FLEURS-Badini: Translation and Recording FLEURS Dataset for Badini Variant of Northern Kurdish

Mohammad Mohammadamini[1], Dilgash Mohammed Salih Tayib[2], Dezheen H. Abdulazeez[2], Barzan Hussein Mohammed[2], Imad Saeed Sadeeq[2], Aveen Jalal Mohammed[2], Amera Ismail Melhum[2], Abuobaida Abdullah Dheyab[2]

[1]LIUM, Le Mans University
[2]University of Duhok

Correspondence: mohammad.mohammadamini@univ-lemans.fr

Abstract

Multilingual speech benchmarks such as the FLEURS benchmark have significantly advanced research across a wide range of languages. However, many languages and dialects, including Badini Kurdish, remain underrepresented, limiting bechmarking in automatic speech recognition (ASR) and speech-to-text translation (S2TT). To address this limitation, this study introduces FLEURS-Badini, a dialect-focused extension designed to support research on Badini variant of Northern Kurdish. The dataset is constructed through a structured process of translation, recording, and validation, resulting in 5,224 utterances paired with their corresponding translated text. The data were collected from 45 speakers. To evaluate the dataset, baseline experiments are conducted using state-of-the-art models for both ASR and S2TT. The results indicate that ASR remains challenging, with the best performance achieved by the W2V-BERT CTC model, reaching a Word Error Rate (WER) of 55% on the test set. Similarly, speech-to-text translation performance is limited, with BLEU scores 6.13 and 5.24 on dev and test sets. The dataset is available under CC-BY-NC-ND-4.0 license [1].

Keywords: Automatic Speech Recognition, Speech Translation, Low-Resource Languages, Badini Kurdish, Multilingual Benchmark.

1 Introduction

The recent advancement in speech technologies—particularly in speech translation and automatic speech recognition—has been largely enabled by the growing availability of large-scale multilingual datasets (Ardila et al., 2020; Baevski et al., 2020; Radford et al., 2023; Vaswani et al., 2017). Nevertheless, this progress has not been evenly distributed, as many low-resource languages and dialects still lack sufficient data to support reliable model development and evaluation. Kurdish, a language with significant dialectal variation and limited resources and standardized datasets, is an ideal example of this limitation.

The development of datasets that integrate both parallel text translations and corresponding speech recordings is essential for enabling multilingual speech processing, especially speech translation tasks. Such datasets provide aligned speech–text pairs that allow models to learn the relationship between spoken input and translated output. In this regard, the FLEURS benchmark was introduced as a large-scale speech dataset including over one hundred languages (Conneau et al., 2022). FLEURS extends the FLORES machine translation benchmark into the speech domain by offering n-way parallel speech data aligned across languages. It creates the conditions for more systematic and comparable evaluation across a range of speech processing tasks, including ASR, S2TT, and speech-to-speech translation (S2ST) (Goyal et al., 2021; Conneau et al., 2022; Popel and Bojar, 2020).

Despite its broad multilingual coverage, several linguistically important language varieties are not included in FLEURS. In the case of Kurdish, resource development has been uneven across its dialects. Central Kurdish has received relatively greater attention in both ASR and speech translation research, supported by curated datasets and multilingual initiatives such as Common Voice (Ardila et al., 2020; Veisi et al., 2022; Zenkel et al., 2020). More recently, FLEURS-Kobani (Jaff and Mohammadamini, 2026) has been introduced to represent Northern Kurdish within the FLEURS; nevertheless, major variants of Kurdish language, particularly Badini, remain absent.

Badini is a widely spoken variant of Northern Kurdish in the Kurdistan Region of Iraq. It is distinguished from other Kurdish varieties by its distinct phonological, lexical, and syntactic features .

[1]https://huggingface.co/datasets/BadiniSpeechNLP/FLEURS-badini

Proceedings of the 23rd International Conference on Spoken Language Translation (IWSLT 2026), pages 119–123

These differences introduce additional challenges for speech processing systems, particularly in low-resource settings where both high-quality speech recordings and reliable translated text data are limited or inconsistent. The development of robust ASR models and, more importantly, speech translation systems for this dialect is severely restricted by the lack of a resources that jointly provides accurate translations and aligned speech.

To address this gap, this paper introduces FLEURS-Badini, a dialect-focused extension of the FLEURS benchmark designed to support both speech recognition and speech translation tasks. The dataset is constructed through a controlled process of translation, recording, and validation, ensuring both linguistic accuracy and audio quality. It consists of 5,224 utterances paired with their corresponding translated text.

1.1 Northern Kurdish (Badini)

Kurdish is an Indo-European language spoken across the Kurdistan region, including Iraq, Iran, Turkey, and Syria, as well as by large diaspora communities. The number of native speakers is commonly estimated to exceed 30 million (Öpengin, 2021). Due to its wide geographic distribution and complex socio-political history, Kurdish exhibits substantial internal variation and is often characterized as a dialect continuum (Eppler and Benedikt, 2017) or a macrolanguage (Sheyholislami, 2015).

Linguistic classifications typically divide Kurdish into several major groups, including Northern Kurdish (Kurmanji, ISO 639-3: KMR), Central Kurdish (Sorani, CKB), and Southern Kurdish (SDH), alongside related varieties such as Zazaki and Hewrami (Haig and Öpengin, 2014). These dialects differ not only in phonology and vocabulary but also in orthographic conventions shaped by geopolitical contexts. Northern Kurdish is generally written using a Latin-based script, whereas Central and Southern Kurdish are predominantly written in Arabic-based scripts (Sheyholislami, 2015).

The focus of this study is the Badini dialect, a regional variety of Northern Kurdish primarily spoken in the Duhok governorate of the Kurdistan Region of Iraq. Although Badini belongs to the Northern Kurdish group, it demonstrates distinct phonetic, lexical, and orthographic characteristics that set it apart from standard Northern Kurdish (Öpengin, 2021). In particular, Badini is commonly written using a modified Arabic script rather than the Latin script typically associated with Northern Kurdish, creating a divergence between spoken forms and written representations.

These differences introduce significant challenges for speech and language technologies. Variations in pronunciation, vocabulary, and script representation reduce the effectiveness of models trained on other Kurdish dialects. Moreover, transliteration between Latin and Arabic scripts often results in inconsistencies and loss of phonetic information, further complicating automatic speech recognition tasks (Mohammadamini et al., 2026).

Despite being widely spoken, Badini remains under-resourced in computational linguistics. Existing resources are limited in both scale and diversity, and there is a lack of standardized datasets tailored specifically to this dialect. This scarcity restricts the development of robust ASR and speech translation systems and highlights the need for dedicated benchmark datasets.

1.2 Related Work

In recent years, several efforts have been made to create multilingual speech corpora. Projects such as Common Voice have made major contributions through crowd-sourced recording and transcription (Ardila et al., 2020; Pratap et al., 2020). The FLEURS benchmark provides a dataset consisting of translated sentences aligned with corresponding speech recordings, enabling systematic evaluation under low-resource and few-shot learning conditions (Conneau et al., 2022).

Despite these advancements, the creation of such datasets remains inconsistent across languages and dialects. Most existing Kurdish efforts have focused on Central Kurdish. A significant step forward in Kurdish speech recognition research is the Asosoft dataset (Veisi et al., 2022). It represents one of the first large-vocabulary ASR systems for Central Kurdish and introduces a phonetically balanced speech corpus along with a pronunciation lexicon. The FLEURS dataset includes Central Kurdish (Conneau et al., 2022), and in (Jaff and Mohammadamini, 2026) has been extended to Northern Kurdish. In (Mohammadamini et al., 2025), the Kuvost dataset for Central Kurdish–English was introduced as a Kurdish counterpart of the CoVoST dataset. In (**?**), a large-scale pseudo-labeled speech dataset was introduced for Central Kurdish. Despite these efforts, the Northern Kurdish branch—and especially the Badini variant—remains largely unexplored. To the best of

our knowledge, the current work is the first effort toward developing speech translation resources for this variant.

2 Dataset Construction and Validation

2.1 Parallel Text Translation

Our dataset construction begins with a structured translation process aimed at creating a parallel English–Badini text corpus. A total of 2,000 English sentences adopted from Flores (Goyal et al., 2021) dataset, were translated into the Badini dialect of Kurdish by 50 students from the Departments of English and Translation at the University of Duhok. The translation was conducted in two stages, after which all outputs were reviewed by faculty members to ensure linguistic accuracy and consistency. All reviewers and translators involved in the dataset creation process are native speakers. The translations were performed directly from English rather than from existing Kurdish versions of FLEURS to ensure consistency with the original FLEURS data creation methodology and to facilitate the use of the dataset for Badini-to-English speech translation.

The translation process involved several linguistic challenges. Many English terms, particularly in specialized domains such as medicine, technology, and politics, do not have direct equivalents in Badini Kurdish. To address this, translators applied strategies such as explanation, borrowing, and transliteration to preserve meaning. Structural differences between English (typically Subject–Verb–Object) and Kurdish (often Subject–Object–Verb) also required careful reorganization of sentence structure to maintain clarity and naturalness.

Additional challenges included handling abbreviations and acronyms, where different strategies were used depending on familiarity, such as retaining the original form, transliteration, or expanding the term in Kurdish. Maintaining coherence across sentences, particularly in the use of pronouns and contextual references, also required careful adjustment due to differences between the two languages.

Overall, the translation process combined literal and adaptive approaches to balance fidelity to the original text with clarity in the target language.

2.2 Speech Data Acquisition

Following the translation phase, speech recordings were collected using a web-based platform. Native speakers of the Badini dialect participated in the recording process, reading the translated sentences under relatively controlled conditions. Participants were instructed to record the sentences clearly in quiet environments using personal devices. At the same time, natural variability in recording conditions—such as differences in microphones, background noise, and speaking styles—was preserved. This approach makes the dataset more representative of real-world scenarios. The resulting recordings form a parallel speech–text corpus, where each audio sample is aligned with its corresponding transcription/translation. This alignment makes the dataset suitable for both automatic speech recognition (ASR) and speech-to-text translation (S2TT) tasks.

Partition	Train	Dev	Test
Number of Utts	2,022	1,165	2,037
Duration	5h47m	3h36m	6h17m
Number of Spks	16	14	15

Table 1: Distribution of utterances and speakers across FLEURS-Badini dataset partitions

2.3 Validation and Data Quality

To ensure data quality, all recordings were manually reviewed for both textual alignment and audio clarity. Each sample was checked to confirm that the spoken content accurately matches the corresponding text. Recordings affected by issues such as lack of intelligibility, strong background noise, incomplete speech, or mismatched content were excluded. In total, 1,078 recordings were rejected during this validation process.

2.4 Data Specifications

The final dataset consists of 5,224 utterances, organized into training, development, and test partitions. The training set contains 2,022 utterances, the development set includes 1,165 utterances, and the test set comprises 2,037 utterances. The overall duration of FLEURS-Badini dataset is 15 hours and 40 minutes. In total, the dataset includes recordings from 45 speakers. There is no overlap of speakers across partitions (i.e. the speakers who are in the training partition of the dataset are distinct from the test and dev partitions). The FLEURS-Badini specification are summarized in Table 1.

3 Experiments and Results

This study evaluates both automatic speech recognition (ASR) and speech-to-text translation (S2TT)

using a range of state-of-the-art models. For the S2TT task, Whisper V3 Large model is used. In the case of ASR, several models are evaluated, including Whisper Large V3 (Radford et al., 2023), W2V-BERT CTC, and Omnilingual LLM model (1B and 3B) (OmnilingualASR et al., 2025).

3.1 ASR Results

As shown in Table 2, the ASR results highlight performance differences across the evaluated models. One of the main objectives of this evaluation is to benchmark previous work. First, the Omnilingual models are evaluated. The Omnilingual LLM 1B model low results, with WER around 60%, while the larger Omnilingual LLM 3B model performs worse, with higher error rates across both datasets. One source of errors in the Omnilingual models is language identification, as we observed that some utterances are not transcribed in Northern Kurdish (Badini) using the Arabic script. Another possible reason is the dialectal variability within Northern Kurdish.

In the second set of experiments, we evaluate the Wav2Vec-BERT CTC and Whisper V3 Large models proposed in (Mohammadamini et al., 2026), which are trained on a transliterated version of Northern Kurdish Common Voice into Arabic script, along with synthetic data. Our experiments show that Wav2Vec-BERT CTC achieves the best results, obtaining a CER of 13.69 on the development set and 14.11 on the test set. The final experiment consists of fine-tuning the Whisper model using the training split of FLEURS-Badini for five epochs. Due to the small size of the dataset, the model converges after the second epoch. With a CER of around 39% on both the development and test sets, this model performs the worst among all evaluated systems.

3.2 Speech-to-Text Translation

Table 3 presents the results for the speech-to-text translation task. In this experiment we fine-tuned Whisper V3 Large model on the train split of FLEURS-Badini. On the development set, the model achieves a BLEU score of 6.13 and a chrF++ score of 29.39. On the test set, the BLEU score decreases slightly to 5.24, while the chrF++ score remains almost unchanged at 29.57. Although the scores are modest, they provide a useful baseline for future research. More importantly, they highlight the need to expand dataset size.

4 Conclusion

This study introduced FLEURS-Badini, a dialect-specific extension of the FLEURS benchmark designed to support automatic speech recognition and speech-to-text translation for the Badini dialect of Kurdish. The dataset was constructed through a structured process of translation, recording, and validation, resulting in a parallel speech–text corpus comprising 5,224 utterances, equivalent to 15 hours and 40 minutes of speech from 45 speakers.

Baseline experiments conducted using state-of-the-art models demonstrate that speech processing for Badini remains challenging, with relatively high error rates in ASR and limited performance in speech-to-text translation. These findings highlight the impact of data scarcity and dialectal variation on model performance. Despite these challenges, the proposed dataset provides a standard benchmark for evaluation of Badini variant speech technologies. A short description of the FLEURS-Badini dataset and its role within low-resource speech translation data creation can be found in the IWSLT 2026 Findings paper (Adelani et al., 2026).

Acknowledgments

The authors would like to thank all contributors involved in data collection and validation.

References

David Ifeoluwa Adelani, Victor Agostinelli, Antonios Anastasopoulos, Luisa Bentivogli, Ondřej Bojar, Sebastien Bratières, Marine Carpuat, Roldano Cattoni, Mauro Cettolo, Lizhong Chen, Marcello Federico, Marco Gaido, Mahendra Gupta, HyoJung Han, Ali Hatami, David Javorský, Yejin Jeon, Marek Kasztelnik, Antoine Laurent, and 33 others. 2026. Speech translation and metrics in 2026: Findings of the iwslt campaign. In *Proceedings of the 23rd International Conference on Spoken Language Translation (IWSLT 2026)*, San Diego, California, US. Association for Computational Linguistics.

Rosana Ardila, Megan Branson, Kelly Davis, Michael Kohler, Josh Meyer, Reuben Morais, Lindsay Saunders, Francis M. Tyers, and Gregor Weber. 2020. Common voice: A massively multilingual speech corpus. In *Proceedings of the 12th Language Resources and Evaluation Conference (LREC)*, pages 4218–4222.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations.

Model	CER (Dev)	WER (Dev)	CER (Test)	WER (Test)
Omnilingual LLM 1B	14.81	60.06	14.75	60.20
Omnilingual LLM 3B	20.97	63.97	22.34	64.47
W2V-BERT CTC	**13.69**	**54.46**	**14.11**	**55.07**
Whisper Large V3	15.71	60.13	16.50	62.02
Baseline Whisper	39.33	70.87	39.05	72.16

Table 2: ASR results across different models

Set	BLEU	chrF++
Dev	6.13	29.39
Test	5.24	29.57

Table 3: S2TT results on development and test sets

In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pages 12449–12460.

Alexis Conneau, Ming Ma, Simran Khanuja, Yu Zhang, Vera Axelrod, Siddharth Dalmia, Marc'Aurelio Ranzato, and Veselin Stoyanov. 2022. Fleurs: Few-shot learning evaluation of universal representations of speech. In *IEEE Spoken Language Technology Workshop (SLT)*.

Eva Eppler and Birgit Benedikt. 2017. Kurdish linguistics. In *The Routledge Handbook of Kurdish Linguistics*. Routledge.

Naman Goyal, Jianfeng Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Saurabh Krishnan, Marc'Aurelio Ranzato, and Angela Fan. 2021. The flores-101 evaluation benchmark for low-resource and multilingual machine translation. In *ACL / TACL*.

Geoffrey Haig and Ergin Öpengin. 2014. Kurdish: A critical research overview. *Iranian Studies*, 47(2):191–225.

Daban Q. Jaff and Mohammad Mohammadamini. 2026. Fleurs-kobani: Extending the fleurs dataset for northern kurdish. *Preprint*, arXiv:2603.29892.

Mohammad Mohammadamini, Daban Jaff, Sara Jamal, Ibrahim Ahmed, Hawkar Omar, Darya Sabr, Marie Tahon, and Antoine Laurent. 2025. Kuvost: A large-scale human-annotated English to Central Kurdish speech translation dataset driven from English common voice. In *Proceedings of the 22nd International Conference on Spoken Language Translation (IWSLT 2025)*, pages 106–109, Vienna, Austria (in-person and online). Association for Computational Linguistics.

Mohammad Mohammadamini, Aveen Jalal Mohammed, Barzan Hussein Mohammed, Dezheen H Abdulazeez, Imad Saeed Sadeeq, Dilgash Mohammed, Salih Tayib, Amera Ismail Melhum, and Abuobaida Abdullah Dheyab. 2026. Exploring the Reusability of Northern Kurdish Resources for Badini Speech Recognition. In *DialRes-LREC26*, Palma De Mallorca, Spain. ELRA Language Resources Association.

OmnilingualASR, Gil Keren, Artyom Kozhevnikov, Yen Meng, Christophe Ropers, Matthew Setzler, Skyler Wang, Ife Adebara, Michael Auli, Can Balioglu, Kevin Chan, Chierh Cheng, Joe Chuang, Caley Droof, Mark Duppenthaler, Paul-Ambroise Duquenne, Alexander Erben, Cynthia Gao, Gabriel Mejia Gonzalez, and 14 others. 2025. Omnilingual asr: Open-source multilingual speech recognition for 1600+ languages. *Preprint*, arXiv:2511.09690.

Martin Popel and Ondřej Bojar. 2020. Transforming machine translation: The flores benchmark. In *ACL*.

Vineel Pratap, Qiantong Xu, Anuroop Sriram, Gabriel Synnaeve, and Ronan Collobert. 2020. Massively multilingual asr: 50 languages, 1 model. In *Interspeech*.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *ICML*.

Jaffer Sheyholislami. 2015. Language varieties of the kurds. In Wolfgang Taucher, Mathias Vogl, and Peter Webinger, editors, *The Kurds: History, Religion, Language, Politics*, pages 30–51. Austrian Federal Ministry of the Interior, Vienna, Austria.

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 *NeurIPS*.

Hadi Veisi, Hawre Hosseini, Mohammad MohammadAmini, Wirya Fathy, and Aso Mahmudi. 2022. Jira: A central kurdish speech recognition system, designing and building speech corpus and pronunciation lexicon. *Language Resources and Evaluation*, 56:917–941.

Thomas Zenkel, Ramon Sanabria, Florian Metze, and Alex Waibel. 2020. Blstm-hmm hybrid systems for low-resource speech recognition. In *Interspeech*.

Ergin Öpengin. 2021. The history of kurdish and the development of literary kurmanji. In Hamit Bozarslan, Cengiz Gunes, and Veli Yadirgi, editors, *The Cambridge History of the Kurds*, pages 603–632. Cambridge University Press.