Kuvost: A Large-Scale Human-Annotated English to Central Kurdish Speech Translation Dataset Driven from English Common Voice

Mohammad Mohammadamini[1], Daban Jaff[2,3], Sara Jamal[3], Ibrahim Ahmed[3], Hawkar Omar[3], Darya Sabir[3], Marie Tahon[1], Antoine Laurent [1]
[1]LIUM, Le Mans University, [2]Erfurt University, [3]Koya University
Correspondence: mohammad.mohammadamini@univ-lemans.fr

Abstract

In this paper, we introduce the Kuvost, a large-scale English to Central Kurdish speech-to-text-translation (S2TT) dataset. This dataset includes 786k utterances derived from Common Voice 18, translated and revised by 230 volunteers into Central Kurdish. Encompassing 1,003 hours of translated speech, this dataset can play a groundbreaking role for Central Kurdish, which severely lacks public-domain resources for speech translation. Following the dataset division in Common Voice, there are 298k, 6,226, and 7,253 samples in the train, development, and test sets, respectively. The dataset is evaluated on end-to-end English-to-Kurdish S2TT using Whisper V3 Large and SeamlessM4T V2 Large models. The dataset is available under the Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License `https://huggingface.co/datasets/aranemini/kuvost`.

1 Introduction

Speech translation is the automatic conversion of audio from a source language into text or audio in a target language (Barrault et al., 2025). Developing a speech-to-text translation system requires large amounts of translated audio; however, most languages lack sufficient data in this area. In (Communication et al., 2023), languages with fewer than 1,000 hours of publicly available transcribed or translated data are classified as low-resource. By this definition, only about a dozen out of 7,000 languages qualify as high-resource. Providing speech translation data—especially for low-resource languages—is therefore crucial for progress in this field.

In this paper, we introduce a speech-to-text translation (S2TT) dataset for Central Kurdish (CKB), which is a low-resource language (Communication et al., 2023). This dataset, called **Kuvost** (Kurdish Common Voice Speech Translation), is derived from the Common Voice 18 dataset. The Kuvost dataset contains 247k unique sentences translated by 230 volunteers and passed through a systematic revision process. Due to multiple recordings for some of the translated sentences in Common Voice, the total audio duration in Kuvost amounts to 1,003 hours.

Extending automatic speech recognition datasets by translating their transcriptions is a common strategy for building speech translation corpora. CoVoST (Wang et al., 2020) and CoVoST 2 (Wang et al., 2021b) are two well-known examples, both derived from Common Voice. CoVoST 2 is currently one of the large-scale publicly available speech translation corpus, including English-to-15 languages and 21-to-English S2TT pairs (Wang et al., 2021b).

Aug-LibriSpeech is a French-translated version of the LibriSpeech corpus, comprising a 236-hour EN→FR S2TT data (Kocabiyikoglu et al., 2018). VoxPopuli is a multi-way speech translation corpus based on European Parliament (EP) event recordings, encompassing 15 European languages (Wang et al., 2021a). TED Talks and TEDx have also been widely used for speech translation. The MUST-C dataset contains English-to-14-language S2TT data derived from TED Talks (Gangi et al., 2019; Cattoni et al., 2021). TEDx includes translations from English to 7 languages, while Indic-TEDST is the third TED-derived corpus, featuring translations from English to 9 Indian languages (Salesky et al., 2021; Sethiya et al., 2024).

The FLEURS dataset is currently the most comprehensive speech translation dataset in terms of the number of covered languages. FLEURS is a multi-way text-to-text and speech-to-speech corpus for 101 languages. It is also the only speech translation dataset that includes the Central Kurdish language—the subject of the current research (Conneau et al., 2023). The goal of this paper is to fill the gap in speech translation data scarcity for the Central Kurdish language.

Proceedings of the 22nd International Conference on Spoken Language Translation (IWSLT 2025), pages 106–109
July 31 - August 1, 2025 ©2025 Association for Computational Linguistics

2 Kurdish language

Kurdish (ISO 639: KUR) is an Indo-European language spoken by more than 30 million native speakers in Kurdistan and among the Kurdish diaspora. Geographic dispersion and socio-political factors have led Kurdish to diversify into several dialects (Matras, 2019; Eppler and Benedikt, 2017). The Kurdish language comprises six dialects: Northern Kurdish (KMR), Central Kurdish (CKB), Southern Kurdish (SDH), Laki (LKI), Zaza (DIQ), and Hawrami (HAQ) (Sheyholislami, 2015). Northern Kurdish and Zazaki are primarily written in a Latin-based script, while the remaining dialects are mainly written in an Arabic-based script.

In this paper, we focus on Central Kurdish, which is spoken by nearly 8 million native speakers (Sheyholislami, 2015). Central Kurdish is a statutory national language in Iraq[1] and a de facto provincial working language in Iran[2]. Although recent years have seen notable progress in data curation and system development for Central Kurdish, the speech translation domain from/to this dialect remains largely unexplored. The goal of this paper is to address this gap.

3 Translation process

The data for translation was sourced from Common Voice 18 [3]. The data creation process consisted of three main steps: (1) announcement and recruitment of volunteers, (2) training and data distribution, and (3) translation and review.

Announcement and Recruitment of Volunteers: An announcement was made in June 2024 to recruit volunteers who study English at the Department of English Language(DENL) at Koya University. A total of 259 volunteers—mostly third- and fourth-year students—signed up. All volunteers were native speakers of Central Kurdish.

Training: A two-week intensive training program was then offered to the volunteers. During this program, participants were introduced to various translation techniques. Additionally, a detailed guideline outlining the rules of translation was provided. After the training workshops, volunteers were given the option to withdraw without providing a reason. At this stage, 17 volunteers dropped out. The remaining participants were divided into three main groups, each supervised by a faculty member. These were further divided into smaller sub-groups of five volunteers. The translation data was distributed via Google Sheets, with access provided to both volunteers and supervisors.

Revision: The review process involved two main stages:

- **Peer Review:** Volunteers reviewed each other's translations within their sub-groups.

- **Professional Review:** Each translation was subsequently reviewed first by a team of Kurdish language experts from Department of Kurdish Language (DKUR) at Koya University and professional supervisor, who provided feedback and suggested edits where necessary.

Furthermore, weekly seminars were also held to address common mistakes and discuss correction strategies. During the revision phase, an additional 12 volunteers dropped out. By the end of the process, a total of 230 volunteers, plus 7 Kurdish language reviewers, had fully or partially completed their tasks, translating 247,373 sentences.

4 Kuvost Statistics

The statistics of the Kuvost dataset are presented in Table 1. The number of unique sentences translated into Kurdish is 247,373. These translated sentences were matched with their corresponding transcriptions and audio in Common Voice 18. We searched for all matching utterances in the validated portion of Common Voice 18, resulting in 786k utterances with Kurdish translations, totaling approximately 1,003 hours of English audio.

The validated and translated utterances were divided into train, development, and test sets according to the original Common Voice 18 partitioning. For each split, we referred to the Common Voice 18 train/dev/test sets and matched the English transcriptions with their corresponding Kurdish translations. The training set includes 298k utterances, equivalent to 417 hours of audio. The development and test sets each contain approximately 9 hours of translated speech. It deserved to be mentioned that all validated examples in the Common Voice are not included in the train/dev/test partitions which leads to lower number of utterances in the partitions.

[1] https://www.ethnologue.com/country/IQ/

[2] https://www.ethnologue.com/country/IR/

[3] https://commonvoice.mozilla.org/en/datasets

Table 1: Kuvost specification and partitions

Part	Train	Dev	Test	Validated
Duration	417h	8h47m	8h55m	1003h
Utterances	298k	6226	7253	786k
Uniq sents	190k	5819	7149	247k
Tokens	1,75m	41k	46k	1.84m

5 Evaluation Systems

The Kuvost dataset is evaluated by fine-tuning two state-of-the-art speech translation models: Whisper V3 (WL V3) Large and SeamlessM4T V2 (SL V2) Large models.

5.1 Whisper Large V3

Whisper is a sequence-to-sequence transformer-based model trained on 680,000 hours of labeled speech data, encompassing tasks such as ASR, S2TT, VAD, and Speaker Recognition (SR) (Radford et al., 2022). Whisper supports more than 80 languages for ASR and S2TT; however, the Kurdish language is not currently supported. We are fine-tuning the Whisper V3 Large model using the AdamW optimizer with a learning rate of 1e-5, a batch size of 16, in 5 epochs.

5.2 SeamlessM4T Large V2

Seamless is a set of models for T2TT, S2TT, S2ST, and ASR. We use the S2TT component, which consists of a Wav2Vec-BERT speech encoder and an NLLB-200 decoder. The model is jointly optimized for ASR and S2TT tasks (Barrault et al., 2025; Communication et al., 2023). We fine-tune the SeamlessM4T V2 Large model using Mel-filter bank (bins = 80) features over 10 epochs, with a batch size of 16 and a learning rate of 1e-4. These hyperparameters are set experimentally.

6 Results and discussion

Throughout the experiments, Kurdish translations were normalized using the Asosoft normalizer (Mahmudi et al., 2019). Key normalization steps included the unification of Unicode characters, standardization of numbers, and normalization of punctuation marks.

The Kuvost dataset was evaluated using two state-of-the-art (SOTA) models: Whisper Large V3 and SeamlessM4T V2 Large. Table 2 presents the results obtained using both models. The first row shows the performance of the fine-tuned Whisper V3 Large model on the training set of Kuvost. This model achieved a BLEU score of 23.76 on the Kuvost test set and 26.01 on the development set. The second row, labeled SL V2, displays the results of the pretrained SeamlessM4T V2 Large model before fine-tuning on the Kuvost training set. This multilingual model supports speech-to-text translation for 101 languages, including Central Kurdish. The baseline model of Seamless achieved a BLEU score of 21.97 on the Kuvost test set. The final row presents results for the fine-tuned version of SeamlessM4T using the Kuvost dataset. In this experiment, the model achieved a significantly improved BLEU score of 35.00 and 32.79 on the dev and test sets respectively. Besides the BLUE score, the ChrF++ is reported for all models. The fine-tuned version of seamless achieves a ChrF++ score of 62.32 on the Kuvost test set.

Table 2: Kuvost evaluation results using Whisper V3 Large (WL V3) and SeamlessM4T V2 Large (SL V2) models. FT stands for fine-tuned model on the train part of Kuvost

Part	Dev		Test	
	BLEU	ChrF++	BLEU	ChrF++
WL V3 FT	26.01	55.14	23.76	51.77
SL V2	22.54	54.18	21.97	53.01
SL V2 FT	35.00	64.10	32.79	62.32

The fine-tuned models on the Kuvost training set were evaluated using the FLEURS benchmark. The results are presented in Table 3. The Whisper model achieved a BLEU score of 7.65, and the Seamless model obtained a BLEU score of 11.17. The baseline SeamlessM4T model (before fine-tuning) achieved a BLEU score of 9.36 on the English→Central Kurdish task. Fine-tuning on the Kuvost training set led to an improvement of nearly 2 BLEU points, reaching 11.17. The marginal improvement in BLEU on the FLEURS dataset is likely due to differences in sentence complexity. Kuvost primarily consists of short and simple sentences, while FLEURS includes more complex syntactic structures. Additionally, domain shift may have contributed to the limited performance gain.

Table 3: The generaliability of Models fine-tuned on Kuvost and evaluated on FLEURS benchmark

Fleurs	BLEU	ChrF++
Whisper V3 FT	7,65	39,69
SeamlessM4T V2 FT	11.17	46,46

7 Conclusion

In this paper, we introduced Kuvost, a large-scale, human-annotated speech translation dataset for Central Kurdish. Kuvost consists of 1,003 hours of English-to-Kurdish speech translation, contributed by 230 volunteers. The dataset is evaluated using state-of-the-art speech translation models. For future work, we plan to record Kurdish translations to extend Kuvost for speech-to-speech translation tasks. Additionally, we aim to expand the dataset to support Kurdish-to-X translation for all languages available in the CoVoST 2 dataset (Wang et al., 2021b).

Acknowledgments

We express our heartfelt gratitude to the 236 volunteers from the English Language Department (DENL) and the Department of Kurdish Language (DKUR) at Koya University for their unwavering dedication and hard work throughout the ten months, from June 2024 to April 2025. Their active involvement has played a crucial role in the success of this project. Daban Q. Jaff thanks the Deutscher Akademischer Austauschdienst for his doctoral research grant. The experiments was performed using HPC resources from GENCI–IDRIS in the framework of DGA RAPID COMMUTE project.

References

Loïc Barrault, Yu-An Chung, Mariano Coria Meglioli, and et. al. 2025. Joint speech and text machine translation for up to 100 languages. *Nature*, 637(8046):587–593.

Roldano Cattoni, Mattia Antonino, Luisa Bentivogli, Matteo Negri, and Marco Turchi. 2021. Must-c: A multilingual corpus for end-to-end speech translation. *Computer Speech and Language*, 66:101155.

Seamless Communication, Loïc Barrault, Yu-An Chung, Mariano Coria Meglioli, and et. al. 2023. Seamless: Multilingual expressive and streaming speech translation. *Preprint*, arXiv:2312.05187.

Alexis Conneau, Min Ma, Simran Khanuja, Yu Zhang, Vera Axelrod, Siddharth Dalmia, Jason Riesa, Clara Rivera, and Ankur Bapna. 2023. Fleurs: Few-shot learning evaluation of universal representations of speech. In *SLT*, pages 798–805.

E. D. Eppler and S. Benedikt. 2017. Contact-induced language change in kurdish. In E. D. Eppler and S. Benedikt, editors, *Languages in Contact: A Comprehensive Guide*, pages 345–362. John Benjamins Publishing Company, Amsterdam.

Mattia A. Di Gangi, Roldano Cattoni, Luisa Bentivogli, Matteo Negri, and Marco Turchi. 2019. Must-c: A multilingual speech translation corpus. pages 2012–2017, Minneapolis, Minnesota. Association for Computational Linguistics.

Ali Can Kocabiyikoglu, Laurent Besacier, and Olivier Kraif. 2018. Augmenting librispeech with french translations: A multimodal corpus for direct speech translation evaluation. In *LREC 2018*, Miyazaki, Japan. European Language Resources Association (ELRA).

Aso Mahmudi, Hadi Veisi, Mohammad MohammadAmini, and Hawre Hosseini. 2019. Automated kurdish text normalization.

Y. Matras. 2019. Kurdish linguistics: A brief overview. In Y. Matras and D. Everhard, editors, *Kurdish Linguistics: Focus on Variation and Change*, pages 1–20. De Gruyter Mouton, Berlin.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. Robust speech recognition via large-scale weak supervision. *Preprint*, arXiv:2212.04356.

Elizabeth Salesky, Matthew Wiesner, Jacob Bremerman, Roldano Cattoni, Matteo Negri, Marco Turchi, Douglas W. Oard, and Matt Post. 2021. Multilingual tedx corpus for speech recognition and translation. In *Proceedings of Interspeech*.

Nivedita Sethiya, Saanvi Nair, and Chandresh Maurya. 2024. Indic-tedst: Datasets and baselines for low-resource speech to text translation. In *LREC-COLING 2024*, pages 9019–9024, Torino, Italia. ELRA and ICCL.

Jaffer Sheyholislami. 2015. *The Kurds: History, Religion, Language, Politics*, chapter Language Varieties of the Kurds. Austrian Federal Ministry of the Interior.

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

Changhan Wang, Morgane Riviere, Ann Lee, Anne Wu, Chaitanya Talnikar, Daniel Haziza, Mary Williamson, Juan Pino, and Emmanuel Dupoux. 2021a. Voxpopuli: A large-scale multilingual speech corpus for representation learning, semi-supervised learning and interpretation. pages 993–1003, Online. Association for Computational Linguistics.

Changhan Wang, Anne Wu, Jiatao Gu, and Juan Pino. 2021b. Covost 2 and massively multilingual speech translation. In *Proc. Interspeech 2021*, pages 2247–2251.