

Automatic Speech Recognition for Documenting Endangered Languages: Case Study of Ikema Miyakoan

Chihiro Taguchi¹, Yukinori Takubo², David Chiang¹

¹University of Notre Dame, ²National Institute for Japanese Language and Linguistics

¹Notre Dame, IN, USA, ²Tokyo, Japan

ctaguchi@nd.edu, ytakubo@ninjal.ac.jp, dchiang@nd.edu

Abstract

Language endangerment poses a major challenge to linguistic diversity worldwide, and technological advances have opened new avenues for documentation and revitalization. Among these, automatic speech recognition (ASR) has shown increasing potential to assist in the transcription of endangered language data. This study focuses on Ikema, a severely endangered Ryukyuan language spoken in Okinawa, Japan, with approximately 1,300 remaining speakers, most of whom are over 60 years old. We present an ongoing effort to develop an ASR system for Ikema based on field recordings. Specifically, we (1) construct a 6.33-hour speech corpus from field recordings, (2) train an ASR model that achieves a character error rate as low as 15%, and (3) evaluate the impact of ASR assistance on the efficiency of speech transcription. Our results demonstrate that ASR integration can substantially reduce transcription time and cognitive load, offering a practical pathway toward scalable, technology-supported documentation of endangered languages.

Keywords: language documentation, automatic speech recognition, language resources

1. Introduction

Language endangerment is a pressing global issue, with thousands of languages at risk of disappearing within the coming decades. Recent advancements in language technologies have opened up new opportunities for language documentation, offering computational tools that can assist researchers in preserving linguistic data more efficiently. In particular, automatic speech recognition (ASR) has emerged as a promising technique to support the transcription of spoken data from endangered languages.

Ikema Miyakoan (Ikema henceforth), a highly endangered Ryukyuan language spoken on the Miyako Islands in Japan, represents one such case. With an estimated speaker population of approximately 1,300 individuals, most of whom are over 60 years old, the language faces severe risk of extinction (Nakama et al., 2025). In this context, leveraging speech technologies for Ikema is both urgent and valuable for documentation and revitalization efforts.

This paper reports on our ongoing work to develop a speech dataset and an ASR system for Ikema as part of a broader language documentation project. We describe the construction of a 6.33-hour speech dataset collected from field recordings, pronounced dictionary entries, and audio books. Then, we report the training of an ASR model that achieves a character error rate (CER) as low as 14.8%. The trained model is then integrated into the linguistic annotation software, and we evaluate how much ASR can facilitate the transcription process. Our findings indicate that ASR-assisted

transcription not only speeds up annotation but also reduces the cognitive burden on annotators, highlighting the practical benefits of integrating ASR into endangered language documentation workflows.

The key contributions of this study are:

- The construction of the first speech dataset for Ikema;
- The development of the first ASR model for Ikema with a CER as low as 14.8%;
- The integration of the ASR model into the linguistic annotation tool (ELAN);
- Presenting the positive evidence for the potential of ASR-assisted transcription in language documentation.

The dataset, the model, and the experimental code developed in this study are publicly released.¹

2. Related Work

2.1. Language

Ikema is an endangered Japonic language spoken in the Miyako Islands of Okinawa, Japan. Its linguistic classification is illustrated in Figure 1. The language is spoken in three villages: Ikema Island, the Nishihara village on Miyako Island, and the Sarahama village on Irabu Island, as shown in Figure 2. A recent study predicts that Ikema will become “critically endangered” according to UNESCO’s criteria within 30 years, when the youngest active speakers

¹https://github.com/ctaguchi/ikema_asr

reach 90 years of age (Nakama et al., 2025). Nevertheless, Ikema is comparatively better studied and documented than other Miyakoan varieties, with an existing descriptive grammar (Hayashi, 2013) and dictionary (Nakama et al., 2025). Yamada et al. (2020) report that Ikema is largely unintelligible to speakers of Japonic languages outside the Southern Ryukyuan subgroup.

Although Ikema currently lacks an official orthography, several writing systems have been proposed to document the language. In linguistic literature on the Ikema grammar, phonemic transcription using either the International Phonetic Alphabet (IPA) or romanization has been commonly employed. In contrast, the speaker community more frequently uses *hiragana* or *katakana* (collectively referred to as *kana*) which are syllabic writing systems used in Japanese (Takubo, 2021a). The existing dictionary (Nakama et al., 2025) defines a phonemic kana orthography for Ikema, incorporating several extensions to represent phonemes absent in Modern Standard Japanese, such as a syllabic devoiced nasal consonant. The phonemic *kana* orthography can be deterministically converted into romanized script. Example (1) shows a comparison of the three writing systems: (1a) for *kana*, (1b) for phonemic romanization (*romaji* henceforth), and (1c) for phonemic IPA. For reference, (1d)² provides the gloss and the translation of the example.

- (1) a. っうゝ あー んゝぬ なうゆどう ほうーたー
 b. vvaɑ Nnu nauyudu huutaa
 c. vvaɑ ɲnu naujudu huutaa
 d. vva-a ɲnu nau-ju-du
 2SG-TOP yesterday what-ACC-FOC
 huutaa
 DO.PROG.PST.Q
 ‘What were you doing yesterday?’

2.2. Speech recognition

Speech recognition technologies have advanced drastically with the development of deep learning, and their applications have increasingly extended to low-resource and endangered languages. Such efforts have led to the development of ASR systems for Nasal (Billings and McDonnell, 2025), Newar and Dzardzongke (O’Neill et al., 2023), Sardinian (Chizzoni and Vietti, 2024), Mvskoke (Mainzinger and Levow, 2024), Kichwa (Taguchi et al., 2024), Japhug (Guillaume et al., 2022), and Khinalug (Li et al., 2024), among others. Notably, several of

²2SG: 2nd person singular pronoun; ACC: accusative case; FOC: focus marker; PROG: progressive aspect; PST: past tense; Q: interrogative; TOP: topic marker

Dataset	#Samples	Duration (sec)	Avg. duration (sec)
Field	11126	16058	1.44
Dictionary	5680	5379	0.95
Audiobooks	285	1348	4.73
Total	17091	22785	1.33

Table 1: Statistics of the datasets.

these studies emphasize the use of ASR systems to accelerate the transcription process, which is an essential yet labor-intensive stage of language documentation. Transcription requires significant time investment and specialized linguistic literacy and is often referred to as the “transcription bottleneck” in language documentation (Seifart et al., 2018). Since documenting endangered languages is inherently time-sensitive, mitigating this bottleneck through ASR support represents a promising direction. Indeed, case studies on Yongning Na (Michaud et al., 2018) and Seneca (Jimerson and Prud’hommeaux, 2018) have demonstrated the utility of ASR integration in transcription workflows, where ASR outputs serve as drafts for human post-editing. While these studies confirm the usefulness of the approach, quantitative evidence on the extent to which ASR accelerates transcription in practice remains limited.

3. Dataset

The dataset constructed in this study is composed of three sources: video recordings collected through nearly twenty years of fieldwork (hereafter, “Field” data), pronounced entries from the Ikema dictionary (Nakama et al., 2025) (hereafter, “Dictionary” data), and audiobooks. A large portion of the Field data consists of semi-spontaneous monologues by a single male speaker. Most of these monologues are planned speech, in which the speaker was either provided with a specific topic or has read a prepared script before recording. The Dictionary data contain scripted speech produced by the same male speaker. Most audio segments in the Dictionary data are shorter than one second, as each recording corresponds to a single dictionary entry (i.e., a word). The audiobook data consist of stories read aloud by a female speaker. See Table 1 for the detailed statistics of the data.

The video recordings were first converted to WAV files, after which the audio was segmented and annotated with transcriptions. Transcriptions were provided in *kana* in a phonetically faithful manner; that is, utterances were transcribed as they were pronounced, including fillers, disfluencies, and repetitions. After the *kana* transcription was completed, it was transliterated into *romaji* using deterministic

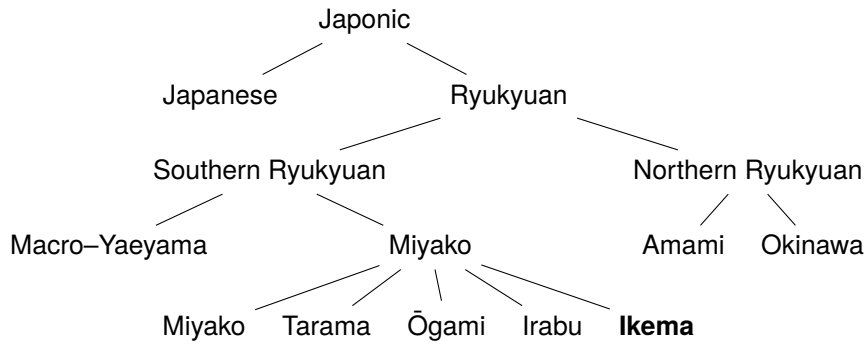


Figure 1: A classification of Japonic languages and Ikema’s position thereof. The classification is largely based on Shimoji (2008) and Pellard (2015).



Figure 2: Map of the Miyako Islands (bottom) in Japan (top). The Ikema-speaking villages are marked with a red circle.

mapping rules. In the dataset, word boundaries are indicated by a half-width space (U+0009).

The transcriptions include several types of tags that provide detailed linguistic information, such as code-switched segments in Japanese (<ja>...</ja>), disfluencies (<dis>...</dis>), songs (<song>...</song>), and personal names (<name>...</name>). In addition, utterances that annotators were unable to understand are annotated with the tag <unsure>...</unsure>. This design allows users who wish to train models on clean transcriptions without disfluencies or repetitions to automatically remove them from the dataset, and also enables the easy exclusion of samples containing personally identifiable information. Given the nature of spontaneous speech, sentence boundaries are often difficult to determine. Therefore, segment boundaries were determined based on pauses between speech regions. In addition, the annotated data include a tier with longer segments that roughly correspond to sentence-level units, as shown in Figure 3.

4. Experiments

4.1. Setup

In our experiments, we train automatic speech recognition (ASR) models on the newly developed Ikema speech dataset. Specifically, we fine-tune pretrained Wav2Vec2 models (Baevski et al., 2020) using a Connectionist Temporal Classification (CTC) decoder layer (Graves et al., 2006). Wav2Vec2 is a self-supervised model that learns speech representations from large amounts of unlabeled audio. Among its multilingual variants, XLS-R (Babu et al., 2021) and MMS (Pratap et al., 2023) have been trained on 128 and 1,406 languages, respectively, enabling robust cross-lingual transfer.

For Ikema, we fine-tune these pretrained models by adding a final CTC layer that maps frame-level features to textual transcriptions. The CTC layer produces a token prediction for each frame, and

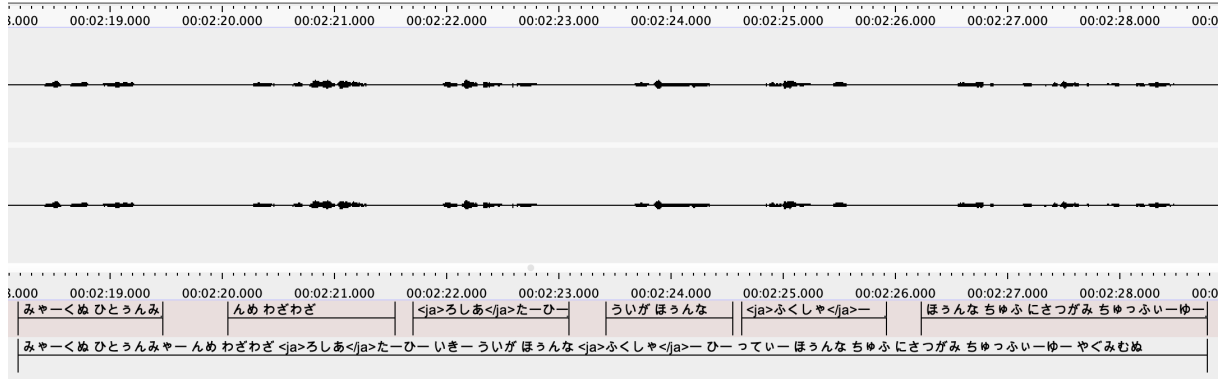


Figure 3: Segmentation and annotation in ELAN. The top transcription tier contains pause-based segments, while the bottom tier contains longer segments that concatenate multiple pause-based segments to approximate sentence-level units.

identical consecutive tokens are collapsed into a single symbol to form the decoded output, typically at the character or grapheme level. Wav2Vec2-based models have been shown to perform effectively even in extremely low-resource settings (Liang and Levow, 2025), making them well-suited for our experiments.

While automatic speech recognition (ASR) models with autoregressive decoders, such as Whisper (Radford et al., 2022), achieve strong performance for high-resource languages, they require large amounts of data to train the decoder to model subword dependencies effectively. In contrast, a model with a standard CTC decoder does not explicitly learn dependencies across tokens, allowing it to directly learn the alignment between the input speech and the corresponding output characters at each time frame. Furthermore, CTC’s per-frame outputs are particularly suitable for our goal of producing phonetically faithful transcriptions, as the transcribed data may contain substantial disfluencies and fillers that should not be omitted in language documentation.

In our experiments, we compare the performance of wav2vec2-xls-r-300m (300M parameters, 128 languages), wav2vec2-xls-r-1b (1B parameters, 128 languages), and mms-1b (1B parameters, 1,406 languages). The tags in the transcription are removed in the preprocessing. Of the dataset, 80% is used as the training data, 10% is reserved for the validation data, and 10% for the test data. We also train models in two writing systems: *kana* and *romaji*. The token vocabulary is constructed based on syllables (or more precisely, morae) for *kana* and phonemes for *romaji*; for example, “じや” /zja/ is counted as one token in the *kana* model, and “zy” /z/ as one token in the *romaji* model.

We report character error rate (CER) and word error rate (WER), defined as the percentage of incorrect characters or words out of 100. It is important to note, however, that WER is less infor-

Model	<i>kana</i>		<i>romaji</i>	
	CER	WER	CER	WER
xls-r-300m	20.94	62.56	14.80	64.99
xls-r-1b	24.07	66.61	20.33	76.73
mms-1b	21.86	63.22	18.98	73.87

Table 2: ASR performances of the trained models.

mative for Ikema, where word boundaries are not well defined and variations appear even in the gold transcriptions.

The learning rate is set to 0.0003. Preliminary experiments showed higher learning rates exhibited to slower and unstable learning curves. The batch size was set to 16 with wav2vec2-xls-r-300m and 4 with wav2vec2-xls-r-1b and mms-1b. The model is trained for 50 epochs. All the models are trained on a single A10 GPU with 16GB RAM.

4.2. Results

Table 2 shows the performance of the trained ASR models under different conditions. Compared to the *kana* transcription, the *romaji* transcription consistently yielded lower CERs. This is likely because the *kana* writing system encodes more phonological information than *romaji*. This finding aligns with the general tendency of Wav2Vec2 models to underperform when transcribing more complex writing systems (Taguchi and Chiang, 2024).

The best performing model was the fine-tuned model based on xls-r-300m in both *kana* and *romaji* transcriptions. On the other hand, the larger model (xls-r-1b) performed the worst among the three models. These results suggest that a larger model size does not necessarily promise a successful training. The lightest model (xls-r-300m) was not only effective in its performance but also advantageous in the training time and the total training steps because of its increased batch size as

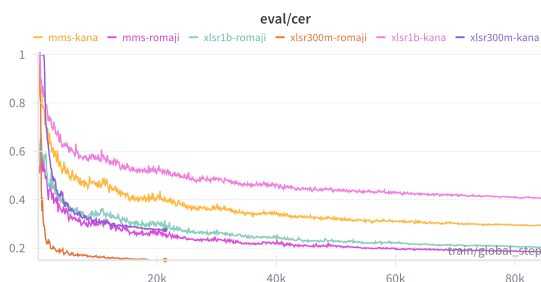


Figure 4: The curves of the CERs on the validation data during the training.

illustrated in Figure 4. In the experiments, the training with `xlsr-300m` completed in approximately 10 hours, while the other 1B models took more than 50 hours to complete. In addition, the lightweight model takes up to approximately 1.2GB as opposed to the 1B models' 3.8GB, and is beneficial for running locally in the integrated annotation tool.

5. Is ASR-assisted transcription helpful?

Whether ASR can truly benefit annotators in the transcription process has been a point of debate. Although many studies have argued for the potential advantages of ASR in language documentation, Prud'hommeaux et al. (2021) reported that members of some speaker communities preferred unassisted transcription without ASR support. To examine this question in the case of Ikema, we conducted a human evaluation to assess the extent to which ASR can actually accelerate the transcription process.

5.1. Setup

For this experiment, the fine-tuned `wav2vec2-xlr-s-300m` model trained on the *kana*-script is integrated into ELAN, linguistic annotation software (Wittenburg et al., 2006), as an extension to transcribe Ikema automatically on the user interface. In this experiment, we choose an unannotated 5-minute recording, segment the spoken parts, and asked two annotators to annotate the same segments on ELAN under the following conditions. The total number of segments is 121, and their total duration is 217.64 seconds. Annotator A is a non-native expert who has worked on the documentation of Ikema for 19 years, and Annotator B is a non-native learner with less than one year of experience. Both are native speakers of Japanese and familiar with the *kana* orthography of Ikema. Annotator A is given an EAF file (an extended XML for ELAN containing the annotation text) in which the first half of the segments (61

	w/o ASR	w/ ASR	Speedup
Annotator A	11:36	9:43	+19.38%
Annotator B	30:49	25:00	+23.27%

Table 3: A comparison of annotation speed, with and without ASR output. The speedup is computed by dividing the time without ASR by the time with ASR.

samples, 110.12 seconds) are blank and the second half (60 samples, 107.52 second) are filled in with the model's prediction output. Conversely, Annotator B is provided with an EAF file where the first half contains the model's transcription and the second half is left blank. Both annotators are asked to transcribe the audio manually, and their annotation times are recorded separately for each half. Note that they are both allowed to have listened to the recording before and know what is talked about.

5.2. Results

As Table 3 demonstrates, both annotators reported a positive speedup in ASR-assisted transcription. The annotators also noted that they felt it somewhat easier to annotate the data with the ASR draft despite its errors. Annotator B noted that segments with a frequently occurring filler and a fixed expression were often transcribed correctly, decreasing the need of manual correction with clicking, typing, and re-listening. The results confirmed the positive effect of ASR outputs in language documentation even when the model's accuracy is not ideal.

As for the inter-annotator agreement between the two annotators, there was a CER of 17.32% and a WER of 46.24%. The CER of the ASR output was 29.61% with Annotator A, and 25.25% with Annotator B. However, when minor orthographic ambiguities are ignored, the CER of the ASR output decreases to 20.34%. The high error rates reflect the non-standardized aspects of the language's writing system, such as post-lexical morphophonological changes (Takubo, 2021b) and the position of whitespaces. Example (2) shows an instance of such cases.³ The words *umatti* 'coming here' and *Nmeera* 'umm' (a filler) are often pronounced as if they are single words. For this reason, Annotator A (2a) transcribed them as single words, while Annotator B (2b) inserted spaces based on their underlying analyzed forms. In addition, while Annotator A often omitted disfluencies and repetitions, Annotator B transcribed the speech faithfully including the speech errors.

For reference, (2c) shows the actual output by

³COND: conditional conjunctive suffix; DSC: discourse sentence-final particle; INF: infinitive form.

the recognizer. While the high error rates may suggest low quality, a closer look at the output reveals that the model’s prediction reflects how the sample sounds. In fact, the suffix *-tigaa* often realizes as phonetically shortened forms *-tiyaa*, *-tiya*, or *-taa*, the filler *Nme* is also pronounced as *Nmya*, and the sentence-final particle *ira* can also realize as *ra*. This phonetic faithfulness is characteristic of the CTC decoder outputting frame by frame without a language model. While the model still mispredicts token boundaries, the errors observed in the outputs are often in fact free variations. This observation is coherent with the discussion provided in the previous study on ASR for Japhug (Guillaume et al., 2022).

- (2)a. うま っ て い う ぐ な い て い が ー ん め ー ら
umatti ugunaaitigaa Nmeera
uma+tti ugunaai-tigaa Nme+ira
 here+come.INF gather-COND well+DSC
 ‘If (they) come here and get together, well’
- b. う ま っ て い う ぐ な い て い が ー ん め い ら
uma tti ugunaaitigaa Nme ira
uma tti ugunaai-tigaa Nme ira
 here come.INF gather-COND well DSC
- c. う ま て い う ぐ な い た ー ん み や ら
uma tiugunaitaaNmyara

6. Conclusion

This study presented an ongoing effort to develop an automatic speech recognition (ASR) system for Ikema Miyakoan, an endangered Ryukyuan language spoken in Okinawa, Japan. We constructed a 6.33-hour speech corpus from recordings through collaborative fieldwork with the speaker community. Based on this dataset, we trained an ASR model achieving a CER as low as 14.8%. Then, we integrated the model into ELAN as an ASR extension and evaluated its potential to mitigate the transcription bottleneck in language documentation. The results indicate that the integration of ASR into transcription workflows can substantially accelerate annotation and reduce the cognitive burden on human transcribers.

7. Ethics statements

This research is grounded in close collaboration with the Ikema-speaking community and adheres to ethical standards for language documentation and computational research on endangered languages. All speakers participated voluntarily with informed consent, and their privacy and data rights were respected throughout data collection and processing.

The speech recordings are used strictly for research and documentation purposes, and any future public release of the dataset will follow community consultation and appropriate licensing to ensure responsible use. Potentially personally-identifiable names that appear in our datasets are annotated with the tags `<name> . . . </name>` so that such samples can be removed automatically.

We recognize that the deployment of language technologies in small and vulnerable communities entails complex ethical and social implications. While ASR systems can effectively accelerate transcription and support revitalization efforts, they may also raise concerns regarding data ownership, representation, and the potential misuse of linguistic data. To mitigate these risks, we emphasize transparent documentation of data ownership and licensing, in close collaboration with the community and through shared decision-making.

Another relevant important consideration is that developing an ASR system for a traditionally unwritten language may privilege a particular writing system that is not necessarily widely supported within the community. While the constructed dataset uses a *kana* (*hiragana*) orthography, some speakers in fact find this orthography difficult to read. This study, as well as the dataset and model developed in this work, does not intend to impose the use of the *kana* orthography on the community.

More broadly, this work contributes to the equitable development of language technologies by extending computational research beyond high-resource languages. By developing resources and methodologies for Ikema, we aim to promote linguistic diversity and inclusivity within NLP, demonstrating that advances in speech technology can serve both scientific and community-centered goals when guided by ethical collaboration and respect for speaker agency.

8. Acknowledgments

The material was based on work supported in part by the US National Science Foundation under Grant Number BCS-2109709 and IIS-2137396. We also thank the Ikema Miyakoan native speakers who helped the authors collect the data and review the model output.

9. Appendix

Table 4 lists the field recordings used in the Field dataset, as well as their basic statistics.

10. Bibliographical References

Recording ID	Style	Duration	#Samples	#Kana	#Words
I0482_usl	Spontaneous	111.03	69	1105	202
I0482_ngi	Spontaneous	100.36	72	885	201
I0482_pic_muclusa	Spontaneous	85.93	56	683	132
I0482_pic_buugii	Spontaneous	72.91	49	534	117
I0482_byuuigassa	Spontaneous	53.51	35	454	98
I0482_bippii	Spontaneous	26.27	15	229	43
I0482_pic_barazan	Spontaneous	30.50	16	324	58
I0482_niguu	Spontaneous	86.81	56	675	148
I0482_pic_takanna	Spontaneous	44.99	34	412	83
I0482_pic_taka	Spontaneous	179.77	108	1698	345
I0412_suuni	Spontaneous	29.76	21	246	48
I0483_0414_satatinpura	Spontaneous	44.30	29	382	77
I0482_pic_zzakugii	Spontaneous	73.37	48	586	122
I0482_gislcl	Spontaneous	107.87	72	934	193
I0482_kami	Spontaneous	197.85	138	1521	332
I0482_gazlhanagii	Spontaneous	47.87	29	426	86
I0482_uGAN	Spontaneous	84.47	61	677	129
I0483_0277_yasaiitame	Spontaneous	19.33	16	114	27
I0483_0276_waa	Spontaneous	16.76	10	143	30
I0483_0405_islusl	Spontaneous	114.51	77	942	203
I0483_0527_suba	Spontaneous	37.64	22	345	69
I0483_0412_waanimumun	Spontaneous	34.99	26	283	61
I0483_0688_mamisuumai	Spontaneous	63.79	40	540	110
I0483_0409_fukyagi	Spontaneous	66.88	45	568	111
I0483_dakyau	Spontaneous	19.40	15	127	27
I0482_pc_taumabasl	Spontaneous	324.87	185	3049	586
I0483_0678_saguna	Spontaneous	118.40	70	1093	220
I0482_kkuclgii	Spontaneous	65.86	46	534	111
I0483_0222_avvansu	Spontaneous	23.82	19	180	39
I0488_しーさー	Spontaneous	100.76	69	832	180
I0490_あまん	Spontaneous	111.11	79	958	214
I0490_池間大橋	Spontaneous	167.68	113	1485	306
I0496_kaichosyokureki	Conversation	912.26	339	8407	1706
I0501_nakasonetuimyya	Spontaneous	305.98	86	2495	461
I0502_harimizuuutaki	Spontaneous	405.51	151	3419	660
I0503_Nevskynohi	Spontaneous	145.16	60	1054	204
I0506_aisatsutehon_Va	Spontaneous	19.18	7	85	19
I0506_aisatsutehon_Vb	Spontaneous	13.60	6	62	19
I0506_aisatsutehon_Vc	Spontaneous	15.88	7	109	21
I0506_aisatsutehon_Vd	Spontaneous	101.15	25	639	131
I0506_aisatsutehon_Ve	Spontaneous	159.49	47	1260	247
I0506_aisatsutehon_Vf	Spontaneous	79.75	17	612	122
I0507_mingukaisetsu_Va	Spontaneous	112.81	67	978	215
I0507_mingukaisetsu_Vb	Spontaneous	126.18	39	942	191
I0508_nakamautaki_V	Spontaneous	367.59	102	2716	507
I0509_zyaagama_V	Spontaneous	254.00	76	1918	347
I0510_kyuukoominkan_Vb	Spontaneous	116.36	22	787	152
I0510_kyuukoominkan_Va	Spontaneous	182.85	39	1387	263

Table 4: The detail of the spontaneous speech dataset (Field). The Audiobook data and the Dictionary data are not shown in this table.

Arun Babu, Changhan Wang, Andros Tjandra,
Kushal Lakhotia, Qiantong Xu, Naman Goyal,

Kritika Singh, Patrick von Platen, Yatharth Saraf,
Juan Pino, Alexei Baevski, Alexis Conneau, and
Michael Auli. 2021. [XLS-R: Self-supervised cross-lingual speech representation learning at](#)

- scale.
- Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. [wav2vec 2.0: A framework for self-supervised learning of speech representations](#).
- Blaine Billings and Bradley McDonnell. 2025. [Connecting automated speech recognition to transcription practices](#). In *Proceedings of the Eighth Workshop on the Use of Computational Methods in the Study of Endangered Languages (ComputEL)*, pages 128–132.
- Ilaria Chizzoni and Alessandro Vietti. 2024. [Towards an ASR system for documenting endangered languages: A preliminary study on Sardinian](#). In *Proceedings of the 10th Italian Conference on Computational Linguistics (CLiC-it 2024)*, pages 214–220.
- Alex Graves, Santiago Fernández, Faustino Gomez, 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*, pages 369–376.
- Séverine Guillaume, Guillaume Wisniewski, Cécile Macaire, Guillaume Jacques, Alexis Michaud, Benjamin Galliot, Maximin Coavoux, Solange Rossato, Minh-Châu Nguyễn, and Maxime Fily. 2022. [Fine-tuning pre-trained models for automatic speech recognition, experiments on a fieldwork corpus of Japhug \(Trans-Himalayan family\)](#). In *Proceedings of the Fifth Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pages 170–178.
- Yuka Hayashi. 2013. *Minami ryuukyuu miyakogo ikema hougen no bunpou*. Ph.D. thesis, Kyoto University.
- Robbie Jimerson and Emily Prud'hommeaux. 2018. [ASR for documenting acutely under-resourced indigenous languages](#). In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.
- Zhaolin Li, Monika Rind-Pawłowski, and Jan Niehues. 2024. [Speech recognition corpus of the khinalug language for documenting endangered languages](#). In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 15171–15180.
- Siyu Liang and Gina-Anne Levow. 2025. [Breaking the transcription bottleneck: Fine-tuning ASR models for extremely low-resource fieldwork languages](#). In *Proceedings of the Fourth Workshop on NLP Applications to Field Linguistics*, pages 26–37.
- Julia Mainzinger and Gina-Anne Levow. 2024. [Fine-tuning ASR models for very low-resource languages: A study on Mvskoke](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop)*, pages 76–82.
- Alexis Michaud, Oliver Adams, Trevor Cohn, Graham Neubig, and Séverine Guillaume. 2018. [Integrating automatic transcription into the language documentation workflow: Experiments with Na data and the Persephone toolkit](#). *Language Documentation & Conservation*, 12:393–429.
- Hiroyuki Nakama, Yukinori Takubo, Shoichi Iwasaki, Yosuke Igarashi, Daniel Wymark, and Natsuko Nakagawa. 2025. *Dictionary of Ikema spoken in Nishihara village, a variety of Miyako Ryukyuan*, 3rd edition. National Institute for Japanese Language and Linguistics.
- Alexander O'Neill, Marieke Meelen, Rolando Coto-Solano, Sonam Phuntsog, and Charles Ramble. 2023. [Language preservation through ASR](#). In *Cambridge Language Sciences Annual Symposium*.
- Thomas Pellard. 2015. [The linguistic archeology of the Ryukyu Islands](#). In Patrick Heinrich, Shinsho Miyara, and Michinori Shimoji, editors, *Handbook of the Ryukyuan languages: History, structure, and use*, pages 13–37. De Gruyter Mouton, Berlin; Boston.
- Vineel Pratap, Andros Tjandra, Bowen Shi, Paden Tomasello, Arun Babu, Sayani Kundu, Ali Elkahky, Zhaoheng Ni, Apoorv Vyas, Maryam Fazel-Zarandi, Alexei Baevski, Yossi Adi, Xiaohui Zhang, Wei-Ning Hsu, Alexis Conneau, and Michael Auli. 2023. [Scaling speech technology to 1,000+ languages](#).
- Emily Prud'hommeaux, Robbie Jimerson, Richard Hatcher, and Karin Michelson. 2021. [Automatic speech recognition for supporting endangered language documentation](#). *Language Documentation & Conservation*, 15:491–513.
- Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. [Robust speech recognition via large-scale weak supervision](#).
- Frank Seifart, Nicholas Evans, Harald Hammarström, and Stephen C. Levinson. 2018. [Language documentation twenty-five years on](#). *Language*, 94(4):e324–e345.

Michinori Shimoji. 2008. *A Grammar of Irabu, a Southern Ryukyuan Language*. Ph.D. thesis, Australian National University.

Chihiro Taguchi and David Chiang. 2024. [Language complexity and speech recognition accuracy: Orthographic complexity hurts, phonological complexity doesn't](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15493–15503.

Chihiro Taguchi, Jefferson Saransig, Dayana Velásquez, and David Chiang. 2024. [Killkan: The automatic speech recognition dataset for kichwa with morphosyntactic information](#). In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 9753–9763.

Yukinori Takubo. 2021a. Hougen wo kana de kaku: Ryuukyuu miyakogo ikemahougen wo rei ni. *Kotoba to Mozi*.

Yukinori Takubo. 2021b. [Morphophonemics of Ikema Miyakoan](#). In *Studies in Asian Historical Linguistics, Philology and Beyond*, chapter 5. Brill, Leiden, The Netherlands.

Peter Wittenburg, Hennie Brugman, Albert Russel, Alex Klassmann, and Han Sloetjes. 2006. [ELAN: a professional framework for multimodality research](#). In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC)*.

Masahiro Yamada, Yukinori Takubo, Shoichi Iwasaki, Celik Kenan, Soichiro Harada, Nobuko Kibe, Tyler Lau, Natsuko Nakagawa, Yuto Ninaga, Tomoyo Otsuki, Manami Sato, Rihito Shirata, Gijs van der Lubbe, and Akiko Yokoyama. 2020. Experimental study of inter-language and inter-generational intelligibility: Methodology and case studies of Ryukyuan languages. In Shoichi Iwasaki, Susan Strauss, Shin Fukuda, Sun-Ah Jun, Sung-Ock Sohn, and Kie Zuraw, editors, *Japanese/Korean Linguistics, Vol. 26*. CSLI Publications, Stanford, CA.