Artificial Intelligence to Accelerate Language Documentation

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

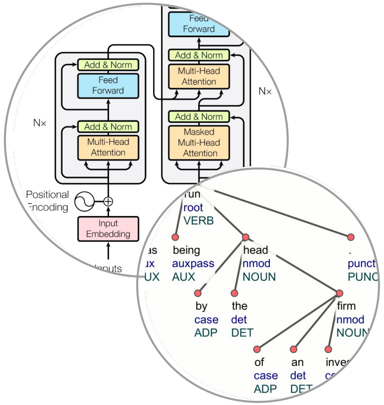
Sofía Flores
Isaac Feldman (NMT)
Veronica Quidore (Parsing)
Annie Tang (Keyboards)
Catharine Herrera (Morphology)
Mien Nguyen (Morphology)

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

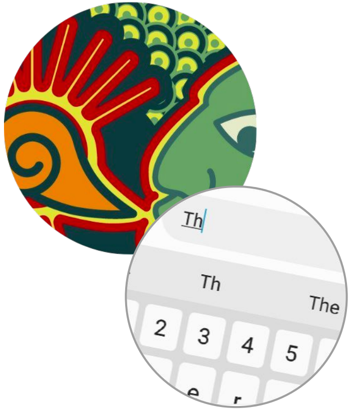
Parts of the talk

NLP, language documentation and revitalization

The Bribri and Cook Islands Māori languages and people



Algorithms for NLP and Indigenous Languages



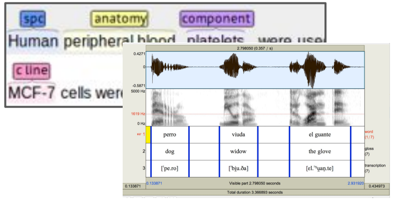
The future: What are we doing this for?

NLP and Language Documentation

Some of our most common tasks involve tasks that are repetitive, but that require very high levels of expertise.

Transcription

Translation

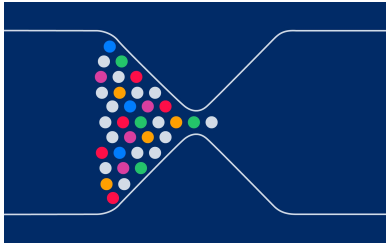


Annotation of corpora

Turning these into learning materials

LangDocumentation: Transcription

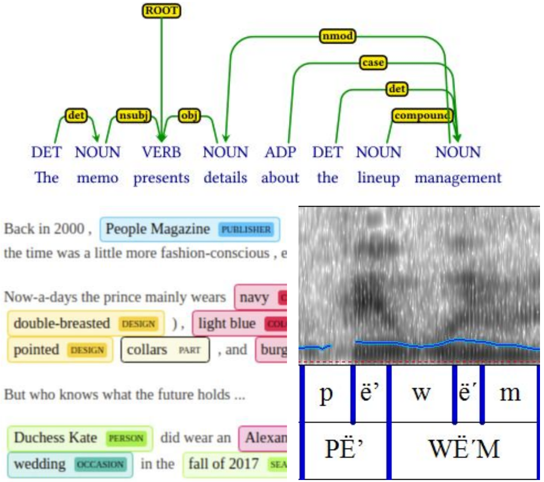
You need 50 hrs of work to transcribe one hour of audio (Shi et al. 2021)



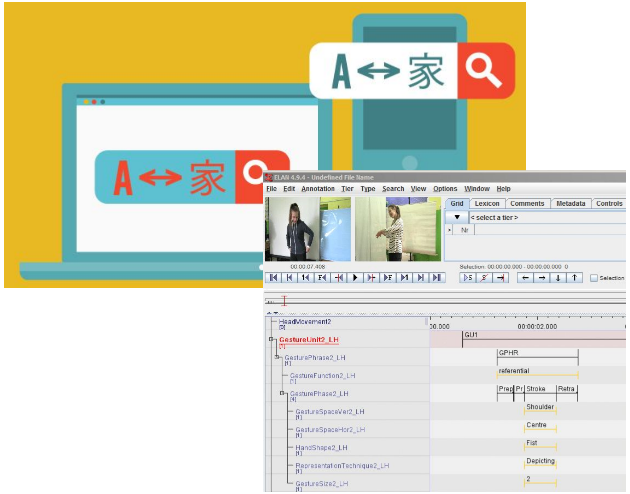
This bottleneck slows down all other analyses.

The technology is far from perfect for English, but it does exist.

LangDocumentation: Analysis



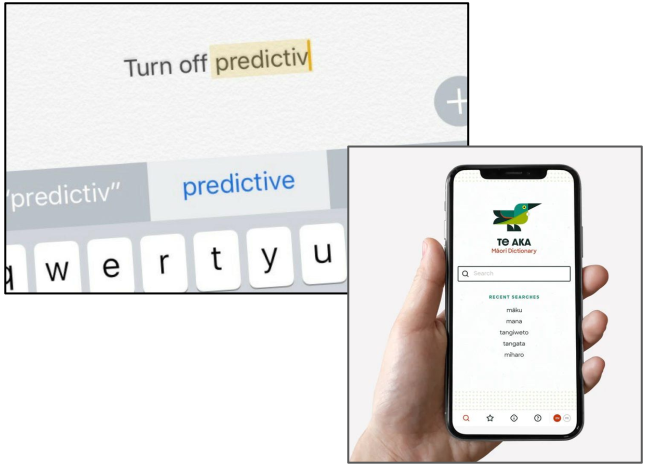
Tagging corpora
(e.g. forced alignment,
taggers and parsers)



Translating corpora
(translation exists for English)

LangDocumentation: Tools for Revitalization

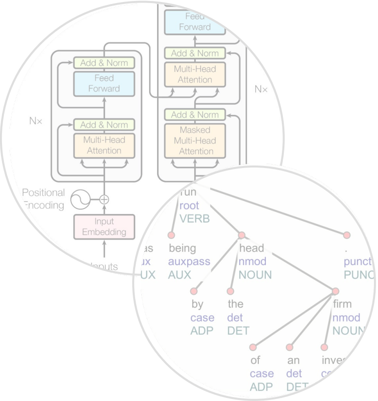
Children’s books, dictionaries, public corpora



Tools (e.g. dictionaries, apps)

NLP, language documentation and revitalization

The Bribri and Cook Islands Māori languages and people

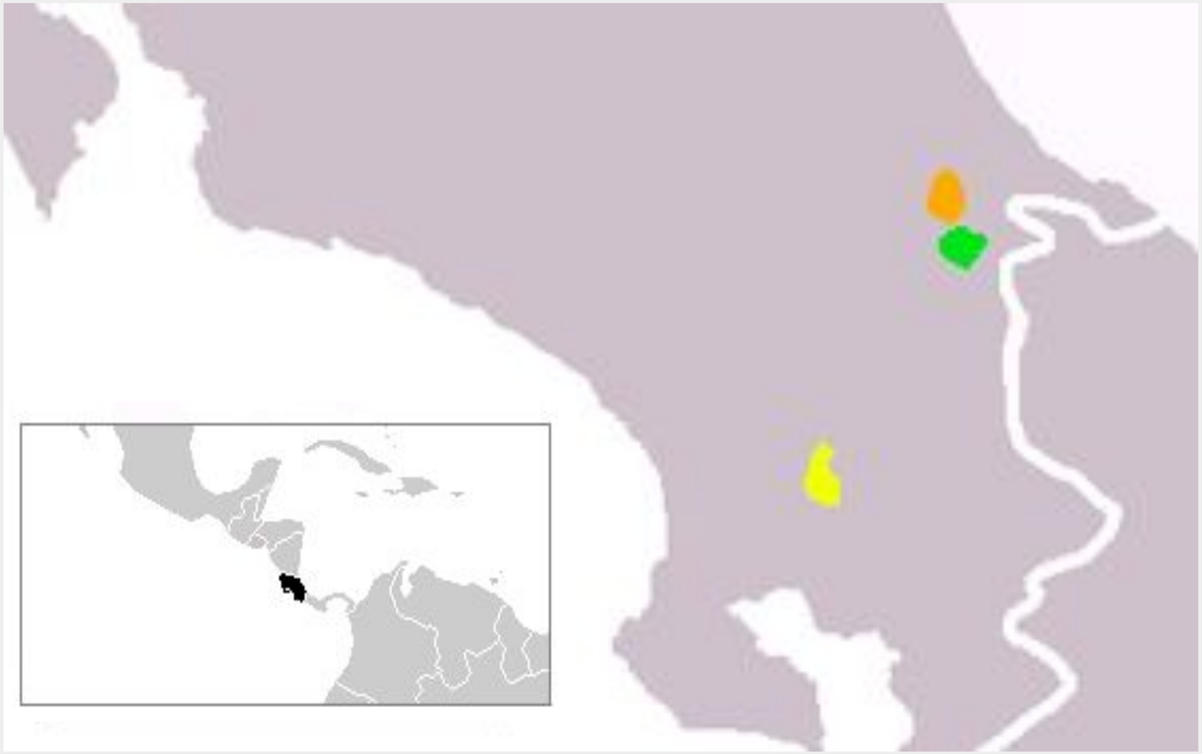


Algorithms for NLP and Indigenous Languages

The future: What are we doing this for?

Bribri

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



Bribri Grammar

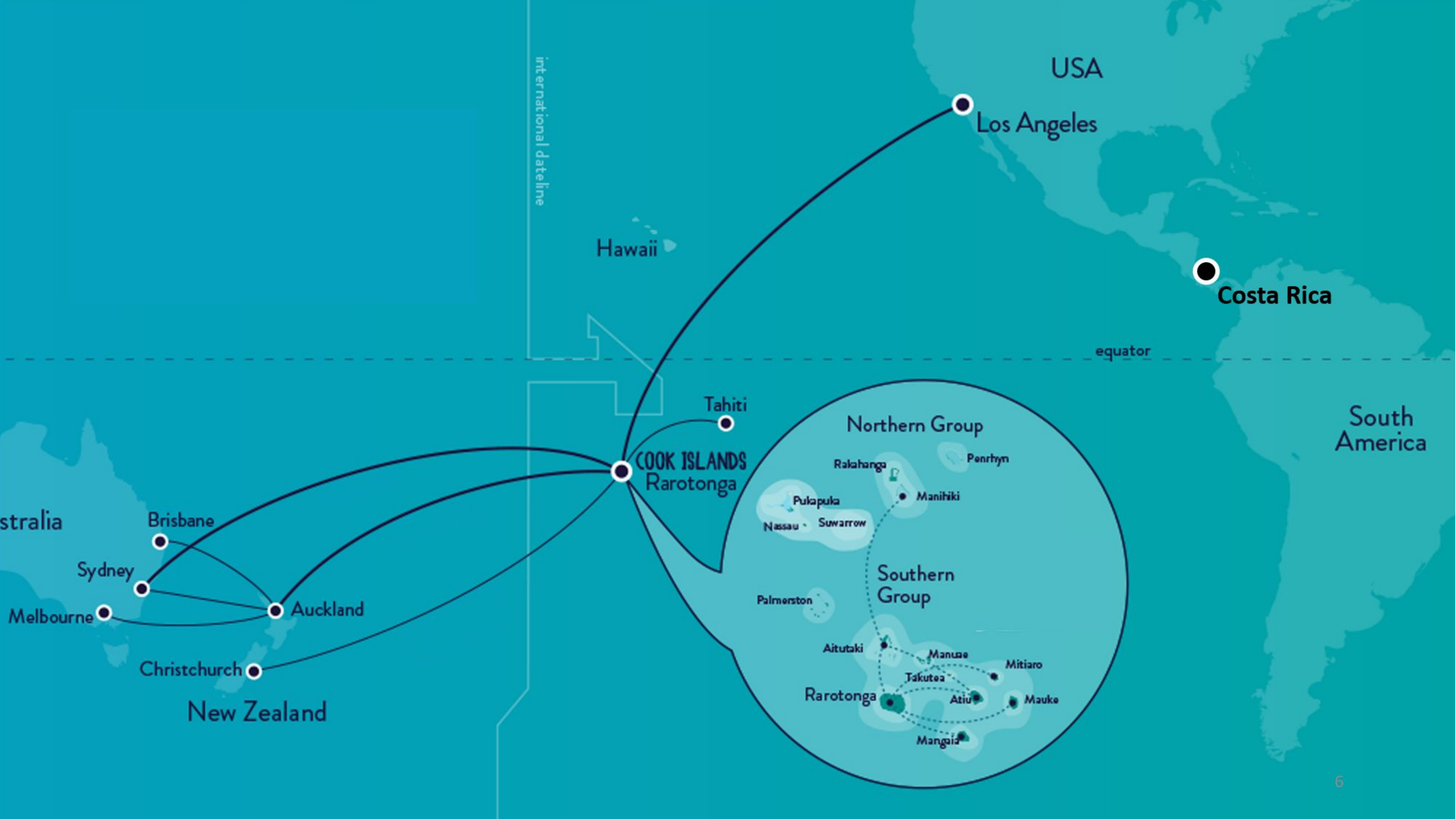
SOV, Ergative	Ye'	**tö**	ù	sú
	I	ERG	house	see-PST.PERF
	I saw the house.			
Inflectional morphology	Ye'	tö	ù	s**awé**
	I	ERG	house	see-PST.IPFV
	I would see the house.			

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]*
Numerical classifiers	dù bö̀**tk**	two-(flat) birds
	alaköl bö́**l**	two-(human) women

Bribri Data Sources

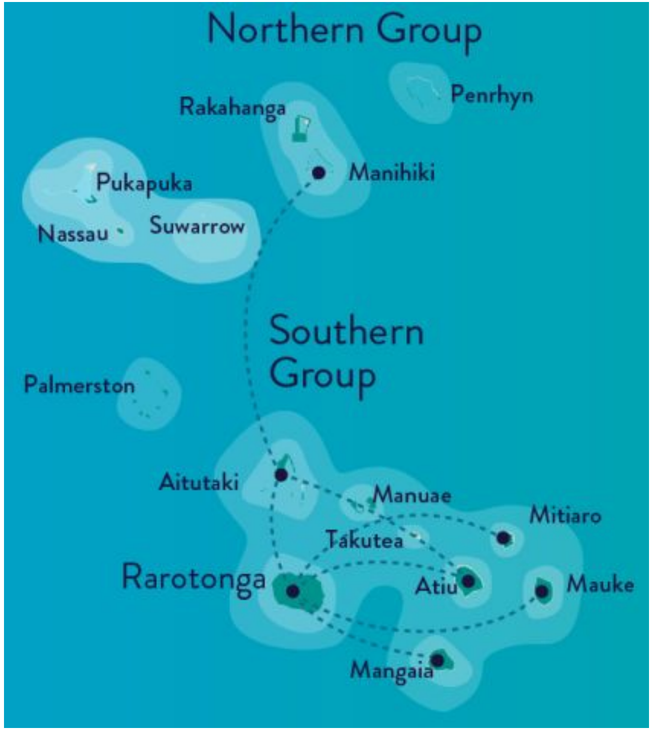
Oral Corpus
Sofía Flores: bribri.net
(~68 minutes of transcribed audio)

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



international dateline
USA
Los Angeles
Hawaii
Costa Rica
equator
Tahiti
COOK ISLANDS
Rarotonga
South America
stralia
Brisbane
Sydney
Melbourne
Auckland
Christchurch
New Zealand
Northern Group
Rakahanga
Penrhyn
Pukapuka
Manihiki
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

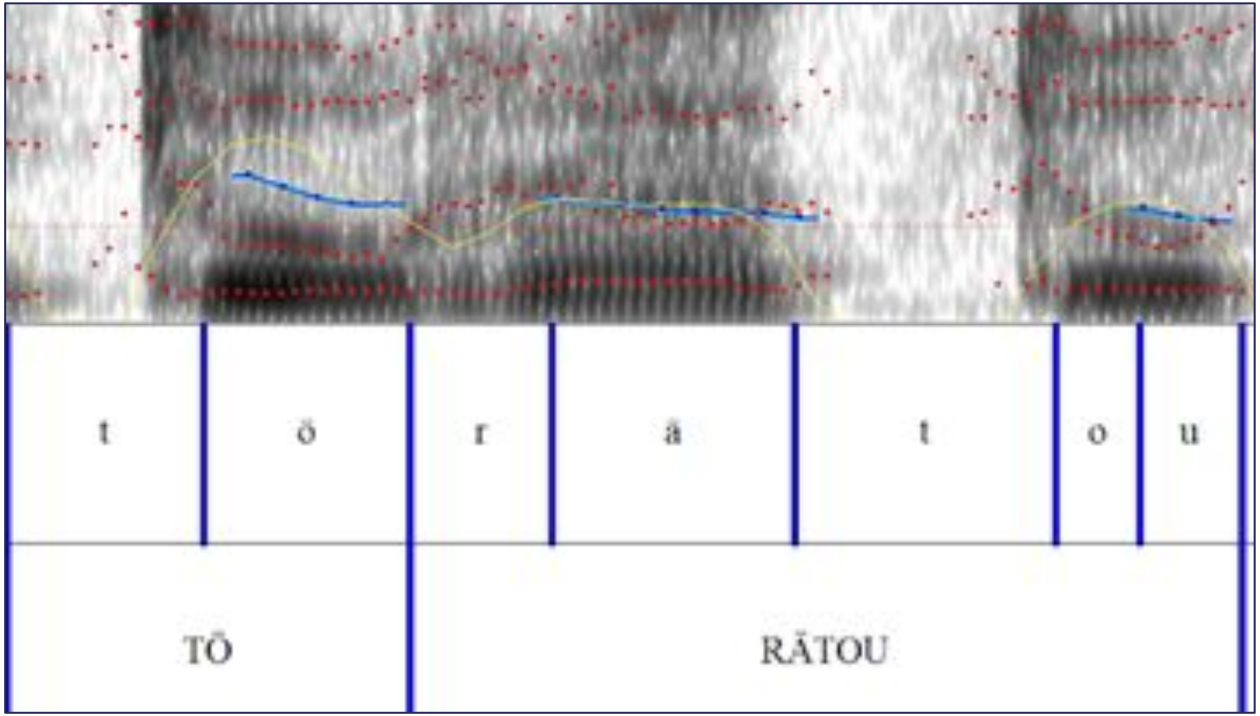
The dogs have left.

Data Source: *Te Vairanga Tuatua*

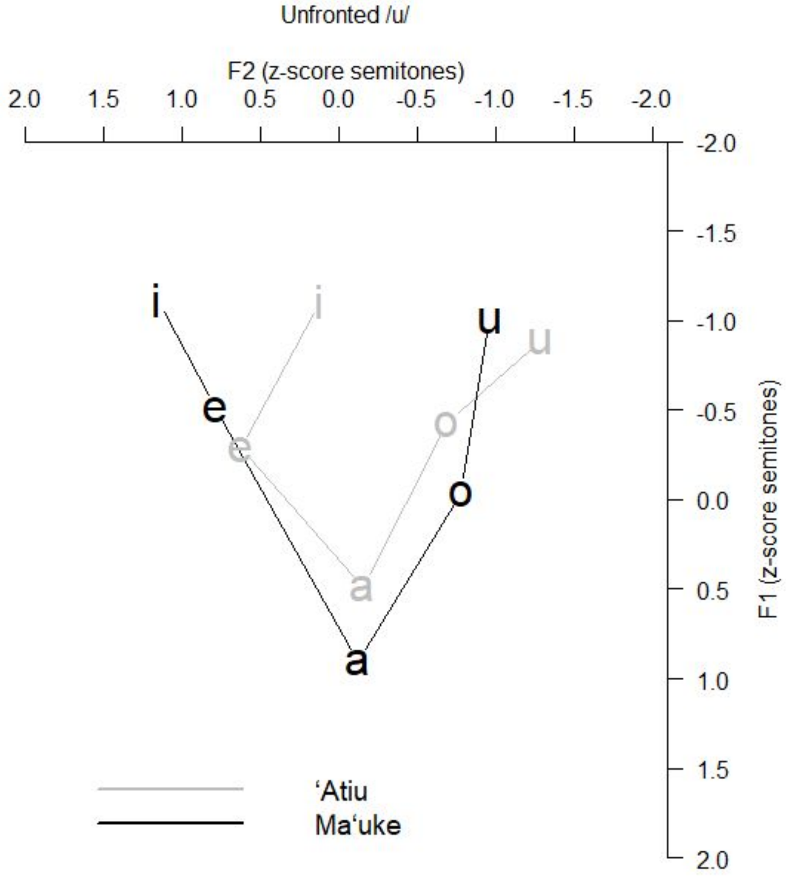
Large (dozens of hours)
Linguistically rich
Little annotation
Transcription is a major bottleneck
~4 transcribed hrs

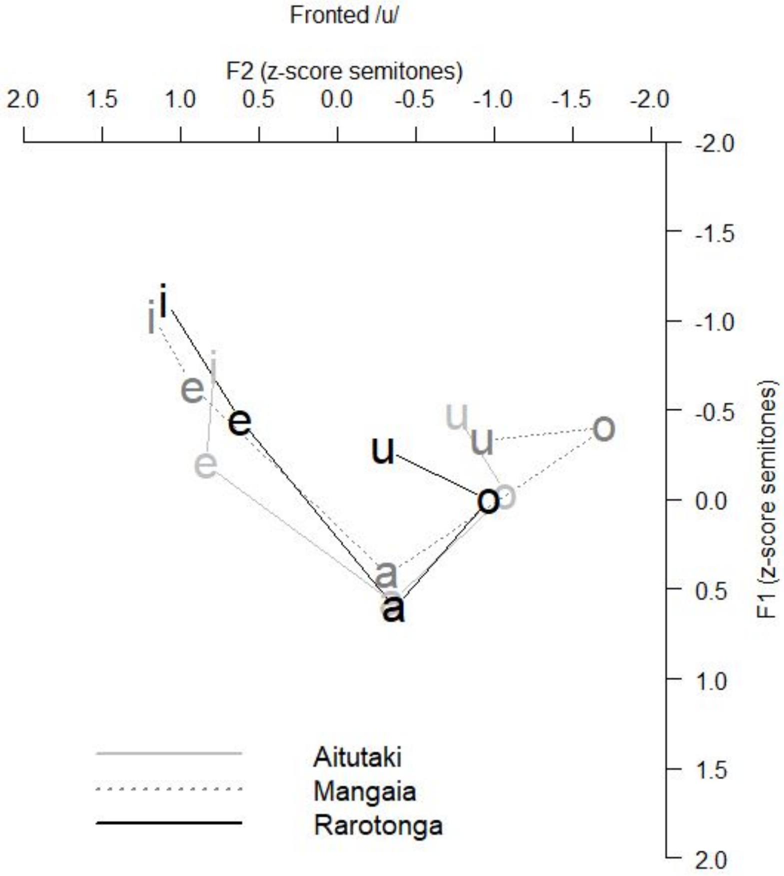
Forced Alignment

Untrained forced alignment
(8% of error when finding the center of the word)

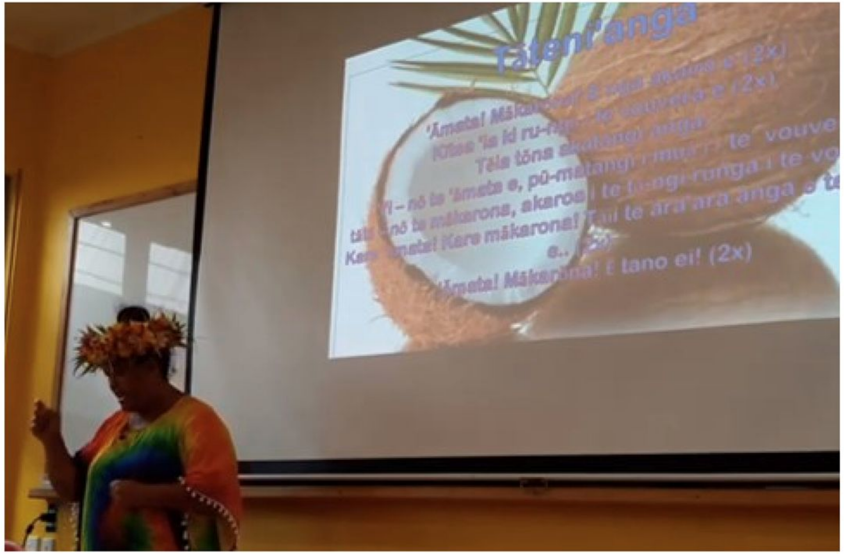


Forced Alignment and Vowels





Forced Alignment and Vowels



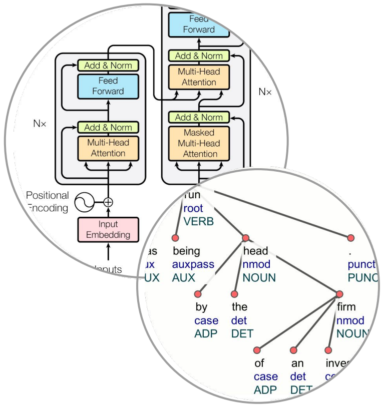
Teacher Tereapii Upokokeu from 'Atiu, singing the “Glottal Stop Song”. USP, January 2019 >>

“I am proud and excited of how complex and sophisticated our language is. They always told me our language was simple and not as good as English, and I can see that that’s not true”

(Creating a virtuous circle in NLP work).

NLP, language documentation and revitalization

The Bribri and Cook Islands Māori languages and people



Algorithms for NLP and Indigenous Languages

The future: What are we doing this for?

NLP for Indigenous Languages

There are fewer data to train systems.

Data are much more difficult (and expensive!!!) to generate

There is orthographic divergence.

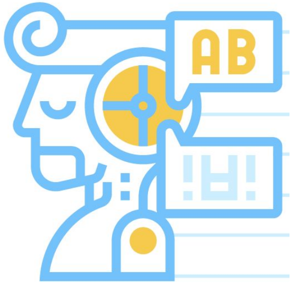
We find complex sociolinguistic environments (e.g. *code-switching*).

English is not very morphologically rich. Languages with rich morphology have many more unique words, and therefore their corpora are more sparse.

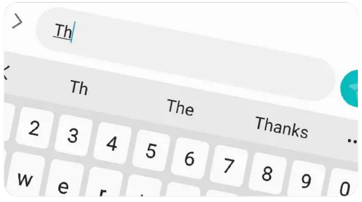
NLP for Indigenous Languages

Speech Recognition

Machine Translation



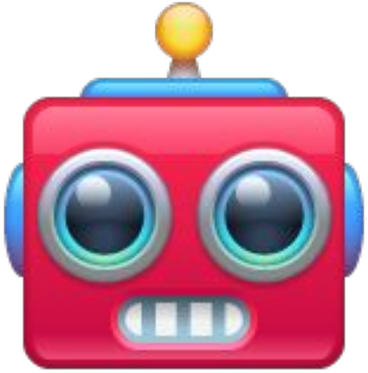
Parsing



Predictive keyboards

Speech Recognition

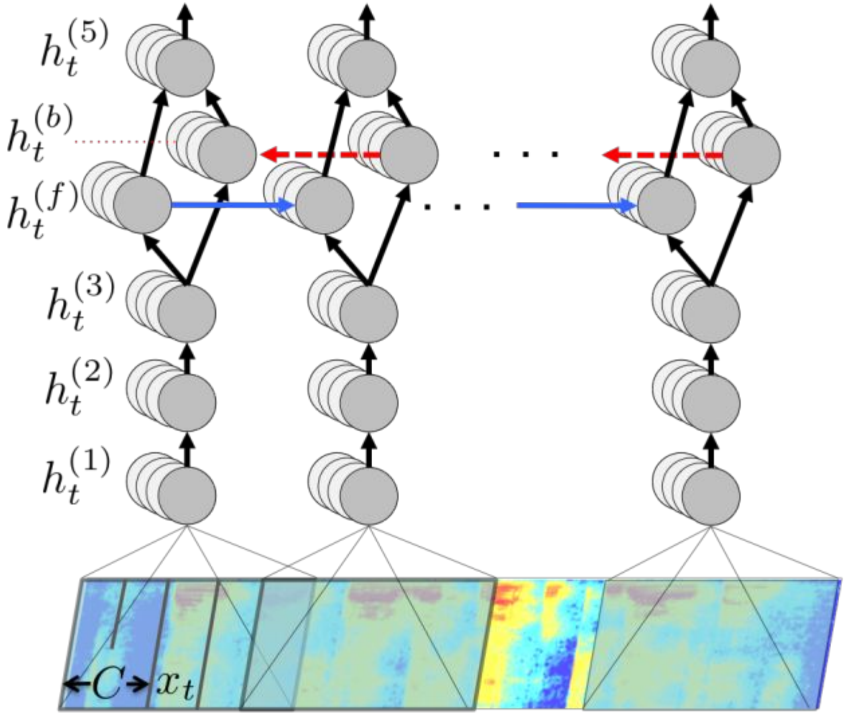
Transcription Bottleneck: You need 50 hrs of work to transcribe one hour of audio (Shi et al. 2021)



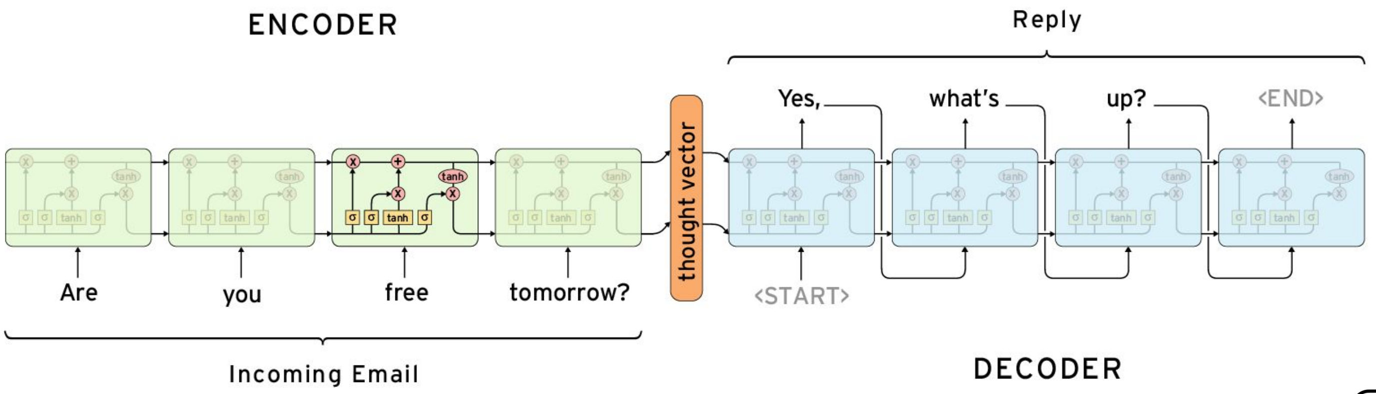
In the last 5 years there have been significant advances in NLP. This can help our documentation work.

Speech Recognition: Algorithms

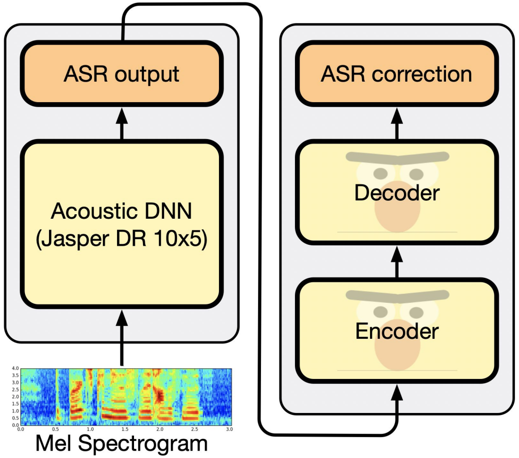
Algorithms based on deep learning (e.g. DeepSpeech) try to classify sections of an audio recording and transform them into characters.



Speech Recognition: Algorithms

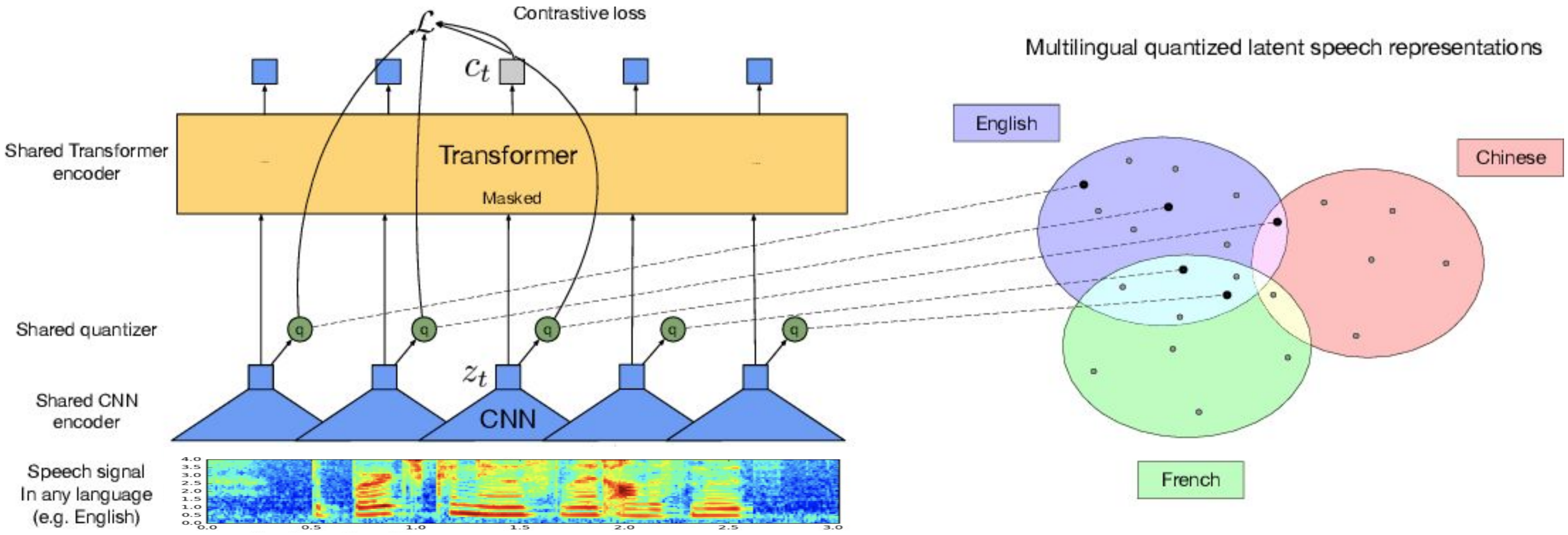


Contemporary Algorithms (e.g. Transformers): The input is codified into an intermediate representation. It is then transformed into an output.



Speech Recognition: Algorithms

Multilingual components (e.g. Wav2Vec2):
The algorithm is pretrained with knowledge from other languages.

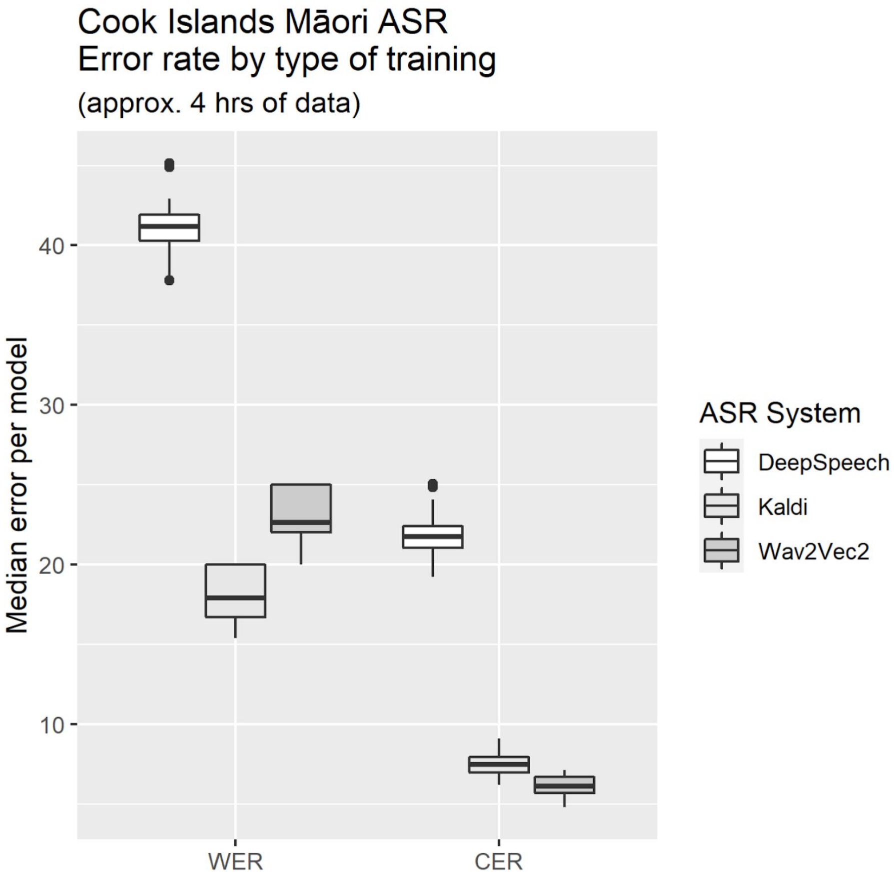


Speech Recognition: Data

237 minutes (~4 hrs), 5033 files
36K 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 6, WER 43)
Original	e' ta be' bák ia tsítsir wake'
Wav2Vec2	**e'ta** be' bák ia **tsítsi** wake'
English	*So he left the place where his house was* (CER 22, WER 67)
Original	e'rö ie' r ë̀ ù ttö́ améat
Wav2Vec2	e'rö ie' **rë́** ù **jtö** améat
English	*Well, you should start telling me why* (CER 65, WER 100)
Original	ma íkëne apàkö́mine tö ì kuéki
Wav2Vec2	**mike na i apàkomie të**

28 speakers 68 minutes

CER: 23±2 WER: 65±3

Speech Recognition: Cabécar Results

English	*Only Kál Kébla brought his log of wood, Jak Kébla brough his stone, the suita stone* (CER: 7)
Original	jíbä kal kébla né wa ijé kalí dëká ják kébla né wa jí jákí ju kä dëlëká rä
Wav2Vec2	**s**ibä kal kébla né wa ijé kalí dëká ják kébla né **y**a jí ják**i** ju kä **r**ëlëká rä
English	*So when he saw it, he turned his face and went to see her; she had the girl in her arms* (CER: 12)
Original	jéra ijé te i suáni ra ijé te jé suá ijé wä**t**káwa tkáu ijé sua ijé wa yaba ka yaba kala
Wav2Vec2	jéra ijé te i suáni ra ijé te jé suá ijé wäkáwa **ká** ijé jé su**á** **jé**ijé wa yaba k**á** yaba kala
English	*They were exterminated, they said... It was not their fault, they were exterminated.* (CER: 31)
Original	ijéwá wä́é̱lë́ ká jíyë́ kú̱na̱ ijéwá **te** i shë́ rä wä́é̱lë́
Wav2Vec2	**ijé wa** wä́é̱l**ä** ká **i yë** kú̱na̱ ijéwá **dishärí** wä́é̱**rä**

12 speakers 53 minutes
CER: 22 WER: 53

Speech Recognition: Held-Out Speakers

Partition	Train-Validation-Test Splits (#files and %)	WER	CER	Test speaker(s)	% total files	% total time
1	4036 - 504 - 493 80% - 10% - 10%	32.9 ± 0.9	8.4 ± 0.2	A K T2 R	3.7 3.6 2 0.5	3.4 4.5 4.5 1.0
2	4007 - 500 - 526 80% - 10% - 10%	40.1 ± 1.9	11.0 ± 0.5	T3 M2	6.9 3.4	7.6 7.2
3	3849 - 481 - 703 76% - 10% - 14%	64.5 ± 3.1	24.5 ± 1.0	M1	14.0	8.0
4	3769 - 419 - 845 75% - 8% - 17%	25.0 ± 0.0	5.9 ± 0.3	B	17.0	18.5
5	3268 - 408 - 1357 65% - 8% - 27%	50.0 ± 0.0	16.4 ± 0.5	J	27	30
6	3532 - 392 - 1109 70% - 8% - 22%	65.9 ± 1.9	23.0 ± 0.2	T1	22	15
Average		46.4 ± 15.6	14.9 ± 7.2			

Speech Recognition: Held-Out Speakers

<u>Partition 5</u>

Meaning: *From morning till night.*

Target: mei te pōpongi mai e pō

Inference: mei te pupongi mai ēpo

(CER=16, WER=50)

<u>Partition 1</u>

Meaning: *When we die we die, when we live we live.*

Target: mē mate tātou kua mate mē ora kua ora

Inference: mē mati tātou kua mate me ora kua ra

(CER=8, WER=33)

Speech Recognition: Future Work

We have a working prototype of an ASR transcription system for CIM.

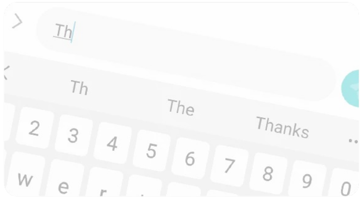
For Bribri and Cabécar we need to transcribe more recordings.

NLP for Indigenous Languages

Speech Recognition

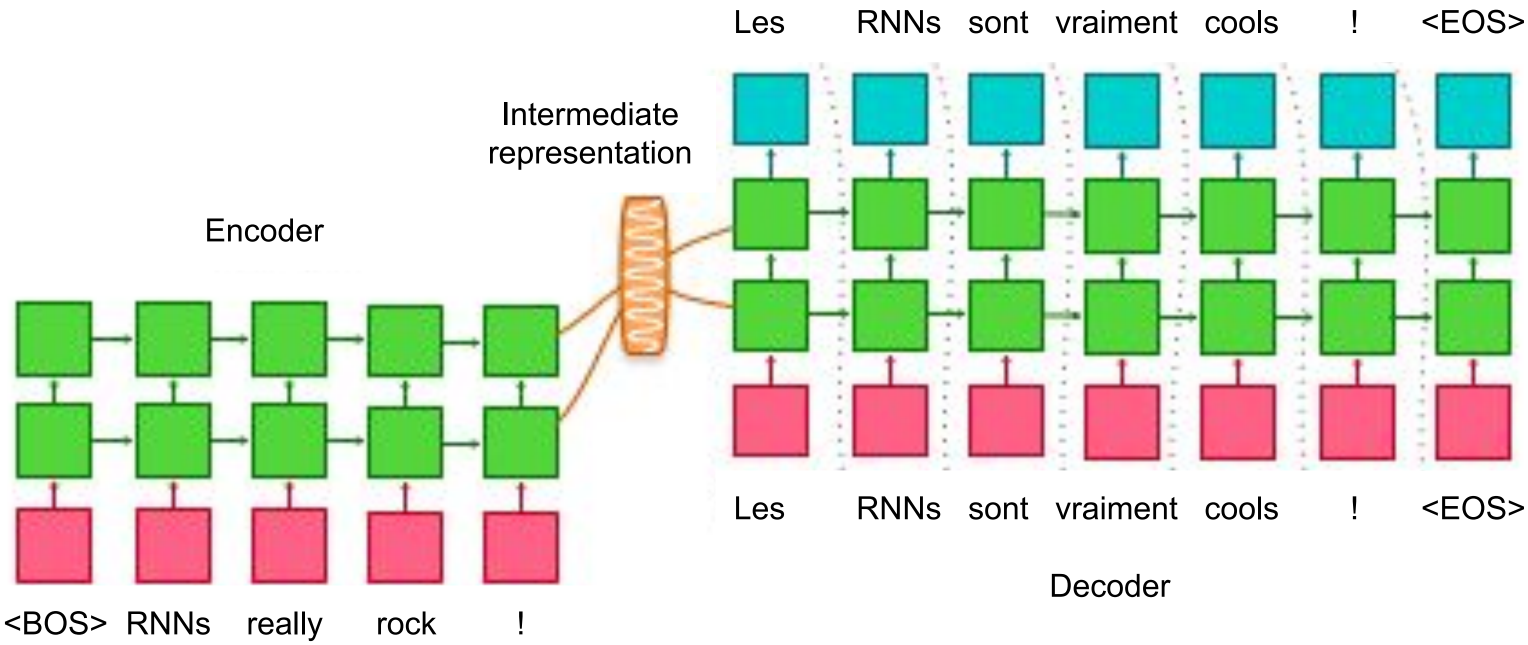
Machine Translation

Parsing



Predictive keyboards

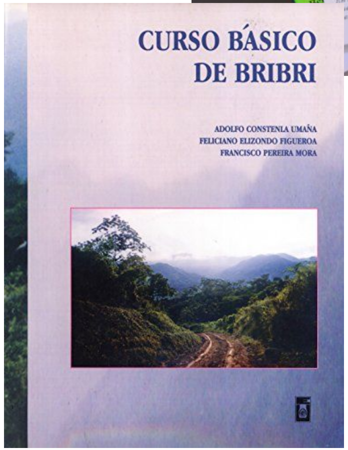
Machine Translation



OpenNMT: Transformer with RNNs
(recurrent neural networks)

Machine Translation: Data

10K Bribri-Spanish
sentence pairs
(~90K words)

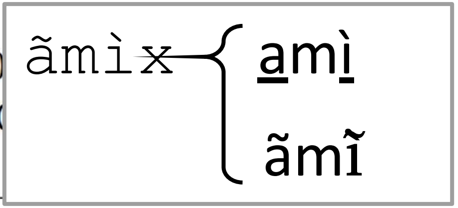


Machine Translation: Data Variation

	Differences
Writing system	ù̱ ‘cooking pot’ (Constenla et al., 2004) ũ̀ (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: mĩ̀ ∼ ãmì ‘mother’
Sociolinguistic and dialectal variation	ña̱là̱ (Amubri) ’road’ (Constenla et al., 2004) ñolõ̀ (Coroma) ’road’ (Jara Murillo, 2018a)
Orthographic variation	(a) ìë’pa rör këképa táìn ë. (MEP, 2017, 18) ie’pa dör akë́këpa ta̱îë. (Equivalent in Constenla et al. (2004)) ‘They are important elders’. (b) E’kũḗk 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

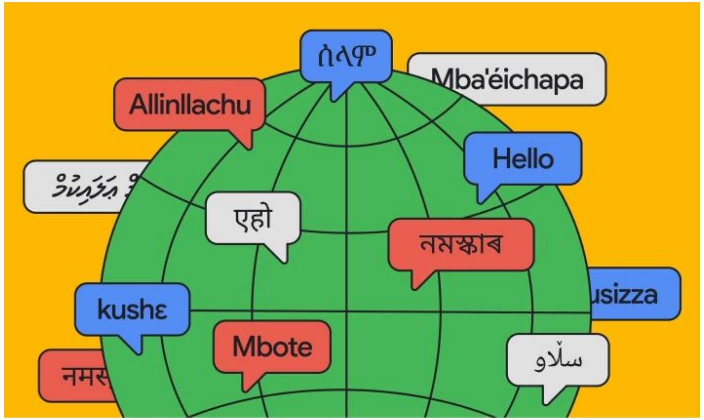
	Differences
Writing system	ù̠ 'cooking pot' (Constenla et al., 2004) ù̀ (Jara Murillo, 2018a), ụ̀ (Margery, 2005)
Diacritic encoding	ù̠ 'cooking pot': comb. grave (U+0300) comb. low line (U+0332) comb. grave (U+0300) comb. minus sign below (U+0320) latin small u with grave (U+00F9) comb. macron (U+0331)
Phonetics and phonology	Nasal assimilation: amì ∼ a̠mì̠ 'mother' Unstressed vowel deletion: mĩ̀ ∼ ãmì 'mother'
Sociolinguistic and dialectal variation	ña̠là̠ (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 ta̠î̠ë. (Equivalent in Constenla et al. (2004)) 'They are important elders'. (b) E'kũẽ́k és ikíe dör (García Segura, 20 E' ku̠é̠ki̠ e's i kie dör. (Equivalent in ('That's why it is called like this'.



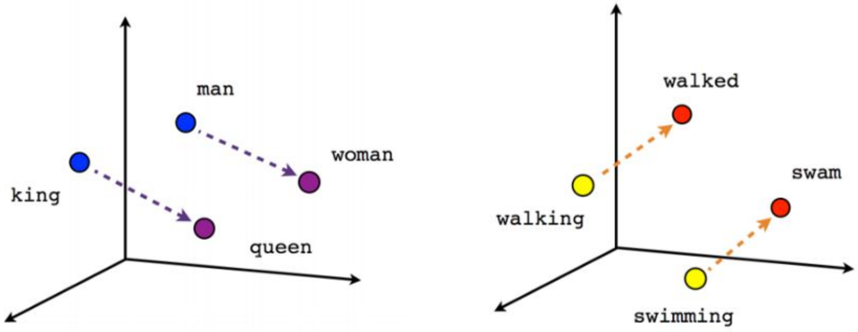
Machine Translation

English	Bribri reference	Bribri translation	Observations
1. The bird is (sitting) on the branch.	Dù **tkër** kàlula ki̱ .	Dù **tkër** kàlula ki̱ .	Correct positional: *tkër*: to be sitting.
2. The dog is (lying down) by the edge of the river.	Chìchi **tër** di' jkö .	Chìchi **tër** ña̱la̱ 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** a̱wíe̱ ye' wa̱.	A̱@@wìe̱ apàio **tër**.	Wrong positional: *a'r*: hang; *tër*: lying down
4. He was (standing) in the house.	Ie' bák **dur** ù a̱ .	Ie' bák ù a̱ .	Missing positional: *dur*: to be standing. Translation means: 'He was in/by the house'

Machine Translation: Future Work



We haven't started this process in CIM.

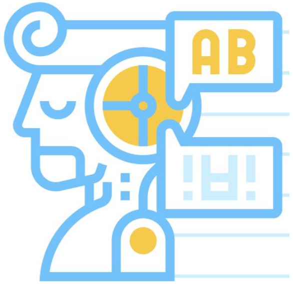


We are testing unsupervised methods to improve Bribri and test Cabécar translation.

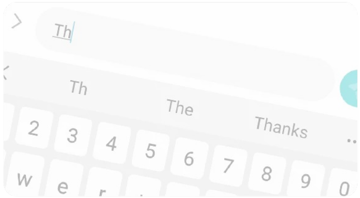
NLP for Indigenous Languages

Speech Recognition

Machine Translation



Parsing

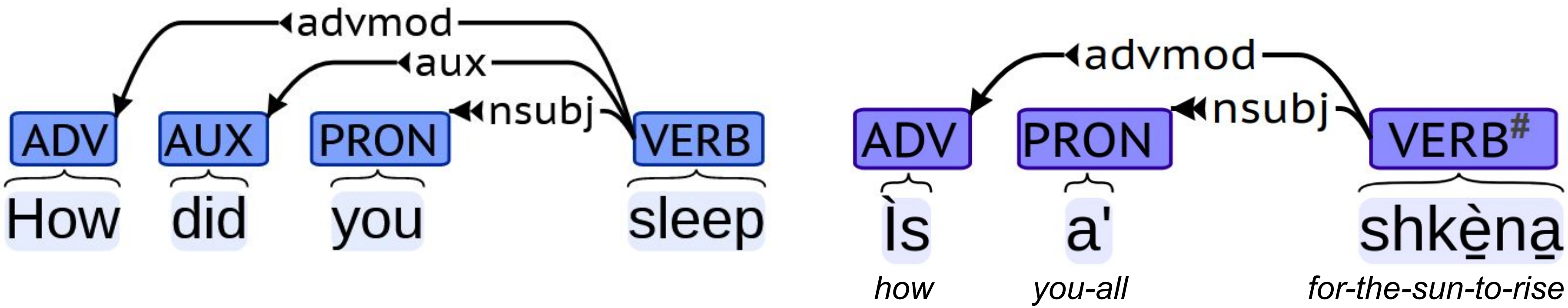


Predictive keyboards

Parsing

Automated syntactic analysis, or **parsing**, is used to create corpora and study the morphosyntax of a language.

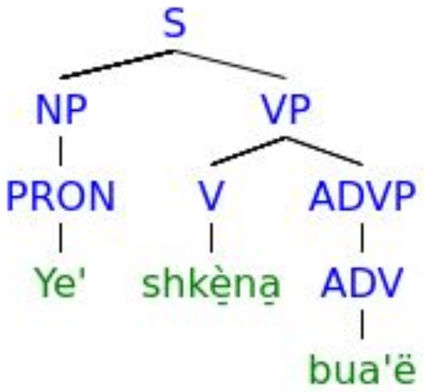
This task includes **Part-of-Speech Tagging**.



Parsing: Corpus

First step: Create a corpus of parsed structures, a **Treebank**.

(1) Assign POS and parse as constituent tree (CFG)



(2) Convert CFG trees into a **dependency** structure.

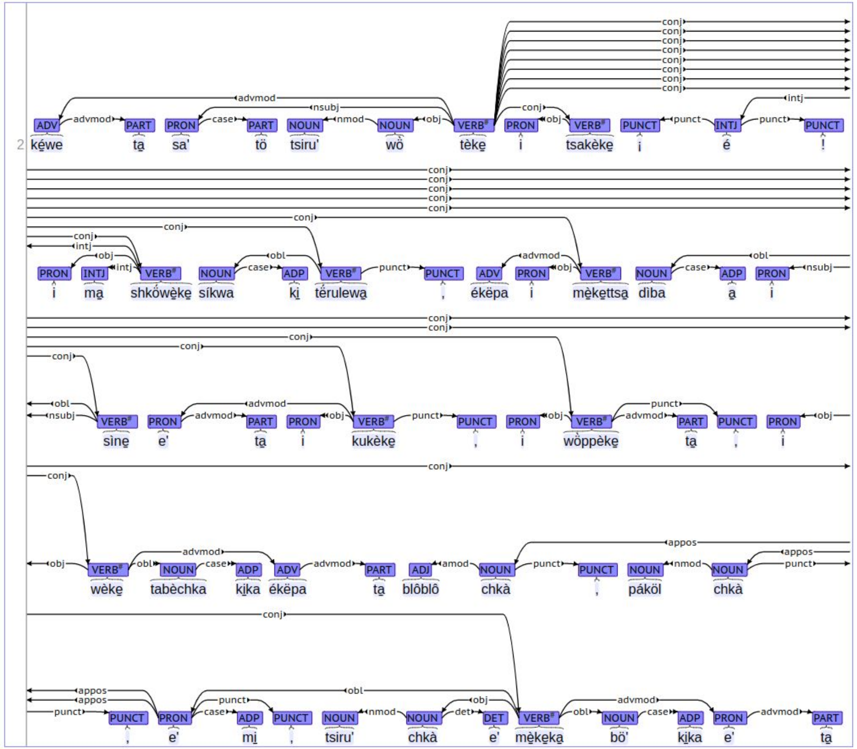
Parsing: Corpus

Corpus created with CFG rules: 1570 words, 315 sentences.

Kéwe ta sa' tö tsiru' wö̀ tèke i tsakèke ¡é! i ma shkö́wèke síkwa ki tḗrulewa , ékëpa i mèkettsa dìba a i sìne e' ta i kukèke , i wö̀ppèke ta , i wèke tabèchka kika ékëpa ta blôblô chkà , páköl chkà , e' mi , tsiru' chkà e' mèkeka bö' kika e' ta

First we cut the cocoa fruit, we split it, right? And we ferment it. You cut it over leaves, put it there and it dries in the sun. Then we toast it, we air it, and then we grind it in this machine. Then [you take] the sweet thing, the sugar, mix it with the cocoa and put it in the fire.

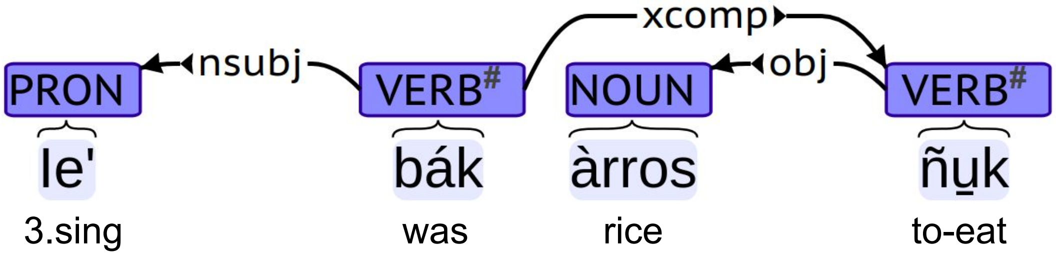
(B09h22m53s05sep2012-01)



Parsing: Evaluation

With the existing data we trained an automated parsing model (based on a multilingual BERT and UDPipe2).

UAS: 100%
LAS: 100%
UPOS: 100%



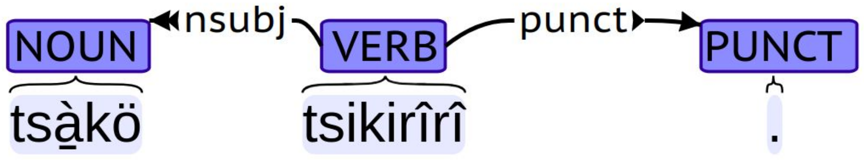
She was eating rice.

UAS: *Unlabelled attachment score*
LAS: *Labelled attachment score*

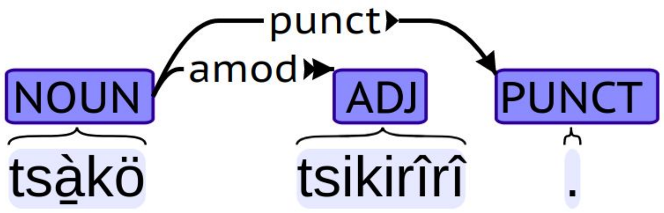
Parsing: Evaluation

With the existing data we trained an automated parsing model (based on a multilingual BERT and UDPipe2).

UAS: 0%
LAS: 0%
UPOS: 66%



hair yellow
Yellow hair



UAS: *Unlabelled attachment score*
LAS: *Labelled attachment score*

Parsing: Bribri Results

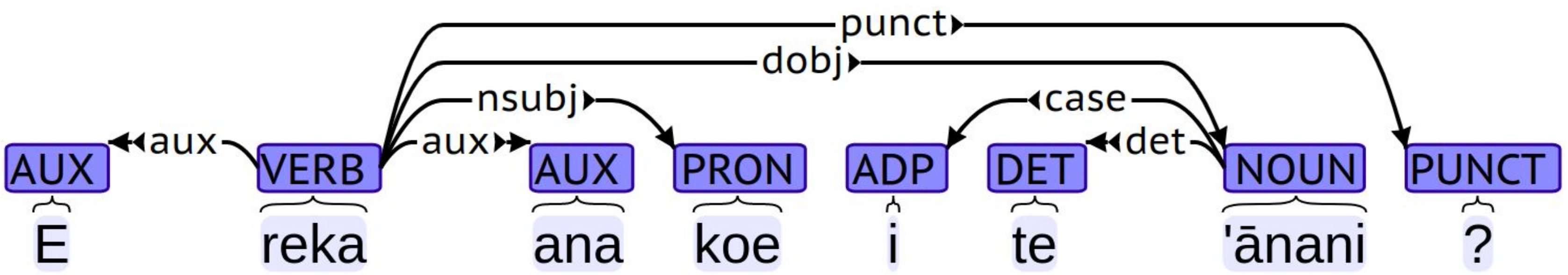
With the existing data we trained an automated parsing model (based on a multilingual BERT and UDPipe2).

Preliminary results:

UAS	85% ± 7%
LAS	81% ± 7%
UPOS	90% ± 3%

Parsing: Cook Islands Māori

We have begun the CIM parsing process (1035 words, 126 sentences). The tagger is about 92% accurate.

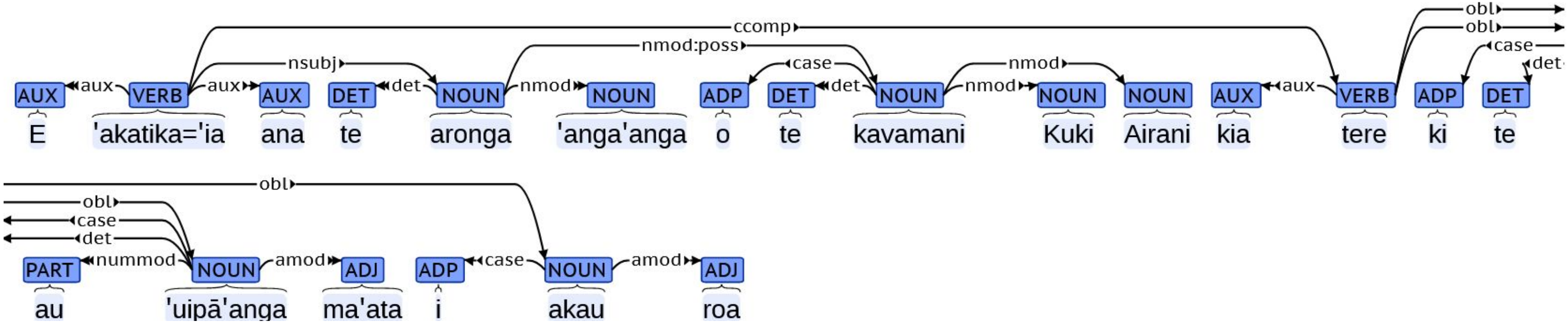


PRES like PRES you OBJ the orange ?

Do you like oranges?

Parsing: Cook Islands Māori

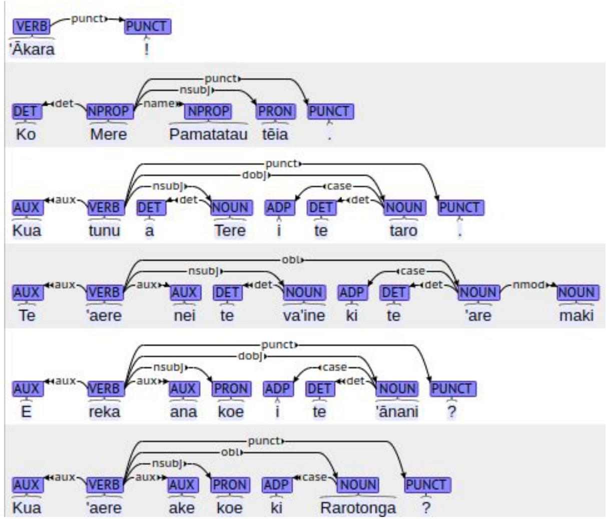
Corpus so far: 126 sentences, 1035 words



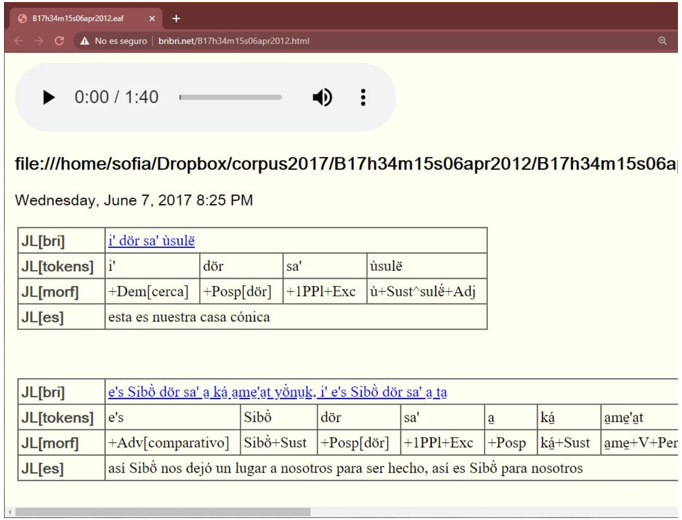
E 'akatika'ia ana te aronga 'anga'anga o te kavamani Kuki Airani kia tere ki te au 'uipā'anga ma'ata i akau roa.

The Cook Islands public servants are permitted to travel to meetings overseas. (Nicholas 2017:366, example 536b)

Parsing: Future Work



We hope to release the CIM treebank in the next 6~9 months.



We are expanding the Bribri treebank to tag the oral corpus.

NLP for Indigenous Languages

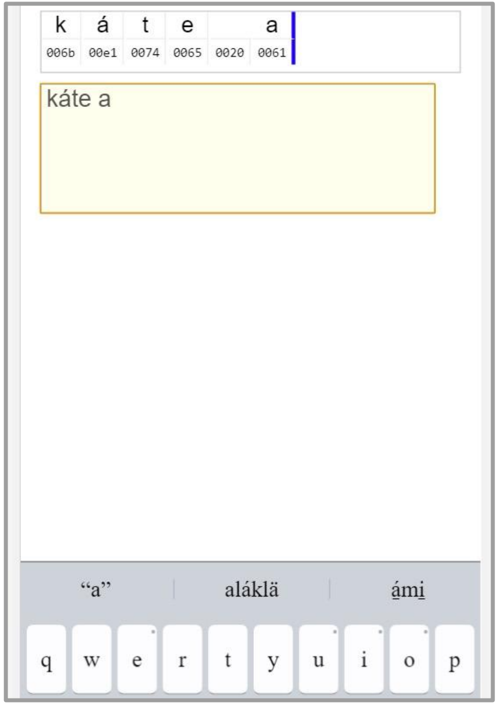
Speech Recognition

Machine Translation

Parsing

Predictive keyboards

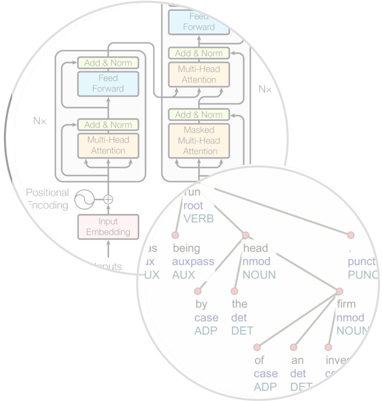
Deploying Predictive Keyboards



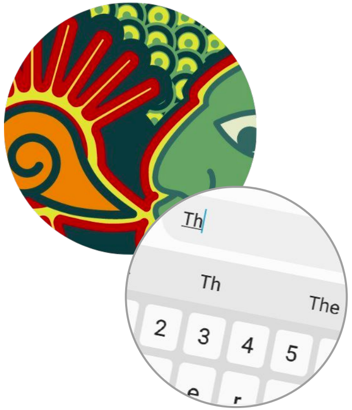
Keyman keyboards have been the necessary tool to deploy them.

NLP, language documentation and revitalization

The Bribri and Cook Islands Māori languages and people



Algorithms for NLP and Indigenous Languages

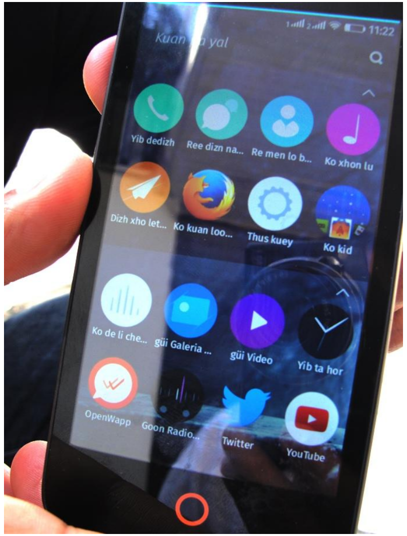


The future: What are we doing this for?

Technology and Revitalization

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

Technology and Revitalization



Incorporating Indigenous languages into technology creates a positive impact, particularly amongst younger generations.

It helps create new usage domains and new communities.

“Use your Voice” Zapotec project (Lillehaugen 2016)

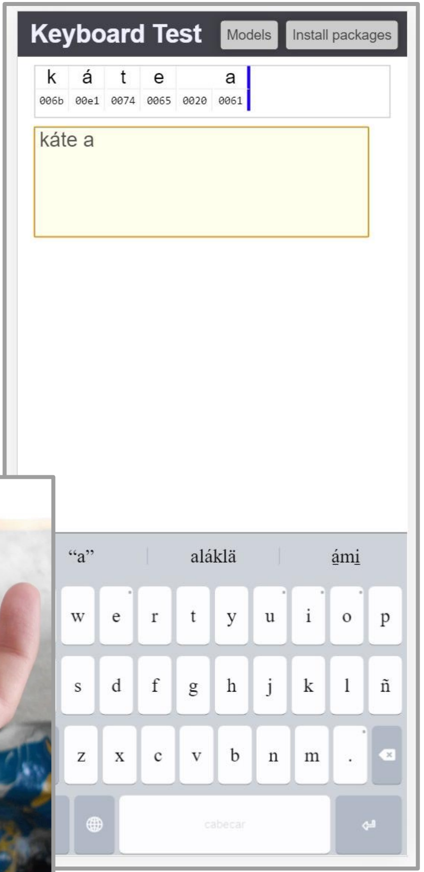
Indigenous Communities and NLP

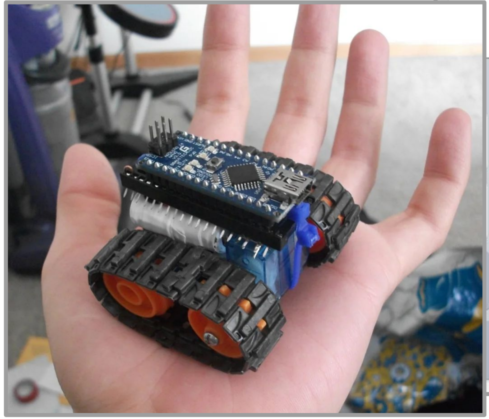
Useful tools: Predictive keyboards
Future tools: ASR Robots

On the Cook Islands, the CS people are working for the community.

In Costa Rica we are still facing this challenge: How can we transfer ownership of these projects to the community?

Example: Data Sovereignty





Meitaki! Wë'ste! Thank you! ¡Gracias!

(rolando.a.coto.solano@dartmouth.edu)