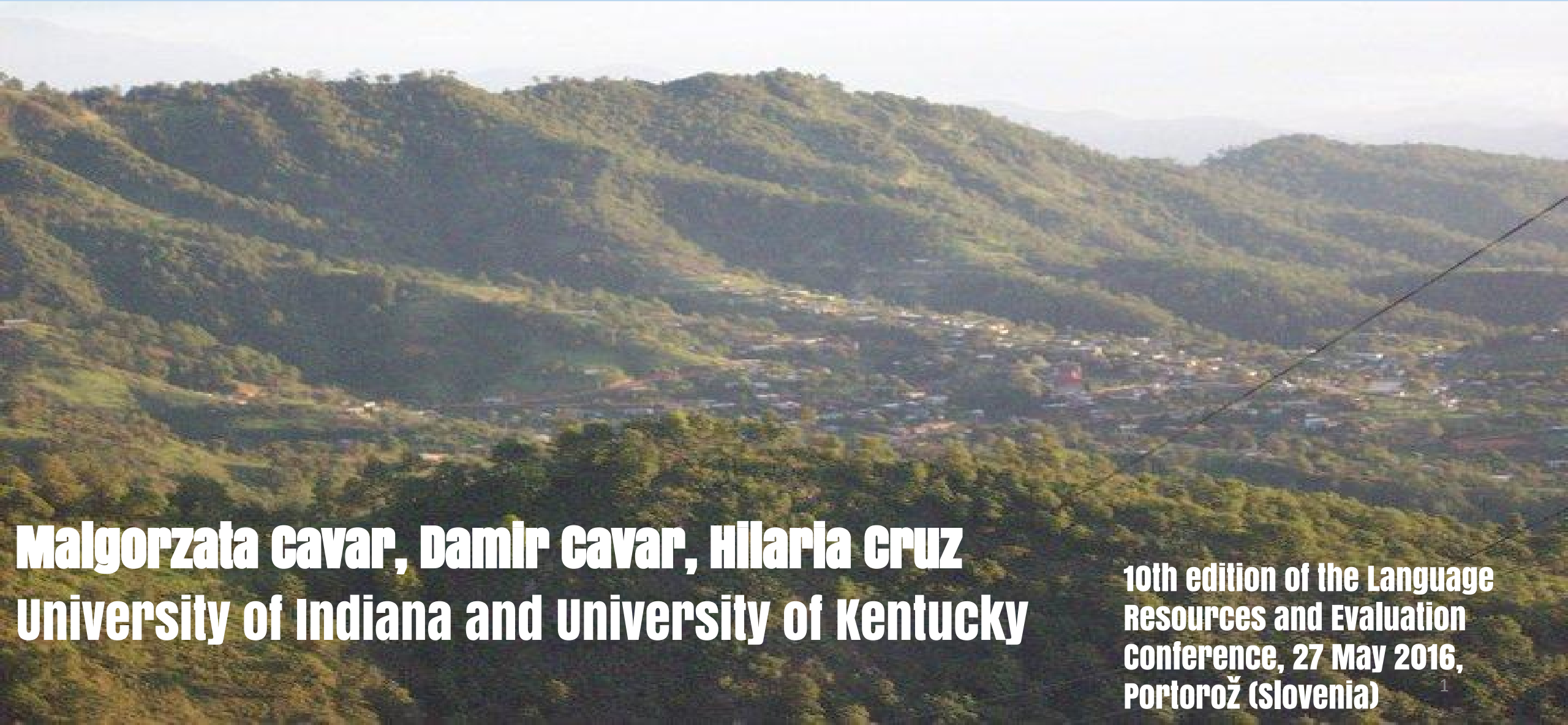
Endangered Language Documentation: Bootstrapping a Chatino Speech Corpus, Forced Aligner, ASR



Malgorzata Cavar, Damir Cavar, Hilaria Cruz
University of Indiana and University of Kentucky

10th edition of the Language Resources and Evaluation Conference, 27 May 2016, Portorož (Slovenia)

Agenda

- Work based on the NSF-funded AARDVARC project (http://info.linguistlist.org/aardvarc/) (NSF #1244713)
- Creating the first San Juan Quiahije Eastern Chatino (CTP) Speech Corpus
- Experiment with Forced Alignment
- Why?

The Problems

- The Language Resource Bottleneck
- Large Data Collections in Archives
- The Transcription Bottleneck

From:
http://fitforprostatesurgery.com/get-fit-get-healthy/establish-an-exercise-routine/

Endangered Language Documentation

- Large collections of recordings over the last decades
  - Limited amount of transcribed, translated, or analyzed data

- Ethnologue lists 7,097 living languages (yesterday)
  - 1% of those languages are well-resourced.
  - Biodiversity and language diversity threatened in a similar way.
  - Large number of languages: unwritten, partially qualitatively documented, etc.
  - Growing gap between the **low 99%** and the **1% highly resourced** languages.

Language Resources

- Audio and Video from documentary linguistic work
  - The Archive of the Indigenous Languages of Latin America at UT Austin (AILLA)
  - The Alaska Native Language Archive
  - DOBES, Max Planck Institute
  - SOAS, London
  - many more
- Lack of transcription, effort
  - 50 to 100 x real-time, i.e. one hour of recording requires 50 to 100 hours of work

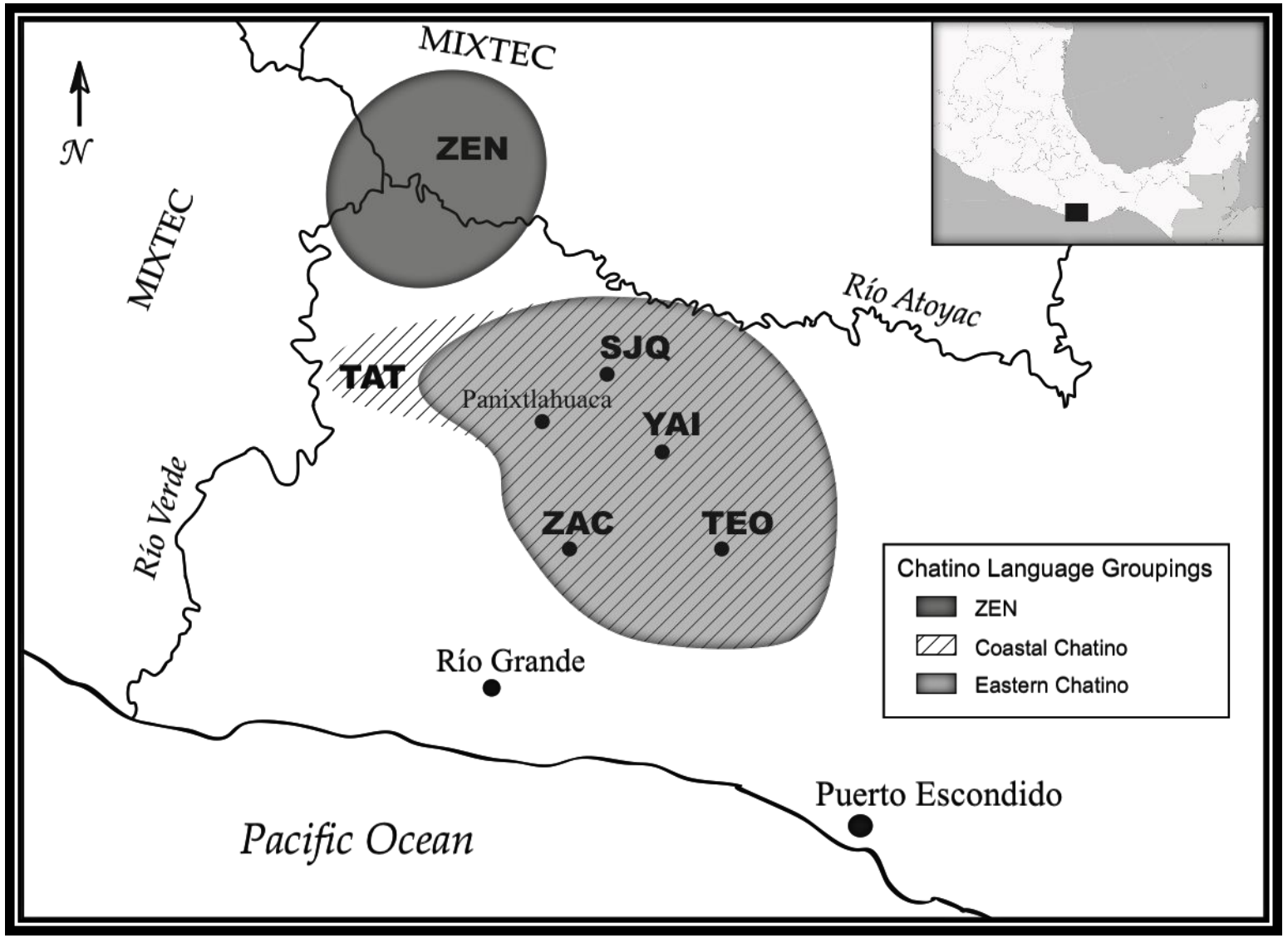
Possible Approach

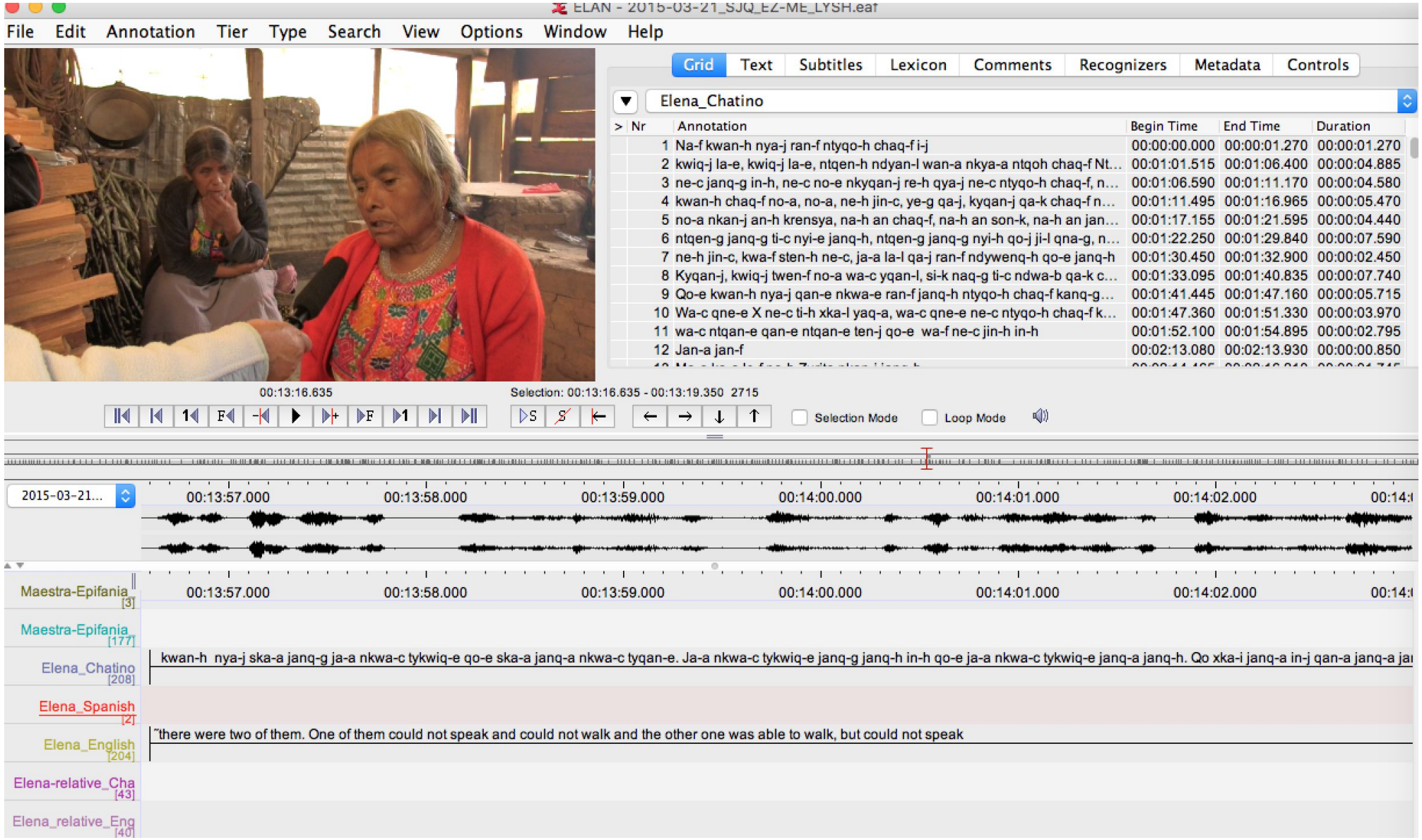
- Speech and video technologies (AARDVARC project)
  - Forced Alignment to speed up the production of speech corpora
  - Automatic Speech Recognition (ASR) for the recordings could speed up the transcription task.
    - Low number of speakers and longer spoken sequences = less signal variation.

- Task to estimate effort for:
  - Initial speech corpus for Forced Aligner (approx. 2 to 5 hours).
  - Effective corpus size for ASR.

The Chatino languages spoken in Oaxaca, Mexico

Chatino is a group of three languages





The shown scene was recorded by Lynn Hou.

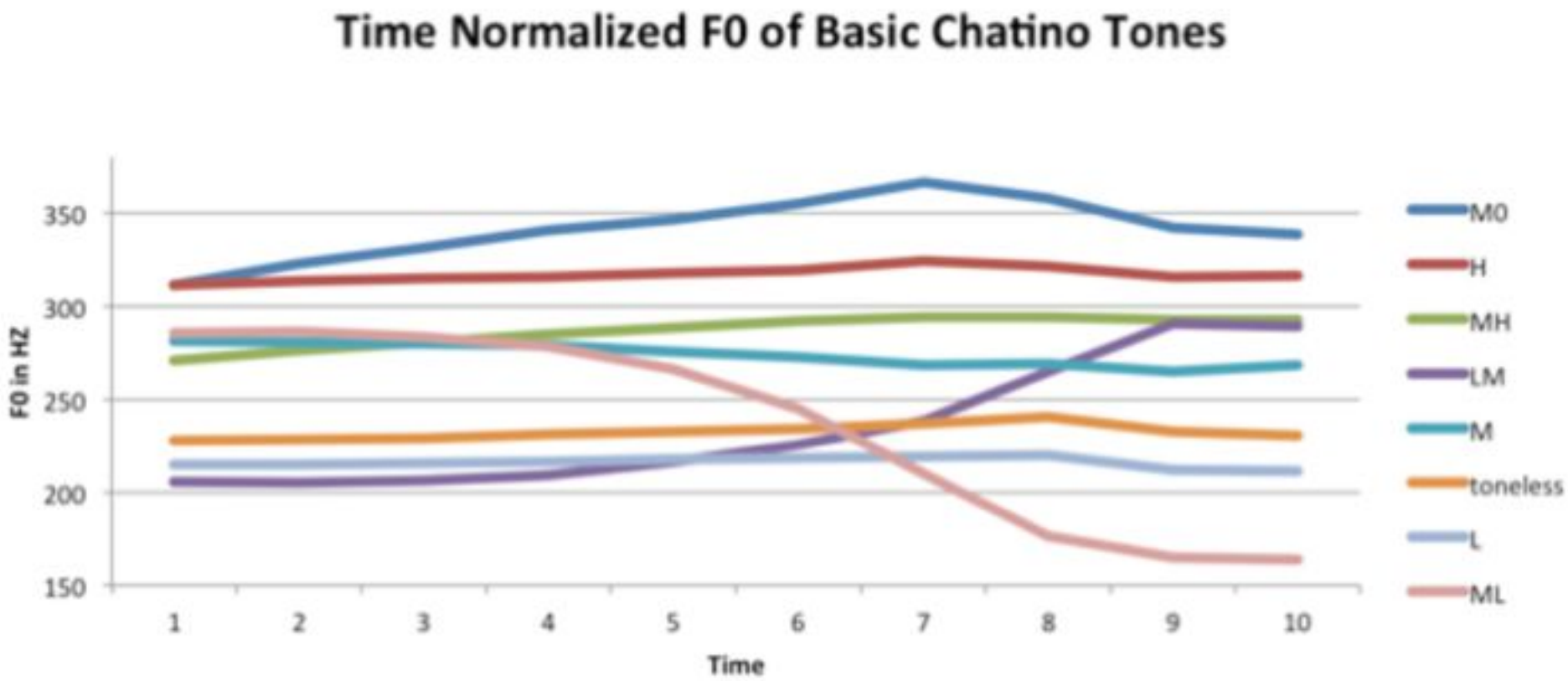
This is how Chatino sounds

Margarita Balthazar Garcia, Cienequilla.

Typological features of the Chatino Languages

- Highly tonal (Especially Eastern Chatino)
- Strongly head-marking
- Head-initial syntax
- VSO word order
- Alienable vs. inalienable possession
- complex verbal inflection classes
- Vigesimal numeral system
- no plural marking on nouns
- Excl vs. Incl. pronouns

TONES


Time Normalized F0 of Basic Chatino Tones
F0 in HZ
350
300
250
200
150
1
2
3
4
5
6
7
8
9
10
Time
M0
H
MH
LM
M
toneless
L
ML

Representation of tone

	Number	**Tonal group**	**gloss**
Level tones	ska1+0	skaK	sugar
	kla1	klaE	Weaving loom
	kla2	klaC	water pool
	kla3	klaF	dream
	kla4	klaA	fish, old, star
descending	kla24	klaJ	twenty
	tqwa14	tqwaB	cold
	tyuq04	tyuqM	term of endearment
ascending	kla42	klaG	you will arrive
	sqen32	sqenI	scorpion
	kla20	klaH	you will sing
	xkwan40	xkwanL	I will throw it

kta^{E} ‘foráneo (foreign)’

kta^{C} ‘harina (flour)’

kta^{F} ‘chepil (chepil *Crotalaria longirostrata*)’

kta^{A} ‘tabaco, se bañará (tabacco, s/he will take a shower)’

kta^{G} ‘lo sebraras, te bañaras(you will take a shower)’

kta^{H} ‘machucar (to bruise)’

kta^{I} ‘se sembrará (it will be planted)

Kta^{B} ‘ ganado (livestock)’

Approach

- Identification of text for reading and recording
- Recording initial speech corpus of 5 hours
- Transcription and time alignment
- Training a Forced Aligner
  - Prosodylab Aligner (PLA, Python module using HTK)
  - ELAN2split (https://bitbucket.org/dcavar/elan2split) corpus creation for PLA
  - Pronunciation dictionary (tokens plus tokenized phonetic transcription)
  - Espeak for Praat: Language model for TXT2Speech
- Extending the speech corpus

Outcome

- Initial speech corpus:
  - Approx. 5 hours speech transcribed and time aligned, PoS-tagged and translated
  - Initial annotation in ELAN (time-alignment correction in Praat)
- Workload (ignoring previous investment in text, transcription schema)
  - Ford Assembly Line approach: bootstrapping initial corpus, FA-training, …
  - 4 to 6 person weeks
  - Estimate $ 24,000 – 50,000
  - If we would do that for 3,500 languages: < $ 84 mil.

Resources

- GORILLA site (http://gorilla.linguistlist.org/)
  - Audio, ELAN and Praat transcription/annotation files
  - Corpus licenses: CC BY-SA, i.e. free for commercial use (donation-ware and copyright free resources)
  - Code and software: Apache 2.0 licensed, i.e. free for commercial use
  - Every resource comes with a paper to cite
  - LLOD-linked
  - CLARIN-linked
  - …

Resources

- Internships at LINGUIST List and Indiana University:
  - Work on LL resources, but also: corpus creation, speech and language technologies, qualitative and quantitative language related research
  - Corpora created with colleagues, students, native speakers and community members:
    - Chatino, Burmese, Turkic (Baharlu, Khorasan, Iran), Croatian, Russian, Spanish, …
  - Welcoming students from all over the world!

WaC xqweF
qwanJ

Hvala!