CoVoST 2 and Massively Multilingual Speech-to-Text Translation

Changhan Wang⋆, Anne Wu⋆, Juan Pino⋆

Facebook AI

{changhan,annewu,juancarabina}@fb.com

Abstract

Speech-to-text translation (ST) has recently become an increasingly popular topic of research, partly due to the development of benchmark datasets. Nevertheless, current datasets cover a limited number of languages. With the aim to foster research in massive multilingual ST and ST for low resource language pairs, we release CoVoST 2, a large-scale multilingual ST corpus covering translations from 21 languages into English and from English into 15 languages. This represents the largest open dataset available to date from total volume and language coverage perspective. Data sanity checks provide evidence about the quality of the data, which is released under CC0 license. We also provide extensive speech recognition, bilingual and multilingual machine translation and ST baselines.

1 Introduction

The development of benchmark datasets, such as MuST-C (Di Gangi et al., 2019), Europarl-ST (Iranzo-Snchez et al., 2020) or CoVoST (Wang et al., 2020), has greatly contributed to the increasing popularity of speech-to-text translation (ST) as a research topic. MuST-C provides TED talks translations from English into 8 European languages, with data amounts ranging from 385 hours to 504 hours, thereby encouraging research into end-to-end ST (Berard et al., 2016) as well as one-to-many multilingual ST (Di Gangi et al., 2019). Europarl-ST offers translations between 6 European languages, with a total of 30 translation directions, enabling research into many-to-many multilingual ST (Inaguma et al., 2019). The two corpora described so far involve European languages that

---

⋆ Equal contribution.

are in general high resource from the perspective of machine translation (MT) and speech. CoVoST is a multilingual and diversified ST corpus from 11 languages into English, based on the Common Voice project (Ardila et al., 2020). Unlike previous corpora, it involves low resource languages such as Mongolian and it also enables many-to-one ST research. Nevertheless, for all corpora described so far, the number of languages involved is limited.

In this paper, we describe CoVoST 2, an extension of CoVoST (Wang et al., 2020) that provides translations from English (En) into 15 languages— Arabic (Ar), Catalan (Ca), Welsh (Cy), German (De), Estonian (Et), Persian (Fa), Indonesian (Id), Japanese (Ja), Latvian (Lv), Mongolian (Mn), Slovenian (Sl), Swedish (Sv), Tamil (Ta), Turkish (Tr), Chinese (Zh)—and from 21 languages into English, including the 15 target languages as well as Spanish (Es), French (Fr), Italian (It), Dutch (Nl), Portuguese (Pt), Russian (Ru). The overall speech duration is extended from 700 hours to 2880 hours. The total number of speakers is increased from 11K to 78K. We make data available at https://github.com/facebookresearch/covost under CC0 license.

2 Dataset Creation

2.1 Data Collection and Quality Control

Translations are collected from professional translators the same way as for CoVoST. We then conduct sanity checks based on language model perplexity, LASER (Artetxe and Schwenk, 2019) scores and a length ratio heuristic in order to ensure the quality of the translations. Length ratio and LASER score checks are conducted as in the original version of CoVoST. For language model perplexity checks, 20M lines are sampled from the OSCAR corpus (Ortiz Suárez et al.,

	Hours (CoVoST ext.)			Speakers (CoVoST ext.)			Src./Tgt. Tokens		
	Train	Dev	Test	Train	Dev	Test	Train	Dev	Test
	X→En								
Fr	180 (264)	22 (23)	23 (24)	2K (2K)	2K (2K)	4K (4K)	2M/2M	145K/143K	143K/141K
De	119 (184)	21 (23)	22 (120)	1K (1K)	1K (1K)	4K (5K)	1M/1M	137K/152K	777K/844K
Es	97 (113)	22 (22)	23 (23)	1K (1K)	2K (2K)	4K (4K)	747K/751K	131K/134K	131K/132K
Ca	81 (136)	19 (21)	20 (25)	557 (557)	722 (722)	2K (2K)	939K/972K	142K/148K	160K/169K
It	28 (44)	14 (15)	15 (15)	236 (236)	640 (640)	2K (2K)	307K/329K	89K/95K	88K/93K
Ru	16 (18)	10 (15)	11 (14)	8 (8)	30 (30)	417 (417)	118K/144K	89K/109K	81K/100K
Zh	10 (10)	8 (8)	8 (8)	22 (22)	83 (83)	784 (784)	131K/85K	91K/60K	88K/57K
Pt	7 (10)	4 (5)	5 (6)	2 (2)	16 (16)	301 (301)	67K/68K	27K/28K	34K/34K
Fa	5 (49)	5 (11)	5 (40)	532 (545)	854 (908)	1K (1K)	307K/313K	67K/73K	244K/271K
Et	3 (3)	3 (3)	3 (3)	20 (20)	74 (74)	135 (135)	23K/32K	19K/27K	20K/27K
Mn	3 (3)	3 (3)	3 (3)	4 (4)	24 (24)	209 (209)	20K/23K	19K/22K	18K/20K
Nl	2 (7)	2 (3)	2 (3)	74 (74)	144 (144)	379 (383)	58K/59K	19K/19K	20K/20K
Tr	2 (4)	2 (2)	2 (2)	34 (34)	76 (76)	324 (324)	24K/33K	11K/16K	11K/15K
Ar	2 (2)	2 (2)	2 (2)	6 (6)	13 (13)	113 (113)	10K/13K	9K/11K	8K/10K
Sv	2 (2)	1 (1)	2 (2)	4 (4)	7 (7)	83 (83)	12K/12K	8K/9K	9K/10K
Lv	2 (2)	1 (1)	2 (2)	2 (2)	3 (3)	54 (54)	11K/14K	6K/7K	8K/10K
Sl	2 (2)	1 (1)	1 (1)	2 (2)	1 (1)	28 (28)	11K/13K	3K/4K	2K/2K
Ta	2 (2)	1 (1)	1 (1)	3 (3)	2 (2)	48 (48)	6K/10K	2K/3K	3K/5K
Ja	1 (1)	1 (1)	1 (1)	2 (2)	3 (3)	37 (37)	20K/9K	12K/5K	12K/6K
Id	1 (1)	1 (1)	1 (1)	2 (2)	5 (5)	44 (44)	7K/8K	5K/5K	5K/6K
Cy	1 (2)	1 (12)	1 (16)	135 (135)	234 (371)	275 (597)	11K/10K	79K/76K	110K/103K
	En→X								
De							3M/3M	156K/155K	4M/4M
Tr							3M/2M	156K/125K	4M/2M
Fa							3M/3M	156K/172K	4M/4M
Sv							3M/3M	156K/143K	4M/3M
Mn							3M/3M	156K/144K	4M/3M
Zh							3M/6M	156K/332K	4M/6M
Cy							3M/3M	156K/168K	4M/4M
Ca	364 (430)	26 (27)	25 (472)	10K (10K)	4K (4K)	9K (29K)	3M/3M	156K/171K	4M/4M
Sl							3M/3M	156K/145K	4M/3M
Et							3M/2M	156K/120K	4M/3M
Id							3M/3M	156K/142K	4M/3M
Ar							3M/2M	156K/133K	4M/3M
Ta							3M/2M	156K/121K	4M/3M
Lv							3M/2M	156K/130K	4M/3M
Ja							3M/8M	156K/444K	4M/9M

Table 1: Basic statistics of CoVoST 2 using original CV splits and extended CoVoST splits (only for the speech part). Token counts on Chinese (Zh) and Japanese (Ja) are based on characters (there is no word segmentation).

2020) for each CoVoST 2 language, except for English, Russian for which pre-trained language models (Ng et al., 2019) are utilized[1]. 5K lines are reserved for validation and the rest for training. BPE vocabularies of size 20K are then built on the training data, with character coverage 0.9995 for Japanese and Chinese and 1.0 for other languages. A Transformer *base* model (Vaswani et al., 2017) is then trained for up to 800K updates. Professional translations are ranked by perplexity and the ones with the lowest perplexity are manually examined and sent for re-translation as appropriate. In the data release, we mark out the sentences that

cannot be translated properly[2].

2.2 Dataset Splitting

Original Common Voice (CV) dataset splits utilize only one sample per sentence, while there are potentially multiple samples (speakers) available in the raw dataset. To allow higher data utilization and speaker diversity, we add part of the discarded samples back while keeping the speaker set disjoint and the same sentence assignment across different splits. We refer to this extension as CoVoST splits. As a result, data utilization is increased from 44.2% (1273 hours) to 78.8% (2270 hours). We by default use CoVoST train split for model

	ASR	X→En					En→X				
		MT	MT+Rev	C-ST	C-ST+Rev	ST	MT	MT+Rev	C-ST	C-ST+Rev	ST
En	30.5										
Fr	20.9	37.9	38.1	26.4	26.5	23.2					
De	24.1	28.2	31.2	19.8	21.3	15.7	29.0	29.1	16.2	16.1	12.5
Es	19.2	36.3	36.2	26.0	26.1	20.2					
Ca	14.6	24.9	31.1	20.7	24.9	17.9	38.8	38.6	21.1	21.1	17.1
It	29.1	19.2	19.0	13.3	13.2	10.7					
Ru	33.9	19.8	19.4	16.6	16.1	14.1					
Zh⋆	41.7	7.6	16.6	6.6	9.5	4.4	35.3	38.9	21.4	23.1	20.0
Pt	55.9	14.6	13.9	8.1	7.7	8.2					
Fa	79.6	2.4	15.1	2.0	4.9	1.6	20.1	20.0	12.1	11.9	9.1
Et	68.6	0.3	13.7	0.2	4.2	0.3	24.0	24.3	12.6	12.6	9.3
Mn	74.5	0.2	5.4	0.1	1.3	0.1	16.8	17.1	9.6	9.7	6.4
Nl	64.3	2.6	2.5	2.2	1.9	2.2					
Tr	62.8	1.1	25.9	1.0	9.8	2.2	20.0	19.7	9.9	9.7	6.7
Ar	82.3	0.1	34.7	0.1	7.2	2.7	21.6	21.6	12.1	12.1	9.1
Sv	82.2	0.2	37.7	0.1	5.2	1.4	39.4	39.2	21.5	21.2	18.1
Lv	71.9	0.2	19.6	0.2	5.7	1.2	22.5	22.9	12.4	12.8	8.7
Sl	65.5	0.1	29.2	0.1	8.1	1.5	29.1	29.4	15.7	15.8	11.6
Ta	109.3	0.0	4.0	0.1	0.2	0.2	22.7	22.2	11.4	11.5	7.4
Ja⋆	58.8	0.0	14.6	0.0	1.7	1.1	42.8	42.2	29.1	29.3	25.6
Id	80.8	0.1	36.7	0.1	6.3	1.0	39.0	38.8	20.1	20.1	15.2
Cy	86.6	0.1	49.2	0.1	4.2	1.7	41.6	41.6	22.3	22.3	18.9

Table 2: Test WER (CER) for monolingual ASR; test BLEU for bilingual MT, cascaded ST (C-ST) and end-to-end ST. All non-English ASR and all ST model encoders are pre-trained on English ASR. ⋆ We use CER and character-level BLEU on Chinese and Japanese (no word segmentation).

training and CV dev (test) split for evaluation. The complementary CoVoST dev (test) split is useful in the multi-speaker evaluation (Wang et al., 2020) to analyze model robustness, but large amount of repeated sentences (e.g. on English and German) may skew the overall BLEU (WER) scores.

2.3 Statistics

Basic statistics of CoVoST 2 are listed in Table 1, including speech duration, speaker counts as well as token counts for both transcripts and translations. As we can see, CoVoST 2 is diversified with large sets of speakers even on some of the low-resource languages (e.g. Persian, Welsh and Dutch). Moreover, they are distributed widely across 66 accent groups, 8 age groups and 3 gender groups.

3 Models

Our speech recognition (ASR) and ST models share the same BLSTM-based encoder-decoder architecture (Bérard et al., 2018), which is similar to the Listen, Attend and Spell (LAS) architecture (Chan et al., 2016; Chiu et al., 2017; Park et al., 2019). Specifically, on the encoder side, audio features $\mathbf{x} \in \mathbb{R}^{T \times d_0}$ are first fed into a two-layer DNN with $tanh$ activations and hidden

sizes d_1 and d_2. Then two 2D convolutional layers with kernel size 3x3 and stride 2x2 are applied to reduce the sequence length to $\frac{T}{4}$. Both convolutional layers have 16 output channels and project the features to $4d_2$ dimensions after flattening. Finally, the features are passed to a stack of l_e bidirectional LSTM layers of hidden size d_3 to form encoder output states $\mathbf{h} \in \mathbb{R}^{T \times 2d_3}$. For the decoder side, a stack of l_d LSTM layers with hidden size $2d_3$ and additive attention (Bahdanau et al., 2014) is applied, followed by a linear projection to size d_o. In the multilingual setting (En→All and All→All), we follow Inaguma et al. (2019) to force decoding into a given language by using a target language ID as the first token.

For MT, we use a Transformer *base* architecture (Vaswani et al., 2017) with l_e encoder layers, l_d decoder layers, 0.3 dropout, and shared embeddings for encoder/decoder inputs and decoder outputs. For multilingual models, encoders and decoders are shared as preliminary experimentation showed that this approach was competitive.

4 Experiments

We provide MT, cascaded ST and end-to-end ST baselines under bilingual settings as well as multilingual settings: All→En (A2E), En→All (E2A)

and All→All (A2A). Similarly for ASR, we provide both monolingual and multilingual baselines.

4.1 Experimental Settings

For all texts, we normalize the punctuation and build vocabularies with Sentence-Piece (Kudo and Richardson, 2018) without pre-tokenization. For ASR and ST, character vocabularies with 100% coverage are used. For bilingual MT models, BPE (Sennrich et al., 2016) vocabularies of size 5k are learned jointly on both transcripts and translations. For multilingual MT models, BPE vocabularies of size 40k are created jointly on all available source and target text. For MT and language pair a-b, we also contrast using only a-b training data and both a-b and b-a training data (we also remove any overlap between training data from b-a and development or test set from a-b; this is also done for the A2A multilingual MT setting). The latter setting is referred to as +Rev subsequently.

We extract 80-channel log-mel filterbank features (windows with 25ms size and 10ms shift) using Kaldi (Povey et al., 2011), with per-utterance cepstral mean and variance normalization applied. We remove training samples having more than 3,000 frames or more than 512 characters for GPU memory efficiency.

For ASR and ST, we set $d_1 = 256$, $d_2 = 128$, $d_3 = 512$ and $d_o = 128$. We use $l_e = 3$ and $l_e = 2$ for bilingual models and $l_e = 5$ and $l_e = 3$ for multilingual models. We adopt SpecAugment (Park et al., 2019) (LB policy without time warping) to alleviate overfitting. To accelerate model training, we pre-train all non-English ASR and all ST models with English ASR model encoder. For MT, we set $l_e = l_d = 3$ for bilingual models and $l_e = l_d = 6$ for multilingual models. All models are implemented in Fairseq (Ott et al., 2019).

We use a beam size of 5 for all models and length penalty 1. We use the best checkpoint by validation loss for MT, and average the last 5 checkpoints for ASR and ST. For MT and ST, we report case-sensitive detokenized BLEU (Papineni et al., 2002) using sacre-BLEU (Post, 2018) with default options, except for English-Chinese and English-Japanese where we report character-level BLEU. For ASR, we report character error rate (CER) on Japanese and Chinese (no word segmentation) and word error rate (WER) on the other languages using VizSeq (Wang et al., 2019). Before calculating WER (CER), sentences are tokenized by sacre-BLEU tokenizers, lowercased and with punctuation removed (except for apostrophes and hyphens).

4.2 Monolingual and Bilingual Baselines

Table 2 reports monolingual baselines for ASR and bilingual MT, cascaded ST (C-ST) and end-to-end ST baselines. As expected, the quality of transcriptions and translations is very dependent on the amount of training data per language pair. The poor results obtained on low resource pairs can be improved by leveraging training data from the opposite direction for MT and C-ST. These results serve as baseline for the research community to improve upon, including methods such as multilingual training, self-supervised pre-training and semi-supervised learning.

4.3 Multilingual Baselines

A2E, E2E and A2A baselines are reported in Table 3 for language pairs into English and in Table 4 for language pairs out of English. Multilingual modeling is shown to be a promising direction for improving low-resource ST.

5 Conclusion

We introduced CoVoST 2, the largest speech-to-text translation corpus to date for language coverage and total volume, with 21 languages into English and English into 15 languages. We also provided extensive monolingual, bilingual and multilingual baselines for ASR, MT and ST. CoVoST 2 is free to use under CC0 license and enables the research community to develop methods including, but not limited to, massive multilingual modeling, ST modeling for low resource languages, self-supervision for multilingual ST, semi-supervised modeling for multilingual ST.

References

Rosana Ardila, Megan Branson, Kelly Davis, Michael Kohler, Josh Meyer, Michael Henretty, Reuben Morais, Lindsay Saunders, Francis Tyers, and Gregor Weber. 2020. Common voice: A massively-multilingual speech corpus. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 4218–4222, Marseille, France. European Language Resources Association.

	Fr	De	Es	Ca	Nl	Tr	Ar	Sv	Lv	Sl	Ta	Ja	Id	Cy
Bi. ST	23.2	15.7	20.2	17.9	2.2	2.2	2.7	1.4	1.2	1.5	0.2	1.1	1.0	1.7
ASR[†]	19.4	21.1	16.3	12.6	42.4	52.3	80.1	76.5	75.6	68.6	92.6	56.8	72.4	66.7
+FT[‡]					36.2	40.8	62.6	68.7	68.2	52.8	86.0	47.9	63.6	53.6
+TS[*]	28.9	31.7	23.8	20.2	44.2	49.8	78.0	72.3	71.4	60.0	100.9	55.9	69.2	60.0
A2E MT[1]	38.0	27.0	38.2	29.8	13.5	9.2	17.3	22.0	10.2	9.3	1.1	6.3	18.7	10.0
A2A MT[2]	40.9	31.7	41.0	32.4	19.0	12.1	17.9	27.0	11.8	9.5	1.0	6.5	23.5	14.1
† + 1	28.4	20.3	28.8	25.1	8.0	6.5	5.5	4.4	3.9	3.5	0.3	1.6	3.7	2.6
‡ + 1					8.7	7.3	8.7	6.4	4.2	5.8	0.6	2.4	6.2	5.1
* + 1	23.4	16.7	25.1	22.3	7.5	6.6	6.0	5.8	4.0	4.1	0.2	1.8	5.6	4.2
† + 2	30.1	23.0	30.6	26.9	10.1	7.1	5.1	4.8	3.6	3.0	0.3	1.8	4.7	3.5
‡ + 2					11.5	9.0	8.7	7.4	4.5	5.1	0.4	2.7	7.5	6.9
* + 2	24.4	19.0	26.8	23.7	9.8	7.2	5.4	6.4	4.1	4.2	0.2	1.7	7.2	5.6
A2E ST	26.6	19.5	26.3	23.5	8.6	2.1	0.3	1.3	0.6	1.4	0.1	0.6	0.3	0.9
+FT					8.3	4.7	1.5	1.5	0.9	2.2	0.2	0.7	1.1	1.7
+TS[*]	26.5	19.7	27.7	25.8	7.0	4.0	1.5	1.3	1.3	1.7	0.1	0.5	0.5	2.5
A2A ST	16.1	9.6	14.5	13.8	1.5	0.3	0.4	0.4	0.4	0.5	0.1	0.2	0.3	0.4
+FT					8.0	4.0	1.2	2.2	1.7	1.3	0.2	0.6	1.1	1.5
+TS[*]	17.1	10.5	17.3	15.7	5.6	2.0	0.9	1.9	1.0	1.3	0.1	0.4	0.8	1.3

Table 3: CV test WER for multilingual ASR, and CV test BLEU for multilingual MT and ST on high-resource (Fr, De, Es and Ca) as well as low-resource X→En. [†] A multilingual model trained on all 22 languages. [‡] Finetuned jointly on all low-resource languages: Nl, Tr, Ar, Sv, Lv, Sl, Ta, Ja, Id and Cy. [*] With temperature-based (T=2) data sampling (Arivazhagan et al., 2019) to improve low-resource directions.

	De	Ca	Zh	Fa	Et	Mn	Tr	Ar	Sv	Lv	Sl	Ta	Ja	Id	Cy
Bi. ST	12.5	17.1	20.0	9.1	9.3	6.4	6.7	9.1	18.1	8.7	11.6	7.4	25.6	15.2	18.9
ASR[†]								27.8							
E2A MT[1]	31.9	41.6	40.9	22.2	27	19.1	21.3	23.5	41.2	26.1	32.2	24.5	45.6	40.9	43.1
A2A MT[2]	30.8	40.2	39.0	21.1	25.7	18.4	20.4	21.9	40.1	24.6	30.2	23.4	44.9	39.9	41.6
† + 1	19.0	24.5	26.1	14.0	15.3	11.6	11.6	14.1	24.4	15.4	18.6	13.5	33.1	22.7	25.0
† + 2	18.4	23.8	25.1	13.6	14.6	11.3	11.2	13.3	23.7	14.6	17.7	13.0	32.6	22.2	24.1
E2A ST	12.6	17.7	22.2	9.1	9.5	6.3	7.3	8.0	18.3	8.9	11.4	7.3	28.2	16.0	19.3
A2A ST	12.4	18.1	22.2	9.5	9.5	6.4	7.2	8.4	18.6	8.8	11.4	7.2	28.3	15.9	19.1

Table 4: CV Test WER for multilingual ASR and CV test BLEU for multilingual MT/ST on En→X (all directions have equal resource). [†] A multilingual model trained on all 22 languages.

Naveen Arivazhagan, Ankur Bapna, Orhan Firat, Dmitry Lepikhin, Melvin Johnson, Maxim Krikun, Mia Xu Chen, Yuan Cao, George Foster, Colin Cherry, Wolfgang Macherey, Zhifeng Chen, and Yonghui Wu. 2019. Massively multilingual neural machine translation in the wild: Findings and challenges.

Mikel Artetxe and Holger Schwenk. 2019. Margin-based parallel corpus mining with multilingual sentence embeddings. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3197–3203, Florence, Italy. Association for Computational Linguistics.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Alexandre Bérard, Laurent Besacier, Ali Can Ko-cabiyikoglu, and Olivier Pietquin. 2018. End-to-end automatic speech translation of audiobooks. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6224–6228. IEEE.

Alexandre Berard, Olivier Pietquin, Christophe Servan, and Laurent Besacier. 2016. Listen and translate: A proof of concept for end-to-end speech-to-text translation. In *Proceedings of the 2016 NeurIPS Workshop on End-to-end Learning for Speech and Audio Processing*.

William Chan, Navdeep Jaitly, Quoc Le, and Oriol Vinyals. 2016. Listen, attend and spell: A neural network for large vocabulary conversational speech recognition. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4960–4964. IEEE.

Chung-Cheng Chiu, Tara N. Sainath, Yonghui Wu, Ro-

hit Prabhavalkar, Patrick Nguyen, Zhifeng Chen, Anjuli Kannan, Ron J. Weiss, Kanishka Rao, Ekaterina Gonina, Navdeep Jaitly, Bo Li, Jan Chorowski, and Michiel Bacchiani. 2017. State-of-the-art speech recognition with sequence-to-sequence models.

M. A. Di Gangi, M. Negri, and M. Turchi. 2019. One-to-many multilingual end-to-end speech translation. In *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 585–592.

Mattia A. Di Gangi, Roldano Cattoni, Luisa Bentivogli, Matteo Negri, and Marco Turchi. 2019. MuST-C: a Multilingual Speech Translation Corpus. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2012–2017, Minneapolis, Minnesota. Association for Computational Linguistics.

H. Inaguma, K. Duh, T. Kawahara, and S. Watanabe. 2019. Multilingual end-to-end speech translation. In *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 570–577.

J. Iranzo-Snchez, J. A. Silvestre-Cerd, J. Jorge, N. Rosell, A. Gimnez, A. Sanchis, J. Civera, and A. Juan. 2020. Europarl-st: A multilingual corpus for speech translation of parliamentary debates. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8229–8233.

Taku Kudo and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium. Association for Computational Linguistics.

Nathan Ng, Kyra Yee, Alexei Baevski, Myle Ott, Michael Auli, and Sergey Edunov. 2019. Facebook FAIR's WMT19 news translation task submission. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 314–319, Florence, Italy. Association for Computational Linguistics.

Pedro Javier Ortiz Suárez, Laurent Romary, and Benoît Sagot. 2020. A monolingual approach to contextualized word embeddings for mid-resource languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1703–1714, Online. Association for Computational Linguistics.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of NAACL-HLT 2019: Demonstrations*.

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 on association for computational linguistics*, pages 311–318. Association for Computational Linguistics.

Daniel S Park, William Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Ekin D Cubuk, and Quoc V Le. 2019. Specaugment: A simple data augmentation method for automatic speech recognition. *arXiv preprint arXiv:1904.08779*.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, et al. 2011. The kaldi speech recognition toolkit. In *IEEE 2011 workshop on automatic speech recognition and understanding*. IEEE Signal Processing Society.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

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

Changhan Wang, Anirudh Jain, Danlu Chen, and Jiatao Gu. 2019. Vizseq: A visual analysis toolkit for text generation tasks. *EMNLP-IJCNLP 2019*, page 253.

Changhan Wang, Juan Pino, Anne Wu, and Jiatao Gu. 2020. CoVoST: A diverse multilingual speech-to-text translation corpus. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 4197–4203, Marseille, France. European Language Resources Association.