KILLKAN: The Automatic Speech Recognition Dataset for Kichwa with Morphosyntactic Information

Chihiro Taguchi[1], Jefferson Saransig[2], Dayana Velásquez[1], David Chiang[1]
[1]University of Notre Dame
Notre Dame, IN, USA
{ctaguchi,dvelasqu,dchiang}@nd.edu
[2]Pontifical Catholic University of Ecuador
Quito, Ecuador
jisaransig@puce.edu.ec

Abstract

This paper presents KILLKAN, the first dataset for automatic speech recognition (ASR) in the Kichwa language, an indigenous language of Ecuador. Kichwa is an extremely low-resource endangered language, and there have been no resources before KILLKAN for Kichwa to be incorporated in applications of natural language processing. The dataset contains approximately 4 hours of audio with transcription, translation into Spanish, and morphosyntactic annotation in the format of Universal Dependencies. The audio data was retrieved from a publicly available radio program in Kichwa. This paper also provides corpus-linguistic analyses of the dataset with a special focus on the agglutinative morphology of Kichwa and frequent code-switching with Spanish. The experiments show that the dataset makes it possible to develop the first ASR system for Kichwa with reliable quality despite its small dataset size. This dataset, the ASR model, and the code used to develop them will be publicly available. Thus, our study positively showcases resource building and its applications for low-resource languages and their community.

Keywords: Kichwa, automatic speech recognition, language resources, low-resource

1. Introduction

Language endangerment has been one of the world's cultural crises, by which many of the world's languages are losing their speakers at an unprecedented pace (Belew and Simpson, 2018). Among efforts to document and revitalize languages, recent years have seen growing attention and work to incorporate digital technologies and media to this end (Jimerson and Prud'hommeaux, 2018; Michaud et al., 2018; Prud'hommeaux et al., 2021; Shi et al., 2021; Tsoukala et al., 2023). In the same spirit, this study presents the KILLKAN[1] corpus, the first dataset for automatic speech recognition (ASR) for the Kichwa language. Though Kichwa is estimated to have a few hundred thousand speakers in Ecuador, it is considered endangered, as the society is undergoing a language shift to Spanish only. In natural language processing (NLP), Kichwa is an extremely low-resource language, as there have been no datasets available for either building language models or conducting computational linguistic research of Kichwa.

Our dataset consists of 4 hours of audio with its orthographic transcription containing 26,544 tokens. Furthermore, each sentence is annotated with its Spanish translation and morphosyntactic information in the CoNLL-U format of Universal Dependencies (UD) (Nivre et al., 2020). To evaluate the utility of the dataset, we train ASR models on it by fine-tuning the pretrained model `wav2vec2-xlsr-53`. The experiment shows that the fine-tuned model's performance was 2.04% Character Error Rate (CER), which is comparable to Wav2Vec2 models fine-tuned on high-resource languages.

In the following section, we provide a linguistic overview of the Kichwa language. Then, Section 3 surveys previous related work done in the field of NLP for Quechuan languages. Section 4 describes the details of our dataset, including the data source, the annotation process, and a brief analysis of the dataset. Section 5 reports the experimental results of training ASR models on our dataset, followed by concluding remarks in Section 6.

Our contributions in this work are the following:

- We publish the first dataset for Kichwa containing manually annotated audio, transcription, its Spanish translation, and morphosyntactic information;
- We develop the first ASR models for Kichwa;
- We present a new UD Treebank for Kichwa incorporated in ELAN annotation;
- Our dataset, the ASR models, and the code used to develop them are publicly available.[2]

[1]KILLKAN stands for ***Ki**chwa uyashkata pay**ll**atak **k**illkak **an**ta* (Kichwa automatic speech recognizer) in Kichwa. The word *killkan* also means "it writes".

[2]The dataset and the code are available in `https://github.com/ctaguchi/killkan`, and the model is available in `https://huggingface.co/ctaguchi/killkan_asr`.

LREC-COLING 2024, pages 9753–9763
20-25 May, 2024.

Figure 1: The distribution of Quechua II languages mentioned in this paper. This map was created with the `lingtypology` library (Moroz, 2017).

2. Background

The Kichwa language. Kichwa is the most widely spoken indigenous language in the Republic of Ecuador, particularly along the Andean mountain range in the middle and the Amazonian region to the east of the country. Though the number of speakers greatly varies among different statistics, the language is estimated to have at least 300,000 speakers (King and Haboud, 2002). Kichwa is classified in the Northern Quechua branch of the Quechua II group in the Quechuan language family. Though the Quechua II group also includes more widely spoken varieties such as Cuzco Quechua and Ayacucho Quechua of Peru, Kichwa shows a number of differences from them in phonology and morphosyntax. For example, Kichwa has lost ejective consonants, possessive suffixes, the inclusive/exclusive distinction in the first-person plural pronoun, has a reduced system of evidentiality (Adelaar, 2021).

Kichwa is in fact an umbrella term that involves several regional varieties of Northern Quechua. The Endangered Language Project (Project, 2023) lists Highland Ecuadorian Kichwa and Lowland Ecuadorian Kichwa, under which several subvarieties are further categorized. See Figure 2 for a summary of the classification of Quechuan varieties, and see Figure 1 for a map of their distribution.

With regard to typological aspects, like other Quechuan varieties, Kichwa is an agglutinative language, where verbal and nominal suffixes and discourse clitics are attached to the root to mark verbal features, cases, and information structure. The example in (1) shows the agglutination of voice, tense, case, and topic morphemes on the verb root *llamka* "to work".

(1) *llamka-naku-nka-kaman=ka*
work-RCP-PROSP-TER=TOP
'Until (someone) works together[3]'

Language contact and code-switching. Since the arrival of Spanish colonizers in the 16th century, Quechuan languages have had language contact with Spanish (Torero, 2007). The centuries of bilingualism in the Andean region have strongly influenced the lexicon of Quechuan languages, and it is common for Kichwa speakers to code-switch to Spanish in daily speech, which is a language variety sometimes referred to as Media Lengua (Deibel, 2019). The spoken samples in our dataset also contain code-switched speech with Spanish. The code-switched segment can range from a morpheme (also called intra-word code-switching (Nguyen and Cornips, 2016)) to a whole phrase; examples from the dataset are shown in (2) and (3), respectively, where code-switched parts are in green and underlined.

(2) *Ñuka=rak* *vacuna-ri-kri-ni.*
1SG=CONT vaccinate-REFL-PROSP-PRS.1SG
'I am going to get vaccinated first.[4]'

(3) *Consulta popular alli=mi ri-ku-n.*
inquiry popular good=FOC go-PROG-PRS.3
'The referendum is going well.[5]'

Most Kichwa speakers are bilingual with Spanish and speak Spanish with non-Kichwa speakers. Though still estimated to have a few hundred thousand speakers, Kichwa is an endangered language that younger generations often do not inherit, speaking only Spanish instead (Acosta Muñoz, 2017). In addition, Kichwa is both politically and socially marginalized, as suggested by the pejorative term for Kichwa, *yanka shimi* "useless language" (Larrea Maldondo et al., 2007; Kowii, 2017). Unlike Quechua in Peru and Bolivia, Kichwa is merely a

[3] RCP: reciprocal voice, PROSP: prospective aspect, TER: terminative case, TOP: topic.

[4] 1SG: first person singular, REFL: reflexive voice, PRS: present tense.

[5] FOC: focus, PROG: progressive aspect, 3: third person.

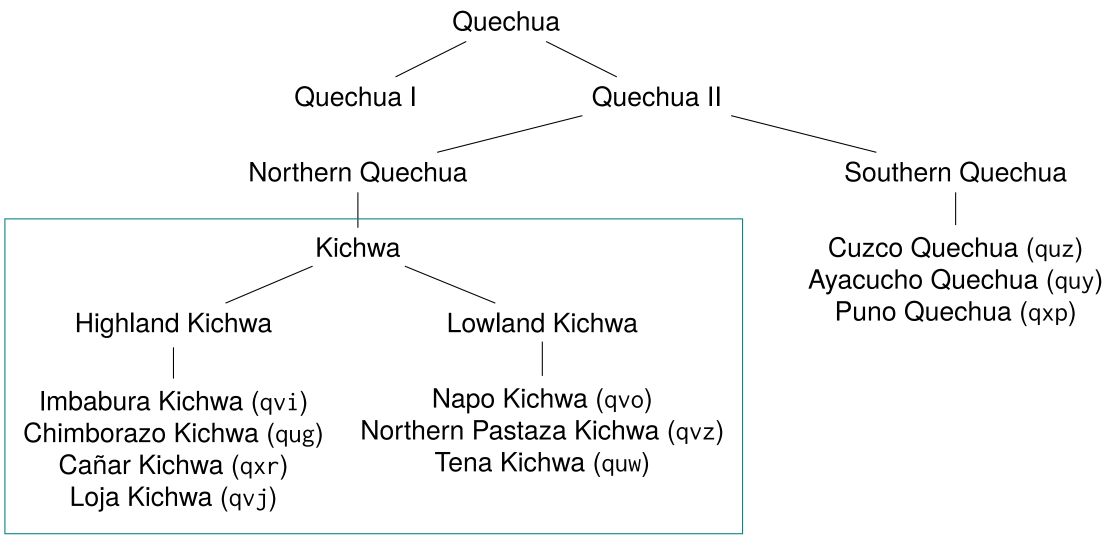

Figure 2: A classification of Quechuan languages with their ISO 639-3 language codes with a focus on the varieties mentioned in this paper. The branches in the box are called Ecuadorian Kichwa.

recognized language and is not granted an official status in Ecuador. These factors add to the ongoing endangerment of the language, and resource building for language technologies is indispensable for both documentation and revitalization of the language. Yet, it is worth mentioning that there are ongoing revitalization activities and Kichwa–Spanish bilingual schooling in Ecuador.

Orthography. The orthography of Kichwa is based on the Latin alphabet. The modern orthographic standardization of Kichwa has undergone two crucial modifications in the late 20th century. The first attempt to standardize the Kichwa orthography was proposed in 1980. This orthography exhibits several influences from the Spanish orthography, such as the use of <c> and <q> for the phoneme /k/ (e.g., <quillca> /kilka/ 'writing') and the use of <hu> to represent the phoneme /w/ (e.g., <huahua> /wawa/ 'child'). In 1998, the orthography was revised again and has been the standard since then (Chasiquiza, 2019). The major modifications are phonology-based simplification, where redundant graphemes such as <qu>/<c> for /k/ and <hu> for /w/ were changed to <k> and <w>, respectively. Though the old orthography is still sometimes informally used, the transcription in our dataset is in the new orthography since the latter orthography is officially and widely used in today's writing.

The modern Kichwa orthography has 18 letters including three digraphs: <a>, <ch>, <h>, <i>, <k>, <l>, <ll>, <m>, <n>, <ñ>, <p>, <r>, <s>, <sh>, <t>, <u>, <w>, <y>. On top of this, two graphemes, <ts> and <z>, may also be used for a small number of words depending on dialects. For code-switched Spanish words, Spanish orthography is used, though Kichwa orthography may also be used for old loanwords such as <ura> 'time' from Spanish *hora*. Though the correspondence between the orthography and pronunciation is more or less regular, there are slight dialectal differences in the actual phonetic value for each grapheme. For example, the word *alli* "good" is pronounced as /ali/ in Imbabura Kichwa and /aʐi/ in Chimborazo Kichwa.

3. Related Work

Although there is no previous dataset for Ecuadorian Kichwa, there have been several efforts to create datasets and NLP applications for other related Quechuan languages, especially Southern Quechua varieties such as Cuzco Quechua of Peru. Rios and Mamani (2014) developed a text normalization pipeline and a morphological analyzer for Cuzco Quechua, to which a machine translation system and a dependency treebank are added in their later work (Rios, 2016). Cardenas et al. (2018) is a speech corpus for Ayacucho (Chanca) Quechua and Puno (Collao) Quechua, which are both Southern Quechuan languages of the Quechua II group spoken in Peru. Ortega et al. (2020) introduces a new parallel text corpus and trains a neural machine translation system for Quechua from Peru and Bolivia, though it does not mention which specific Quechuan variety the text is written in. Since Quechuan languages are highly agglutinative, they have been sometimes used in morphology-related tasks in NLP. For example, Chen and Fazio (2021) investigates the effect of morphology-aware segmentation instead of Byte-

Pair Encoding (BPE) on Quechua.[6] More recently, another speech dataset for Peruvian indigenous language was released that includes ∼180 hours of Southern Quechua audio (Zevallos et al., 2022). All in all, NLP research and applications for Quechuan languages have centered around the varieties spoken in Peru and Bolivia, and other varieties like Ecuadorian Kichwa have yet to be included in language technologies.

4. Dataset

This section describes the details of our dataset.

4.1. Source

The source of the audio in the dataset is a radio program "Jaboneropak Ayllullaktapi" (In the neighborhood of Jabonero) provided by Radialistas Apasionadas y Apasionados,[7] an Ecuador-based non-profit radio station. It is a compilation of fictional stories related to life during the COVID-19 pandemic. The program is published under a Creative Commons BY-SA license, permitting re-use and re-distribution of the work. The acted characters include male and female with various voice qualities and with both adult and child roles. Though the detailed demographic information of the voice actors is unavailable, it is certain that the speech contains several regional varieties of Highland Kichwa. The radio program contains 20 episodes in total, each of which has a length of ∼12 minutes approximately. The total audio length of the whole dataset is ∼234.86 minutes (∼3.91 hours). The dataset contains 3,928 samples, where each audio sample corresponds to a sentence. The average length of a sample is ∼3.59 seconds. The transcription contains 26,544 tokens, and the average length of a token was ∼6.12 characters. The average sentence length was ∼6.76 tokens.

4.2. Annotation

The annotation of the dataset contains the following elements: time-aligned sentence-level transcriptions, their translation in Spanish, and morphosyntactic annotation compatible with UD. All of these annotations were done in ELAN (The Language Archive, 2023), and the annotated data are saved as XML-based EAF (ELAN Annotation Format) files. ELAN is software commonly used to annotate spoken audio and video clips collected during linguistic fieldwork. A screenshot of the annotation interface for the dataset building in this study is shown in Figure 3. To create an ASR dataset containing pairs of an audio sample and its transcription, the original audio files were segmented into sentence-level audio files based on the timestamps logged in the EAF files. To process the annotation document with Python, the annotated EAF files were parsed into Python objects by the `pympi` library (Lubbers and Torreira, 2013–2021). Though there are UD treebanks that were converted from ELAN-native annotation (Östling et al., 2017), our dataset is the first attempt to directly incorporate UD annotation in the CoNLL-U format into ELAN to our knowledge.

4.2.1. Transcription

The source website provides transcriptions in Kichwa for each episode. However, there were three problems in using the provided transcriptions. First, words that are actually said by the actors often differ from the transcriptions (token-level inconsistencies). Second, the actors often insert short sentences or interjections that do not appear in the transcriptions (utterance-level inconsistencies). Third, the provided transcriptions have inconsistencies in the orthography (orthographic inconsistencies). Table 1 summarizes the errors that the original transcription had compared to manually corrected transcriptions. The metrics used in the Table, Character Error Rate (CER), Word Error Rate (WER), and Word Information Lost (WIL), are defined as follows:

$$\text{CER, WER} = \frac{S + D + I}{N}$$

$$\text{WIL} = 1 - \frac{C}{N} + \frac{C}{P},$$

where S, D, I are the numbers of necessary substitutions, deletions, and insertions, respectively, to match the reference text, and N is the number of characters (for CER) or words (for WER and WIL). P is the number of words in the prediction, and C is the number of correctly predicted words. As the Table shows, the original transcription had 22.7% CER compared to the corrected transcriptions, meaning that approximately one in five characters was either missing, wrong, or unnecessary, and 54.6% WER, meaning that more than half of the originally transcribed words required some correction. Furthermore, 7.2% of the actual utterances was missing from the original transcriptions. Because these discrepancies make it difficult to automatically align the transcriptions with the audio segments, every sentence was manually checked and aligned.

4.2.2. Translation

The Spanish translations of the transcriptions are also given by Radialistas. However, they tend to

[6] The paper does not mention what variety of Quechua was used in their experiments. Their dataset description implies that some Peruvian varieties were used.

[7] https://radialistas.net.

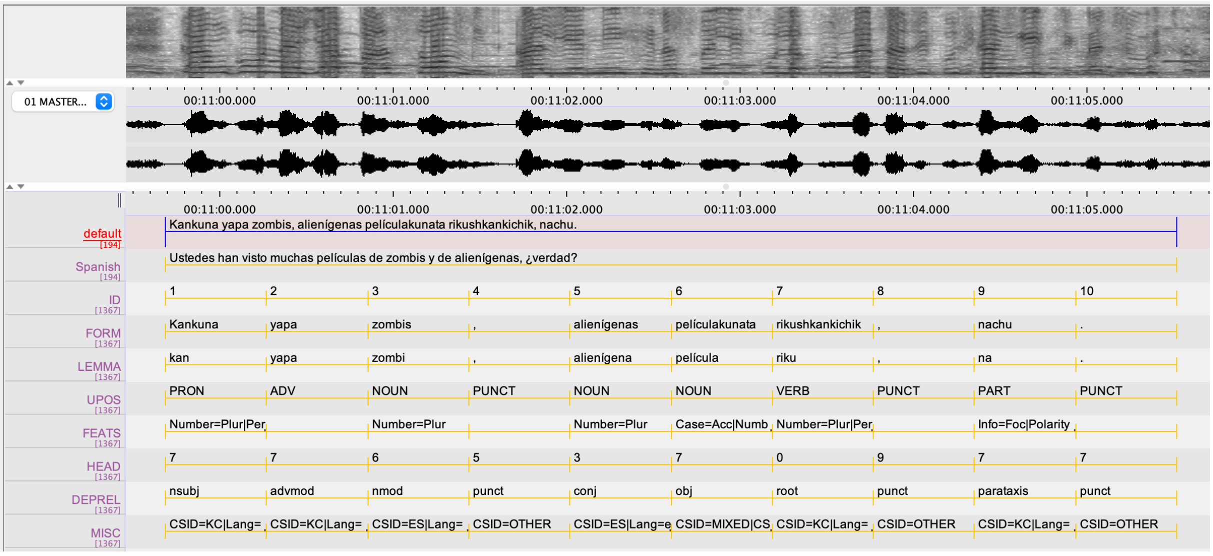

Figure 3: A screenshot of annotating a transcription, its Spanish translation, and UD-style morphosyntactic information in ELAN.

Metrics	Raw (%)	Normalized (%)
CER	22.70	20.76
WER	54.61	43.63
WIL	74.13	55.86
Empty ratio	7.22	

Table 1: A summary on how correct the original transcriptions are with respect to what is actually said in the audio. The "Raw" column shows a comparison with no preprocessing to the transcriptions, while the "Normalized" column shows the results after applying lowercasing and removing punctuation. "Empty ratio" refers to the ratio of the number of uttered sentences that were not in the original transcriptions out of the total number of sentences.

be free translations that depend on the surrounding contexts and sometimes deviate from the information expressed in Kichwa. For this reason, the translations were manually checked by a Kichwa–Spanish bilingual speaker and were corrected if necessary.

4.2.3. Morphosyntactic annotation

The morphosyntactic annotation of this dataset follows the CoNLL-U format of UD that annotates the lemma (`LEMMA`), part-of-speech (`UPOS`), morphological features (`FEATS`), syntactic head (`HEAD`), and dependency relation (`DEPREL`) for each token.

Since Kichwa is highly agglutinative and employs a number of suffixes to express functional meanings, there are several morphological features that are absent in the standard UD guidelines and are newly introduced in this dataset. A list of newly introduced morphological features are summarized in Table 2. The feature `Deixis=Ven` stands for the ventive (cislocative) morpheme that expresses the "coming" motion in the action expressed by the verb. The feature key `Focus=` corresponds to focus-sensitive morphemes. Kichwa has the additive focus marker *=pash* "also" and the restrictive focus marker *=lla* "only". The feature key `Switch=` is used to mark the switch-reference features (Finer, 1985) that co-occur with converbs[8]. Switch-reference in Kichwa specifies whether the subject of the subordinate clause is the same as or different from that of the main clause. For example, in (4 a), the subject is the same, while in (4 b), the subject is different:

(4) (a) *miku-nkapak* *muna-ni.*
eat-CNV.PRP.**SS** want-PRS.1SG
'I want to eat.'

(b) *miku-chun* *muna-ni.*
eat-CNV.PRP.**DS** want-PRS.1SG
'I want (somebody else) to eat.[9]'

Another significant modification from the standard UD guidelines is that this dataset annotates topic and focus in Kichwa as morphological features. In current UD, morphological features cannot express grammatically marked topic and focus, because the guidelines do not have any features for them. One reason for this treatment is that

[8] A converb is "a nonfinite verb form whose main function is to mark adverbial subordination" (Haspelmath, 1995).

[9] CNV: converb, PRP: purposive mood, SS: same subject, DS: different subject

Feature	Morpheme	Description
`Deixis=Ven`	*mu*	Ventive (cislocative)
`Focus=Addit`	*pash*, *pish*	Additive focus-sensitive marker
`Focus=Restr`	*lla*	Restrictive focus-sensitive marker
`Info=Top`	*ka*	Topic in the information structure
`Info=Foc`	*mi*, *chu*, *tak*	Focus in the information structure
`State=Cont`	*rak*	Continuative state
`Switch=Same`	*shpa*, *nkapak*	Switch reference with the same subject
`Switch=Diff`	*kpi*, *chun*	Switch reference with a different subject

Table 2: A list of newly introduced morphological features.

markers like topic and focus are syntactically less selective and can be attached to both nominal and verbal expressions. Because UD's morphological features only allow for lexical, nominal (e.g., case), and verbal features (e.g., tense), features that have to do with the information structure cannot fit in the framework. Indeed, unlike canonical affixes, the morphemes listed in Table 2 have less syntactic restrictions as to which syntactic category they can be attached to; therefore, previous studies call them *morfemas independientes* "independent morphemes" (Chasiquiza, 2019) or *enclíticos* "enclitics" (Catta, 1994).

In the standard UD guidelines, clitics are usually treated as independent tokens and are not represented as morphological features of the head token. However, in this approach, it is impossible to annotate the topic and focus features as morphological features, and therefore information structure remains underrepresented in current UD for topic-prominent languages (Li and Thompson, 1976) that mark topic and focus morphologically like Kichwa. For this reason, we tentatively added those information-structural features, which can be automatically converted to separate tokens if necessary.

Given the frequent code-mixing with Spanish in spoken Kichwa, the annotation in the dataset also includes the language code and the intra-word code-switching boundary for each token. The code-switching annotation is listed in the `MISC` column, following the format in other code-switching UD treebanks (Çetinoğlu and Çöltekin, 2022).

4.3. Analysis

This subsection provides a brief analysis of our dataset with a focus on agglutinativity and code-switching of Kichwa.

Morphological complexity. Table 3 reports the morphological complexity scores of our Kichwa dataset based on the measures proposed in Çöltekin and Rama (2022). The table demonstrates that the morphological complexity of the Kichwa dataset is the highest for MSP (mean size of paradigm), WS (information in word structure), IS (inflectional synthesis), MFH (morphological feature entropy). This shows the extremely high agglutinativity of Kichwa morphology, because MSP, IS, and MFH are calculated based on the diversity of inflected forms and morphological features. On the other hand, our dataset did not show a high degree of complexity in terms of TTR (type–token ratio), WH (word entropy), and LH (lemma entropy). This implies that there is not much diversity in the vocabulary of the dataset, since the dataset consists of a series of stories and has common topics and characters throughout the radio program.

Measure	Kichwa	min	max
TTR	0.24	0.17 (vi)	0.59 (kor.kai)
MSP	**3.73**	0.99 (kor.kai)	2.52 (chu)
WS	**1.05**	0.16 (urd)	0.62 (lat.itt)
WH	10.70	8.94 (afr)	12.84 (kor.gsd)
LH	8.11	7.99 (chu)	12.85 (kor.gsd)
IS	**34.31**	0.00 (jpn)	19.13 (eus)
MFH	**5.20**	1.03 (kor.gsd)	4.04 (ces.fic)

Table 3: The morphological complexity scores of our Kichwa dataset and its comparison to the minimum and maximum scores reported in Çöltekin and Rama (2022). The codes in parentheses refer to specific UD datasets, and the measures are type–token ratio (TTR), mean size of paradigm (MSP), information in word structure (WS), word entropy (WH), lemma entropy (LH), inflectional synthesis (IS), and morphological feature entropy (MFH); see Çöltekin and Rama (2022) for details. The boldfaced scores in Kichwa mean that they are higher than any other reported scores.

Code-switching. Table 4 shows the distribution of languages in the dataset. Code-switched tokens comprise ∼11.19% of the entire dataset, and approximately half of them are word-internally code-

Language		Ratio (%)
Kichwa-only		64.15
Code-switched	Spanish-only	5.83
	Spanish–Kichwa	5.59

Table 4: A summary of the ratios of code-switched tokens. 'Spanish–Kichwa' shows the ratio of tokens with intra-word code-switching. Other tokens are punctuation symbols.

	CER	WER	WIL
KILLKAN, all	**2.94**	**19.94**	**34.29**
KILLKAN, 2k	3.57	23.28	39.30
KILLKAN, 1k	4.96	32.58	51.82
KILLKAN, 500	7.35	47.12	69.31
Huqariq	28.73	—	—

Table 5: The results of Kichwa ASR on the test set. Huqariq (Zevallos et al., 2022) shows the result of Southern Quechua ASR fine-tuned on 144-hour data with pretrained Spanish Wav2Vec2. Note that their results are from their test set in Southern Quechua and not from our test set in Kichwa.

switched tokens. It is empirically known that agglutinative languages in language contact tend to derive morpheme-level code-switching such as in Turkish–German (Çetinoğlu and Çöltekin, 2019), and the code-switching distribution in Kichwa also follows this tendency. As Deibel (2019) pointed out, code-switched Spanish words appear either in an uninflected form (root) or in a fully inflected form, which can be followed by Kichwa morphemes, and, on the contrary, Kichwa stems are not followed by Spanish morphemes. In terms of the selectivity of parts-of-speech, various syntactic categories can be code-switched. Though open-class categories such as nouns and verbs commonly exhibit code-switching, closed-class categories like conjunctions also employ Spanish words, particularly in spoken varieties, as exemplified in the underlined word in (5). Other colored segments are open-class Spanish words.

(5) *kay=ka* *gasto=chu* *o*
this=TOP expense=FOC.PLQ or
inversión=chu *ka-n.*
investment=FOC.PLQ be-PRS.3
'Are these expenses or investments?[10]'

5. Experiments: ASR

This section reports the results of training the first ASR models for Kichwa based on our proposed dataset.

5.1. Setup

We developed a Kichwa ASR model by fine-tuning `wav2vec2-xlsr-53` with the Kichwa dataset. Wav2Vec2 is a framework for pretraining a self-supervised ASR model that learns contextualized speech representations (Baevski et al., 2020). Wav2Vec2 first segments the raw speech input into frames with 16kHz sampling rate and encodes into 512-dimensional features through 7 convolution blocks. The feature vectors are then quantized into discrete values using Gumbel-softmax (Jang et al., 2017), and these quantized values are used as the labels later during the pretraining. For the training step, some parts of the input values are masked, and the same feature vectors are fed into Transformer layers to predict the discrete labels of the masked frames, through which the model learns generalized speech representations. In this way, Wav2Vec2 does not require manually labeled datasets for training and is able to be flexibly fine-tuned to a wide range of speech-related downstream tasks. In particular, its offset pretrained model Wav2Vec2-XLSR-53 is trained on 53 languages, and it has been empirically shown that it has a strong adaptability to various languages by fine-tuning with small datasets (Conneau et al., 2020).

For our purpose, the training, validation, and test sets were generated by an 8:1:1 split, respectively. During the preprocessing, we removed samples shorter than 1 second and longer than 15 seconds to ensure that frame masking is correctly done and to prevent the out-of-memory error, respectively. The learning rate was set to 10^{-4}. We also trained models with smaller training sizes with 500, 1k, and 2k samples to imitate various degrees of low-resource settings. The training was run for 30 epochs on 1 NVIDIA A10 GPU with a 24GB RAM. The training took about 6 hours to complete, and the average power usage during the training was about 102W. For the evaluation metrics, we used CER, WER, and WIL.

5.2. Results

The experimental results are shown in Table 5. It compares four ASR models trained on different numbers of training samples: all samples (3,128),

[10] TOP: topic, PLQ: polar question

Spanish	Reference Prediction	Shuk **periodista ecuatoriano** rurashka. Shuk **periodiste cuatoriano** rurashka.
Code-switching	Reference Prediction	Ama kayapa alcaldíata **visita**nkapak sakishun. Ama kayapa alcaldíata **wisita**nkapak sakishun.
Kichwa	Reference Prediction	Ñukanchikpa tarpushkataka yalli **mishki** kan. Ñukanchikpa tarpushkataka yalli **nishki** kan.
Spacing	Reference Prediction	**Shuk kalluka** rurashallami ninkapak. **Shukkalluka** rurashallami ninkapak.
Punctuation	Reference Prediction	Mana pitapash llakichik**?** Mana pitapash llakichik**.**
Alternative spelling	Reference Prediction	**Kikinpak** warmi muspa ñawi mana pinkay niwarka. **Kikinpa** warmi muspa, ñawi mana pinkay niwarka.
Interjection	Reference Prediction	Paykunapa kawsaykunaka, **uff**, ninan llakipimi kan. Paykunapa kawsaykunaka, ninan llakipimi kan.

Table 6: Examples of errors in the predicted transcriptions for the dev set. Errors are in bold-faced type.

2k samples, 1k samples, and 500 samples, with the same hyperparameters. The best model was the one trained with the most training data, which conforms with the general trend in machine learning.

For comparison, Table 5 also lists the CER score of the Southern Quechua ASR model (Huqariq) reported in Zevallos et al. (2022); Huqariq was fine-tuned on Spanish monolingual Wav2Vec2 with 144-hour Southern Kichwa training data. Though the test datasets and the pretrained models are different between our studies and theirs, the clear contrast in CER (28.73% and 2.94%) shows the relatively successful performance of the Kichwa ASR model that was only trained on less than 3% of the Southern Quechua training data. Importantly, even the extremely low-resource scenario with only 500 training samples achieved 7.35% CER. Note that WER in Kichwa can be higher than WER in analytic languages like English, as tokens in Kichwa tend to consist of more characters with multiple agglutinated suffixes. For example, the average length of English tokens in the GUM corpus (Zeldes, 2017) is 4.08 while that of Kichwa tokens in our dataset is 6.04.

5.3. Error analysis

For an error analysis, we prepared seven error types (`Spanish`, `Code-switching`, `Kichwa`, `Spacing`, `Punctuation`, `Alternative spelling`, `Interjection`) and categorized the errors found in the dev set. `Spanish`, `Code-switching`, and `Kichwa` are errors in transcribing tokens in those languages. `Spacing` is an error where unnecessary spacing is inserted or a necessary spacing is omitted. `Punctuation` is an error in choosing a punctuation symbol or capitalization. `Alternative spelling` is an error where the spellings in both the reference and prediction texts are acceptable. In other words, this type of error is not a wrong transcription in practice. `Interjection` is an error in transcribing an interjection tokens. Table 5.2 lists an actual prediction given by the model for each error type.

Error	Ratio (%)
`Punctuation`	34.32
`Kichwa`	27.39
`Alternative spelling`	12.21
`Spanish`	10.23
`Code-switching`	10.23
`Spacing`	4.95
`Interjection`	0.66

Table 7: The distribution of each transcription error type in the dev set.

Table 7 provides the distribution of the transcription error types found in the dev set. The most common errors were punctuation errors, which took up more than one-third of the errors. Given the fact that 67.81% of the dataset is Kichwa tokens and 8.19% either Spanish or code-switched as shown in Table 4, it can be observed that Spanish and code-switching tokens tend to cause errors relatively more often than Kichwa tokens. Because code-switched Spanish words tend to be either technical words, proper nouns, or other relatively uncommon words, it is difficult to train the model to be able to predict such corner cases correctly in this monolingual fine-tuning. The prediction examples also exhibit the model's confusion in different

	CER	WER	WIL
KILLKAN, all	**2.04**	**13.41**	**23.27**
KILLKAN, 2k	2.69	16.96	29.03
KILLKAN, 1k	3.79	24.20	39.45
KILLKAN, 500	6.13	39.26	59.98

Table 8: The results of Kichwa ASR on the test set after normalizing texts by lowercasing and removing punctuation.

Spanish spellings that share the same phoneme, such as <ho>/<o> and <v>/<b>. Investigation of the methods to improve the transcription of low-frequency code-switched segments is beyond the scope of this study and is left for future work.

Considering the fact that the most common errors were mere punctuation errors and capitalization errors, we also measured the metrics after normalizing texts by lowercasing and removing punctuation. As summarized in Table 8, without casing and punctuation errors, CER was 2.04% and WER 13.41% for the best performing model.

6. Conclusion

This study presented KILLKAN, the first linguistic dataset for Kichwa. It contains speech and manually annotated transcription, Spanish translation, and morphosyntactic parsing information in UD's CoNLL-U format. Our dataset also annotates morpheme-level code-switching with Spanish, which enabled us to conduct linguistic analyses related to code-switching such as measuring code-switching frequency.

Our study showcased the process of resource building and ASR model development for an extremely low-resource language. The experimental results demonstrated 2.04% CER for the speech recognition task by the ASR model trained on less than 4 hours of audio data from our KILLKAN dataset. Though this is a promising result for the extremely low-resource language, the analysis of the predicted output highlighted the difficulty for the model to correctly predict uncommon code-switched words. Since code-switching is a common linguistic activity found across all over the world, especially among endangered languages in contact with other prestige languages, it is an important remaining task to improve prediction accuracy of code-switched words. Also, the experimental results suggested that having more training samples is likely to contribute to improving the performance of Kichwa ASR, calling for more active resource building for low-resource languages.

7. Ethical Considerations

As our dataset has been developed only based on publicly available audio data, there is no direct concern of copyright infringement in this work. However, there are several potential ethical concerns pertaining to technologies for low-resource languages in general.

Accessibility. Though our dataset and model are publicly available, the mode of the distribution is primarily in English, which might be an obstacle for the non-English-speaking users. We will try to mitigate the disproportionate accessibility by adding descriptions in Kichwa and Spanish.

Demand by the community. Although our project was positively regarded by several native speakers during the first author's fieldwork in Quito, it does not mean that the technology should be embraced unconditionally by all speakers.

Language standardization. As described in Section 2, Ecuadorian Kichwa has a number of subdialects that have slightly different vocabulary, phonology, and morphology from each other. Since our dataset and our ASR model are based on the standardized writing system, they might become an implicit force to use linguistic expressions of standardized Kichwa. While this could be a positive effect on the literacy, it could also negatively affect the linguistic diversity of the Kichwa-speaking world.

8. Acknowledgements

This material is based upon work supported by the National Science Foundation under Grant No. BCS-2109709 and by the University of Notre Dame under the Summer Language Abroad Grant (Quechua). We are grateful to Luis Santillán and Lourdes Perugachi for the feedback on the project and the support during the linguistic fieldwork. We also thank the feedback given at the VI Seminario Internacional Revitalizando Ando and at the Tecnologías Digitales y Lenguas Indígenas Workshop.

9. Bibliographical References

Felipe Esteban Acosta Muñoz. 2017. *Shunguhuan Yuyai: The battle for Kichwa language and culture revitalization in Ecuador as thinking-feeling and performance*. Honors thesis, University of North Carolina at Chapel Hill.

Willem F.H. Adelaar Adelaar. 2021. Morphology in Quechuan languages. In *The Oxford Encyclopedia of Morphology*. Oxford University Press.

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

Anna Belew and Sean Simpson. 2018. The status of the world's endangered languages. In Kenneth L. Rehg and Lyle Campbell, editors, *The Oxford Handbook of Endangered Languages*, Oxford Handbooks, pages 21–47. Oxford University Press.

Ronald Cardenas, Rodolfo Zevallos, Reynaldo Baquerizo, and Luis Camacho. 2018. Siminchik: A speech corpus for preservation of Southern Quechua. In *Proceedings of the Workshop on Improving Social Inclusion using NLP: Tools, Methods, and Resources (ISINLP2)*, pages 21–26.

Javier Catta. 1994. *Gramática del quichua ecuatoriano*. Ediciones Abya-Yala.

Özlem Çetinoğlu and Çağrı Çöltekin. 2019. Challenges of annotating a code-switching treebank. In *Proceedings of the 18th International Workshop on Treebanks and Linguistic Theories (TLT, SyntaxFest 2019)*, pages 82–90.

Luis Montaluisa Chasiquiza. 2019. *La estandarización ortográfica del quichua ecuatoriano: Consideraciones históricas, dialectológicas y sociolingüísticas*. Universidad Politécnica Salesiana.

William Chen and Brett Fazio. 2021. Morphologically-guided segmentation for translation of agglutinative low-resource languages. In *Proceedings of the 4th Workshop on Technologies for MT of Low Resource Languages (LoResMT2021)*.

Alexis Conneau, Alexei Baevski, Ronan Collobert, Abdelrahman Mohamed, and Michael Auli. 2020. Unsupervised cross-lingual representation learning for speech recognition. arXiv:2006.13979.

Isabel Deibel. 2019. Adpositions in media lengua: Quichua or Spanish? – Evidence of a lexical-functional split. *Journal of Language Contact*.

Daniel L. Finer. 1985. The syntax of switch-reference. *Linguistic Inquiry*, 16(1):35–55.

Martin Haspelmath. 1995. The converb as a cross-linguistically valid category. In Martin Haspelmath and Ekkehard König, editors, *Converbs in Cross-Linguistic Perspective: Structure and Meaning of Adverbial Verb Forms - Adverbial Participles, Gerunds*, pages 1–56. De Gruyter Mouton.

Eric Jang, Shixiang Gu, and Ben Poole. 2017. Categorical reparameterization with Gumbel-Softmax. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

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)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Kendall A. King and Marleen Haboud. 2002. Language planning and policy in Ecuador. *Current Issues in Language Planning*, 3(4):359–424.

Ariruma Kowii. 2017. Runa shimi, Kichwa shimi wiñaymanta. *Americanía: Revista de Estudios Latinoamericanos*, pages 153–174.

Carlos Larrea Maldondo, Fernando Montenegro Torres, Natalia Greene López, and María Belén Cevallos Rueda. 2007. *Pueblos indígenas, desarrollo humano y discriminación en el Ecuador*. Ediciones Abya-Yala.

Charles N. Li and Sandra A. Thompson. 1976. Subject and topic: A new typology of language. In Charles N. Li, editor, *Subject and Topic*, pages 457–489. Academic Press, New York.

Mart Lubbers and Francisco Torreira. 2013–2021. pympi-ling: a Python module for processing ELANs EAF and Praats TextGrid annotation files. `https://pypi.python.org/pypi/pympi-ling`. Version 1.70.

Alexis Michaud, Oliver Adams, Trevor Anthony 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.

George Moroz. 2017. *lingtypology: easy mapping for Linguistic Typology*.

Dong Nguyen and Leonie Cornips. 2016. Automatic detection of intra-word code-switching. In *Proceedings of the 14th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 82–86.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Jan Hajič, Christopher D. Manning, Sampo Pyysalo, Sebastian Schuster, Francis

Tyers, and Daniel Zeman. 2020. Universal Dependencies v2: An evergrowing multilingual treebank collection. In *Proceedings of the Twelfth Language Resources and Evaluation Conference (LREC)*, pages 4034–4043.

John E. Ortega, Richard Castro Mamani, and Kyunghyun Cho. 2020. Neural machine translation with a polysynthetic low resource language. *Machine Translation*, 34(4):325–346.

Robert Östling, Carl Börstell, Moa Gärdenfors, and Mats Wirén. 2017. Universal Dependencies for Swedish Sign Language. In *Proceedings of the 21st Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 303–308.

Emily Prud'hommeaux, Robbie Jimerson, Richard Hatcher, and Karin Michelson. 2021. Automatic speech recognition for supporting endangered language documentation. *Language Documentation & Conservation*, 15.

Annette Rios. 2016. A basic language technology toolkit for Quechua. *Procesamiento de Lenguaje Natural*, 56:91–94.

Annette Rios and Richard Castro Mamani. 2014. Morphological disambiguation and text normalization for Southern Quechua varieties. In *Proceedings of the First Workshop on Applying NLP Tools to Similar Languages, Varieties and Dialects (VarDial)*.

Jiatong Shi, Jonathan D. Amith, Rey Castillo García, Esteban Guadalupe Sierra, Kevin Duh, and Shinji Watanabe. 2021. Leveraging end-to-end ASR for endangered language documentation: An empirical study on yolóxochitl Mixtec. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1134–1145, Online. Association for Computational Linguistics.

The Language Archive. 2023. ELAN (version 6.6) [computer software].

Alfredo Torero. 2007. *El quechua y la historia social andina*. Fondo Editorial del Pedagógico San Marcos.

Chara Tsoukala, Kosmas Kritsis, Ioannis Douros, Athanasios Katsamanis, Nikolaos Kokkas, Vasileios Arampatzakis, Vasileios Sevetlidis, Stella Markantonatou, and George Pavlidis. 2023. ASR pipeline for low-resourced languages: A case study on pomak. In *Proceedings of the Second Workshop on NLP Applications to Field Linguistics*, pages 40–45, Dubrovnik, Croatia. Association for Computational Linguistics.

Amir Zeldes. 2017. The GUM corpus: Creating multilayer resources in the classroom. *Language Resources and Evaluation*, 51(3):581–612.

Rodolfo Zevallos, Luis Camacho, and Nelsi Melgarejo. 2022. Huqariq: A multilingual speech corpus of native languages of Peru for Speech recognition. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference (LREC)*, pages 5029–5034.

Özlem Çetinoğlu and Çağrı Çöltekin. 2022. Two languages, one treebank: building a Turkish–German code-switching treebank and its challenges. *Language Resources and Evaluation*, pages 1–35.

Çağrı Çöltekin and Taraka Rama. 2022. What do complexity measures measure? Correlating and validating corpus-based measures of morphological complexity. *Linguistics Vanguard*, 9(s1):27–43.

10. Language Resource References

The Endangered Languages Project. 2023. *Catalogue of Endangered Languages*. University of Hawaii at Manoa.

Appendix A. Glossing abbreviations

Gloss	Function
1, 2, 3	1st, 2nd, 3rd person
CNV	converb
DS	different subject
FOC	focus
PLQ PROG	progressive aspect
PROSP	prospective aspect
PRP	purposive mood
PRS	present tense
RCP	reciprocal voice
REFL	reflexive voice
SG	singular
SS	same subject
TER	terminative case
TOP	topic

Table 9: A list of glossing abbreviations used in the paper.