



# Kichwa meets language technologies

Automatic speech recognition for Kichwa

Chihiro Taguchi + Dayana Velásquez, Jefferson Saransig  
University of Notre Dame

## Chihiro

- **Notre Dame hatun yachanawasipi yachakuk (doctorado)**  
Estudiante de doctorado en la Universidad de Notre Dame
  - **Ingeniería, lingüística**
  - **Imashinatak inteligencia artificial nishkawan shimikunata kawsachina, allichina**  
Investigo sobre cómo podemos proteger las lenguas minorizadas con la inteligencia artificial
  - **Kichwa shimita yachakuk**  
Estudio y investigo la lengua kichwa
  - **Japón mamallaktamanta**  
De Japón
-

# Kay presentaciónmanta

---

Sobre esta presentación

Kay proyecto: kichwapa “reconocimiento automático del habla” (RAH)

Reconocimiento automático del habla...

- rimashkata killkak anta

Transcribir lo que decimos

Alli tecnologíakunaka hatun shimikunapaklla

Las tecnologías útiles sólo admiten idiomas ampliamente hablados

- Apple-pa Siri
- YouTube-pa “subtítulos automáticos”

“Quechua” shimikunapakka mana tiyanchu.

No existe tal herramienta para los idiomas quechuas.



Shuyuka kay pankamantami:  
<https://developer.nvidia.com/blog/how-to-build-domain-specific-automatic-speech-recognition-models-on-gpus/>

# Kichwa shimimanta

Sobre el idioma kichwa

“Quechua” shimikunataka Ecuador, Perú,  
Bolivia llaktakunapi rimankuna.

Los idiomas quechuas se hablan principalmente en Ecuador, Perú, y Bolivia.

Kichwa shimitaka Ecuador llaktapimi rimankuna.

El quechua hablado en Ecuador se llama kichwa.

Ecuador mamallaktapa shimika mishu (español)  
shimimi kan, kichwataka mana alli yanapan.

El idioma oficial del Ecuador es el español, y el kichwa no recibe suficiente apoyo del gobierno.



# Ima nishpatak kichwa shimi?

¿Por qué el kichwa?

Notre Dame hatun yachanawasipika quechua shimita yachachin. Chaypi Otavalo llaktamanta yachachikkunawan kichwa shimita yachakurkani.

En Notre Dame se enseña el idioma quechua cada año. Yo estudié el kichwa con los profes de Otavalo.

Ashtawan kichwata yachakunkapak Ecuador llaktamanpash rirkani.

También fui a Ecuador para seguir estudiando el kichwa.

Ñukaka ingeniería yachakuk kashkamanta, yachachikkunawan kichwapak tecnológiata rurankapak llamkay kallarikani.

Como estudiante de ingeniería, comencé a trabajar en la creación de tecnologías lingüísticas para el kichwa.



Ñuka ayllu runakunawan (familia anfitrióna) — Quito, junio 2023

# Imashinatak ruranchik?

---

Imashina: “aprendizaje automático” (machine learning)

- Kay pachapi tukuy shimi tecnologíataka kashna rurankuna.  
Hoy en día muchas tecnologías lingüísticas se construyen de esta manera.

Shinapash...

Sin embargo...

- Aprendizaje automático nishkaka achka datostami mutsurin.  
El aprendizaje automático requiere muchos datos.
  - Kichwa shimipa datoska mana yapa tiyanchu.  
No hay muchos datos disponibles en kichwa.
  - Chaymantami mushuk datosta rurana kanchik.  
Por eso necesitamos crear un nuevo conjunto de datos.
-

# Datoska maymantatak?

---

## Reconocimiento automático del habla nishkapa datoska ishkey layata mutsurin:

Los datos para el reconocimiento automático del habla requieren dos cosas:

- **Rimaypa archivo**  
Los archivos de voz
- **Rimayta killkashka**  
La transcripciones de los archivos

## Maymantatak chay datosta apamuna?

¿De dónde obtenemos los datos?

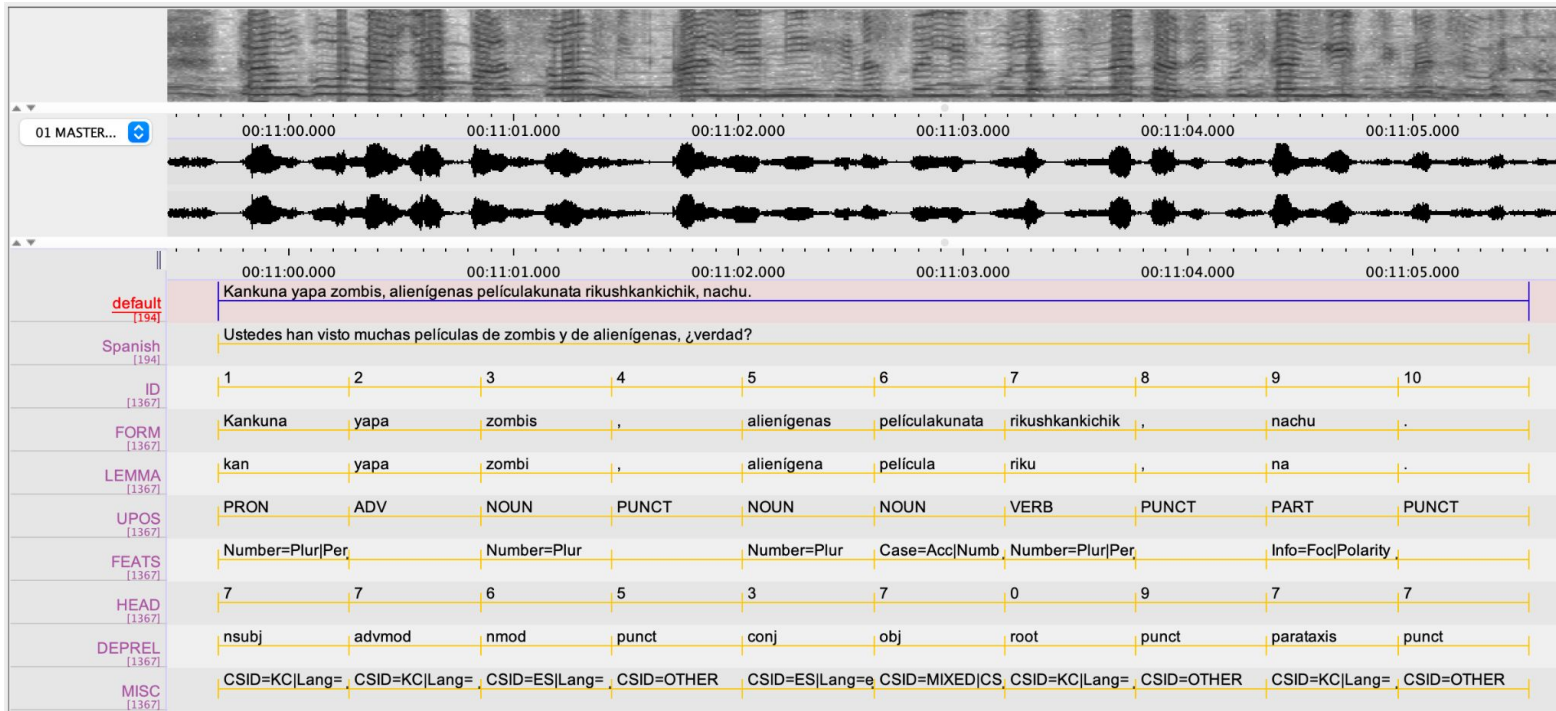
- “Radialistas: Apasionadas y Apasionados” <https://radialistas.net>
  - **Shuk radionovela: “Jaboneropak ayllullaktapi”**  
Una radionovela: “En el barrio del Jabonero”
  - **12 chiniku \* 20 episodiokuna = ~240 chiniku datos**  
12 minutos \* 20 episodios = ~240 minutos de datos
  - **Guíonpash killkashka**  
Con el guión
-

# Imashinatak datosta allichina?

¿Cómo preparamos los datos?

“ELAN” nishka softwarepi datosta allichirkanchik.

Preparamos los datos con el software ELAN



The screenshot displays the ELAN software interface. At the top, there is an audio waveform for a file named '01 MASTER...'. Below the waveform, a timeline shows the audio segments. The main part of the interface is a table with linguistic annotations for the sentence: "Kankuna yapa zombis, alienígenas peliculakunata rikushkankichik, nachu." The table includes columns for ID, FORM, LEMMA, UPOS, FEATS, HEAD, DEPREL, and MISC, with corresponding values for each word in the sentence.

	1	2	3	4	5	6	7	8	9	10
default [194]	Kankuna yapa zombis, alienígenas peliculakunata rikushkankichik, nachu.									
Spanish [194]	Ustedes han visto muchas películas de zombis y de alienígenas, ¿verdad?									
ID [1367]	1	2	3	4	5	6	7	8	9	10
FORM [1367]	Kankuna	yapa	zombis	,	alienígenas	peliculakunata	rikushkankichik	,	nachu	.
LEMMA [1367]	kan	yapa	zombi	,	alienígena	película	riku	,	na	.
UPOS [1367]	PRON	ADV	NOUN	PUNCT	NOUN	NOUN	VERB	PUNCT	PART	PUNCT
FEATS [1367]	Number=Plur Per		Number=Plur		Number=Plur	Case=Acc Numb	Number=Plur Per		Info=Foc Polarity	
HEAD [1367]	7	7	6	5	3	7	0	9	7	7
DEPREL [1367]	nsubj	advmod	nmod	punct	conj	obj	root	punct	parataxis	punct
MISC [1367]	CSID=KC Lang=	CSID=KC Lang=	CSID=ES Lang=	CSID=OTHER	CSID=ES Lang=e	CSID=MIXED CS	CSID=KC Lang=	CSID=OTHER	CSID=KC Lang=	CSID=OTHER



# Modelota imashinatak rurana?

¿Cómo desarrollamos el modelo?

## Wav2Vec2.0 nishka arquitecturawan modelota ruranchik

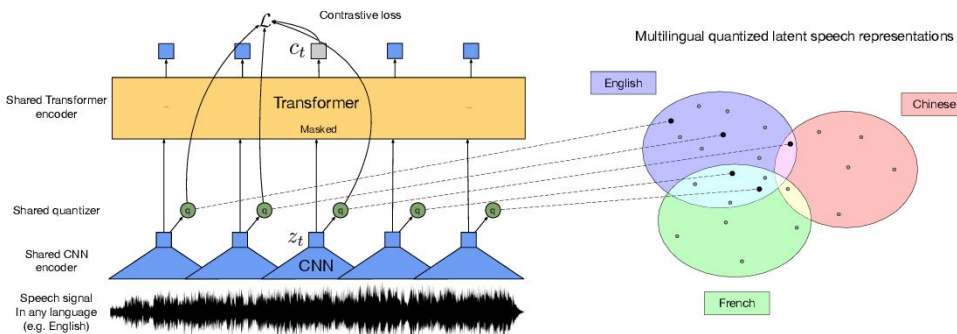
Desarrollamos el modelo con la arquitectura que se llama Wav2Vec2.0

- Ashalla datoswanpash alli RAH modelota ruray ushanchik

Se sabe que funciona bien con una pequeña cantidad de datos

Kay modeloka multilingüemi kan. Kichwa shimipaklla modelota rurankapakka kichwa datoswan entrenanami kanchik.

El modelo original es multilingüe; hay que ajustarlo al kichwa con los datos preparados.



Shuyuka kay pankamanta: <https://huggingface.co/facebook/wav2vec2-large-xlsr-53>

# Modeloman rimashun

---

¡Vamos a hablar al modelo!

**Modelota kay enlacemanta rikuy ushankichik:**

Pueden probar el modelo desde el siguiente enlace:

[https://huggingface.co/ctaguchi/killkan\\_asr](https://huggingface.co/ctaguchi/killkan_asr)

---

# Alli antata rurankapak...

---

Para mejorar el modelo...

**Modeloka ña alli funcionakun.**

El modelo ya funciona bien.

**Shinapash, wakinpika pantarin...**

Sin embargo, a veces se equivoca los palabras españoles y en un ambiente ruidoso

- **Mishu shimikuna (code-switching)**
- **Mana kasilla kuskapi**

**Modelota ashtawan allichinkapakka ashtawan datosta mutsurin!**

Para mejorar el modelo, necesitamos más datos y más personas que hablan y escriben en kichwa

- **ashtawan kichwapi rimakkuna, killkakkuna**
-

# Kichwa shimipak

---

Para el idioma kichwa

Kay proyectopi rurashka modelota, datosta, códigotaka pipash rikuy ushan.

El modelo, los datos, y los códigos creados en este proyecto son disponible públicamente.

- Datos, código: <https://github.com/ctaguchi/killkan>
- Modelo: [https://huggingface.co/ctaguchi/killkan\\_asr](https://huggingface.co/ctaguchi/killkan_asr)

Kichwa shimikunatapash tecnologíapi yaykuchina kanchik.

Debemos incluir los idiomas quechuas en las tecnologías lingüísticas.

Kikinkunapash kashna proyectopi llamkankapak munashpaka llamkashunchik!

Podemos trabajar juntos en este proyecto si ustedes están interesados.

---

Yupaychani! Tapuyta charinkichikchu?

¡Gracias! ¿Tienen preguntas?  
Thank you! Any questions?

correo: [ctaguchi@nd.edu](mailto:ctaguchi@nd.edu)

---