ChoCo: A Multimodal Corpus for the Choctaw Language

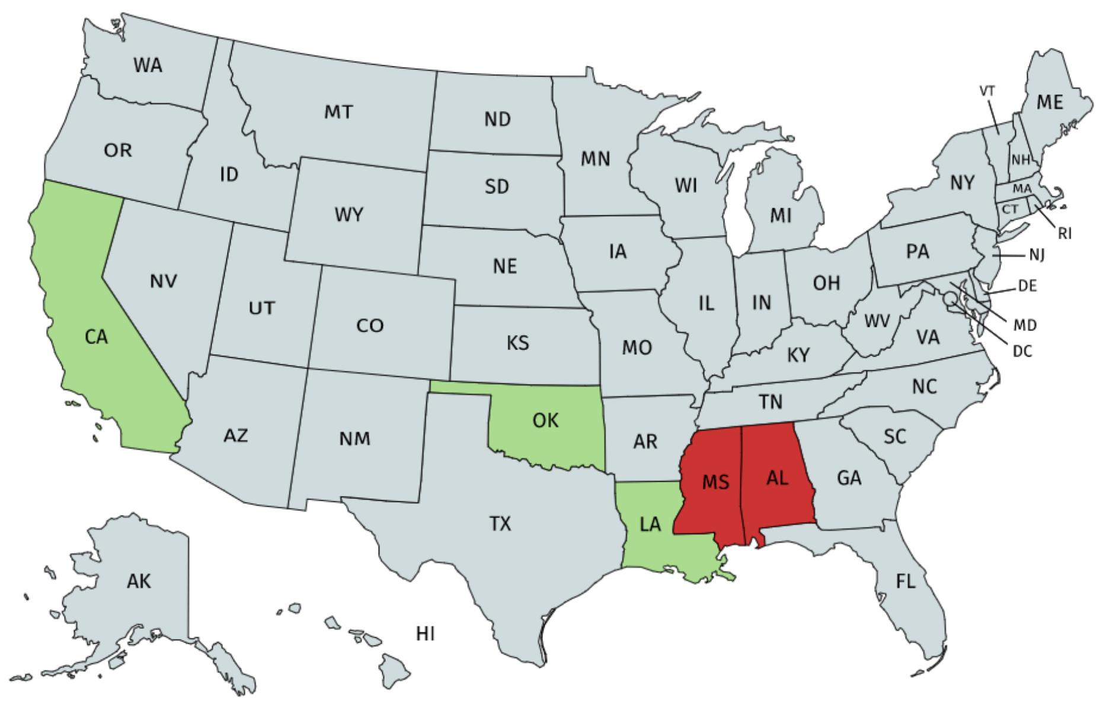
USC University of Southern California

Jacqueline Brixey David Traum

USC Institute for Creative Technologies

Choctaw = Native American tribe and language

- ~ 10,000 fluent speakers
- Threatened language status
- 3 federally recognized tribes
- Variants/dialects but little phonetic difference between groups:
  - Mississippi (MS)
  - Oklahoma (OK)
  - Mississippi Choctaw of Oklahoma (MSOK)

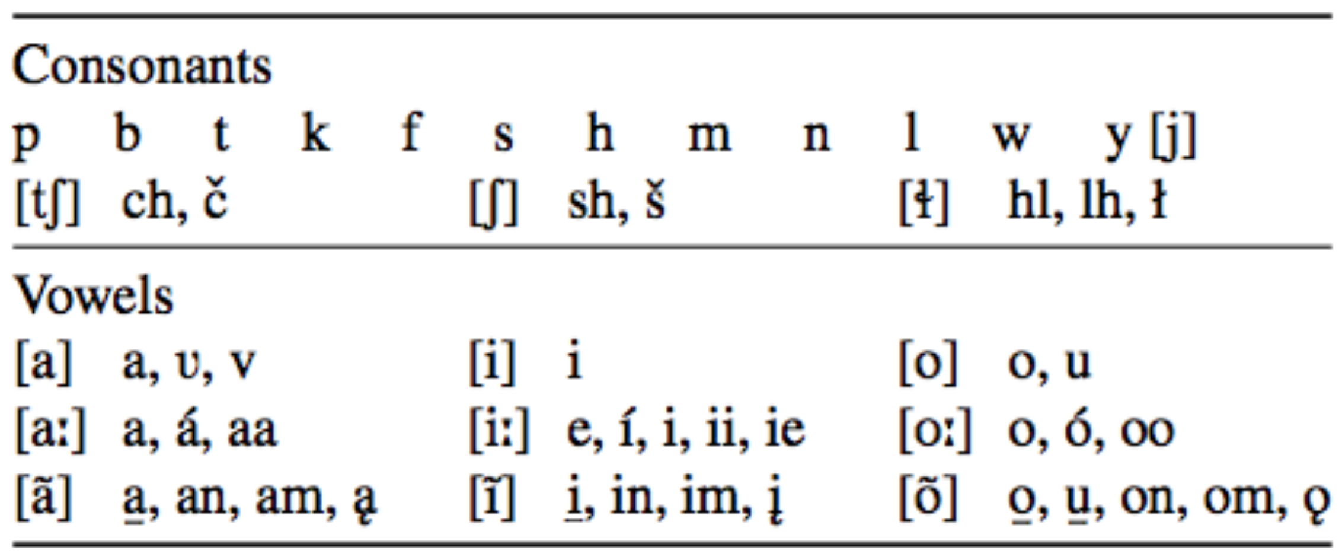


Documentation

- First documented in the 1800s by a missionary, Cyrus Byington
  - Grammar
  - Dictionary
- Other grammars by Nicklas (1972), Ulrich (1986), Broadwell (2006)
- Dictionaries by Byington (1915), Wright (1880), Choctaw Nation of Oklahoma (2016)

Challenges

- Not fully standardized writing system



Consonants											
p	b	t	k	f	s	h	m	n	l	w	y [j]
[tʃ]	ch, č				[ʃ]	sh, š			[ɬ]	hl, lh, ł	

Vowels					
[a]	a, ʋ, v	[i]	i	[o]	o, u
[aː]	a, á, aa	[iː]	e, í, i, ii, ie	[oː]	o, ó, oo
[ã]	a̱, an, am, ą	[ĩ]	i̱, in, im, į	[õ]	o̱, u̱, on, om, ǫ

For example, “possum” can be written as:

shokhata
shukhʋta
shukhata
shokkata

Choctaw Language

- In the Muskogean language family
  - Other members: Chickasaw, Alabama, Muscogee
- Subject-Object-Verb order
- Complex morphology
  - Highly inflectional language

Surface level	I̲kana ʊt aiittʊtoba ia tuk.
Morphology	I̲-kana ʊt aiitʊtoba ia tuk.
Gloss	His-friend the store go PAST
English	‘His friend went to the store.’
Surface level	Ashekonopa ilʊppʊ ish tahakchi.
Morphology	Ashekonopa ilʊppʊ ish ta<ha>kchi.
Gloss	knot this 2SG tie<quickly>
English	‘You tie this knot quickly.’
Surface level	Chik impo.
Morphology	Chik imp-o.
Gloss	not.2SG eat-NEG
English	‘You are not eating.’

- Orthographic difference in spacing of morphemes
- To write ‘I will tie’:
  - Takchi + li + achin (future marker)
    - Takchi la chi̱
    - Takchi la chin
    - Takchilachi̱
    - Takchilachin

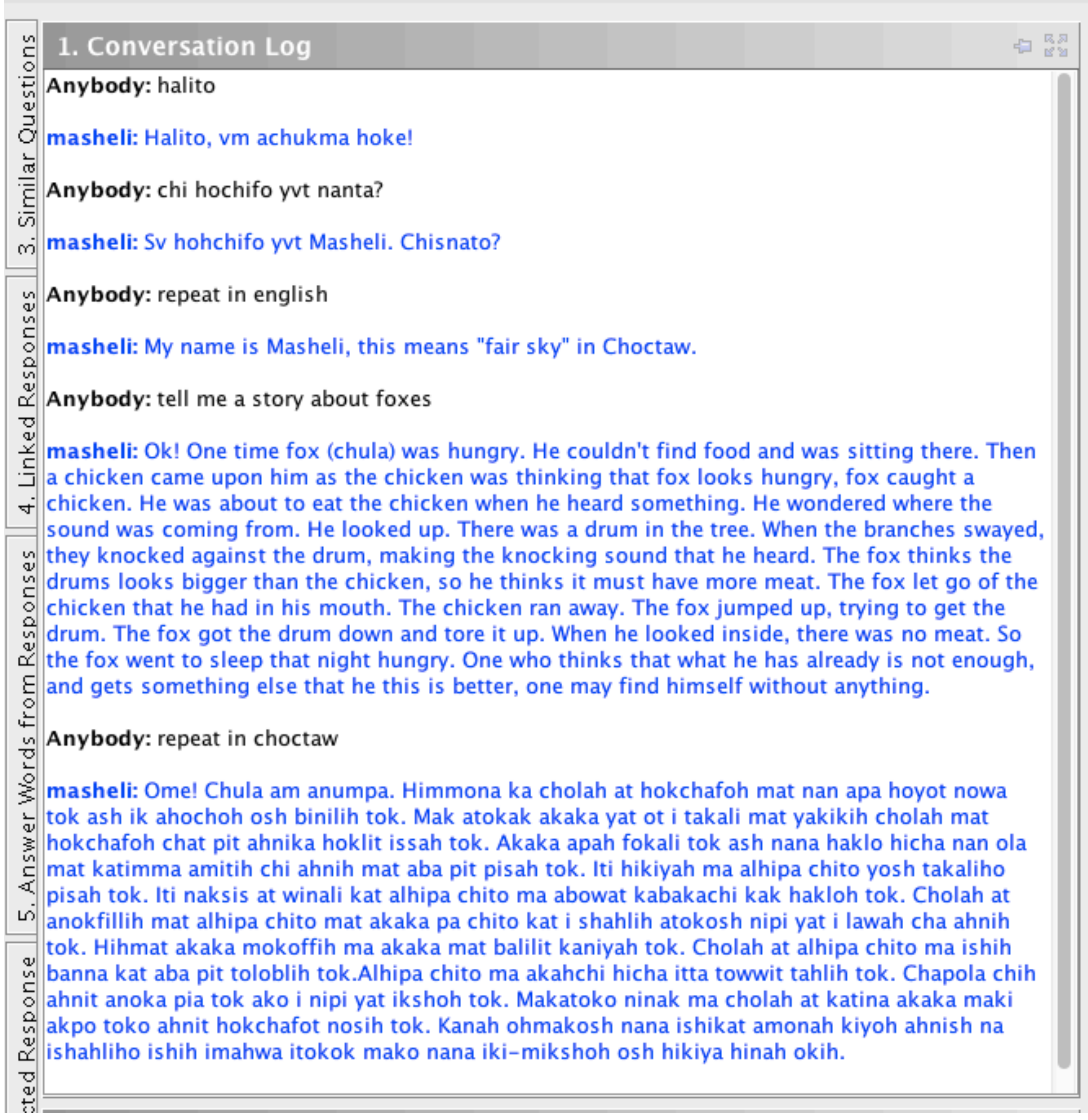
Data Set

- 30 videos (105 minutes 38 seconds)
- Audio (256 minutes 40 seconds)
- Collected sparse existing texts
- All text entries stored in text files
- Each entry has reference for originating source

Type	# Tokens in Oklahoma	# Tokens in Mississippi
Short Stories	5954	1693
Phrases	12711	331
Poetry	243	0
Correspondence	159	0
Religious Texts	222	30010
Examples from Scholarly Sources	589	12
Dictionaries	42381	0
Total Word Tokens	62259	32046

One Use Case

- Masheli chat bot
  - NPC Editor backend
  - Used bilingual stories from corpus
  - Can answer English/Choctaw user input
  - Responds in same language as input
  - Can repeat last utterance in other language if requested



Anybody: halito

masheli: Halito, vm achukma hoke!

Anybody: chi hochifo yvt nanta?

masheli: Sv hohchifo yvt Masheli. Chisnato?

Anybody: repeat in english

masheli: My name is Masheli, this means "fair sky" in Choctaw.

Anybody: tell me a story about foxes

masheli: Ok! One time fox (chula) was hungry. He couldn't find food and was sitting there. Then a chicken came upon him as the chicken was thinking that fox looks hungry, fox caught a chicken. He was about to eat the chicken when he heard something. He wondered where the sound was coming from. He looked up. There was a drum in the tree. When the branches swayed, they knocked against the drum, making the knocking sound that he heard. The fox thinks the drums looks bigger than the chicken, so he thinks it must have more meat. The fox let go of the chicken that he had in his mouth. The chicken ran away. The fox jumped up, trying to get the drum. The fox got the drum down and tore it up. When he looked inside, there was no meat. So the fox went to sleep that night hungry. One who thinks that what he has already is not enough, and gets something else that he this is better, one may find himself without anything.

Anybody: repeat in choctaw

masheli: Ome! Chula am anumpa. Himmona ka cholah at hokchafoh mat nan apa hoyot nowa tok ash ik ahochoh osh binilih tok. Mak atokak akaka yat ot i takali mat yakikih cholah mat hokchafoh chat pit ahnika hoklit issah tok. Akaka apah fokali tok ash nana haklo hicha nan ola mat katimma amitih chi ahnih mat aba pit pisah tok. Iti hikiyah ma alhipa chito yosh takaliho pisah tok. Iti naksis at winali kat alhipa chito ma abowat kabakachi kak hakloh tok. Cholah at anokfillih mat alhipa chito mat akaka pa chito kat i shahlih atokosh nipi yat i lawah cha ahnih tok. Hihmat akaka mokoffih ma akaka mat balilit kaniyah tok. Cholah at alhipa chito ma ishih banna kat aba pit toloblih tok.Alhipa chito ma akahchi hicha itta towwit tahlih tok. Chapola chih ahnit anoka pia tok ako i nipi yat ikshoh tok. Makatoko ninak ma cholah at katina akaka maki akpo toko ahnit hokchafot nosih tok. Kanah ohmakosh nana ishikat amonah kiyoh ahnish na ishahliho ishih imahwa itokok mako nana iki-mikshoh osh hikiya hinah okih.

Current and Future Work

- Corrections to OCR texts (New Testament; Books of Genesis; Choctaw Hymn Book; The Book of Questions of the Gospel of Luke)
- Processing audio recordings of Choctaw speakers that were recently collected
- Morphology generator
  - Will inflect verbs for tense, pronominal agreement, and aspect
  - Input list of verbs from Byington dictionary
  - Will be validated by comparing to corpus and with fluent speakers