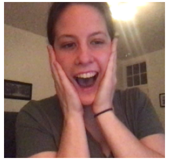
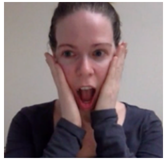
A Corpus For Large-Scale Phonetic Typology



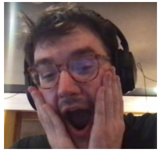
Elizabeth Salesky



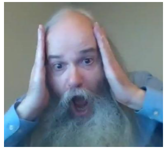
Eleanor Chodroff

Tiago Pimentel

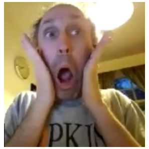
Matthew Wiesner



Ryan Cotterell

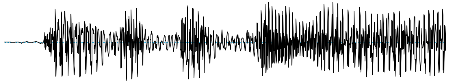


Alan W Black



Jason Eisner

VoxClamantis in deserto:
"a voice crying out in
the wilderness"

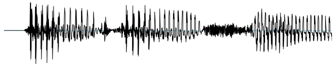


‘in the beginning’

English

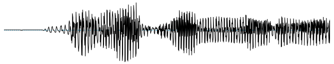
‘ipeuhcan’

Nahuatl



‘am Anfang’

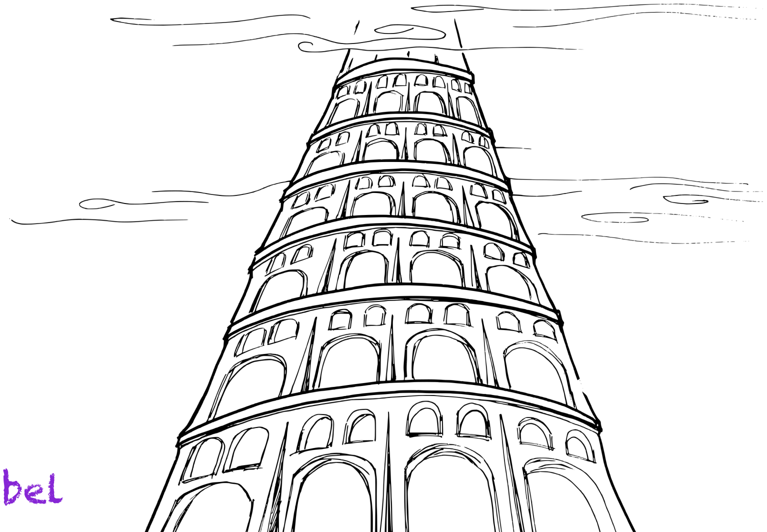
German



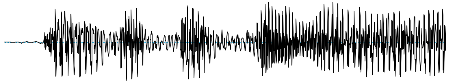
‘በመጀመሪያ’

Amharic

In the beginning, there was **SPEECH**



Tower of Babel

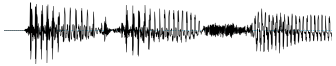


‘in the beginning’

English

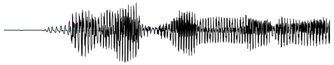
‘ipeuhcan’

Nahuatl



‘am Anfang’

German



‘በመጀመሪያ’

Amharic

In the beginning, there was **SPEECH**

Then the linguist asked:

How do speech and language vary?

↳ **prior cross-linguistic phonetic studies have relied on reported [language-aggregate] measurements**

We create our new corpus, VoxClamantis v1.0, to answer this question!

- ✓ spoken readings of the Bible
- ✓ >600 languages
- ✓ time-aligned phonemic transcriptions
- ✓ phonetic measures for vowel and sibilant **tokens**

This talk

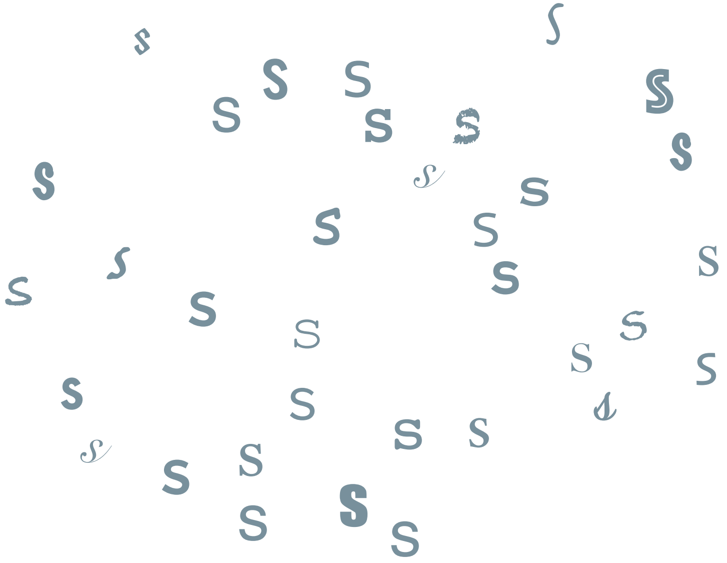
① **WHY** we want this data

② **HOW** we create it

③ **CASE STUDIES** validating the corpus & illustrating two possible uses

Why?

Variation in and across languages



⑤ Spanish	⑦ Romanian
/i/	/i/
/u/	/u/
/o/	/o/
/e/	/e/
/a/	/a/
	/ɨ/
	/ə/

We know phonetic variation within a language,
but what are its range and limits?

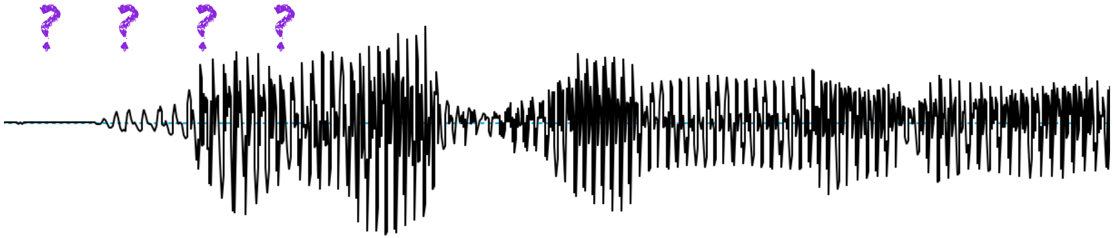
How does the number and set of phonemic
categories influence their realizations?

How?

Resources Needed

① speech
② transcripts
③ phonemic labels

Amharic



bəmədʒəmərija

በመጀመሪያ

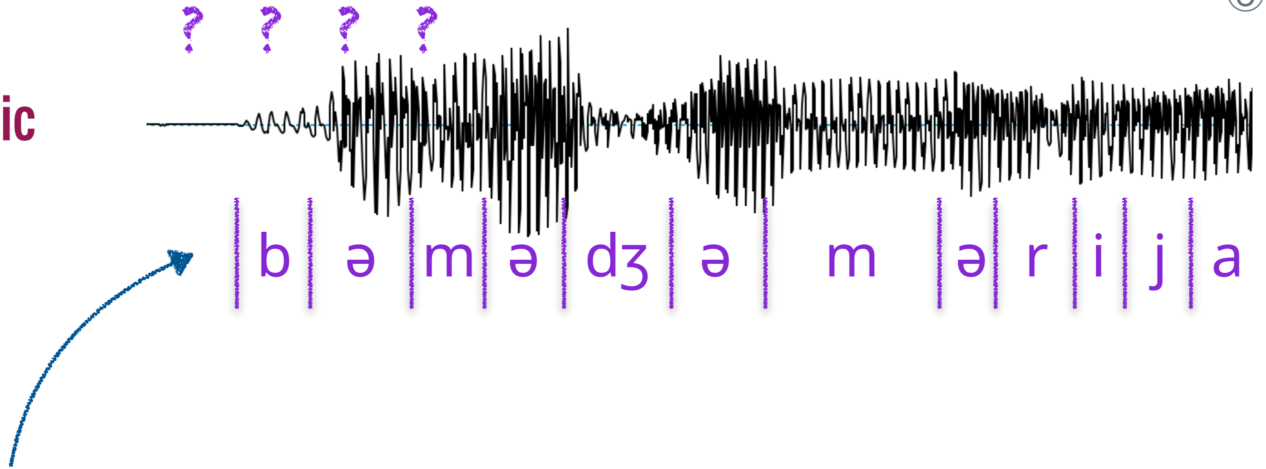
Grapheme-to-Phoneme

(G2P)

Resources Needed

① speech
② transcripts
③ phonemic labels
④ time alignments
⑤ phonetic measures

Amharic



Forced alignment

(HMM acoustic model)

በመጀመሪያ

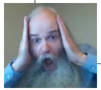
Phonetic measures (R or Praat):

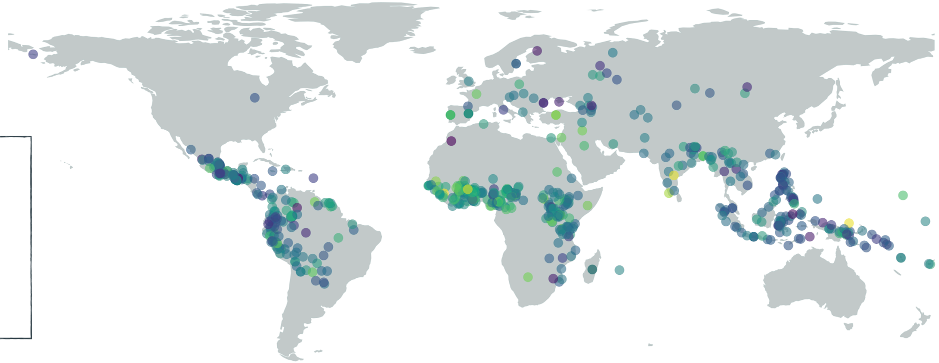
Formant frequencies, mid-frequency peak, duration…

Extraction Process

① speech
② transcripts

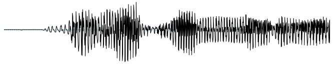
CMU Wilderness (2019)





699 Bible readings!

with ① speech!



and ② transcripts!

‘በመጀመሪያ’

Amharic

>1TB 😱

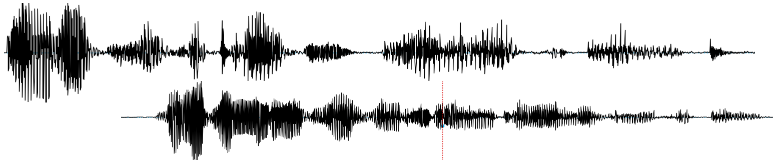
>6 years of CPU compute 😱

Extraction Process

① speech
② transcripts

CMU Wilderness dataset

Chapter:

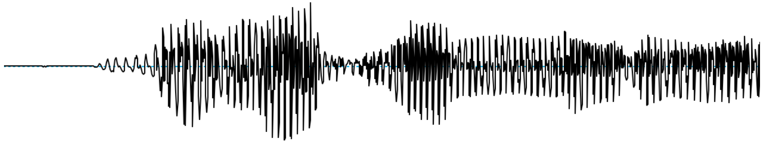


~30min

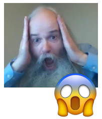
1 የፍጥረት አጀማመር በመጀመሪያ እግዚአብሔር (ኤሎሂም) ሰማያትንና ምድርን ፈጠረ። 2 ምድርም ቅርጽ የለሽና ባዶ ነበረች።※ የምድርን
ጥልቅ ስፍራ ሁሉ ጨለማ ውጦት ነበር። የእግዚአብሔርም (ኤሎሂም) መንፈስ በውሆች ላይ ይረብብ ነበር። 3 ከዚያም
እግዚአብሔር (ኤሎሂም) "ብርሃን ይሁን" አለ፤ ብርሃንም ሆነ። 4 እግዚአብሔርም (ኤሎሂም) ብርሃኑ መልካም እንደሆነ አየ፤ ብርሃኑን ከጨለማ
ለየ። 5 እግዚአብሔርም (ኤሎሂም) ብርሃኑን "ቀን"፣ ጨለማውን "ሌሊት" ብሎ ጠራው። መሸ፤ ነጋም፤ የመጀመሪያ ቀን። 6 እግዚአብሔር (ኤሎሂም)፣
"ውሃን ከውሃ የሚለይ ጠፈር በውሆች መካከል ይሁን" አለ። 7 ስለዚህ እግዚአብሔር (ኤሎሂም) ጠፈርን አድርጎ ከጠፈሩ በላይና ከጠፈሩ በታች ያለውን ውሃ ለየ፤
እንዳለውም ሆነ። 8 እግዚአብሔር (ኤሎሂም) ጠፈርን "ሰማይ" ብሎ ጠራው። መሸ፤ ነጋም፤ ሁለተኛ ቀን። 9 ከዚያም እግዚአብሔር (ኤሎሂም)፣ "ከሰማይ በታች ያለው
ውሃ በአንድ.

...

Utterance:



<30s

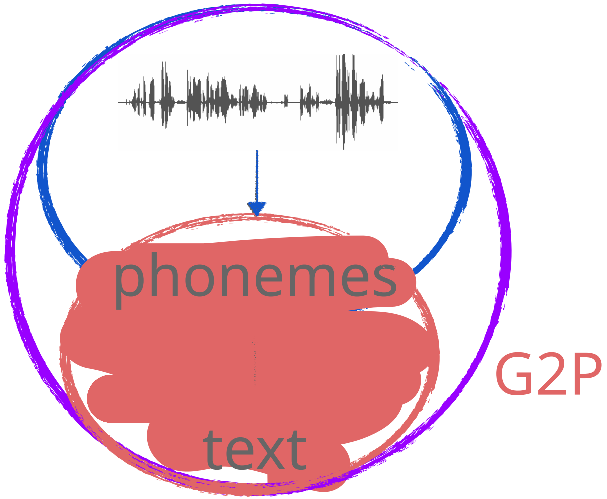


በመጀመሪያ

Extraction Process

① speech
② transcripts
③ phonemic labels

Which phonemes are present?



~~/ɹɛt/~~

/ɹɛd/

read

/ɛ/

read

/i/

Phoneme “Transcriptions”– Grapheme-to-Phoneme

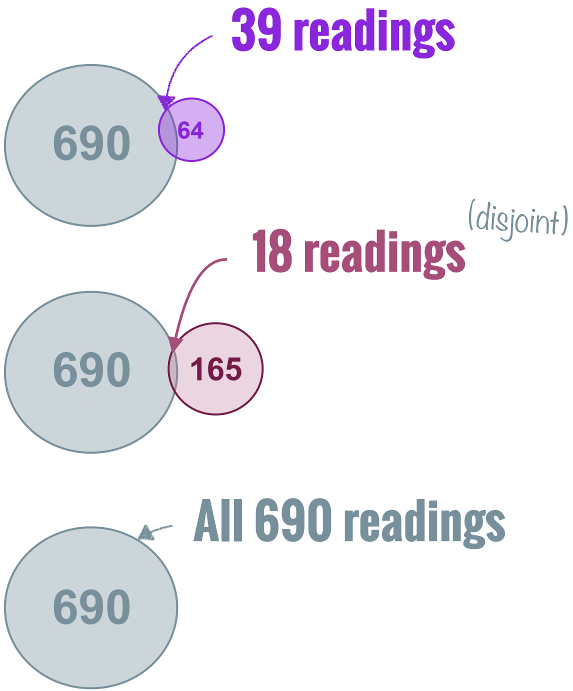
① **Linguist-created rules** *(Epitran)*

② **Wisdom of Crowds** *(Wiktionary/WikiPron)*

+ our own WFST-models (Phonetisaurus 🦕)

③ **Naïve baseline** *(Unitran)*

😱 “first-pass transcription”

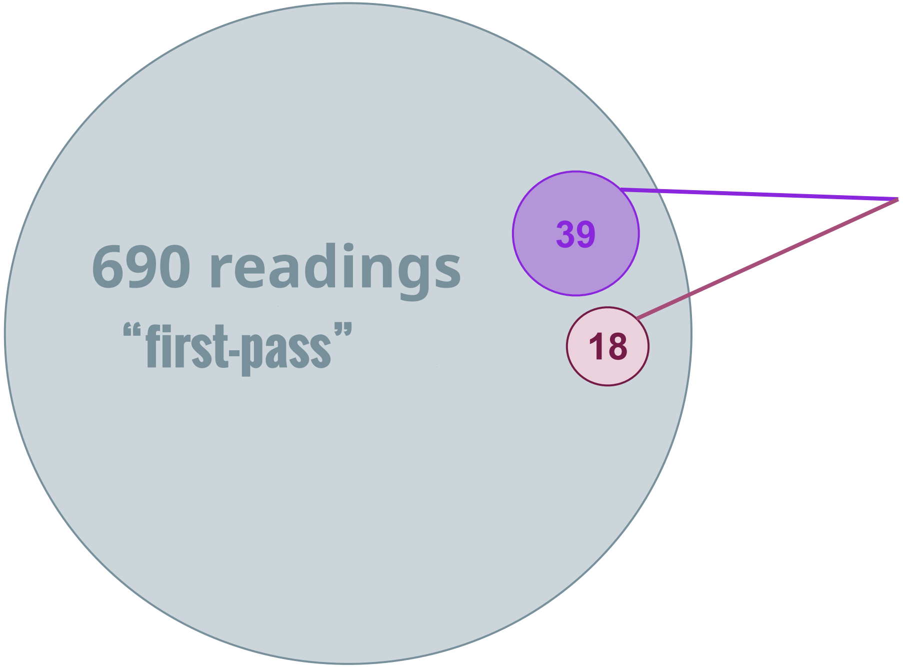


39 readings

18 readings (disjoint)

All 690 readings

G2P Summary



57 readings
“High-resource (HR)”

ALL 690 readings
“First-pass (FP)”

🤔 why provide **FP** alignments for languages with **HR** ? We’ll come back to that 😉

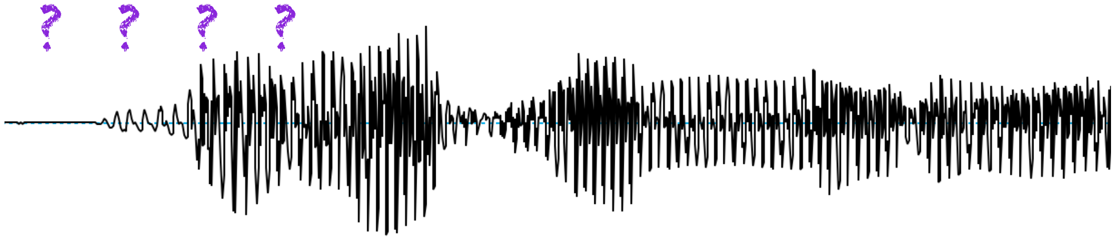
Extraction Process

① speech

② transcripts

③ phonemic labels

Amharic



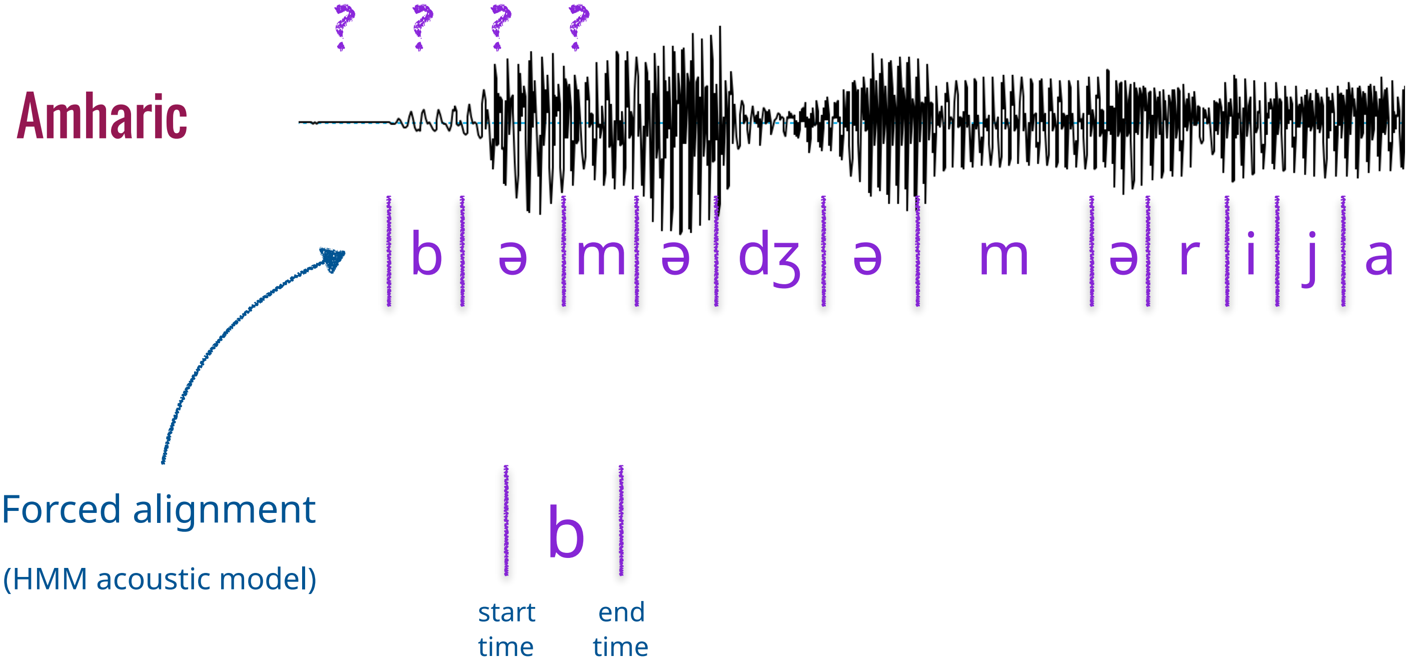
bəmədʒəmərija

Forced alignment

(HMM acoustic model)

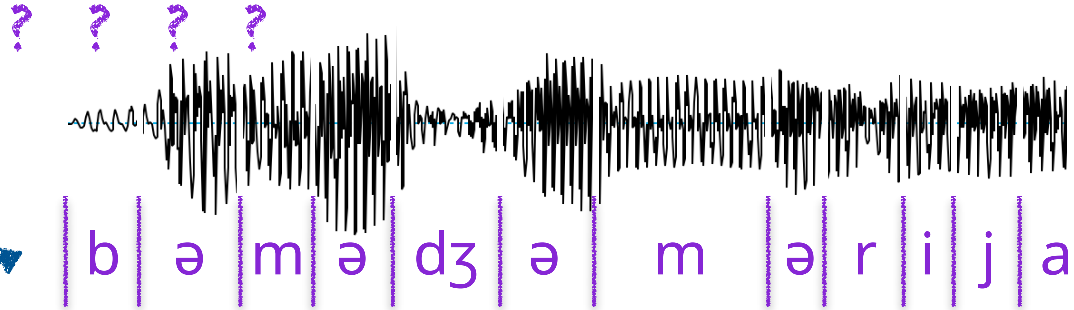
Extraction Process

① speech
② transcripts
③ phonemic labels
④ time alignments



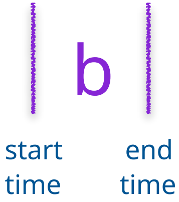
Extraction Process

① speech
② transcripts
③ phonemic labels
④ time alignments



Forced alignment

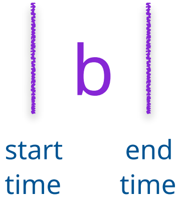
(HMM acoustic model)



Extraction Process

① speech
② transcripts
③ phonemic labels
④ time alignments

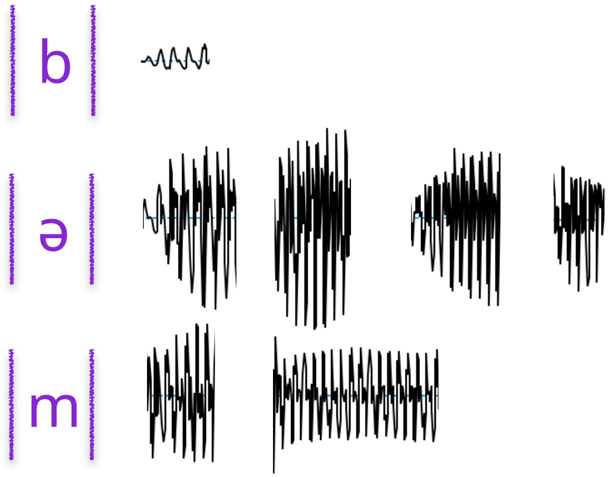
Amharic


b

start time

end time

Phoneme tokens:


b

ə

m

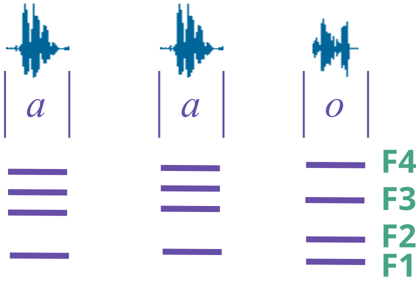
...

Extraction Process

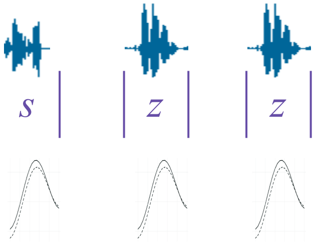
Phonetic Measures

① speech
② transcripts
③ phonemic labels
④ time alignments
⑤ phonetic measures

VOWELS

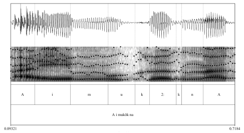


SIBILANTS



eg high-amplitude frequencies

Formants



PRAAT TEXTGRID

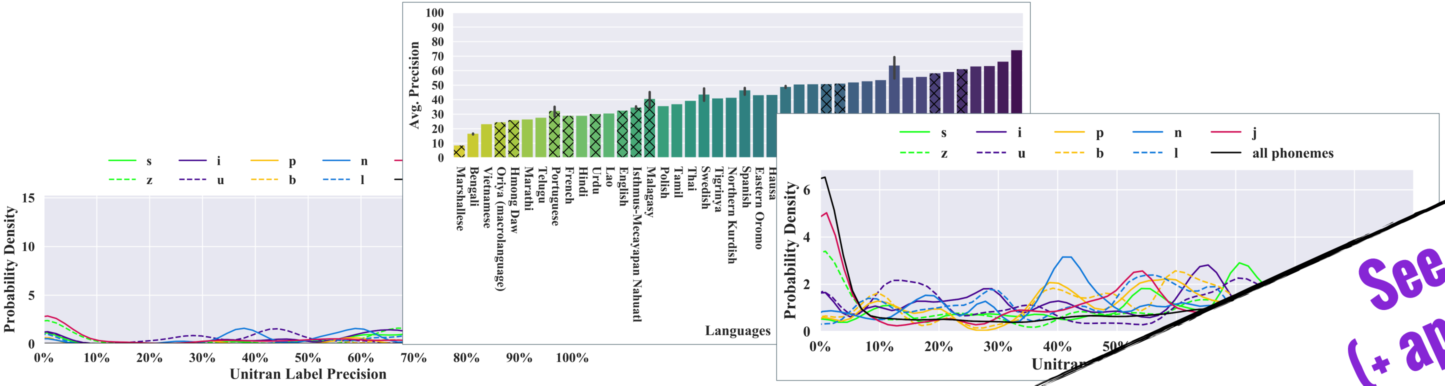
Spectral peak, COG, Duration, ...

Evaluation

🤔 Why provide both Unitran and High-Resource alignments?

Use multiple sets of alignments to assess Unitran alignment quality

- How much does quality vary across languages?
- Are certain phonemes more accurate than others?
- What about time alignment accuracy?



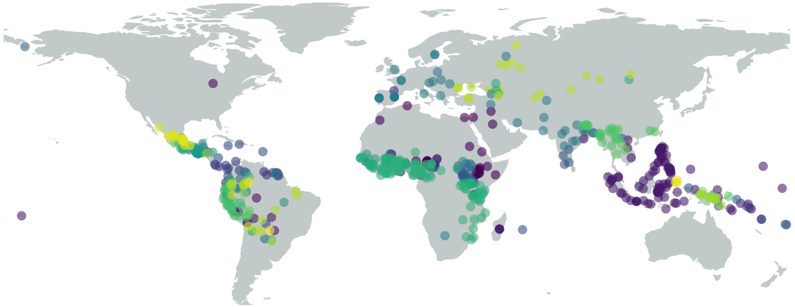
See paper!
(+ appendices)

Corpus Summary

VoxClamantis v1.0 provides **tokens** of **phoneme-level measurements** in hundreds of languages!

- 690 recorded readings of the Bible
- 635 languages (ISO 639-3)
- 70 language families

- >400 million aligned phoneme-level segments
- Subsequent phonetic measures for all vowels and sibilants



Case Studies

Case studies with VoxClamantis v1.0

Vowels

~50 phonemes

Sibilants

/s/ /z/

48 High-Resource Readings

① Reproduction of previous results validates resource

② Research at scale suggests general cross-linguistic principles

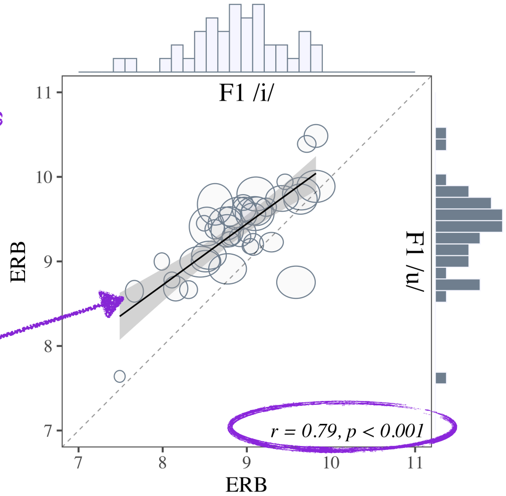
Phonetic Uniformity

Are shared characteristics realized uniformly within languages?

(eg: vowel height, POA) (eg: measures strongly correlated)

Formants: Vowels

/i/, /u/: high vowels

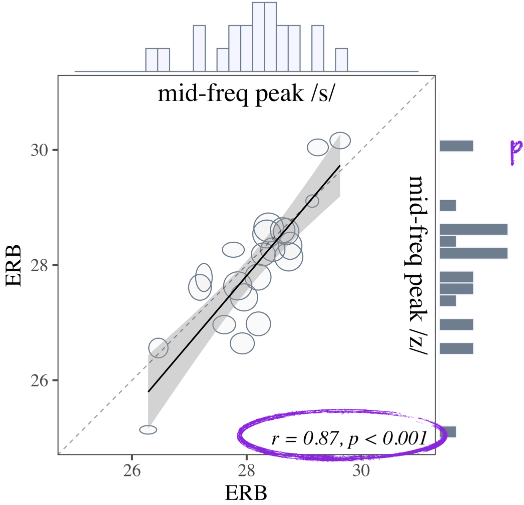


(eg: language)

While variation exists across languages, within language F1 strongly correlated

Mid-Freq Peak: Sibilants

/s/, /z/: alveolar place of articulation



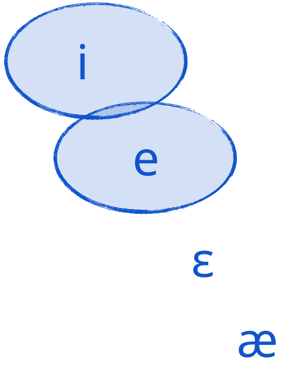
Supports hypothesis that this may be a universal principle

Reproduce previous results, but with many more languages

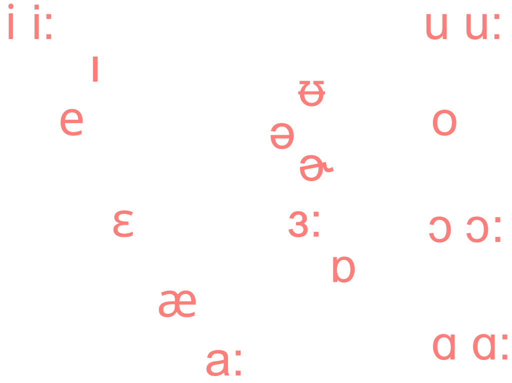
Phonetic Dispersion

Is inventory size correlated with articulatory precision?

VOWELS **4 vowels** → **20 vowels**



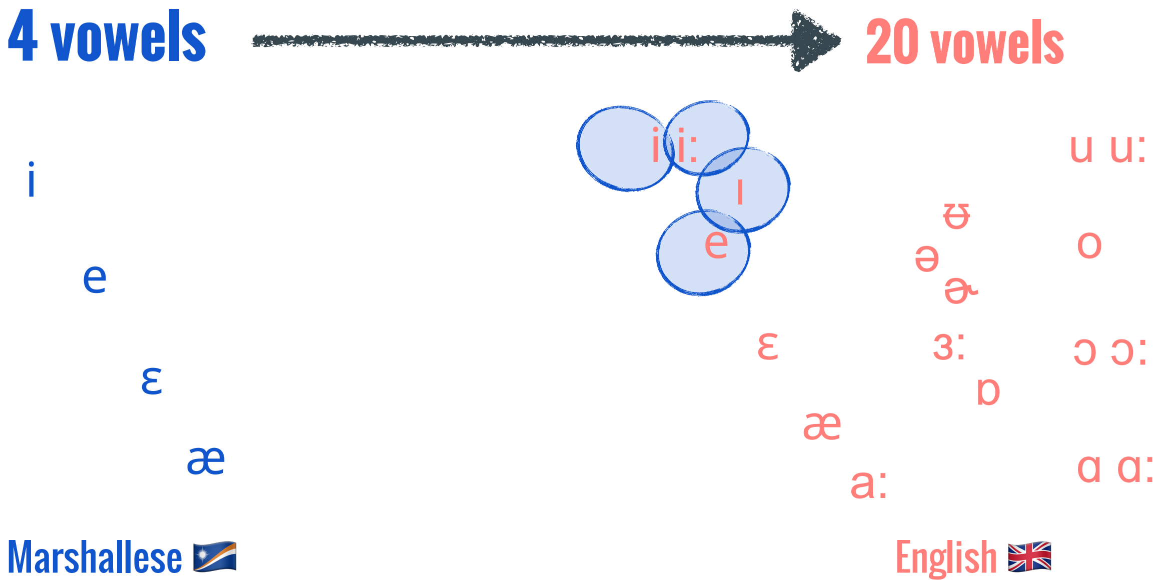
Marshallese 🇲🇭



English 🇬🇧

Phonetic Dispersion

Is inventory size correlated with articulatory precision?

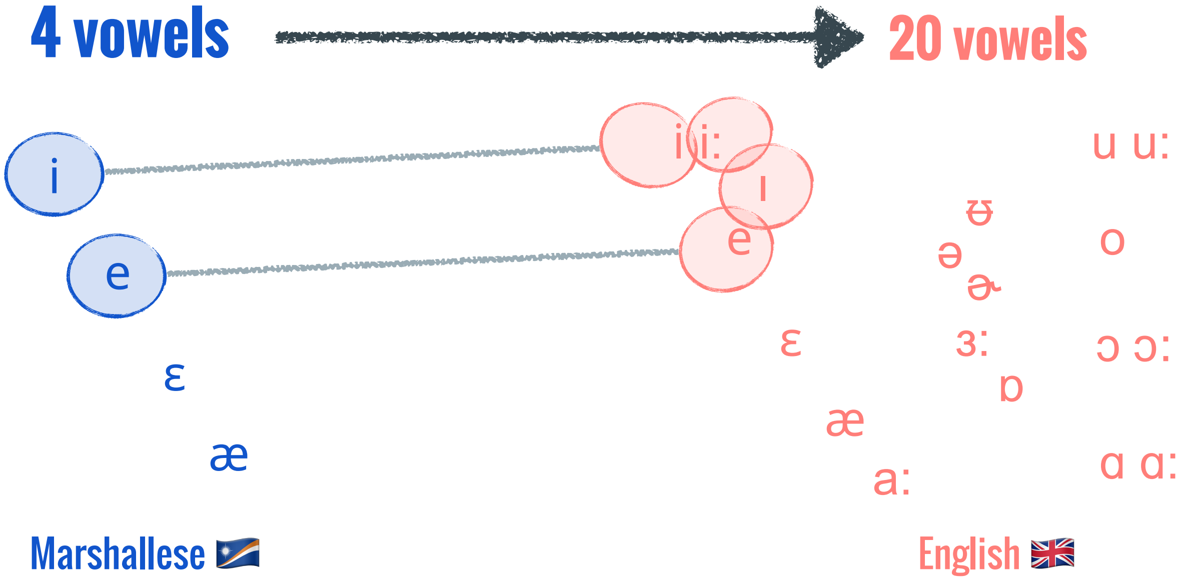


Phonetic Dispersion

Is inventory size correlated with articulatory precision?

No

(Spearman ρ = 0.11, p = 0.44;
Pearson r = 0.11, p = 0.46)


4 vowels → 20 vowels

Marshallese 🇲🇭

English 🇬🇧

Previously shown,

but not possible to study at scale

Supports hypothesis that this may [not] be a universal principle

B+	Utterance alignment	Filter -- in future, realign!
A- D+	Automatic phoneme labels	Better G(+A)2P
A 0% 😱	Alignment assessment!	Curate more resources!
B	Corpus representation (e.g. speakers)	Curate more resources!

Summary

Conclusion

VoxClamantis v1.0 corpus:

voxclamantisproject.github.io

aligned phoneme-level segments in hundreds of languages

57 high-resource, 690 first-pass

😱 methodology is not perfect – version 1.0!

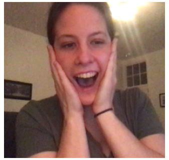
⬇️ download 🥳 use for research ⬆️ contribute to v2.0!

Contact Us!

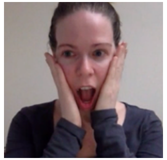
Questions!
Comments!
Contributions!

voxclamantisproject.github.io

voxclamantisproject@gmail.com



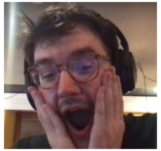
Elizabeth Salesky



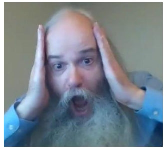
Eleanor Chodroff

Tiago Pimentel

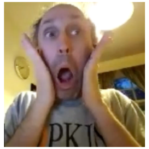
Matthew Wiesner



Ryan Cotterell



Alan W Black



Jason Eisner

VoxClamantis in deserto:
"a voice crying out in
the wilderness"