Deep learning and severely under-resourced languages

How much can the model actually learn?

Rolando Coto Solano. Dartmouth College

CLASP Research Seminar Series, University of Gothenburg. March 2023

Meitaki! Wë'ste! Thank you! ¡Gracias!

Cook Islands Team

Sally Akevai Nicholas
Jean Tekura Mason
Teachers USP@Raro
Teachers Ma’uke School
Tyler Peterson
Piripi Wills
Liam Koka'ua
Emma Ngakuravaru Powell

Samiha Datta (ASR)
Victoria Quint (Keyboards)
Jessica Cheng (OCR)
Syed Tanveer (ASR)
Sarah Karnes (Parsing)
Ryan Dudak (Alignment)
Caroline Conway (Morphology)
Hermilla Fentaw (Morphology)

Chibchan Team

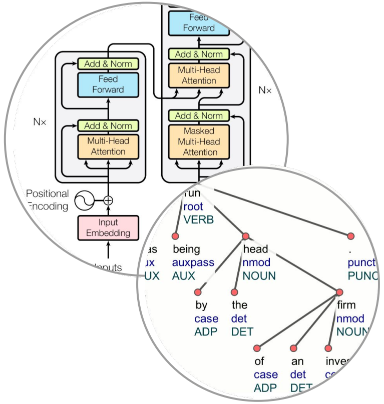
Sofía Flores
Isaac Feldman (NMT)
Veronica Quidore (Parsing)
Annie Tang (Keyboards)
Catharine Herrera (Morphology)
Mien Nguyen (Morphology)

Sharid Loáiciga (Parsing)
Guillermo González
Tai Wan Kim (ASR)
Freddy Obando
Franklin Morales
Alex Jones (NMT)

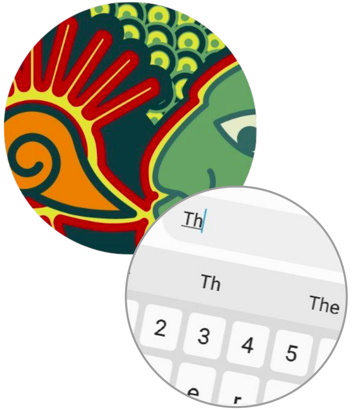
Parts of the talk

Bribri, Cabécar and Cook Islands Māori languages and people

What works and what doesn’t



Problems with simulating under-resourced datasets



What’s valuable to the users

Chibchan Languages: Bribri and Cabécar

Bribri has 7K speakers. Cabécar has 11K speakers.
They are both **vulnerable**.

Bribri Grammar

Relatively complex phonology

7 oral vowels, 5 nasal vowels
19 consonants
5 tones

SOV, Ergative

Ye'	**tö**	ù	sú
I	ERG	house	see-PST.PERF

I saw the house.

Inflectional morphology

Ye'	tö	ù	s**awé**
I	ERG	house	see-PST.IPFV

I would see the house.

Complex demonstratives

dù **e'**	*that bird*
dù **aí**	*that bird [up there, nearby]*
dù **dià**	*that bird [down there, far away]*
dù **se'**	*that bird [that you can hear]*

Bribri Grammar

Numerical classifiers

dù bö̀**tk**	two-(flat) birds
alákölbö́**l**	two-(human) women
dawás bö̀**k**	two-(round) years
awà bö̀**töm**	two-(long) otters

Head-internal relative clause

I built **the house**$_i$ that [you saw ~~the house~~$_i$]

Ye' tö [be' tö **ù**$_i$ sú] ~~ù~~$_i$ (e') yö'

I ERG you ERG house.ABS see.PRF DEM build.PRF

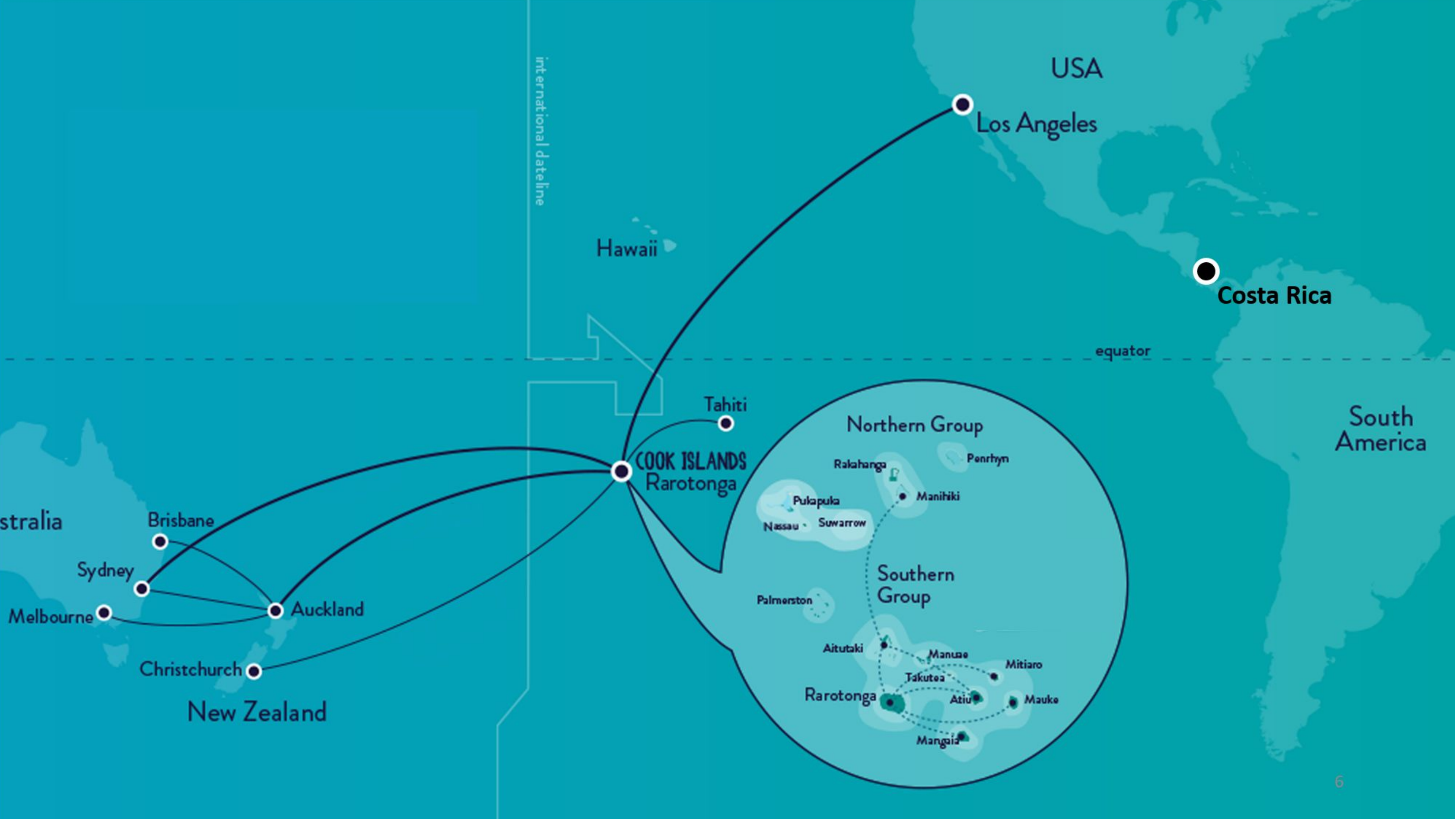
Bribri Data Sources

Oral Corpus
Sofía Flores: bribri.net
(~68 minutes of transcribed audio)

Existing publications
(from Costa Rican universities)
Total: ~90K words

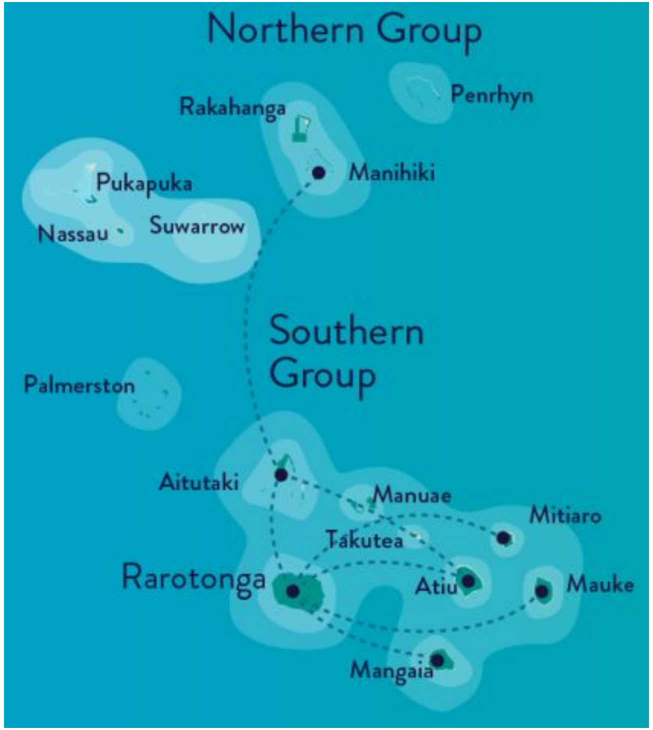
Cabécar Data Sources

Existing publications
(from Costa Rican universities)
Total: ~26K words



USA
Los Angeles
Costa Rica
international dateline
Hawaii
equator
Tahiti
COOK ISLANDS
Rarotonga
South America
stralia
Brisbane
Sydney
Melbourne
Auckland
Christchurch
New Zealand
Northern Group
Rakahanga
Penrhyn
Manihiki
Pukapuka
Nassau
Suwarrow
Southern Group
Palmerston
Aitutaki
Manuae
Mitiaro
Takutea
Rarotonga
Atiu
Mauke
Mangaia

Cook Islands Māori



13K speakers
+8K in NZ and AUS

Endangered in Rarotonga

Vulnerable in the other islands

Cook Islands Māori

Relatively few phonemes	5 vowels:	a e i o u
	9 consonants:	k m n ŋ p r t v ʔ

Isolating morphology

Kua	tunu	au	i	te	taro
PRF	plant	I	ACC	the	taro

I planted the taro.

Kua	'akaruke	atu	te	au	kurī
PRF	leave	away	the	PL	dog

The dogs have left.

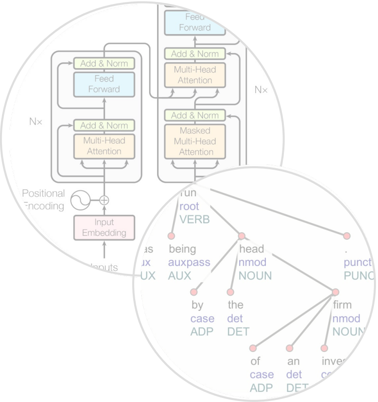
Data Source: *Te Vairanga Tuatua*

Large (dozens of hours)
Linguistically rich
Little annotation
Transcription is a major bottleneck
~4 transcribed hrs (37K words)

Parts of the talk

Bribri, Cabécar and Cook Islands Māori languages and people

What works and what doesn't



Problems with simulating under-resourced datasets

What's valuable to the users

Let’s get it over with: ChatGPT doesn’t know Bribri

RO

Write a brief conversation in Bribri, in the style of https://bribri.net/

Manuel: Kekölma, pü'sa?

(Salutations, how are you?)

Lucía: Amalira, mae. ¿Tü'sa kela?

(Very good, and you? How's it going?)

Manuel: Wína kela. Kólisba yamön melö?

(Things are going well. Did you hear about the festival?)

Lucía: Kólisba kün. Pü'sa ba yamön melö?

(It sounds good. What festival is it?)

Manuel: Yakö yaöyöl nöka, mae. Töka wíknöl weya näkö we'nö, nya'aka kën.

(It's the Yakö Yaöyöl festival. We dance all day and eat traditional food.)

Let's get it over with: ChatGPT doesn't know Bribri

RO

Write a brief conversation in Bribri, in the style of https://bribri.net/

Manuel: Kekölma, pü'sa?

(Salutations, how are you?)

Lucía: Amalira, mae. ¿Tü'sa kela?

(Very good, and you? How's it going?)

Manuel: Wína kela. Kólisba yamön melö?

(Things are going well. Did you hear about the festival?)

Lucía: Kólisba kün. Pü'sa ba yamön melö?

(It sounds good. What festival is it?)

Manuel: Yakö yaöyöl nöka, mae. Töka wíknöl weya näkö we'nö, nya'aka kën.

(It's the Yakö Yaöyöl festival. We dance all day and eat traditional food.)

Manuel: Shkè̠na̠ bua'ë?

Lucía: Bua'ë. Ìs be' shkè̠na̠?

Manuel: Bua'ë. Be' tö i-stsé tö kalö̀te tso?

Lucía: Tö́. Kalö̀te e' dör wé̠?

Manuel: I' dör "Yakö yaöyöl" kalö̀te. Kalö̀tè̠ne̠ ñî̠ e̠na̠ ñà̠ne̠ ta̠î̠.

Even if you try to teach it Bribri, it has trouble learning

RO

In Bribri, the phrase "My name is Rolando" is "Ye' kie Rolando". The phrase "My name is Sharid" is "Ye' kie Sharid". How would you say "My name is Juan"?

In Bribri, "My name is Juan" would be "Ye' kie Juan".

RO

In Bribri, the phrase "How are you all doing?" is "Ìs a' shkè̱na̱?". The phrase "We are doing very well" is "Sa' shkè̱na̱ bua'ë". The phrase "How are you doing?" is "Ìs be' shkè̱na̱?". How would you say "I am doing very well".

In Bribri, "I am doing very well" would be "Ni' shkè̱na̱ bua'ë".

Even if you try to teach it Bribri, it has trouble learning

In Bribri, "I am doing very well" would be "Ni' shkè̱na̱ bua'ë".

RO

In Bribri, the phrase "I saw you" is "Ye' tö be' sú̱". The phrase "I saw her" is "Ye' tö ie' sú̱". How would you say "I am doing very well"?

I apologize, my previous response was incorrect.

In Bribri, "I am doing very well" would be "Ni' bua'ë shkèna".

Even if you try to teach it Bribri, it has trouble learning

I apologize, my previous response was incorrect.

In Bribri, "I am doing very well" would be "Ni' bua'ë shkè̱na̱".

RO

The mistake is that the phrase "I am doing well" should be "Ye' shkèna bua'ë". The word "I" is "ye'" in that context. So, tell me again: How would you say "I am doing well"?

I apologize for the confusion. Thank you for correcting me.

In Bribri, "I am doing well" would be "Ye' shkèna bua'ë".

It knows really bad Cook Islands Māori

RO

Write a brief conversation in Cook Islands Maori, in the style of http://cookislandsdictionary.com/

Kia orana! Te mea koe e kite atu nei i te kakara o te puakaoa?
(Kia orana! Have you noticed the scent of the frangipani flowers?)

E kare, ka rava te kakara o te puakaoa i tenei ra!
(No, I haven't, the scent of frangipani is abundant today!)

E mea ana koe i te rave mai i te aere ki te tautai?
(Are you planning to go fishing?)

Aere atu au ki te tautai i teia po. Kua akarongo au i te rave atu i te aere ki te tautai ki te aroaro o te po.
(I'm going fishing tonight. I've heard that it's good to go fishing at night.)

Do you see forth the smell of the dog?

No, it is enough the smell of the dog toDaY.

Do you mean to do forth the going to fishing?

I go fishing tonight. I have heard that to do forth going to fishing in presence of the night.

It knows really bad Cook Islands Māori

RO

Write a brief conversation in Cook Islands Maori, in the style of http://cookislandsdictionary.com/

Kia orana! Te mea koe e kite atu nei i te kakara o te puakaoa?
(Kia orana! Have you noticed the scent of the frangipani flowers?)

E kare, ka rava te kakara o te puakaoa i tenei ra!
(No, I haven't, the scent of frangipani is abundant today!)

E mea ana koe i te rave mai i te aere ki te tautai?
(Are you planning to go fishing?)

Aere atu au ki te tautai i teia po. Kua akarongo au i te rave atu i te aere ki te tautai ki te aroaro o te po.
(I'm going fishing tonight. I've heard that it's good to go fishing at night.)

Kia orana! Kua 'ongi koe i te kakara o te au tīpani?

Kāre rāi 'e kī te 'enua i te kakara o te tīpani i tēia rā.

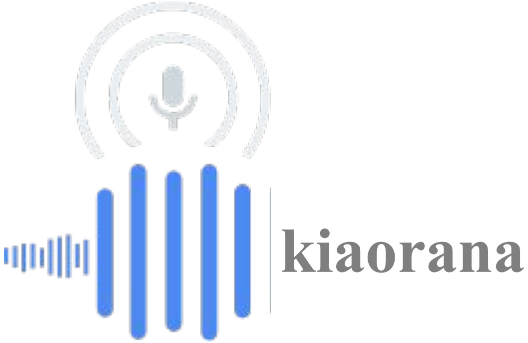
Ka tautai koe ā tēia rā?

Ka tautai au i tēia pō. Tāku i kite ko te pō te tūātau mako nō te tautai.

Algorithms we have tried

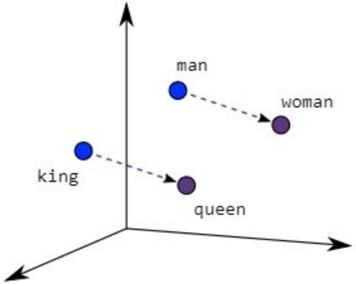
	CIM	Bribri	Cabécar
Speech recognition (ASR)	✅ (4 hrs)	⚙️ (68 mins)	⚙️ (54 mins)
Dependency Parsing Treebanks	✅ (1000 words)	✅ (1500 words)	
Automatic Translation (NMT)		⚙️ (BLEU=14~20)	
Embedding Analysis		✅	
Force-aligned corpus	⚙️	⚙️	⚙️
Predictive keyboards	⚙️	⚙️	⚙️
OCR	⚙️		
Morphological Analysis	⚙️	⚙️	
TTS (speech synthesis; text-to-speech)	⚙️		
NLI (Natural Language Inference)		⚙️	

NLP for Indigenous Languages



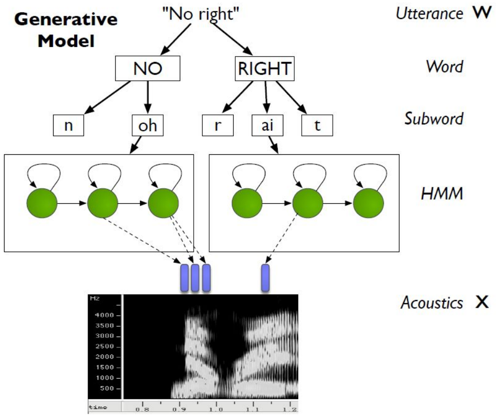
Speech
Recognition

Machine
Translation

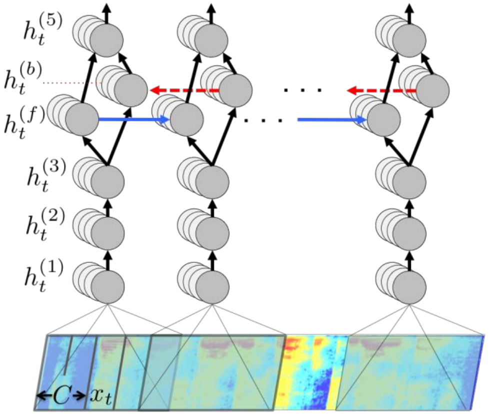


Embeddings

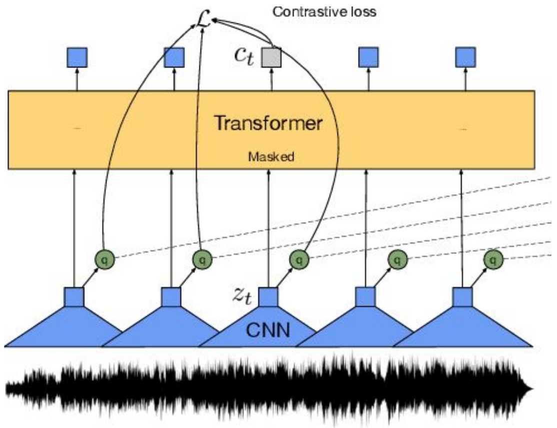
Speech Recognition: Algorithms



HMM/GMM
(Kaldi)



CTC + RNN
(DeepSpeech)



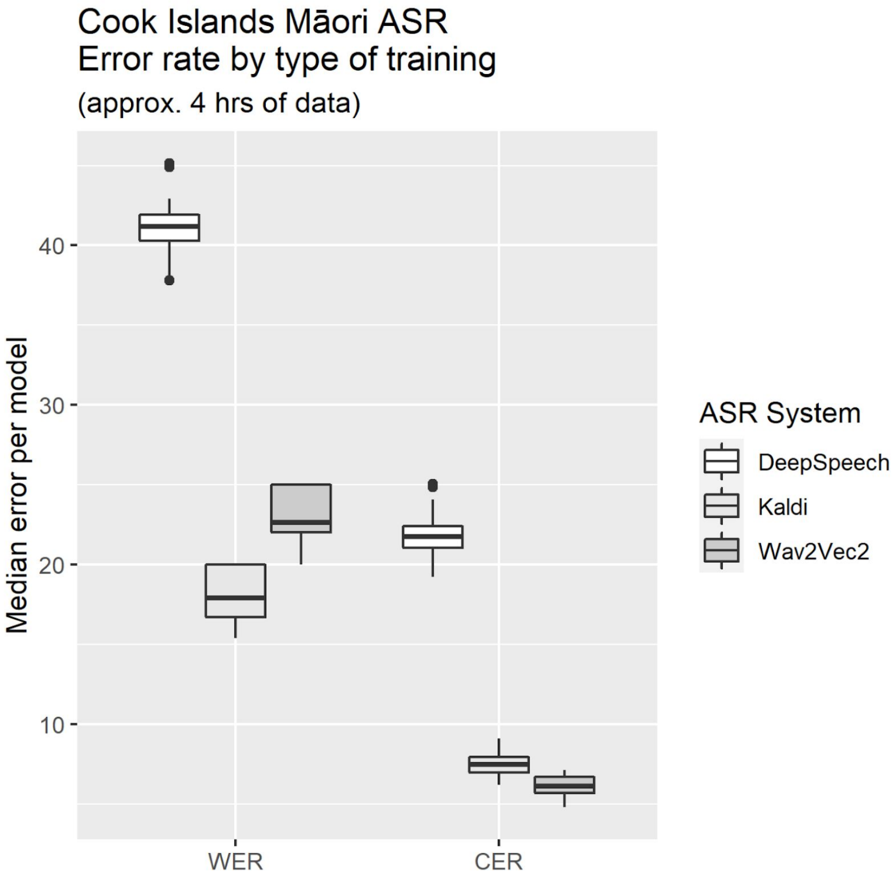
Multilingual Transformer
(Wav2Vec2)

Speech Recognition: Data

237 minutes (~4 hrs), 5033 files
37K total words, 2362 unique words
10 speakers (30-75 years old)
4 islands (Rarotonga, Tongareva, Ma'uke, 'Atiu)

Speaker 1 Māori Transcription	Ana Andrew	29.126	31.067	1.941	I runga i te 'enua ko Tupuaki,
Speaker 1 Māori Transcription	Ana Andrew	31.635	32.731	1.096	i te tuātau ta'ito,
Speaker 1 Māori Transcription	Ana Andrew	33.202	37.468	4.266	tē no'o ra tēta'i māpū māro'iro'i, ko Rū tōna ingoa.
Speaker 1 Māori Transcription	Ana Andrew	38.356	39.477	1.121	Kāre ia i te ariki,
Speaker 1 Māori Transcription	Ana Andrew	39.932	42.371	2.439	ē kāre katoa aia i te tamaiti nā te ariki,
Speaker 1 Māori Transcription	Ana Andrew	42.617	43.383	0.766	inārā,

Speech Recognition: CIM Results



	WER	CER
Kaldi	**17.9** ± **1.7**	7.5 ± 0.8
DeepSpeech	41.1 ± 2.0	21.9 ± 1.6
Wav2Vec2	22.9 ± 2.0	**6.1** ± **0.6**

Speech Recognition: CIM Results

English	*One day I was just sitting in my car*		
Target	i tēta'i rā tē no'o 'ua ara au i roto i tōku motoka	WER	CER
Kaldi	ki tēta'i rā tē no'o 'ua ara 'oki i roto i tōku motoka	15	9
DeepSpeech	i tēta'i a te no'o ara i roto i tōku motoka	31	18
Wav2Vec2	i tēta'i rā tē no'o 'ua ara au i roto i tōku moutakā	8	5

English	*I was sure that it was the pig who had rooted (it up)*		
Target	kua kite ra 'oki au ē nā te puaka i ketu	WER	CER
Kaldi	kua kite rā 'oki au e nā te puaka i ketu	18	5
DeepSpeech	kite rāi koe i nā te puaka i ki	55	38
Wav2Vec2	kua kite rā 'aki au ē nā te puaka i kit	27	10

English	*Absolutely, it will get mixed up*		
Target	āe 'oki ka iroiro atu	WER	CER
Kaldi	'aere ka'iro i roa atu	80	50
DeepSpeech	āe ki ka'iro 'oki roa te	100	50
Wav2Vec2	āe 'oki kā'iro'i roa atu	40	23

Speech Recognition: Bribri Results

English	*So, you were young anyways, right?*	CER	WER
Reference	e' ta be' bák ia tsítsir wake'		
Kaldi	**e'ta** be' bák ia tsítsir wake'	**4**	**29**
DeepSpeech	e' ta **ie'** **i**	65	86
Target	**e'ta** be' bák ia **tsítsi** wake'	6	43
English	*So he left the place where his house was*		
Original	e'rö ie' r ȅ ù ttő améat		
Kaldi	**e'r** **ie'r** ȅ ù tt**ȍ** améat	**12**	**57**
DeepSpeech	**e'**	89	100
Wav2Vec2	e'rö ie' **rế** ù **jtö** améat	22	67

	28 speakers	68 minutes
Kaldi	CER: 33	WER: 50
DS:	CER: 70	WER: 86
W2V2:	CER: 23	WER: 65

Speech Recognition: Bribri Results

English	*Well, you should start telling me why*	CER	WER
Original	ma íkëne apàkö́mine tö ì kuéki		
Kaldi	**mítkene pàkane**	**64**	**100**
DeepSpeech	**e'**	65	86
Wav2Vec2	**mike na i apàkomie të**	65	100

Btw, transfer from English/Mandarin didn't make DeepSpeech work (WER: 100~130).

Speech Recognition: Cabécar Results

English	Only Kál Kébla brought his log of wood, Jak Kébla brought his stone, the *suita* stone	CER	WER
Target	jíbä kal kébla né wa ijé kalí dëká ják kébla né wa jí jákí ju kä dëlëká rä		
Kaldi	**i wa** kal kébla né wa ijé kalí dëká ják kébla né wa **i yína** jákí ju kä dëlëká rä	13	22
DS	**jé jé jé jé jé jé**	83	100
W2V2	**s**ibä kal kébla né wa ijé kalí dëká ják kébla né **ya** jí **jáki** ju kä **rëlëká** rä	**6**	**22**
English	So when he saw it, he turned his face and went to see her; she had the girl in her arms		
Original	jéra ijé te i suáni ra ijé te jé suá ijé wätkáwa tkáu ijé sua ijé wa yaba ka yaba kala		
Kaldi	jéra ijé te i suáni ra ijé te jé suá ijé wätkáwa **ká** ijé **jé suá jé rä** ijé wa yaba **ká jé** yaba kala	18	33
DS	**jé jé jé jé jé jé**	82	95
W2V2	jéra ijé te i suáni ra ijé te jé suá ijé **wäkáwa ká** ijé **jé suá jéijé** wa yaba **ká** yaba kala	**11**	**29**

	12 speakers	53 minutes
Kaldi	CER: 20	WER: 33
DS:	CER: 75	WER: 95
W2V2:	CER: 22	WER: 53

Speech Recognition: Summary

With very little data (53~68 minutes)
the Multilingual Transformer **can match** the performance of the HMM.

With more data (~4 hrs)
the Multilingual Transformer can **surpass** the performance of the HMM.

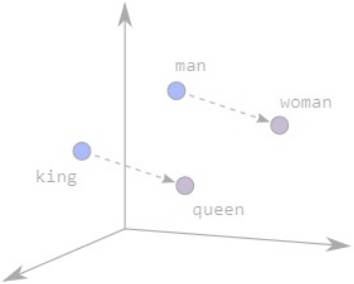
Without the **multilingual** element,
Deep Learning is always **inferior** to the HMMs.

Transfer learning without the transformer never worked.
(Will Transformers work without the multilingual knowledge? I bet they won't).

NLP for Indigenous Languages

Speech
Recognition

Machine
Translation



Embeddings

Machine Translation: Data

10.8K Bribri-Spanish
sentence pairs
(~86K words)

4.2K Cabécar-Spanish
sentence pairs
(~26K words)

Machine Translation: Data Variation

	Differences
Writing system	ù̱ 'cooking pot' (Constenla et al., 2004) ȕ (Jara Murillo, 2018a), ụ̀ (Margery, 2005)
Diacritic encoding	ù̱ 'cooking pot': comb. grave (U+0300) comb. low line (U+0332) comb. grave (U+0300) comb. minus sign below (U+0320) latin small u with grave (U+00F9) comb. macron (U+0331)
Phonetics and phonology	Nasal assimilation: amì ∼ a̱mì̱ 'mother' Unstressed vowel deletion: mȉ ∼ ãmì 'mother'
Sociolinguistic and dialectal variation	ña̱là̱ (Amubri) 'road' (Constenla et al., 2004) ñolȍ (Coroma) 'road' (Jara Murillo, 2018a)
Orthographic variation	(a) ìë'pa rör këképa tâìn ë. (MEP, 2017, 18) ie'pa dör akë́këpa ta̱îë. (Equivalent in Constenla et al. (2004)) 'They are important elders'. (b) E'kũẽ́k és ikíe dör (García Segura, 2016, 11) E' ku̱éki̱ e's i kie dör. (Equivalent in Constenla et al. (2004)) 'That's why it is called like this'.

Machine Translation: Data Variation

	Differences
Writing system	ù̲ 'cooking pot' (Constenla et al., 2004) ū̀ (Jara Murillo, 2018a), ụ̀ (Margery, 2005)
Diacritic encoding	ù̲ 'cooking pot': comb. grave (U+0300) comb. low line (U+0332) comb. grave (U+0300) comb. minus sign below (U+0320) latin small u with grave (U+00F9) comb. macron (U+0331)
Phonetics and phonology	Nasal assimilation: amì ∼ a̲mì̲ 'mother' Unstressed vowel deletion: mĩ̀ ∼ ãmì 'mother'
Sociolinguistic and dialectal variation	ña̲là̲ (Amubri) 'road' (Constenla et al., 2004) ñolõ̀ (Coroma) 'road' (Jara Murillo, 2018a)
Orthographic variation	(a) ìë'pa rör këképa tâìn ë. (MEP, 2017, 18) ie'pa dör akḗkëpa ta̲îë. (Equivalent in Constenla et al. (2004)) 'They are important elders'. (b) E'kũḗk és ikíe dör (García Segura, 20 E' ku̲éki̲ e's i kie dör. (Equivalent in ('That's why it is called like this'.

ãmìx a̲mì̲
ãmĩ

MT: First Results for Bilingual Models

Model	Training pairs	Validation pairs	Testing pairs	BRI→SPA		SPA→BRI	
				Avg	Max	Avg	Max
1.5K	1184	148	148	6.3 ± 2.0	9.9	8.7 ± 0.6	9.5
3K	2368	296	296	11.2 ± 1.9	14.9	11.1 ± 0.7	12.0
6K	4737	593	593	16.9 ± 1.7	19.8	14.2 ± 2.7	18.9

(Transformer-based *OpenNMT*. Results in BLEU)

MT: Examples of translations

English	Bribri reference	Bribri translation	Observations
1. The bird is (sitting) on the branch.	Dù **tkër** kàlula ki .	Dù **tkër** kàlula ki .	Correct positional: *tkër*: to be sitting.
2. The dog is (lying down) by the edge of the river.	Chìchi **tër** di' jkö .	Chìchi **tër** ñalà jkö .	Correct positional: *tër*: to be lying down. Translation means: 'The dog is (lying down) by the edge of the road'.
3. The shirt is (hanging) over there.	Apàio **a'r** awíe ye' wa.	A@@wìe apàio **tër**.	Wrong positional: *a'r*: hang; *tër*: lying down
4. He was (standing) in the house.	Ie' bák **dur** ù a .	Ie' bák ù a .	Missing positional: *dur*: to be standing. Translation means: 'He was in/by the house'

MT: Synthetic data and back translation helped

Real pairs (base)	**Synth Pairs**	**Domain of SPA for Synth Bribri**	**Δ from original model (BLEU)**
2961	2961	In domain	1.0 ± 0.9
5923	5923	Out of domain	-1.9 ± 0.5

(Transformer-based *OpenNMT*)

MT: Multilingual models still work

	μ_4	cjp → spa	spa → cjp	bzd → spa	spa → bzd
		Bilingual			
Cabécar+Spanish (4000 steps)	–	22.5	26.4	–	–
+bilingual lexicon data	–	21.4	29.0	–	–
Bribrí+Spanish (4000 steps)	–	–	–	**30.8**	28.6
		Trilingual			
Trilingual baseline (4000 steps)	24.2	21.8	28.8	18.9	27.3
8000 steps	26.0	24.2	29.3	20.5	**29.8**
12000 steps	25.1	24.2	28.3	19.6	28.2
Baseline+<4src> tagging	25.9	**26.4**	**30.9**	19.1	27.3
Baseline+joint denoising training	22.0	20.2	[illegible]	18.8	23.3
Baseline+joint denoising training, MT finetuning	**26.1**	22.1	29.5	25.1	27.7
Baseline+joint MASS training (replace span)	11.1	9.0	14.7	11.5	9.3
Baseline+joint MASS training (replace token)	8.6	6.7	9.5	9.8	8.5

MT: Denoising probably works

	μ_4	cjp → spa	spa → cjp	bzd → spa	spa → bzd
		Bilingual			
Cabécar+Spanish (4000 steps)	–	22.5	26.4	–	–
+bilingual lexicon data	–	21.4	29.0	–	–
Bribrí+Spanish (4000 steps)	–	–	–	**30.8**	28.6
		Trilingual			
Trilingual baseline (4000 steps)	24.2	21.8	28.8	18.9	27.3
8000 steps	26.0	24.2	29.3	20.5	**29.8**
12000 steps	25.1	24.2	28.3	19.6	28.2
Baseline+<4src> tagging	25.9	**26.4**	**30.9**	19.1	27.3
Baseline+joint denoising training	22.0	20.2	25.5	18.8	23.3
Baseline+joint denoising training, MT finetuning	**26.1**	22.1	29.5	25.1	27.7
Baseline+joint MASS training (replace span)	11.1	9.0	14.7	11.5	9.3
Baseline+joint MASS training (replace token)	8.6	6.7	9.5	9.8	8.5

MT: MASS didn’t work at all

	μ_4	cjp → spa	spa → cjp	bzd → spa	spa → bzd
		Bilingual			
Cabécar+Spanish (4000 steps)	–	22.5	26.4	–	–
+bilingual lexicon data	–	21.4	29.0	–	–
Bribrí+Spanish (4000 steps)	–	–	–	**30.8**	28.6
		Trilingual			
Trilingual baseline (4000 steps)	24.2	21.8	28.8	18.9	27.3
8000 steps	26.0	24.2	29.3	20.5	**29.8**
12000 steps	25.1	24.2	28.3	19.6	28.2
Baseline+<4src> tagging	25.9	**26.4**	**30.9**	19.1	27.3
Baseline+joint denoising training	22.0	20.2	25.5	18.8	23.3
Baseline+joint denoising training, MT finetuning	**26.1**	22.1	29.5	25.1	27.7
Baseline+joint MASS training (replace span)	11.1	9.0	14.7	11.5	9.3
Baseline+joint MASS training (replace token)	8.6	6.7	9.5	9.8	8.5

MT: Weird things happen with BPE

Both Bribri and Spanish are morphologically complex languages:

“I did not sing”

Ye' kȅ sts-**ȅ-ne**
I not sing-**PAST-MID**
Yo no cant-**é**

“I saw two snakes”

Ye' tö tkabȅ bȍt+**öm** sa-**wé**
I **ERG** snake two-**CL:LONG** see-**PAST**
Yo v-**i** dos culebra-**s**.

MT: Weird things happen with BPE

NMT systems benefit from tokenization
(Nießen and Ney 2004, Pinnis et al., 2017):

Ye' kẽ̀ sts@ ẽ̀@ ne ‘I didn’t sing’
Yo no cant@ é

MT: Weird things happen with BPE

But in low-resource environments, some tokenizations can rip morphemes apart (Sennrich & Zhang 2019):

A@@ fuera hay cinco tigres .
"O@@ utside there are five tigers"

Mi apellido es L@@ o@@ pez .
"My last name is L@@ o@@ pez"

MT: Weird things happen with BPE

But in low-resource environments, some tokenizations can rip morphemes apart (Sennrich & Zhang 2019):

Ye' s<u>è</u>rk<u>e</u> Tal@@ a@@ ma@@ nca .
“I live in Tal@@ a@@ ma@@ nca”

Correct division:
Ye' s<u>è</u>-**r**-**k<u>e</u>** Talamanca .
“I live in Talamanca”

MT: Weird things happen with BPE

There are many ways to tokenize sentences:

Language-independent	BPE (Shibata et al. 1999, Sennrich et al. 2015) WordPiece (Schuster et al. 2012) SentencePiece (Kudo et al. 2018)
Morphologically-oriented	Goldsmith (Goldsmith 2000) Morfessor 2 (Virpioja et al. 2013)
Rule-based system	214 rules (based on Flores 2017) 1 syntactic 11(num) + 172(verbs) + 10(adv) + 12(nouns) + 9(adj)

MT: Weird things happen with BPE

	Tokenizer	ChrF++ (Popović 2017)
Tokenized Bribri	Morfessor	39.5
	Rule-based	38.8
	Goldsmith	38.3
	SentencePiece	37.8
	WordPiece	37.7
	BPE	35.0
Control	No tokenization	38.1

Average over x30 runs
Transformer-based *OpenNMT*

Machine Translation Summary

Deep Learning can learn NMT from very small corpora.

<u>Some techniques still work</u>

- Multilingual models
- Synthetic data and back translation
- Adding tokens for translation direction
- Joint denoising (descrambling sequences) ??

<u>Some techniques that should have worked didn’t</u>

- Joint MASS training (demasking sequences)
- BPE-type subword tokenization

<u>Some linguistically-inspired techniques might work</u>

- Linguistically-informed subword tokenization

Quick detour: Natural Language Inference

We created a resource called AmericasNLI (Ebrahimi et al. 2021).

Language	ISO	Family	Dev	Test
Aymara	aym	Aymaran	743	750
Asháninka	cni	Arawak	658	750
Bribri	bzd	Chibchan	743	750
Guarani	gn	Tupi-Guarani	743	750
Náhuatl	nah	Uto-Aztecan	376	738
Otomí	oto	Oto-Manguean	222	748
Quechua	quy	Quechuan	743	750
Rarámuri	tar	Uto-Aztecan	743	750
Shipibo-Konibo	shp	Panoan	743	750
Wixarika	hch	Uto-Aztecan	743	750

Table 1: The languages in AmericasNLI, along with their ISO codes, language families, and dataset sizes.

Language	Premise	Hypothesis
en	And he said, Mama, I'm home.	He told his mom he had gotten home.
es	Y él dijo: Mamá, estoy en casa.	Le dijo a su madre que había llegado a casa.
aym	Jupax sanwa: Mamita, utankastwa.	Utar purinxtwa sasaw mamaparux sanxa
bzd	Ẽnã ie' iche: ãmìx, ye' tso' ù ã.	I ãmìx ã iché irir tö ye' démĩnẽ ù ã.
cni	Irori ikantiro: Ina, nosaiki pankotsiki.	Ikantiro iriniro yaretaja pankotsiki.
gn	Ha ha'e he'i: Mama, aime ógape.	He'íkuri isýpe oĝuahêhague hógape.
hch	metá mik+ petay+: ne mama kitá nepa yéka.	yu mama m+pa+ p+ra h+awe kai kename yu kitá he nuakai.
nah	huan yehhua quiihtoh: Nonantzin, niyetoc nochan	quiilih inantzin niehcoquia
oto	xi nydi biênâ: maMe dimi an ngû	bimâbi o ini maMe guê o ngû
quy	Hinaptinmi pay nirqa: Mamay wasipim kachkani.	Wasinman chayasqanmanta mananta willarqa.
shp	Jara neskata iki: tita, xobonkoriki ea.	Jawen tita yoiaia iki moa xobon nokota.
tar	A'lí je aníli échiko: ku bitichí ne atíki Nana	Iyéla ku ruyéli, mapu bitichí ku nawáli.

Table 3: A parallel example in AmericasNLI with the label *entailment*.

Quick detour: Natural Language Inference

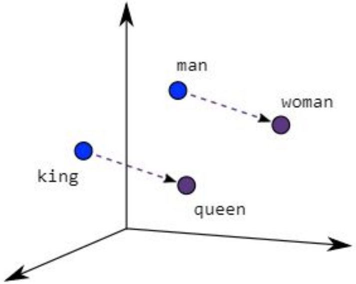
NLI results are better than random. They use an architecture called **Translate-Train**, where back-translated examples are used to help the learning process.

System	aym	bzd	cni	gn	hch	nah	oto	quy	shp	tar	Avg.
Baseline	49.33	**52.00**	42.80	55.87	41.07	54.07	36.50	59.87	52.00	43.73	48.72
Helsinki-5	**57.60**	48.93	**55.33**	**62.40**	**55.33**	**62.33**	**49.33**	**60.80**	**65.07**	**58.80**	57.59
NRC-CNRC-1	-	-	-	57.20	50.40	58.94	-	-	-	53.47	55.00*

NLP for Indigenous Languages

Speech
Recognition

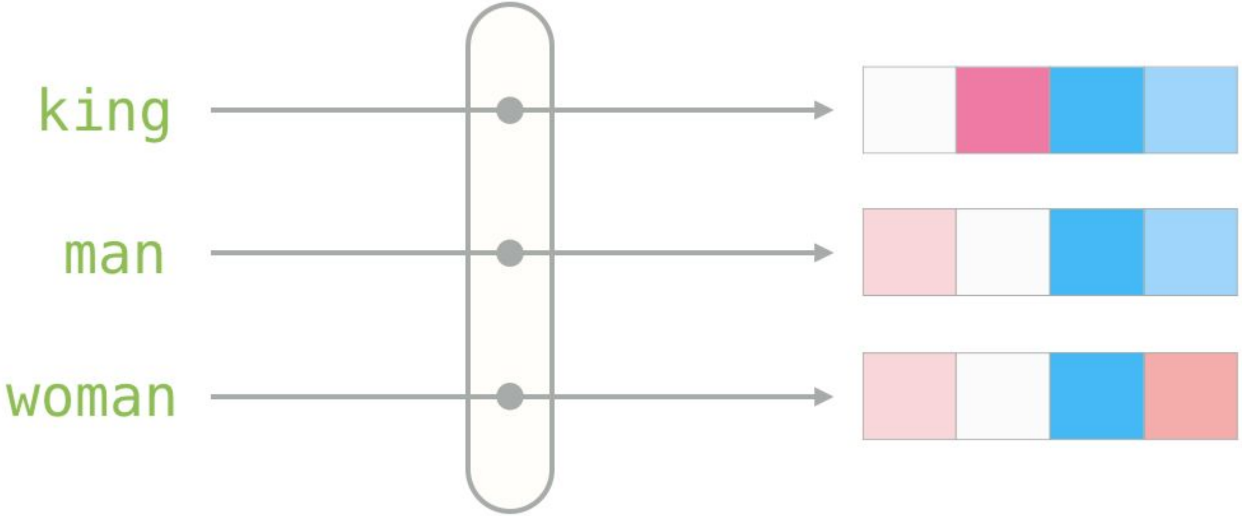
Machine
Translation



Embeddings

Embeddings

Embeddings are a critical component of deep learning. How well do they encode information in extremely low-resource use cases?



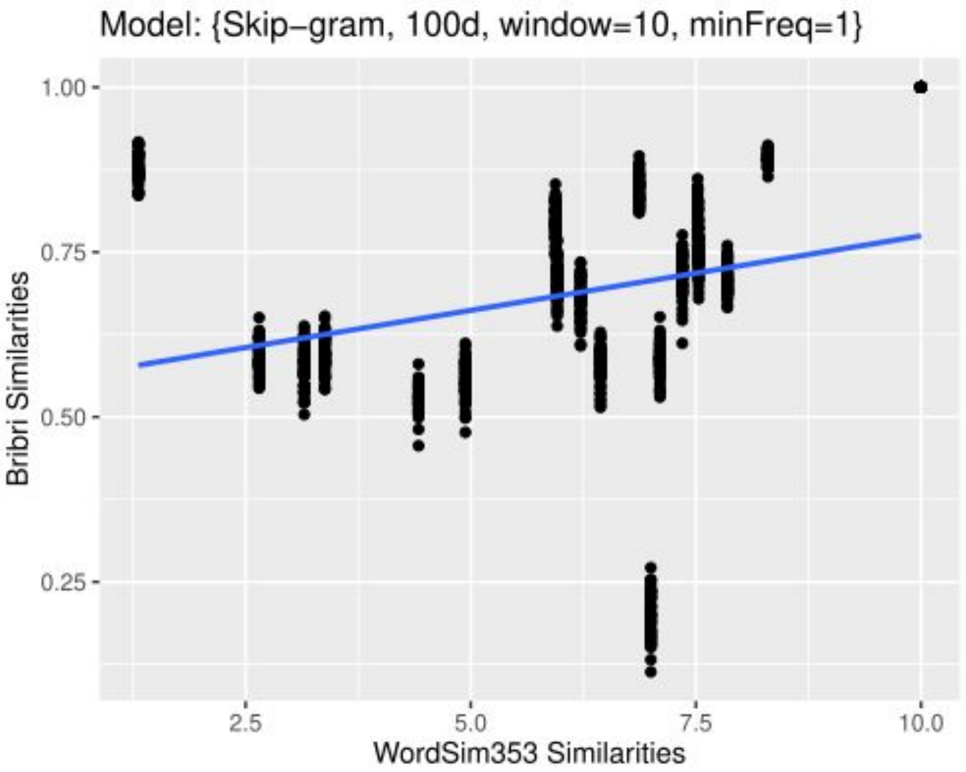
How to test Bribri embeddings

(1) **WordSim353**: Similarities between words.
Many sets are culturally specific (Ex. *Harvard* / *Yale*).
19 pairs could be translated into Bribri (Ex. tiger / cat)

(2) **Odd-One-Out**: Find “odd” word in a triad.
Ex. woman / man / TIGER
20 semantic and 20 structural triads were constructed

(3) **Analogies**: Perform vector algebra.
Based on BATS. Many are culturally specific (euler/mathematician)
Ex. father - man + woman = MOTHER
20 semantic and 20 structural quartets were constructed

WordSim353 (word similarity)



Top models are between **r=0.28** and **r=0.33**

Odd-One-Out testing

Semantic: **~75% accuracy**

Ex. aláköl / wë́m / NAMÙ
woman / man / TIGER

Structural: **~70% accuracy**

Ex. e'köl / bö́l / MAÑÀTÖM
one.human / two.human / THREE.LONG

Top models have large windows (10), use skip-grams, use all the data (minFreq=1) and have medium dimensionality (100d).

Analogies testing

How often is the target word found amongst the top 25 results of the analogy?

Semantic: **~59%** of the time

Ex. wə́m : alákӧl :: yə́ :: AMÌ
man : woman :: father :: MOTHER

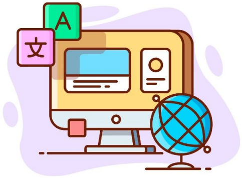
Structural: **~11%** of the time

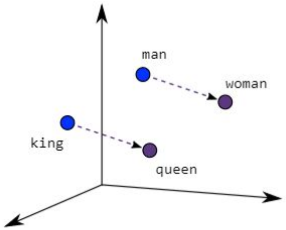
Ex. të' : tèke :: yö' : YAWÈKE
hit : hitting :: made : MAKING

There is relatively little structural learning!

Deep Learning Summary

kiaorana





We only have:

86K words in Bribri
26K words in Cabécar
37K words in Cook Islands Māori

These were enough to create functioning ASR tools and embeddings, and prototypes of NMT.

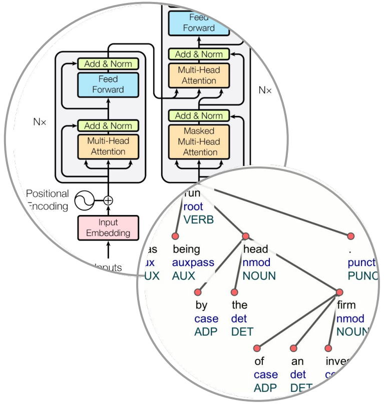
Many techniques improve learning:

- Transfer learning from multilingual knowledge
- Back translation
- Helping the LM learn (e.g. with denoising)

Parts of the talk

Bribri, Cabécar and Cook Islands Māori languages and people

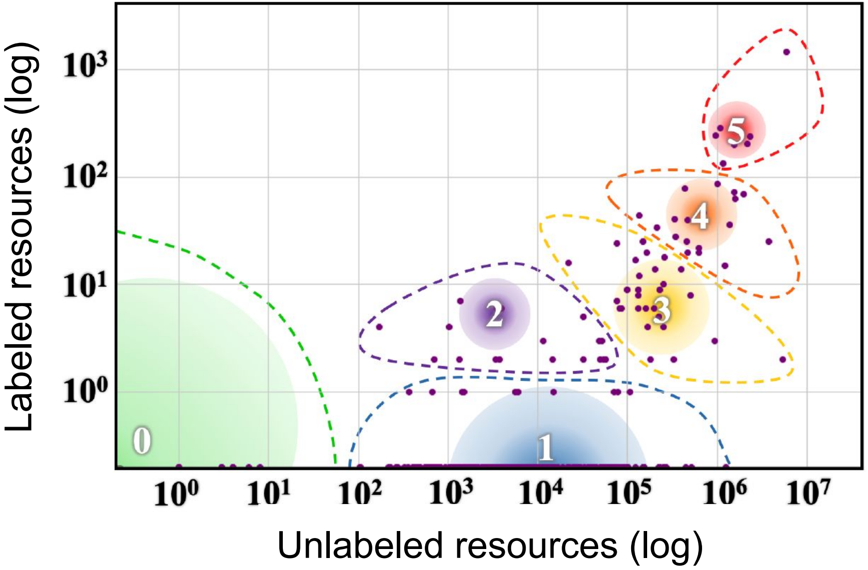
What works and what doesn't



Problems with simulating under-resourced datasets

What's valuable to the users

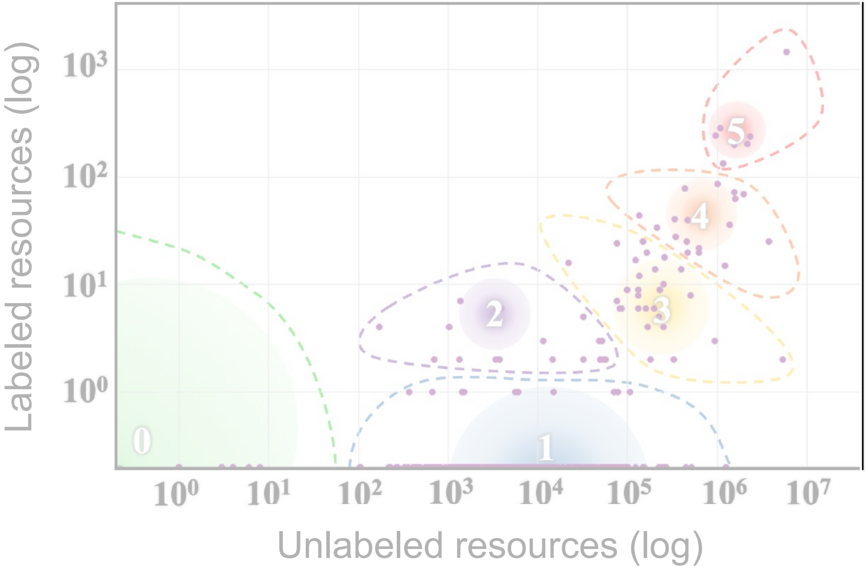
What is an under-resourced language?



Class	**Examples**	**% Langs**
0: Left bys	Warlpiri, Popoluca, Wallisian	88%
1: Scraping bys	Cherokee, Fijian, Navajo	9%
2: Hopefuls	Zulu, Maltese	0.76%
3: Rising Stars	Indonesian, Ukrainian, Hebrew	1.13%
4: Underdogs	Russian, Hungarian, Vietnamese, Korean	0.72%
5: Winners	English, Spanish, Japanese, French	0.28%

Joshi et al. (2020)

What is an under-resourced language?



Labels like “low resource” and “under-resourced” are used so often that it’s difficult to know if results will apply to your language!

Class	Examples	% Langs
0: Left bys	Warlpiri, Popoluca, Wallisian	88%
1: Scraping bys	Cherokee, Fijian, Navajo	9%
2: Hopefuls	Zulu, Maltese	0.76%
3: Rising Stars	Indonesian, Ukrainian, Hebrew	1.13%
4: Underdogs	Russian, Hungarian, Vietnamese, Korean	0.72%
5: Winners	English, Spanish, Japanese, French	0.28%

Joshi et al. (2020)

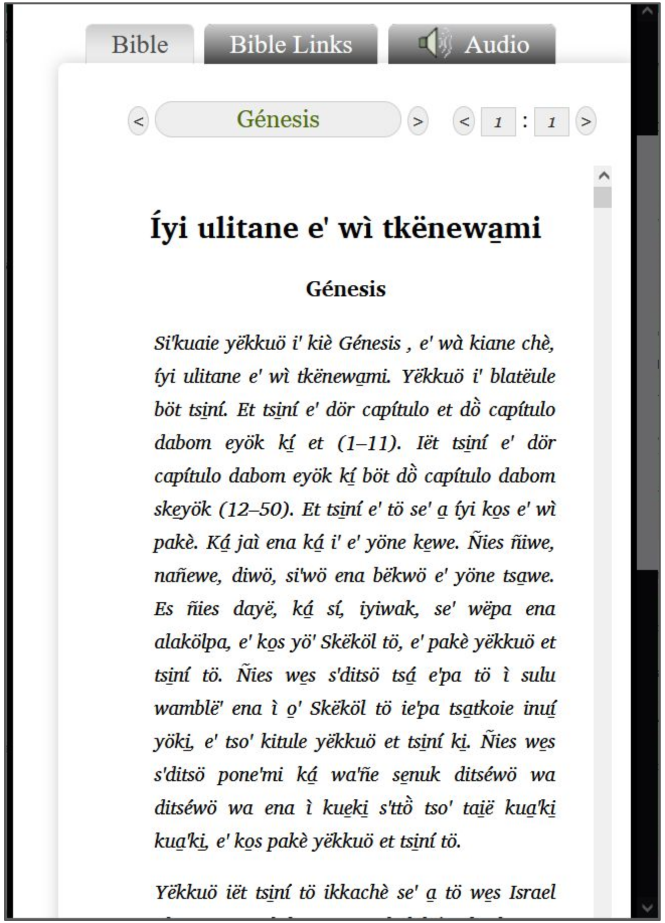
It’s not just about mass of data

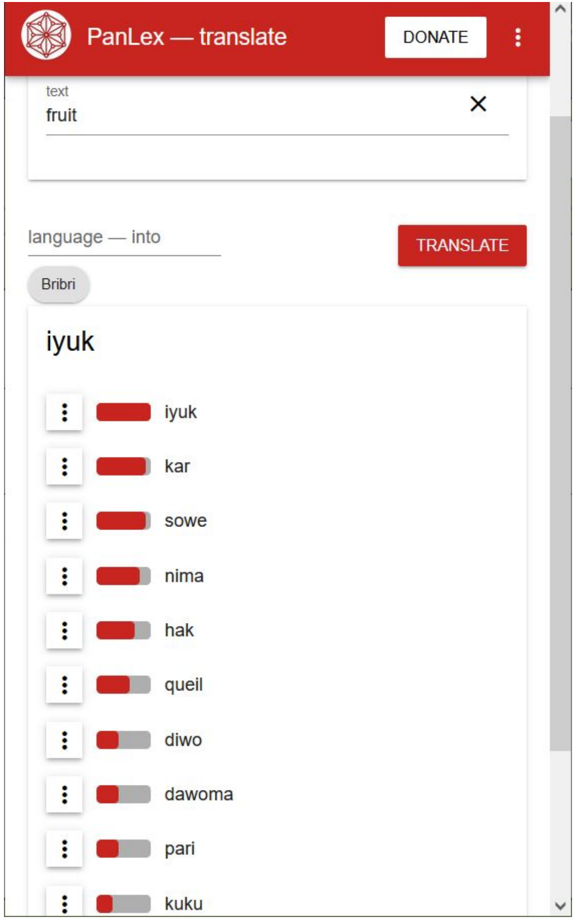
People have thought that being “under-resourced” is a problem of how much data you have.

This has led people to **simulate** low-resource conditions using large languages.

This is **unrealistic** and can be **unhelpful**.

Most of what's on the Internet is garbage





Data is difficult to get

The actual data is not just **expensive** to get.

The people who can read and write Bribri are school teachers, the busiest people on Earth.

Social contexts of data

This data will be much more multilingual, and have much more **code-switching** than data from English.

Issues surrounding the data

The data will be surrounded by **political** and **ethical** issues.

Orthographic variation is of a very different nature.

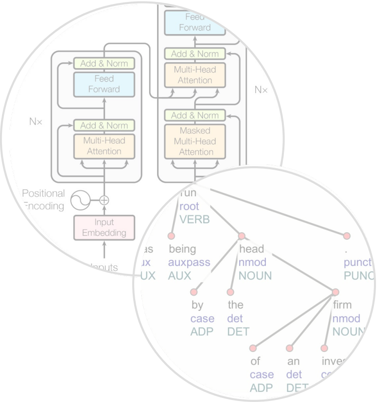
Summary

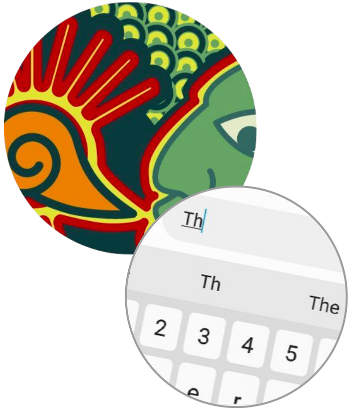
Working on low-resource languages has complications beyond just lack of data.

NLP people should be aware of these in order to make impactful contributions.

Please don’t simulate low resource conditions using English.

Parts of the talk





Bribri, Cabécar and Cook Islands Māori languages and people

What works and what doesn’t

Problems with simulating under-resourced languages

What’s valuable to the users

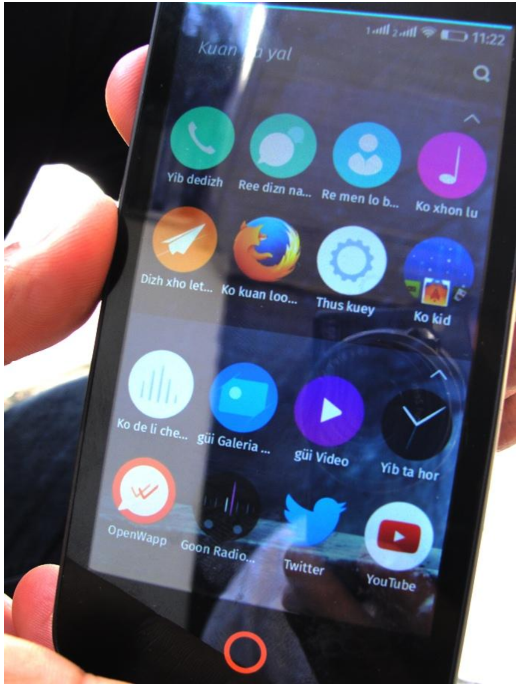
Technology and Revitalization

A computer that knows the language will **NOT** revitalize the language.

Simple things

Simple things can be **transformative**, particularly to young people.

They need to understand that their language is valid as a way to interact with technology.



Conclusions

We have plenty of work ahead of us
(e.g. what do BERT-like LMs learn small languages).

But Deep Learning algorithms can definitely learn
under-resourced languages, even in extreme circumstances
like those of Bribri, Cabécar and Cook Islands Māori.


Meitaki! Wë'ste! Thank you! ¡Gracias!
(rolando.a.coto.solano@dartmouth.edu)

How to test Bribri embeddings

Parameters tested:

(i) Window size:	{2,5,10}
(ii) Training:	{Skip-gram, CBOW}
(iii) MinFreq:	{1,2}
(iv) Dimensions:	{10, 25, 50, 100, 150, 200, 250, 300}

Each of the 96 permutations was trained 60 times using Word2Vec.

Trilingual NMT Input Examples

Source	Target
<4bzd> <2spa> Ìs be' shkèna ?	How are you?
<2denoise> <2bzd> ? be' shkèna Ìs	Ìs be' shkèna?

Source	Target
<4bzd> <2spa> Ìs be' shkèna ?	How are you?
<2mass> <2bzd> Ìs be' [MASK] ?	Ìs be' shkèna?

It knows really bad Cook Islands Māori

RO

Write a brief conversation in Cook Islands Maori, in the style of http://cookislandsdictionary.com/

Kia orana! Te mea koe e kite atu nei i te kakara o te puakaoa?
(Kia orana! Have you noticed the scent of the frangipani flowers?)

E kare, ka rava te kakara o te puakaoa i tenei ra!
(No, I haven't, the scent of frangipani is abundant today!)

E mea ana koe i te rave mai i te aere ki te tautai?
(Are you planning to go fishing?)

Aere atu au ki te tautai i teia po. Kua akarongo au i te rave atu i te aere ki te tautai ki te aroaro o te po.
(I'm going fishing tonight. I've heard that it's good to go fishing at night.)

Ka aere tahi au ki te tautai i teia po. Ka aere mai koe maa'u ia ki te kai.
(I'll join you fishing tonight. Bring back some fish for me to eat.)

Ka mou te ngaru moana, ka mou te ngaru tai.
(When the ocean waves are calm, so are the waves on the shore.)

Kare e mou te patu i te kope.
(You can't beat experience.)

Kia orana! Kua 'ongi koe i te kakara o te au tīpani?

Kāre rāi 'e kī te 'enua i te karaka
o te tīpani i tēia rā.

Ka tautai koe ā tēia rā?

Ka tautai au i tēia pō. Tāku i kite ko te pō te tūātau mako nō te tautai.

RO

Write a brief conversation in Bribri, in the style of https://bribri.net/

Manuel: Kekölma, pü'sa?

(Salutations, how are you?)

Lucía: Amalira, mae. ¿Tü'sa kela?

(Very good, and you? How's it going?)

Manuel: Wína kela. Kólisba yamön melö?

(Things are going well. Did you hear about the festival?)

Lucía: Kólisba kün. Pü'sa ba yamön melö?

(It sounds good. What festival is it?)

Manuel: Yakö yaöyöl nöka, mae. Töka wíknöl weya näkö we'nö, nya'aka kën.

(It's the Yakö Yaöyöl festival. We dance all day and eat traditional food.)

Lucía: Amala. Kün kune yakö nöka kelölba?

(Sounds interesting. When is the Yakö Yaöyöl festival?)

Manuel: Yükra wíknöl weya näkö we'nö, nya'aka kën.

(It's on the first day of the new moon.)

Lucía: Bétari. Nöka kune kela.

(Alright. I'll be there.)

Manuel: Pura vida, mae.

(Excellent, my friend.)

RO

Write a brief conversation in Cook Islands Maori, in the style of http://cookislandsdictionary.com/

Kia orana! Te mea koe e kite atu nei i te kakara o te puakaoa?
(Kia orana! Have you noticed the scent of the frangipani flowers?)

E kare, ka rava te kakara o te puakaoa i tenei ra!
(No, I haven't, the scent of frangipani is abundant today!)

E mea ana koe i te rave mai i te aere ki te tautai?
(Are you planning to go fishing?)

Aere atu au ki te tautai i teia po. Kua akarongo au i te rave atu i te aere ki te tautai ki te aroaro o te po.
(I'm going fishing tonight. I've heard that it's good to go fishing at night.)

Ka aere tahi au ki te tautai i teia po. Ka aere mai koe maa'u ia ki te kai.
(I'll join you fishing tonight. Bring back some fish for me to eat.)

Ka mou te ngaru moana, ka mou te ngaru tai.
(When the ocean waves are calm, so are the waves on the shore.)

Kare e mou te patu i te kope.
(You can't beat experience.)

Chibchan Languages: Bribri and Cabécar

The Bribri language has 7K speakers in Costa Rica. It is vulnerable.
Cabécar has 11K speakers. It also vulnerable.

https://lifeculturetravelcostarica.com
/bri-bri-dialect-class/

https://www.facebook.com/DIREXTUNED/photos/escritora-e-historiadora-bribri-juanita-s%C3%A1nchez-da-clases-de-cultura-y-de-bibri/1726461820984146/?paipv=0&eav=Afa4TdPNavCry0rk4pCL1C5fnQTFizWXks0ZEn2rMyp87GyImKwuKv86PYIwf-KBoOo&_rdr

Escritora e historiadora bribri Juanita Sánchez da clases de cultura y de Bribri