Southern Kurdish Speech Recognition Resources and Benchmarking

Mohammad Mohammadamini, Marie Tahon
LIUM, Le Mans University
{first.last}@univ-lemans.fr

Abstract

This article introduces a dedicated speech recognition dataset for Southern Kurdish, which is a threatened variant of Kurdish macrolanguage. We present 30 hours of validated read speech for training and an evaluation benchmark for Southern Kurdish Automatic Speech Recognition (ASR). Both the training data and evaluation benchmark are read speech recorded by crowdsourcing campaigns. Besides a detailed description of the provided resources, we provide the ASR baselines using Whisper-turbo and wav2vec-bert CTC architectures. We achieved a 4.09 CER and 24.26 WER on our benchmark using wav2vec-bert model. We also provide a categorization of errors to support further improvements in future studies.The resources and trained models are released under the CC BY-NC-ND 4.0 license and are publicly available at Southern Kurdish ASR Corpus.

Keywords: Southern Kurdish, Speech Recognition, Low-resource, Threatened language

1. Introduction

Kurdish is a macrolanguage used for a dialect continuum spoken by more than 35 million native speakers (Sheyholislami, 2015). Due to the significant diversity of Kurdish dialects, the speech technologies developed for one dialect rarely work with an agreeable performance in other dialect. Therefore providing dialect specific speech technologies and resources is required. In this paper, we provide and evaluate 30 hours of validated read speech for Southern Kurdish and also we provide a speech recognition evaluation benchmark for this dialect of Kurdish language.

Although Northern Kurdish and Central Kurdish are the two main dialects of the Kurdish language, Southern Kurdish is considerably less represented in the media (Sheyholislami, 2008) and is not used in formal education. Consequently, this dialect exists in a severely low-resource context, which poses significant challenges for corpus development. Moreover, since Southern Kurdish never used for education, many native speakers have limited experience with reading or writing in this language, making it difficult to recruit participants for data collection such as speech recording. Likewise, obtaining written materials that can be used in various stages of developing ASR models for this dialect remains a major challenge.

The research in the Kurdish language ASR has emerged in the recent years, but the majority of works mainly focused on Central Kurdish. In Veisi et al. (2022) a phonetically balanced speech corpus and evaluation benchmark is introduced. This dataset used in later studies shows significant results for ASR (Mohammadamini et al., 2025). In Ahmadi et al. (2024) a dataset curated for Central Kurdish sub-dialects. While main commercial ASR systems rarely support Kurdish dialects, some major research projects such as Seamless model include Central Kurdish (Barrault et al., 2025). Common Voice is the largest available dataset which includes three Kurdish dialects, and in its last version (Common Voice 22[1]) includes 136 hours of Central Kurdish, 71 hours of Northern Kurdish and 2 hours of Zazaki validated speech (Ardila et al., 2020). The only effort that considers Southern Kurdish ASR is Hameed et al. (2025), which introduced 7 hours of read speech, evaluated by fine-tuning Whisper v2 base and small ASR models.

However other main Kurdish dialects and particularly Central Kurdish has substantial ASR resources, there are significant phonological, morphological and syntactical differences between Southern Kurdish and Central Kurdish (Mirmukri et al., 2019), which makes the Central Kurdish ASR systems less effective for Southern Kurdish (see Section 6). The goal of our research is to fill this gap by providing dedicated speech recognition resources for Southern Kurdish, reliable enough to develop ASR system with state-of-the-art models.

When dealing with low resource languages with dialect continuum, spontaneous speech collection for ASR might lead to dead ends due to data/speaker scarcity and high phonological diversity. Therefore, we decided to collect read speech so that we can rely on native linguists to be sure that the target dialect is the one indeed spoken. In the current study, we first provide a Southern Kurdish text collection from published magazines and books covering several domains. A list of sentences is extracted from the provided resources.

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

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 5538–5544
11–16 May 2026.

The extracted sentences are revised intensely by professional Southern Kurdish editors. The revised sentences are read and recorded using dedicated tools by native speakers. Second, we provide a multi-domain ASR evaluation benchmark for Southern Kurdish. Both training dataset and evaluation benchmark are validated and filtered manually for any possible issues.

The paper is organized as follows: In section 2 a description of Southern Kurdish is presented. In section 3 the method of providing the data is described. Section 4 describes the characteristics of the provided resources. In section 5 the models are described, and section 6 presents the results on the provided benchmark using wav2vec-bert CTC and Whisper-turbo models and a deep analysis of the transcription errors.

2. Southern Kurdish

Southern Kurdish (ISO 639 sdh) is a threatened variant of Kurdish (ISO 639 kur). It is spoken by around 6 million people (with declining number of speakers) in the southern part of Kurdistan (more precisely Ilam, Kermanshah, Hamedan and Kurdistan Provinces in Iran, Eastern parts of Diyala and Wasit provinces in Iraq) [2] (Belelli, 2019). There is also a significant population of Southern Kurdish speakers known as Feyli Kurds living in Baghdad. The region where Southern Kurdish is the predominant language is shown on Figure 1.

Southern Kurdish is considered to have several linguistic subgroups, including Garrusi, Kolyā'i, Kermānshāhi, Kalhori, Malekshāhi, Badre'i, and Kordali (Fattah, 2000). However, some researchers question the strict boundaries of this classification, especially in larger cities where speakers of different subgroups are more mixed (Belelli, 2019). Although there are some phonological and morphological differences (*e.g.*, between Kalhori and Malekshāhi) (Azadi, 2022), these variants are largely mutually intelligible (Belelli, 2019). Another variant of Kurdish language that is closely related to Southern Kurdish is Laki and its classification as Southern Kurdish remains disputed (Belelli, 2021). Therefore, it is not considered in the current research (see Figure 1).

Similar to other Kurdish dialects, the Southern Kurdish orthography is almost phonemic (*i.e.* there is one-to-one relationship between letters and phonemes) but there are some exceptions that can play an important role in the ASR domain which are as follows:

- The <ح> and <ع> letters stand respectively for /ħ/ and /ʕ/ phonemes in Arabic, but are pronounced as /h/ and /ʔ/ in almost all vernaculars of Southern Kurdish. This letters are used mainly in Arabic loanwords.
- Similar to other Kurdish dialects, the letter <ۆ> stands for both /ʊ/ and /w/ phonemes, and letter <ی> stands for both /i/ and /j/ phonemes (Veisi et al., 2022).
- The /y/ vowel exists in Southern Kurdish and is represented by the <ێ> character (Azin and Ahmadi, 2021). This vowel does not exist in Central and Northern Kurdish.
- Similar to all other dialects of Kurdish written in Arabic script, there is no letter for /ə/ phoneme (Veisi et al., 2022).
- In some vernaculars of Southern Kurdish, the /ɣ/ phoneme changes to /q/. Also in Kermānshāh city and Mandali variants dark /ɬ/ sometimes replaced by /l/ (Belelli, 2019).

Figure 1: Regions of predominant Southern Kurdish speakers (Belelli, 2019)

3. Data design and collection

3.1. Training corpus

As the first step towards the development of an ASR system from read speech, we need a pool of sentences for recording. In order to provide an inclusive sentence pool that covers different phonological and morphological aspects of Southern Kurdish, we asked several native writers to provide us a raw text. In the end, we received the published text from 5 writers and 15 published volumes of a social science magazine. Approximately 400,000 space separated tokens were collected in total. The data was mainly from folk books, novels, anecdotes,

[2] https://www.ethnologue.com/language/sdh/

and social science[3]. The collected text is then processed as follows to design our training corpus:

- **Sentence selection:** From the collected text data, around 16,000 unique sentences containing between 3 and 15 tokens were randomly selected.
- **Text revision:** The selected sentences were revised by four professional native editors of the dialect. All sentences from other dialects, such as Central Kurdish or Persian, were excluded. Code-switched or loanwords were written in the Kurdish script.
- **Audio recording:** The readers were recruited for recording sentences using crowdsourcing tools. Approximately five hours of data were recorded using a Telegram bot, while the remaining data were collected through a web-based tool.

3.2. Evaluation benchmark

A second resource developed during our study is a multi-domain Southern Kurdish ASR benchmark. To create this benchmark, we extended the Central Kurdish Asosoft test set (Veisi et al., 2022) to Southern Kurdish by translating all sentences into the target dialect. The benchmark consists of 100 sentences covering a variety of domains. The sentences were originally written in Central Kurdish and then provided to a professional native writer of Southern Kurdish, who rewrote them in the target dialect. The final dataset was recorded by eight native speakers representing different vernaculars of Southern Kurdish. We ensured that the speakers involved in the benchmark did not participate in recording the training data.

3.3. Data validation

The recorded sentences from both training and evaluation data are revised manually by two native speakers. A recording is removed when one of the following problems occurs:

- **Intelligibility :** If a recording is not intelligible.
- **Partial recording:** If a part of the sentence is missing at the beginning or end of the recording.
- **Noise dominance:** Samples for which the background noise dominates the main speaker.
- **Miscue:** Samples having a miscue or errors in reading.

[3]In (Azin and Ahmadi, 2021), a text collection was introduced for Southern Kurdish, in which only a small portion (around 10k tokens) comes from the subgroups that constitute the majority of Southern Kurdish speakers (5.1 million out of a total of 6 million speakers). The rest of the data presented in that research comes from a website called *shafaq.com*, which was incorrectly attributed to the Feyli vernacular of Southern Kurdish. Although we had access to this resource, the editors decided not to use it in our research because it diverged significantly from the accepted orthography and lexicon of Southern Kurdish.

4. Data specifications

4.1. Training data

The training data were recorded using a Telegram bot and a web application developed for this task. All recordings are in mono channel, 24,000 Hz, WAV format. After validation, the total number of recordings is 18,636 files, comprising 15,273 unique utterances. We aimed to cover a broader linguistic space by recording a larger number of unique sentences. The dataset includes 208 speakers, aged between 18 and 49 years. Due to the difficulty of recruiting speakers, some participants were asked to record additional sentences. The highest number of recordings belongs to one speaker, who contributed 4,000 utterances. Although this introduced some imbalance, it was unavoidable as we rely on a small number of native volunteers. More detailed information about the training data is presented in Table 1. To facilitate reference to this corpus, we refer to it as the Bestun Corpus[4].

Item	Value
Sentences	18,636
Speakers	208
Age range	(18,49)
Unique sentences	15,273
Unique tokens	23,976
Duration	30 hours
Total tokens	1,554,994

Table 1: Overall Bestun (training) corpus characteristics.

4.2. Evaluation benchmark

The evaluation benchmark sentences are recorded by 8 native speakers from different Southern Kurdish speaking regions. Each speaker were asked to record all 100 unique sentences. After discarding the sentences according to our validation rules (section 3) the total number of sentences remained

[4]Bestun (Version 1.0) is named after a mountain located in Kirmashan, the largest city where Southern Kurdish is spoken.

in the benchmark is 773 for a duration of 86 minutes. The number of sentences and the region of speakers are listed in Table 2. One speaker (spk-E) recorded 94 sentences but only 88 sentences remain after validation. One weakness of our evaluations benchmark is gender imbalance having only two female speakers in the test set.

Speaker	G	Vernacular	Sent.
spk-A	F	Kermānshāhi	98
spk-B	M	Kermānshāhi	99
spk-C	F	Kalhori	99
spk-D	M	Kalhori	99
spk-E	M	Kalhori	88
spk-F	M	Malekshāhi	100
spk-G	M	Garrusi	95
spk-H	M	Kolyā'i	95
Total		**773 sentences (86.74 min)**	

Table 2: Southern Kurdish evaluation benchmark statistics by speaker, vernacular, and number of sentences.

5. Evaluation systems

5.1. Whisper

The first model used for evaluating the curated resources is Whisper Turbo (Radford et al., 2022). Whisper is an encoder-decoder transformer architecture trained using a weakly supervised approach. It supports ASR for more than 80 languages and performs speech-to-text translation from a language X to English. The original Whisper models published by Open AI does not include any Kurdish dialects. We expand the Whisper Turbo 3 tokenizer to include Kurdish tokens, and trained it for 10 epochs with a learning rate starting at 1e-5. We reserved 500 samples from the training corpus for validation, and the remaining data were used for training.

5.2. Wav2vec2-bert CTC

Wav2Vec-BERT is a self-supervised speech representation model that combines contrastive learning and masked language modeling objectives. The contrastive loss discretizes the speech signal, while the masked language model loss learns contextualized speech representations (Chung et al., 2021). In our research, we use Wav2Vec-BERT2, which is trained on 4.5 million hours of speech from 143 languages (Barrault et al., 2023). The details of the data used for pretraining are not reported, so it is unclear whether any data from Kurdish dialects or closely related languages were included. During fine-tuning, we optimized a Connectionist Temporal Classification (CTC) loss function (Graves et al., 2006). The model is trained for 10 epochs with an initial learning rate of 1e-5. We reserve 500 samples for validation, and the remaining data is used for training.

6. Results

In this section, we present the results obtained by both the Whisper-Turbo and Wav2Vec-BERT models and a deep analysis of automatic transcription errors. Table 3 presents the overall results obtained on the provided benchmark. The Whisper-Turbo model achieved a Character Error Rate (CER) of 6.75 and a Word Error Rate (WER) of 34.43, while the Wav2Vec-BERT model achieved a CER of 4.09 and a WER of 24.26. The two possible reasons why Wav2Vec-BERT performs better might be: (1) the larger amount of data seen by its encoder and the greater number of languages it was trained on, and (2) the use of character-level prediction units in its CTC model, which can yield better results with limited resources.

The only prior work on Southern Kurdish is from Hameed et al. (2025) which is indicated as baseline in Table 3. Using our evaluation setup and the published model from that research, the model achieved a CER of 22.46 and a WER of 82.20 on our evaluation benchmark. When comparing the results we obtained from the baseline ones, we conclude that the presented dataset in our study brings significant improvements of Southern Kurdish ASR.

In another experiment, we evaluate the Southern Kurdish benchmark on a Whisper model fine-tuned for Central Kurdish which obtains 23.11 CER and 84.23 WER (see Table 3), while the same model gives 2.65 CER and 13.40 WER on the Central Kurdish version of the benchmark. This experiment reaffirm the necessity of paying special attention to Southern Kurdish and the limitation of models developed for other dialects of Kurdish language in the context of Southern Kurdish.

ASR model	WER (%)	CER (%)
baseline	82.20	22.46
whisper-turbo-SDH	34.43	6.75
wav2vec-bert2-SDH	**24.26**	**4.09**
whisper-turbo-CKB	84.23	23.11

Table 3: Overall ASR performance on the evaluation benchmark when models are fine-tuned on Southern Kurdish (SDH) or Central Kurdish (CKB). Baseline from (Hameed et al., 2025).

A detailed evaluation per speaker is presented in Figure 2. The best-performing speaker reaches

No	Hypothesis	Reference	CER (%)	WER (%)
1	وە بونەی درسکردن کەش زانستی و بێز چەوگرتن ڕەوش ئابووری هەریم هەوەجەس ک وەشلخی کردن دۆر بگریەی	وە بونەی درس کردن کەش زانستی و بێز چەو گرتن ڕەوش ئابووری هەریم هەوەجەس ک وەشلخی کردن دۆر بگریەێ	3.19	26.32
2	ئیمەی بەرپرس و ئەورێخەر پروژەگە ئڕای دابین کردن شون دەفتەر داوای پاڵپشتی لە بەڕێزد دێریمن	ئیمەی بەرپرس و ئەورێخەر پروژەگە ئڕای دابینکردن شون دەفتەر داوای پاڵپشتی لە بەڕێزد دیریمن	2.27	21.43
3	لەوایێ خانم نۆسەر و ڕوشنهۆر کوردێگ چۆن نورِیدە چالاکی ئەو ژنەیلەو ک تینە مەیدان نۆسین	لەوای خانم نۆسەر و ڕوشنهۆر کوردێگ چۆن نورِیدە چالاکی ئەو ژنەیلەو ک تیەنە مەیدان نۆسین	3.57	20.00
4	لە ڕهو گیانیانا هاوار کەم بڕی کورد و کوردستان تەنانەت ئەگەر وەگولەی دژمنیل گەل دڵم بوسێدەو	لە ڕح و گیانیانا هاوار کەم بڕی کورد و کوردستان تەنانەت ئەگەر وەگولەی دژمنەیل گەل دڵم بوسێدەو	4.35	23.53
5	لیدان خوازیاریمن وەپاوگورەی پرۆژەیل حکومەتی هەشت لەسەد ژاژ دیاریکریای حەقدەس پرۆژەیل ناوبریا بەینە پیمان	لیدان خوازیاریمن وەپاوگورەی پرۆژەی حکومەتی هەشت لەسەد ژاژ دیاری کریای حەقدەس پرۆژەی ناوبریا بەینە پێمان	3.85	33.33
6	ئەمەڵیات هێزەیل ئەفقان دژ وەداعش دریژە دێرێ	عەمەڵیات هێزەیل ئەفغان دژ وەداعش دریژە دێرێ	4.65	28.57

Table 4: Hypothesis given by wav2vec-bert vs. Reference. *blue*: standardization error; *green:* phonetic and character mismatch; *brown:* vernacular diversity.

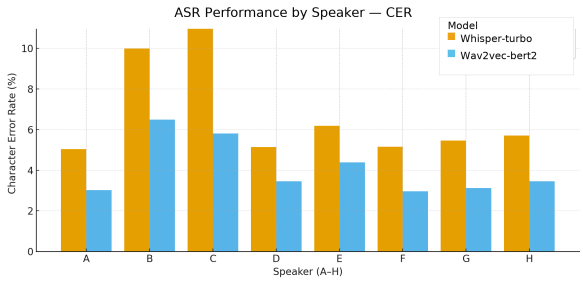

Figure 2: CER per speaker

a CER of 3.02, while the worst-performing speaker, on the same set of sentences, reaches a CER of 6.49 which is more than twice higher than the best one. To explain this difference, we explore some possible acoustic artifacts such as noise and reverberation. However, we did not observe any clear correlation between acoustic variations (*e.g.*, noise or reverberation) and the performance for each speaker. Since all speakers read the same text (minimizing linguistic variation), the significant differences in performance may stem from speaker-related characteristics.

In all experiments, we observed a large difference between the CER and WER. In low-resource languages such as Kurdish language, the lack of standardization can lead to multiple ways of writing the same word. Since Kurdish is a phonemic language with several regional vernaculars, a considerable number of words are written in different forms (Veisi et al., 2022). This phenomenon tends to spread small errors over many words, which dramatically increases the WER relative to the CER. To have a better analysis on the nature of transcription errors we present some samples in Table 4. We observed three main types of errors:

- **Standardization errors:** A main group of mismatches between the reference and hypothesis comes from having several forms of writing. Such type of errors are shown in $blue$. For instance in the first in example, three words are recognized erroneous and in each word one character cause this error, but both the predictions and reference exist in the Southern Kurdish text.

- **Phonetic and character mismatch:** As it was mentioned in Section 2, the ح and ع letters that stands for /ħ/ and /ʕ/ phonemes exist in the Southern Kurdish writing, while these phonemes are pronounced as /h/ and /ʔ/ respectively. In example 4, the ح (/ħ/) is recognized as ه (/h/) same as its pronunciation. In the example 6, the ع (/ʕ/) is recognized wrongly as ئ (/ʔ/). These types of errors are shown in $green$.

- **Vernacular diversity:** The interchangeable use of /ɬ/ for /l/ and /ɣ/ for /q/ is reflected in the errors. These types of errors are shown in $brown$.

7. Conclusion

In this paper, we introduce the first dedicated speech recognition corpus and evaluation benchmark for the Southern Kurdish language. The overall size of the training corpus comprises 30 hours of manually verified read speech. We show that Wav2Vec-BERT model fined-tuned on the training data achieves better results than a Whisper-turbo ASR model. We also provide a detailed analysis with a typology of transcription errors (standardization, phonetic and character mismatch, and vernacular diversity). The system's robustness is also analyzed with respect to each of the benchmark speakers. The provided resources can be a step stone in developing multi-dialect and more comprehensive ASR models for Kurdish language. The

curated resources has some limitations that can be addressed in the future works such as lack of spontaneous speech.

8. Acknowledgements

We appreciate the voluntarily contribution of Jiar Jahanfard and Afshin Ghomali, Saro Khosrawi and Farhad Jahanbeigi for providing the raw text material and participating in revising the text before recording. Also we deeply appreciate the participation of Mahsa Zarei, Yosra Izadi, Farshad Weliyan, Karzan Senaie and Idris Gushbor in recording process and networking. We appriciate the participation of all volunteers in recording. Without their contribution the curation of the dataset was impossible. Finally, we appreciate the participation of Mahrokh Modiri in the validation of recordings. Special thanks to Navid Dabaghi for reformatting the map of Southern Kurdish speakers in a colored format. This research is done at LIUM, Le Mans University and the experiments are done using LIUM computing resources.

References

Sina Ahmadi, Daban Jaff, Md Mahfuz Ibn Alam, and Antonios Anastasopoulos. 2024. Language and speech technology for Central Kurdish varieties. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 10034–10045, Torino, Italia. ELRA and ICCL.

Rosana Ardila, Megan Branson, Kelly Davis, Michael Henretty, Mike 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 2020)*, pages 4211–4215.

Sakina Azadi. 2022. A comparison between feyli and kalhori verbs of southern kurdish. *Regional language and literature of Iran*, 36(ISSUE):1–28. Accessed 9 October 2025.

Zahra Azin and Sina Ahmadi. 2021. Creating an electronic lexicon for the under-resourced southern varieties of kurdish language. In *Proceedings of the Seventh Biennial Conference on Electronic Lexicography (eLex 2021)*, pages 479–488.

Loïc Barrault, Yu-An Chung, Mariano Cora Meglioli, David Dale, Ning Dong, Paul-Ambroise Duquenne, Hady Elsahar, Hongyu Gong, Kevin Heffernan, John Hoffman, Christopher Klaiber, Pengwei Li, Daniel Licht, Jean Maillard, Alice Rakotoarison, Kaushik Ram Sadagopan, Guillaume Wenzek, Ethan Ye, Bapi Akula, Peng-Jen Chen, Naji El Hachem, Brian Ellis, Gabriel Mejia Gonzalez, Justin Haaheim, Prangthip Hansanti, Russ Howes, Bernie Huang, Min-Jae Hwang, Hirofumi Inaguma, Somya Jain, Elahe Kalbassi, Amanda Kallet, Ilia Kulikov, Janice Lam, Daniel Li, Xutai Ma, Ruslan Mavlyutov, Benjamin Peloquin, Mohamed Ramadan, Abinesh Ramakrishnan, Anna Sun, Kevin Tran, Tuan Tran, Igor Tufanov, Vish Vogeti, Carleigh Wood, Yilin Yang, Bokai Yu, Pierre Andrews, Can Balioglu, Marta R. Costa-jussà, Onur Celebi, Maha Elbayad, Cynthia Gao, Francisco Guzmán, Justine Kao, Ann Lee, Alexandre Mourachko, Juan Pino, Sravya Popuri, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, Paden Tomasello, Changhan Wang, Jeff Wang, and Skyler Wang. 2023. Seamlessm4t: Massively multilingual & multimodal machine translation. arXiv preprint arXiv:2308.11596.

Loïc Barrault, Yu-An Chung, Mariano Coria Meglioli, David Dale, Ning Dong, Paul-Ambroise Duquenne, Hady Elsahar, Hongyu Gong, Kevin Heffernan, et. all., and SEAMLESS Communication Team. 2025. Joint speech and text machine translation for up to 100 languages. *Nature*, 637(8046):587–593.

Sara Belelli. 2019. Towards a dialectology of southern kurdish: Where to begin? In Songül Gündoğdu, Ergin Öpengin, Geoffrey Haig, and Erik Anonby, editors, *Current issues in Kurdish linguistics*, pages 73–92. University of Bamberg Press, Bamberg.

Sara Belelli. 2021. *The Laki Variety of Harsin: Grammar, Texts, Lexicon*. Number 2 in Bamberg Studies in Kurdish Linguistics. University of Bamberg Press, Bamberg.

Yu-An Chung, Yu Zhang, Wei Han, Chung-Cheng Chiu, James Qin, Ruoming Pang, and Yonghui Wu. 2021. W2v-bert: Combining contrastive learning and masked language modeling for self-supervised speech pre-training. arXiv preprint arXiv:2108.06209.

Ismaïl Kamandâr Fattah. 2000. *Les dialectes kurdes méridionaux : étude linguistique et dialectologique*. Number 37 in Acta Iranica. Peeters.

Alex Graves, Santiago Fernández, Faustino Gómez, and Jürgen Schmidhuber. 2006. Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks. In *Proceedings of the 23rd International Conference on Machine Learning (ICML 2006)*.

Razhan Hameed, Sina Ahmadi, Hanah Hadi, and Rico Sennrich. 2025. Automatic Speech Recognition for Low-Resourced Middle Eastern Languages. In *Interspeech 2025*, pages 733–737.

Manijeh Mirmukri, Gholamhossein Karimi-Doostan, Yadgar Karimi, and Vahid Gholami. 2019. Mutual intelligibility between central and southern kurdish dialects (case study: Mahabadi and badrei varieties). *Journal of Linguistics, Islamic Azad University, Sanandaj*. Reçu le 24 janvier 2019 ; accepté le 24 septembre 2019.

Mohammad Mohammadamini, Aghilas Sini, Marie Tahon, and Antoine Laurent. 2025. Scaling pseudo-labeling data for end-to-end low-resource speech translation (the case of Kurdish language). In *Interspeech 2025*, pages 898–902.

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

Jaffer Sheyholislami. 2008. *Identity, Discourse, and the Media: The Case of the Kurds*. Phd dissertation, Carleton University. Accessed 9 October 2025.

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

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(3):917–941.