96	0.28
texi1237	Texistepec Popoluca (Wichmann, 2022)	BY	0.84	0.26
komn1238	Komnzo (Döhler, 2022)	BY	0.73	0.42
arap1274	Arapaho (Cowell, 2022)	BY	0.56	0.88
goro1270	Gorwaa (Harvey, 2022)	BY	0.52	0.45
teop1238	Teop (Mosel, 2022)	BY	0.52	0.52
nngg1234	N||ng (Güldemann et al., 2022)	BY	0.52	0.33
sumi1235	Sümi (Teo, 2022)	BY	0.40	0.40
jeju1234	Jejuan (Kim, 2022)	BY	0.38	0.65
bora1263	Bora (Seifart, 2022)	BY	0.23	1.44
apah1238	Yali (Apahapsili) (Riesberg, 2022)	BY-NC-SA	0.18	0.27
port1286	Daakie (Krifka, 2022)	BY	0.14	0.75
savo1255	Savosavo (Wegener, 2022)	BY	0.10	1.20
trin1278	Mojeño Trinitario (Rose, 2022)	BY	-	1.56
sout2856	Nafsan (South Efate) (Thieberger, 2022)	BY-NC-SA	-	1.55
pnar1238	Pnar (Ring, 2022)	BY-NC	-	0.91
kaka1265	Kakabe (Vydrina, 2022)	BY	-	0.90
Multi-CAST				
vera1241	Vera'a (Schnell, 2015)	BY	1.02	0.97
tond1251	Tondano (Brickell, 2016)	BY	0.22	0.67
taul1251	Tulil (Meng, 2016)	BY	-	1.18
arta1239	Arta (Kimoto, 2019)	BY	-	0.91
nort2641	Northern Kurdish (Haig et al., 2015)	BY	-	0.86
tehr1242	Persian (Adibifar, 2016)	BY	-	0.82
taba1259	Tabasaran (Bogomolova et al., 2021)	BY	-	0.79
sanz1248	Sanzhi Dargwa (Forker and Schiborr, 2019)	BY	-	0.67
kach1280	Jinghpaw (Kurabe, 2021)	BY	-	0.66
mand1415	Mandarin (Vollmer, 2020)	BY	-	0.66
sumb1241	Sumbawa (Shiohara, 2022)	BY	-	0.63
kara1499	Kalamang (Visser, 2021)	BY	-	0.59
COCOON				
slav1254	Slavomolisano (Breu et al., 2018)	BY-NC	1.01	0.96
balk1252	Balkan Romani (Adamou, 2015)	BY-NC-SA	-	0.35
INEL				
dolg1241	Dolgan (Däbritz et al., 2022)	BY-NC-SA	11.64	1.23
kama1378	Kamas (Gusev et al., 2019)	BY-NC-SA	9.91	1.15
selk1253	Selkup (Brykina et al., 2021)	BY-NC-SA	1.70	1.15
even1259	Evenki (Däbritz and Gusev, 2021)	BY-NC-SA	1.54	1.13
NINJAL				
ainu1240	Ainu (Nakagawa et al., 2021)	BY-SA	7.12	1.13
	FIELDWORK Total	BY-NC-SA	41.79	29.56

Table 1: Overview of languages included in the FIELDWORK dataset. All licenses are CC with the specific restrictions for each language listed. We also show the hours of training data and combined dev and test data available for each language.

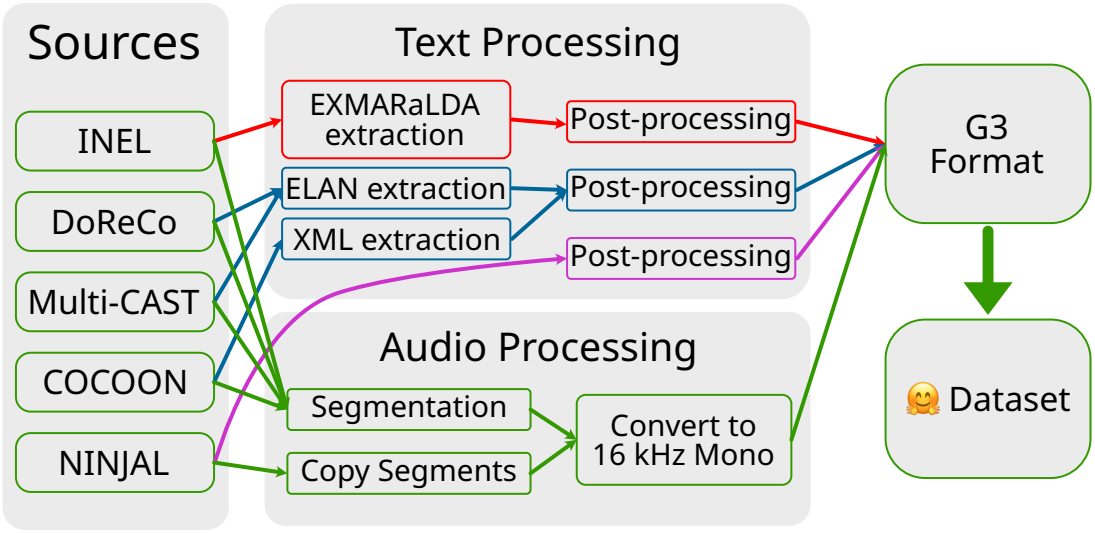


Figure 2: A visualization of the building of the FIELDWORK dataset.

utterances, all of them are assigned to the test set; if there are between 200 and 1,000 utterances, we assign 25% of the data to the dev set, and the rest to the test set; if there are more than 1,000 utterances, we assign 250 utterances to the dev set, 750 utterances to the test set, and the rest to the training set. With those partitions determined, we use a knapsack solver[6] to optimally assign documents to splits based on the number of utterances within each document.[7] We then apply final text cleaning by normalizing punctuations and removing special symbols on the transcriptions and translations, and convert the corpora into a Hugging Face dataset which can be readily used to train and test speech-to-text models.

In the following paragraphs, we present information on the specific data sources and unique processing required for each.

DoReCo **Do**cumentation **Re**ference **Co**rpus (Paschen et al., 2020) provides time-aligned transcriptions for 51 under-resourced languages. Paschen et al. processed each language corpus through a pipeline with consistency checking, multiple rounds of audio-text alignment, and manual corrections. We selected only data that is marked as fully (vs. partially or not) glossed. The annotations in the DoReCo dataset are in the ELAN format and organized such that transcriptions, translations, and utterance IDs share the same time span at the utterance level, and the underlying forms and glosses share the time span at the morpheme level within utterance time spans. After extracting the annotations, we drop words or utterances that consist only of the pause marker `<p:>`, and replace each span marked as `<label<text>>` with its `text` component.

Multi-CAST **Multi**lingual **C**orpus of **A**nnotated **S**poken **T**exts is another collection of annotated time-aligned speech (Haig and Schnell, 2022). Similar to DoReCo, Multi-CAST contains annotated speech for 18 languages, with robust annotation guidelines (Haig and Schnell, 2015). The annotations are also stored in the ELAN format; thus, data extraction is similar to that for DoReCo. One notable difference is Multi-CAST's designation of non-speech annotations, which start with `#`, `0`, or `%`. We delete text tokens marked with these non-speech symbols from the underlying form tier and the gloss tier.

INEL Grammars, Corpora and Language Technology for **I**ndigenous **N**orthern **E**urasian **L**anguages is an ongoing project at the Academy of Sciences and Humanities in Hamburg and the University of Hamburg that focuses on gathering resources for indigenous languages and language varieties of Northern Eurasia.[8] We use all four of the languages released so far (see Table 1) and use all annotated speech available. The datasets are annotated with EXMARaLDA (Schmid and

[6] https://github.com/google/or-tools

[7] Because many of our datasets contained only 1-2 speakers, it was impossible to avoid speaker overlap between splits. However, this is not as serious concern for our application as it may be for others, since language documentation is almost always conducted with a small number of speakers.

[8] https://www.slm.uni-hamburg.de/en/inel.html

Wörner, 2014), which is structurally similar to ELAN.

COCOON **Co**llections de **Co**rpus **O**raux **N**umériques is a large repository of field linguistic data that contains a variety of data types from a wide range of researchers.[9] To narrow down the data that are of interest to us, we first obtained a list of annotation files within the archive through the OLAC Aggregator,[10] targeting the Pangloss DTD format used specifically within COCOON. After the list of annotation files was obtained, we retrieved them along with associated media files, and sorted the results by language. We then used a simple heuristic—checking that there were multiple levels of word-level annotation and that morpheme-level annotation was available—to select a subset of data that likely contained speech with IGT. We then did a first round of manual verification to make sure the license information was available and suitable, and then a second round of more detailed checks for IGT quality. Two languages remained after the process, as shown in Table 1. Though a third language, Kakabe, would also fit our criteria, we chose to use its DoReCo corpus because COCOON data are generally noisier and are not automatically aligned and manually checked as DoReCo data are.

NINJAL Ainu Folklore Ainu is a nearly extinct language spoken in Hokkaido, Japan. The **NINJAL** Glossed Audio Corpus of **Ainu Folklore** (Nakagawa et al., 2021) contains recordings of 38 traditional Ainu folktales by two Ainu speakers, along with their transcriptions (in Latin script and occasionally Japanese script), English and Japanese translations, and underlying and surface gloss forms in English and Japanese. We used the Latin transcriptions and English translation/glosses. We scraped data via the corpus's web interface[11] and stored them in their native JSON format. We communicated with the authors and obtained permission to share them.

[9] https://cocoon.huma-num.fr

[10] Accessible through http://www.language-archives.org/cgi-bin/olaca3.pl?verb=Document. Our data is up-to-date as of Feb 23, 2023. In case the aggregator is not available, metadata can also be obtained through the CLARIN OAI harvester: https://github.com/clarin-eric/oai-harvest-manager.

[11] https://ainu.ninjal.ac.jp/folklore/en/

3 Experiments

We provide benchmarks for automatically generating IGT by fine-tuning pre-trained speech and text models. We choose commonly used models, methods, and finetuning settings with their corresponding codebases to provide a solid baseline for future research. See Appendix A for details on models sizes and hyper-parameter settings.

3.1 End-to-End models

Three of our four tasks (transcription, underlying form, and IGT prediction) are monotonic sequence-to-sequence tasks, similar to automatic speech recognition (ASR). Hence, we employ standard ASR training methods for prediction of each annotation (transcription, underlying form, gloss, and translation). Even though translation is not monotonic with respect to time as the other tasks are, we use the same training scheme, since previous work has found that multi-head attention-based networks can implicitly model non-monotonicity (Yan et al., 2023). Meanwhile, by using the same scheme, we can provide a more straightforward comparison between the performance of different tasks and approaches.

For end-to-end approaches, we use ESPnet (Watanabe et al., 2018) to employ two representative families of state-of-the-art pre-trained speech models: self-supervised and semi-supervised. The first self-supervised model we employ is WavLM Large (Chen et al., 2022), which achieves state-of-the-art performance on the SUPERB benchmark, a leaderboard for various speech-related tasks (Yang et al., 2021; Feng et al., 2023). We also use XLS-R-300M (Babu et al., 2021), a model specifically trained for cross-lingual capabilities, which has shown superior performance on the ML-SUPERB benchmark in multilingual tasks (Shi et al., 2023a,b). WavLM and XLS-R are also in the family of HuBERT- and wav2vec2.0-like models (Hsu et al., 2021; Baevski et al., 2020), respectively, which are commonly used self-supervised models. Similar to Chen et al. (2023), while we freeze the self-supervised models to preserve acoustic knowledge, we attach a conformer encoder and transformer decoder (Guo et al., 2021) to support the four different annotations we infer. Both the encoder (50M parameters) and decoder (26M parameters) have six blocks and eight attention heads, with the addition of the 315M frozen parameters from each of the pre-trained models, bringing the total

parameter count to 391M. WavLM was pretrained on English only, while XLS-R was pretrained on multiple languages, two of which—Mandarin and Persian—are also in FIELDWORK. We train the model with CTC-Attention loss (Kim et al., 2017), where CTC and attention loss are applied to the encoder and decoder, respectively. We train the language model from scratch and employ character-level tokenization with added language and task tokens. For details, see our public source code.[12]

On the other hand, supervised speech models, such as Whisper (Radford et al., 2023) or OWSM (Peng et al., 2023), show reasonably robust performance in various tasks, especially ASR. Hence, we fine-tune the OWSM-v3.1-base model (Peng et al. 2024, 101M parameters), which is an open-sourced reproduction of Whisper with public training software and datasets. We use OWSM in our experiments because its open-source nature is desirable for reproducibility and thorough scientific analysis. These supervised models already contain a predefined vocabulary and the corresponding tokenizer. OWSM uses a BPE tokenizer with 50k vocabulary, trained on a random subset of its training data. We add a specific token per language and two task tokens for gloss and underlying forms. For transcription and translation, we utilize the existing task tokens. The pretraining corpus of OWSM is also multilingual, with Mandarin and Persian being the only two overlapping languages with FIELDWORK. Just as with self-supervised models, we fine-tune with the CTC-Attention loss (the same approach as in OWSM pre-training). Similar to Rouditchenko et al. (2023), we fully fine-tune the OWSM model. The source code for fine-tuning OWSM is available in a public repository.[13]

We train a small number of OWSM models monolingually for transcription in each language, for comparison in Section 5. The remainder we train in a multilingual manner, including all the languages in FIELDWORK. During training, we evaluate the models using the sample-wise average accuracy on the dev set after each epoch, and keep the checkpoint with the highest accuracy. We experiment with both single-task and multi-task models: in the single-task paradigm, we train an individual model for each of the output forms (transcription, underlying, gloss, translation); in the multi-task paradigm, we train a single multi-task model that predicts different output forms based on the task token. We compare the performance in Section 4.

3.2 Cascaded model

From our preliminary experiments, we see that predicting glosses from speech is much more challenging than predicting transcriptions (i.e. ASR). The 2023 SIGMORPHON shared task on interlinear glossing (Ginn et al., 2023) showed that text-to-gloss models can achieve over 90% gloss prediction accuracy in certain languages, given enough training data. Therefore, we evaluate a cascaded approach where we take the best performing ASR models in the end-to-end setting, and use their transcription outputs as inputs to a text-to-gloss model. We use two text models initialized from ByT5-base (Xue et al. 2022, 582M parameters), one trained only on FIELDWORK for the underlying form, gloss, and translation tasks; the other fine-tuned first on ODIN (Lewis and Xia, 2010) and then fine-tuned on FIELDWORK for gloss and translation, similar to the approach of He et al. (2023). All text models are single-task, meaning we train a separate model for each task within each setting. While the ByT5 model was not the best performing in the shared task, it can easily support multilingual training and does not require the inputs to be segmented, and therefore is more suitable for the unsegmented outputs of the ASR systems.

4 Results

We compare average model performance on seen languages (the 22 languages with training sets) and unseen languages (the 15 languages with only dev and test sets) in Table 2. We report character error rates (CER, lower is better) for transcription, underlying, and glossing; and character F-scores (chrF++, Popović 2017, higher is better) for translation. We also evaluate translation outputs with BLEU (Papineni et al., 2002), BLEURT (Sellam et al., 2020), and BERTScore (Zhang et al. 2020, see Appendix B). All reported scores are macro-averaged across languages. We observe that multi-task models are worse across all tasks except glossing, where each of the two multi-task self-supervised models outperforms its single-task counterpart. Of the single-task end-to-end speech models, the two self-supervised models share similar performance across tasks, with the XLS-R based model performing best for transcription and underlying form prediction on seen languages. The

[12] https://github.com/juice500ml/espnet/tree/wav2gloss/egs2/wav2gloss/asr1

[13] https://github.com/juice500ml/finetune_owsm

Model	Transcription CER ↓		Underlying CER ↓		Gloss CER ↓		Translation chrF++ ↑	
	Seen	Unseen	Seen	Unseen	Seen	Unseen	Seen	Unseen
Multi-task								
WavLM E2E	76.9	77.8	66.3	75.0	78.8	**78.7**	7.2	7.6
XLS-R E2E	66.6	80.3	74.3	81.1	78.2	80.5	8.1	9.5
OWSM E2E	53.6	78.5	60.7	92.1	81.0	117.1	14.0	11.3
Single task								
WavLM E2E	38.1	**59.2**	45.9	**64.5**	84.8	88.3	8.4	7.9
XLS-R E2E	**36.8**	59.6	**44.0**	66.8	85.6	90.3	9.2	8.5
OWSM E2E	48.2	67.7	54.8	80.0	**75.0**	102.9	13.7	**11.6**
Cascade								
XLS-R + ByT5	-	-	48.5	70.6	86.7	124.1	16.0	11.0
XLS-R + ByT5 w/ ODIN	-	-	-	-	85.5	120.8	**16.6**	10.6
Ground truth text								
ByT5	-	-	16.0	28.1	55.2	157.0	22.0	12.2
ByT5 w/ ODIN	-	-	-	-	47.7	137.2	23.0	12.2

Table 2: Results from multilingual experiments. The languages are split into two groups: "seen", where the languages are in the training set, and "unseen", where the languages are not in the training set. Each number represents an average of that metric across the languages in that group (macro-averaging). WavLM and XLS-R models have pretrained encoders while OWSM have pretrained encoder and decoder.

OWSM model, on the other hand, is better at generating gloss and translation.

Since XLS-R is our best model for transcription, we employ it for the cascade setting so that its transcription outputs become the inputs for the text annotation model. The cascade approach produces better translation than all end-to-end models, but fails to improve on the underlying or glossing tasks. Pretraining the text model on ODIN slightly improves glossing performance, but not enough to surpass end-to-end.

Unsurprisingly, models generally perform better on seen than unseen languages. The overall performance of the models is low in absolute numbers across most of the tasks, with ASR being the easiest task, and gloss and translation the hardest. Qualitatively, from inspecting the models' outputs on a few languages, it seems that they are not able to generate coherent and relevant translations. This highlights the challenges with building NLP resources for low-resource languages, with minimal data spread across many languages. Even models that were pretrained with multilingual datasets are hard to adapt to languages in FIELDWORK.

5 Discussion

Some aspects of our experimental results highlight trends that may assist further development of WAV2GLOSS technologies.

Pre-trained vocabulary aids glossing and translation One notable difference between OWSM and the other two end-to-end speech models is that it includes a decoder pre-trained for transcription, translation, and language identification. Because the references for our gloss and translation tasks are in high-resource languages that were likely included in OWSM's large, multilingual training set, its BPE tokenizer has likely been exposed to many of their tokens. The references for our transcription and underlying prediction, however, are in low-resource languages likely not included in OWSM's training data. This phenomenon of tokenization is likely why OWSM performs comparatively well for glossing and trasciption, and comparatively poorly for transcription and underlying forms.

Single-task beats multi-task We observe that single-task models are generally much better across all tasks, potentially because the tasks' diversity

causes interference in multi-task objective settings (Yu et al., 2020). However, among pre-trained models, OWSM shows the smallest degradation from single- to multi-task performance, possibly because it was pre-trained with a multi-task objective and can thus extend better to new tasks.

Our cascaded models do not fully realize text-based potential From the results of the models trained on ground truth text in Table 2, it is clear that annotation inference is easier from text than from speech. However, the advantage is not enough to overcome error propagation introduced by a cascade approach, at least for glossing. One appeal of text-input models is training data availability, since text IGT data is far more plentiful than speech. Our own results—indicating that fine-tuning on noisy IGT (ODIN) improves glossing and translation—demonstrate some of this potential. For machine translation in particular, there are specialized pre-trained models with much higher multilingual MT performance that could be deployed as part of the pipeline. Novel approaches to this task should take this into account, and perhaps employ a multimodal model accepting both speech and text inputs.

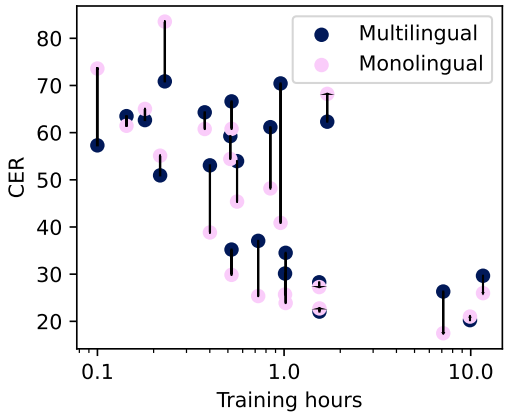


Figure 3: Comparing the seen language performance of multilingual transcription OWSM E2E model with monolingual models, trained separately with individual languages. Multilingual and monolingual performance is denoted as dark blue and light pink dots, respectively. Same language is connected with a black line.

Multilinguality degrades performance except on the lowest-resource languages Multilingual training can vitally boost model performance for low-resource languages in some settings (Chen et al., 2019). However, multilinguality can also cause performance degradation for some languages, a phenomenon known as the "curse of multilinguality" (Conneau et al., 2020). Given this theoretical tension, we analyze monolingual versus multilingual training directly. We can only fairly compare OWSM models in this analysis, since self-supervised models rely on language models trained from scratch, and hence have different vocabulary sizes in monolingual vs. multilingual settings. Transcription error rates from end-to-end single-task OWSM models using monolingual versus multilingual training are shown in Figure 3. We observe that CER improves (decreases) as the dataset size increases. Generally, multilingual models perform better for small datasets, but this trend inverts as data increase, indicating multilingual performance degradation. This performance degradation is less pronounced for the highest-resource languages (Dolgan, Kamas, and Ainu; on the far right in Figure 3) than for other languages. We suspect that multilinguality is less harmful for these languages due to our checkpoint saving strategy. (See Section 3.1.) Because we keep the checkpoint with the best dev set accuracy in multilingual training, the checkpoints are biased towards languages with the highest dev set representation.

6 Related Work

Our study is informed by a richness of previous work in automatic glossing, including a number of proposed systems to predict IGT from segmented or unsegmented transcriptions. In the recent SIGMORPHON shared task on interlinear glossing, the best performing models included hard attention, transformer, and LSTM (Ginn et al., 2023). He et al. (2023) showed that ByT5 models were not best at glossing, but had higher tolerance to noise and could benefit from noisy fine-tuning data such as ODIN (informing our decision to use them here).

FIELDWORK is made possible by the work of field linguists. Previously, there have been many text-only IGT datasets from linguistic field works available for machine learning use, such as ODIN (Lewis and Xia, 2010), IMTVault (Nordhoff and Krämer, 2022), or the SIGMORPHON IGT shared task data (Ginn et al., 2023). However, to our knowledge, FIELDWORK represents the first multilingual machine readable dataset focused on speech and interlinear gloss.

Our work is also informed by prior study in low-resource ASR. Previous low-resource ASR methodologies include fine-tuning high-resource ASR models (Cho et al., 2018) or self-supervised

speech models (Baevski et al., 2020; Babu et al., 2021). Later researchers built on these methods by continuous pre-training (Tian et al., 2022; Chen et al., 2020; Metze et al., 2015; Sakti and Titalim, 2023), model adaptation (Yu et al., 2023; Samarakoon et al., 2018), and data augmentation (Khare et al., 2021; Robinson et al., 2022). These prior works informed our use of pre-trained models; however our work is the first to explore their adaptability to predict morphological, glossing, and translation annotations along with transcription. Since many low-resource languages need these annotations, FIELDWORK can provide a valuable benchmark to facilitate future ASR and IGT prediction.

7 Conclusion and Future Work

In this work, we make a number of strides to advance technology-assisted documentation needed by language communities. We define a new task, WAV2GLOSS, which is to predict IGT annotations directly from speech. We present FIELDWORK, the first dataset for this task, comprised of audio files and expert annotations that were cleaned and formatted. We provide benchmarks across various training methodologies for transcription, underlying form prediction, glossing, and translation. And we analyze various prominent trends from experimental results. These data, benchmarks, and preliminary experimental insights provide a strong foundation for future `wav2gloss` breakthroughs, to expedite creation of needed language resources for communities of dying languages.

This work may be continued in a number of ways. The models presented in this work represent baseline approaches, and we hope this work will spur on more experimentation on the optimal hyperparameters for this task. In particular, the choice of model size and the tokenization method can be further explored. Moreover, we think multimodal setups such as Barrault et al. (2023) where both speech and text inputs can be fed into the model at the same time can be very promising for WAV2GLOSS. Another possible avenue of research is the use of cascades consisting of more than two models, for example, ASR into morpheme segmentation into glossing.

Future researchers may also further normalize IGT labels, to expand FIELDWORK to more diverse languages and phenomena. This is labor intensive, as previous research (List et al., 2021) and we ourselves observed, given inconsistent use of labels and language-specific phenomena across IGT collections. Expansion of FIELDWORK to more languages may also come through community-driven projects to benefit low-resource language communities and academics. Researchers may also focus on improving our work's modeling capabilities—e.g., adapting models to zero-shot performance. In our own work, all models perform notably better on seen than on unseen languages. Future work may mitigate this by mapping all transcriptions to a shared vocabulary, such as IPA, to minimize superficial orthographic differences.

Limitations

While we expect our contributions to be of significant value to the research community, we wish to acknowledge significant limitations to consider. Most notably to start, as is apparent from the scores in Table 2, our models' performance in any of the four subtasks is not sufficient to render useful outputs for application. This highlights the challenging nature of working with low-resource languages and the novel WAV2GLOSS task. By the release of our dataset and benchmarks, we mean to spur future iterations that improve upon our results and move towards applicable solutions for low-resource language communities.

We acknowledge also that, since we did not perform linguistic annotations ourselves of the IGT datasets we include, our dataset's quality is tied to the accuracy of others' linguistic expertise and annotations. Since errors in annotation for training data, such as misalignments and mislabels, can cause mild to severe errors in machine learning outputs, we encourage users of our data to proceed with caution. We ensure our data's quality via cleaning and filtering, in addition to random manual inspection finding no severe errors, but we encourage continued vigilance in this vein.

We hope this work can apply to other speech data from low-resource languages, since FIELDWORK contains a diverse collection of languages, though we have not covered all possible writing systems and the effectiveness of the models with respect to rarer systems such as Cyrillic, Chinese, Arabic, etc. is untested. We consider the systems we propose in this paper for research purposes only and have not tested generated texts for potential harmful or offensive contents.

People familiar with language resource archives will perhaps note the conspicuous absence of

three of the largest archives of field linguistics – TLA, ELAR, and PARADISEC. We forego these archives for this work mostly due to time and resource constraints, and will hopefully include them in the future. The challenges shared by the expansion to the three archives are mostly due to their sheer size and the work required to filter, extract, and obtain rights for the data. Most of this work will have to be done on a language-by-language basis to ensure quality.

Ethics Statement

We wish to emphasize that any work in technologies for low-resource language communities should be approached with a high level of care for ethical practices. Many of these communities have particular needs and interests. And many have socioeconomic disadvantages that could be either helped or exacerbated by technological advancements.

To be forthcoming about these important considerations, we wish to acknowledge some of the ethical concerns involved in the particular substance of our current work. We first acknowledge that data involved in this study were predominantly collected with the assumption that they would be used for language documentation and to support communities. While these are also our own end goals in developing the resources presented here, we recognize that researchers may use our open-source NLP technologies for a variety of purposes.

We also wish to be straightforward about some potential ethical concerns with our data. While these data were collected with consent, in accordance with any pertinent legal and institutional protocols, they still contain sensitive materials. We did not manually anonymize the data, and therefore we urge caution to any users of FIELDWORK to respect the rights and privacy of all individuals concerned as much as possible. We acknowledge that some low-resource language speakers may not see these technologies as a benefit and that they may be concerned about the potential effects of technology on their communities. We also express that we do not necessarily condone any opinions, worldviews, or assertions expressed within the transcriptions or recordings of our dataset. It is possible that some of their material may be offensive in some contexts. This is a common concern in building multilingual and multicultural datasets, since statements or references considered benign in one culture may be seen as offensive in another. We therefore caution users accordingly.

Despite these potential concerns, which should be considered in all seriousness, we intend our work to have a highly positive effect, from an ethical standpoint. We hope our efforts can contribute to serving communities that have otherwise been left behind by many technological advancements, and to assist efforts to preserve valuable languages and cultures across the world, regardless of their socioeconomic privilege.

Acknowledgements

This work is supported by the US National Science Foundation grant #2211951. We also thank the linguists whose resources we used and their collaborators. Without them, this work would not have been possible. For the computational resources, we used the Bridges2 system at PSC and Delta system at NCSA through allocation CIS210014 from the Advanced Cyberinfrastructure Coordination Ecosystem: Services & Support (ACCESS) program, which is also supported by the US National Science Foundation grants #2138259, #2138286, #2138307, #2137603, and #2138296. Finally, we thank Paola Garcia, the anonymous reviewers, and the Area Chair for their valuable feedback to the present paper.

References

Evangelia Adamou. 2015. A corpus-driven analysis of romani in contact with turkish and greek. In *Language Variation - European Perspectives V*, volume 17. John Benjamins Publishing Company, The Netherlands.

Shirin Adibifar. 2016. Multi-CAST Persian.

Jonathan D. Amith, Jiatong Shi, and Rey Castillo García. 2021. End-to-end automatic speech recognition: Its impact on the workflowin documenting yoloxóchitl Mixtec. In *Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas*, pages 64–80, Online. Association for Computational Linguistics.

Arun Babu, Changhan Wang, Andros Tjandra, Kushal Lakhotia, Qiantong Xu, Naman Goyal, Kritika Singh, Patrick von Platen, Yatharth Saraf, Juan Pino, et al. 2021. Xls-r: Self-supervised cross-lingual speech representation learning at scale. *arXiv preprint arXiv:2111.09296*.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33:12449–12460.

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. 2023. Seamless: Multilingual expressive and streaming speech translation. *arXiv preprint arXiv:2312.05187*.

Natalia Bogomolova, Dmitry Ganenkov, and Nils N. Schiborr. 2021. Multi-CAST Tabasaran.

Walter Breu, Giovanni Piccoli, Mia Barbara Mader Skender, Catherine Thornton, and Stuart Cunningham. 2018. *Slavische Mikrosprachen Im Absoluten Sprachkontakt: Glossierte Und Interpretierte Sprachaufnahmen Aus Italien, Deutschland, Österreich Und Griechenland. Teil I: Moliseslavische Texte Aus Acquaviva Collecroce, Montemitro Und San Felice Del Molise*, 1 edition. Harrassowitz Verlag.

Timothy Brickell. 2016. Multi-CAST Tondano.

Hennie Brugman and Albert Russel. 2004. Annotating multi-media/multi-modal resources with ELAN. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, Lisbon, Portugal. European Language Resources Association (ELRA).

Maria Brykina, Svetlana Orlova, and Beáta Wagner-Nagy. 2021. INEL Selkup Corpus.

Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, et al. 2022. Wavlm: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing*, 16(6):1505–1518.

William Chen, Brian Yan, Jiatong Shi, Yifan Peng, Soumi Maiti, and Shinji Watanabe. 2023. Improving massively multilingual asr with auxiliary ctc objectives. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Xilun Chen, Ahmed Hassan Awadallah, Hany Hassan, Wei Wang, and Claire Cardie. 2019. Multi-source cross-lingual model transfer: Learning what to share. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3098–3112, Florence, Italy. Association for Computational Linguistics.

Yi-Chen Chen, Jui-Yang Hsu, Cheng-Kuang Lee, and Hung yi Lee. 2020. DARTS-ASR: Differentiable Architecture Search for Multilingual Speech Recognition and Adaptation. In *Proc. Interspeech 2020*, pages 1803–1807.

Jaejin Cho, Murali Karthick Baskar, Ruizhi Li, Matthew Wiesner, Sri Harish Mallidi, Nelson Yalta, Martin Karafiat, Shinji Watanabe, and Takaaki Hori. 2018. Multilingual sequence-to-sequence speech recognition: architecture, transfer learning, and language modeling. In *2018 IEEE Spoken Language Technology Workshop (SLT)*, pages 521–527. IEEE.

Bernard Comrie, Martin Haspelmath, and Balthasar Bickel. 2008. The leipzig glossing rules: Conventions for interlinear morpheme-by-morpheme glosses.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Édouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451.

Andrew Cowell. 2022. Arapaho DoReCo dataset.

Thi-Ngoc-Diep Do, Alexis Michaud, and Eric Castelli. 2014. Towards the automatic processing of Yongning Na (Sino-Tibetan): developing a 'light' acoustic model of the target language and testing 'heavyweight' models from five national languages. In *4th International Workshop on Spoken Language Technologies for Under-resourced Languages (SLTU 2014)*, pages 153–160, St Petersburg, Russia.

Chris Lasse Däbritz and Valentin Gusev. 2021. INEL Evenki Corpus.

Chris Lasse Däbritz, Nina Kudryakova, and Eugénie Stapert. 2022. INEL Dolgan Corpus.

Christian Döhler. 2022. Komnzo DoReCo dataset.

Tzu-hsun Feng, Annie Dong, Ching-Feng Yeh, Shu-wen Yang, Tzu-Quan Lin, Jiatong Shi, Kai-Wei Chang, Zili Huang, Haibin Wu, Xuankai Chang, et al. 2023. Superb@ slt 2022: Challenge on generalization and efficiency of self-supervised speech representation learning. In *2022 IEEE Spoken Language Technology Workshop (SLT)*, pages 1096–1103. IEEE.

Diana Forker and Nils N. Schiborr. 2019. Multi-CAST Sanzhi Dargwa.

Michael Ginn, Sarah Moeller, Alexis Palmer, Anna Stacey, Garrett Nicolai, Mans Hulden, and Miikka Silfverberg. 2023. Findings of the SIGMORPHON 2023 shared task on interlinear glossing. In *Proceedings of the 20th SIGMORPHON workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 186–201, Toronto, Canada. Association for Computational Linguistics.

Pengcheng Guo, Florian Boyer, Xuankai Chang, Tomoki Hayashi, Yosuke Higuchi, Hirofumi Inaguma, Naoyuki Kamo, Chenda Li, Daniel Garcia-Romero, Jiatong Shi, et al. 2021. Recent developments on espnet toolkit boosted by conformer. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5874–5878. IEEE.

Valentin Gusev, Tiina Klooster, and Beáta Wagner-Nagy. 2019. INEL Kamas Corpus.

Tom Güldemann, Martina Ernszt, Sven Siegmund, and Alena Witzlack-Makarevich. 2022. Nng DoReCo dataset.

Geoffrey Haig and Stefan Schnell. 2015. Annotations using graid:(grammatical relations and animacy in discourse): Manual version 7.0.

Geoffrey Haig and Stefan Schnell, editors. 2022. *Multi-CAST*. University of Bamberg, Bamberg. Version 2211.

Geoffrey Haig, Maria Vollmer, and Hanna Thiele. 2015. Multi-CAST Northern Kurdish.

Andrew Harvey. 2022. Gorwaa DoReCo dataset.

Taiqi He, Lindia Tjuatja, Nathaniel Robinson, Shinji Watanabe, David R. Mortensen, Graham Neubig, and Lori Levin. 2023. SigMoreFun submission to the SIGMORPHON shared task on interlinear glossing. In *Proceedings of the 20th SIGMORPHON workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 209–216, Toronto, Canada. Association for Computational Linguistics.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. 2021. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3451–3460.

Shreya Khare, Ashish Mittal, Anuj Diwan, Sunita Sarawagi, Preethi Jyothi, and Samarth Bharadwaj. 2021. Low Resource ASR: The Surprising Effectiveness of High Resource Transliteration. In *Proc. Interspeech 2021*, pages 1529–1533.

Soung-U Kim. 2022. Jejuan DoReCo dataset.

Suyoun Kim, Takaaki Hori, and Shinji Watanabe. 2017. Joint ctc-attention based end-to-end speech recognition using multi-task learning. In *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 4835–4839. IEEE.

Yukinori Kimoto. 2019. Multi-CAST Arta.

Manfred Krifka. 2022. Daakie DoReCo dataset.

Keita Kurabe. 2021. Multi-CAST Jinghpaw.

William D. Lewis and Fei Xia. 2010. Developing ODIN: A Multilingual Repository of Annotated Language Data for Hundreds of the World's Languages. *Literary and Linguistic Computing*, 25(3):303–319.

Johann-Mattis List, Nathaniel A. Sims, and Robert Forkel. 2021. Toward a Sustainable Handling of Interlinear-Glossed Text in Language Documentation. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 20(2):1–15.

Chenxi Meng. 2016. Multi-CAST Tulil.

Florian Metze, Ankur Gandhe, Yajie Miao, Zaid Sheikh, Yun Wang, Di Xu, Hao Zhang, Jungsuk Kim, Ian Lane, Won Kyum Lee, et al. 2015. Semi-supervised training in low-resource asr and kws. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4699–4703. IEEE.

Boyd Michailovsky and Michel Jacobson. 2001. Pangloss archive DTD.

David R. Mortensen, Ela Gulsen, Taiqi He, Nathaniel Robinson, Jonathan Amith, Lindia Tjuatja, and Lori Levin. 2023. Generalized glossing guidelines: An explicit, human- and machine-readable, item-and-process convention for morphological annotation. In *Proceedings of the 20th SIGMORPHON workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 58–67, Toronto, Canada. Association for Computational Linguistics.

Ulrike Mosel. 2022. Teop DoReCo dataset.

Hiroshi Nakagawa, Anna Bugaeva, Miki Kobayashi, and Yoshimi Yoshikawa. 2021. A glossed audio corpus of ainu folklore.

Sebastian Nordhoff and Thomas Krämer. 2022. IMT-Vault: Extracting and enriching low-resource language interlinear glossed text from grammatical descriptions and typological survey articles. In *Proceedings of the 8th Workshop on Linked Data in Linguistics within the 13th Language Resources and Evaluation Conference*, pages 17–25, Marseille, France. European Language Resources Association.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Ludger Paschen, François Delafontaine, Christoph Draxler, Susanne Fuchs, Matthew Stave, and Frank Seifart. 2020. Building a time-aligned cross-linguistic reference corpus from language documentation data (DoReCo). In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 2657–2666, Marseille, France. European Language Resources Association.

Yifan Peng, Jinchuan Tian, William Chen, Siddhant Arora, Brian Yan, Yui Sudo, Muhammad Shakeel, Kwanghee Choi, Jiatong Shi, Xuankai Chang, et al. 2024. Owsm v3. 1: Better and faster open whisper-style speech models based on e-branchformer. *arXiv preprint arXiv:2401.16658*.

Yifan Peng, Jinchuan Tian, Brian Yan, Dan Berrebbi, Xuankai Chang, Xinjian Li, Jiatong Shi, Siddhant Arora, William Chen, Roshan Sharma, et al. 2023. Reproducing whisper-style training using an open-source toolkit and publicly available data. In *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 1–8. IEEE.

Maja Popović. 2017. chrF++: words helping character n-grams. In *Proceedings of the Second Conference on Machine Translation*, pages 612–618, Copenhagen, Denmark. Association for Computational Linguistics.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *International Conference on Machine Learning*, pages 28492–28518. PMLR.

Sonja Riesberg. 2022. Yali (Apahapsili) DoReCo dataset.

Hiram Ring. 2022. Pnar DoReCo dataset.

Nathaniel Robinson, Perez Ogayo, Swetha Gangu, David R Mortensen, and Shinji Watanabe. 2022. When is tts augmentation through a pivot language useful? In *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, volume 2022, pages 3538–3542.

Françoise Rose. 2022. Mojeño Trinitario DoReCo dataset.

Andrew Rouditchenko, Sameer Khurana, Samuel Thomas, Rogerio Feris, Leonid Karlinsky, Hilde Kuehne, David Harwath, Brian Kingsbury, and James Glass. 2023. Comparison of Multilingual Self-Supervised and Weakly-Supervised Speech Pre-Training for Adaptation to Unseen Languages. In *Proc. INTERSPEECH 2023*, pages 2268–2272.

Sakriani Sakti and Benita Angela Titalim. 2023. Leveraging the multilingual indonesian ethnic languages dataset in self-supervised models for low-resource asr task. In *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 1–8. IEEE.

Lahiru Samarakoon, Brian Mak, and Albert YS Lam. 2018. Domain adaptation of end-to-end speech recognition in low-resource settings. In *2018 IEEE Spoken Language Technology Workshop (SLT)*, pages 382–388. IEEE.

Thomas Schmid and Kai Wörner. 2014. Exmaralda. In *Handbook on Corpus Phonology*, pages 402–419. Oxford University Press, Oxford.

Thomas Schmidt and Kai Wörner. 2014. EXMARaLDA. In *The Oxford Handbook of Corpus Phonology*. Oxford University Press.

Stefan Schnell. 2015. Multi-CAST Vera'a.

Frank Seifart. 2022. Bora DoReCo dataset.

Frank Seifart, Nicholas Evans, Harald Hammarström, and Stephen C. Levinson. 2018. Language documentation twenty-five years on. *Language*, 94(4):e324–e345.

Frank Seifart, Ludger Paschen, and Matthew Stave. 2022. Language Documentation Reference Corpus (DoReCo) 1.2.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. BLEURT: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892, Online. Association for Computational Linguistics.

Jiatong Shi, Jonathan D. Amith, Rey Castillo García, Esteban Guadalupe Sierra, Kevin Duh, and Shinji Watanabe. 2021a. Leveraging end-to-end ASR for endangered language documentation: An empirical study on yolóxochitl Mixtec. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1134–1145, Online. Association for Computational Linguistics.

Jiatong Shi, Jonathan D. Amith, Xuankai Chang, Siddharth Dalmia, Brian Yan, and Shinji Watanabe. 2021b. Highland Puebla Nahuatl speech translation corpus for endangered language documentation. In *Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas*, pages 53–63, Online. Association for Computational Linguistics.

Jiatong Shi, Dan Berrebbi, William Chen, En-Pei Hu, Wei-Ping Huang, Ho-Lam Chung, Xuankai Chang, Shang-Wen Li, Abdelrahman Mohamed, Hung yi Lee, and Shinji Watanabe. 2023a. ML-SUPERB: Multilingual Speech Universal PERformance Benchmark. In *Proc. INTERSPEECH 2023*, pages 884–888.

Jiatong Shi, William Chen, Dan Berrebbi, Hsiu-Hsuan Wang, Wei-Ping Huang, En-Pei Hu, Ho-Lam Chuang, Xuankai Chang, Yuxun Tang, Shang-Wen Li, et al. 2023b. Findings of the 2023 ml-superb challenge: Pre-training and evaluation over more languages and beyond. In *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 1–8. IEEE.

Asako Shiohara. 2022. Multi-CAST Sumbawa.

Amos Teo. 2022. Sümi DoReCo dataset.

Nick Thieberger. 2022. Nafsan (South Efate) DoReCo dataset.

Jinchuan Tian, Jianwei Yu, Chunlei Zhang, Yuexian Zou, and Dong Yu. 2022. LAE: Language-Aware Encoder for Monolingual and Multilingual ASR. In *Proc. Interspeech 2022*, pages 3178–3182.

Martine Vanhove. 2022. Beja DoReCo dataset.

Eline Visser. 2021. Multi-CAST Kalamang.

Maria Vollmer. 2020. Multi-CAST Mandarin.

Kilu von Prince and Sebastian Nordhoff. 2020. An empirical evaluation of annotation practices in corpora from language documentation. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 2778–2787, Marseille, France. European Language Resources Association.

Alexandra Vydrina. 2022. Kakabe DoReCo dataset.

Shinji Watanabe, Takaaki Hori, Shigeki Karita, Tomoki Hayashi, Jiro Nishitoba, Yuya Unno, Nelson Enrique Yalta Soplin, Jahn Heymann, Matthew Wiesner, Nanxin Chen, et al. 2018. Espnet: End-to-end speech processing toolkit. *Interspeech 2018*.

Claudia Wegener. 2022. Savosavo DoReCo dataset.

Søren Wichmann. 2022. Texistepec Popoluca DoReCo dataset.

Alena Witzlack-Makarevich, Saudah Namyalo, Anatol Kiriggwajjo, and Zarina Molochieva. 2022. Ruuli DoReCo dataset.

Linting Xue, Aditya Barua, Noah Constant, Rami Al-Rfou, Sharan Narang, Mihir Kale, Adam Roberts, and Colin Raffel. 2022. Byt5: Towards a token-free future with pre-trained byte-to-byte models. *Transactions of the Association for Computational Linguistics*, 10:291–306.

Brian Yan, Siddharth Dalmia, Yosuke Higuchi, Graham Neubig, Florian Metze, Alan W Black, and Shinji Watanabe. 2023. Ctc alignments improve autoregressive translation. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1615–1631.

Shu-Wen Yang, Po-Han Chi, Yung-Sung Chuang, Cheng-I Jeff Lai, Kushal Lakhotia, Yist Y. Lin, Andy T. Liu, Jiatong Shi, Xuankai Chang, Guan-Ting Lin, Tzu-Hsien Huang, Wei-Cheng Tseng, Ko tik Lee, Da-Rong Liu, Zili Huang, Shuyan Dong, Shang-Wen Li, Shinji Watanabe, Abdelrahman Mohamed, and Hung yi Lee. 2021. SUPERB: Speech Processing Universal PERformance Benchmark. In *Proc. Interspeech 2021*, pages 1194–1198.

Tianhe Yu, Saurabh Kumar, Abhishek Gupta, Sergey Levine, Karol Hausman, and Chelsea Finn. 2020. Gradient surgery for multi-task learning. *Advances in Neural Information Processing Systems*, 33:5824–5836.

Zhongzhi Yu, Yang Zhang, Kaizhi Qian, Cheng Wan, Yonggan Fu, Yongan Zhang, and Yingyan Celine Lin. 2023. Master-asr: achieving multilingual scalability and low-resource adaptation in asr with modular learning. In *International Conference on Machine Learning*, pages 40475–40487. PMLR.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

A Parameter Counts and Hyper-parameters

Model	Total	Trainable
WavLM E2E	391M	76M
XLS-R E2E	391M	76M
OWSM E2E	101M	101M
ByT5	582M	582M

Table 3: Parameter counts of models used in this work.

We show the parameter counts of models used in this work in Table 3. All experiments are done with NVIDIA RTX A6000 GPUs. End-to-end models are trained with 4 GPUs and text models are trained with a single GPU. We used 1,605 GPU hours in total for both training and evaluation.

We do not perform extensive hyper-parameter searches. The settings we used across the experiments are shown in Table 4.

	Conformer	OWSM	ByT5
Optimizer	Adam	AdamW	Adafactor
LR	2e-3	1e-3	5e-5
Warm up Steps	25k	-	-
Epochs	30	10	10

Table 4: Hyper-parameter settings used in this study. "Conformer" includes both WavLM and XLS-R E2E models which share the same settings.

B Additional Evaluation Metrics for Translation

See Table 5. Results are obtained using the Evaluate[14] package.

[14] https://github.com/huggingface/evaluate

	BLEU ↑		BLEURT ↑		BERTScore ↑	
Model	Seen	Unseen	Seen	Unseen	Seen	Unseen
Multi-task						
WavLM E2E	0.0	0.0	-1.35	-1.32	0.66	0.66
XLS-R E2E	0.0	0.0	-1.43	-1.47	0.67	0.67
OWSM E2E	2.1	0.1	-1.26	-1.36	0.72	0.69
Single task						
WavLM E2E	0.0	0.0	-1.57	-1.66	0.69	0.68
XLS-R E2E	0.0	0.0	-1.60	-1.57	0.69	0.68
OWSM E2E	2.0	0.1	-1.29	-1.43	0.72	0.69
Cascade						
XLS-R + ByT5	2.7	0.0	-1.27	-1.52	0.73	0.69
XLS-R + ByT5 w/ ODIN	2.9	0.0	-1.27	-1.52	0.74	0.69
Ground truth text						
ByT5	6.6	0.2	-1.00	-1.48	0.77	0.70
ByT5 w/ ODIN	6.7	0.3	-1.00	-1.48	0.77	0.70

Table 5: Additional metrics for translations.