Generating speech tools and corpora for unwritten endangered languages: Chatino

Damir Cavar
Indiana University

Audio data without transcription

- Example:
  - Chatino in AILLA
    - files: 130, restricted files: 5%
    - audio recordings: 26, length of audio: 7:53:5
    - video recordings: 72, length of video 4:26:4
    - digital texts: 12, pages: 156
    - resources that include transcriptions: 2%
    - Collection "Chatino Documentation of Hilaria Cruz" in AILLA

Issues

- Licensing and availability of resources
- Quality of resources
- Transcription effort
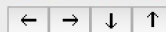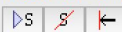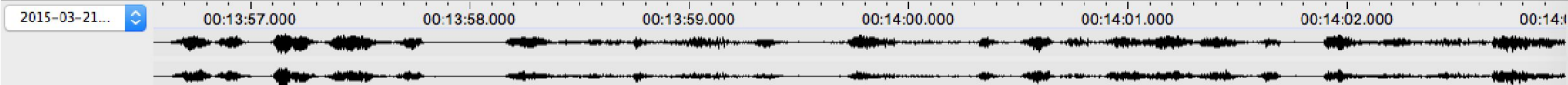- Available transcriptions

Chatino

- Transcriptions and texts (using particular transcription schema, Woodbury & Cruz)
  - Collected by Hilaria Cruz
- Recording of Hilaria under studio conditions
  - 96 kHz, 24 bit, WAV using Zoom H6
- Transcription reduced to:
  - Copy and paste
  - Time alignment at utterance level in ELAN
- Corpus licensed:
  - CC BY-SA (free for commercial and other use)

File   Edit   Annotation   Tier   Type   Search   View   Options   Window   Help

| Grid | Text | Subtitles | Lexicon | Comments | Recognizers | Metadata | Controls |

Elena_Chatino

>	Nr	Annotation	Begin Time	End Time	Duration
	1	Na-f kwan-h nya-j ran-f ntyqo-h chaq-f i-j	00:00:00.000	00:00:01.270	00:00:01.270
	2	kwiq-j la-e, kwiq-j la-e, ntqen-h ndyan-l wan-a nkya-a ntqoh chaq-f Nt...	00:01:01.515	00:01:06.400	00:00:04.885
	3	ne-c janq-g in-h, ne-c no-e nkyqan-j re-h qya-j ne-c ntyqo-h chaq-f, n...	00:01:06.590	00:01:11.170	00:00:04.580
	4	kwan-h chaq-f no-a, no-a, ne-h jin-c, ye-g qa-j, kyqan-j qa-k chaq-f n...	00:01:11.495	00:01:16.965	00:00:05.470
	5	no-a nkan-j an-h krensya, na-h an chaq-f, na-h an son-k, na-h an jan...	00:01:17.155	00:01:21.595	00:00:04.440
	6	ntqen-g janq-g ti-c nyi-e janq-h, ntqen-g janq-g nyi-h qo-j ji-l qna-g, n...	00:01:22.250	00:01:29.840	00:00:07.590
	7	ne-h jin-c, kwa-f sten-h ne-c, ja-a la-l qa-j ran-f ndywenq-h qo-e janq-h	00:01:30.450	00:01:32.900	00:00:02.450
	8	Kyqan-j, kwiq-j twen-f no-a wa-c yqan-l, si-k naq-g ti-c ndwa-b qa-k c...	00:01:33.095	00:01:40.835	00:00:07.740
	9	Qo-e kwan-h nya-j qan-e nkwa-e ran-f janq-h ntyqo-h chaq-f kanq-g ...	00:01:41.445	00:01:47.160	00:00:05.715
	10	Wa-c qne-e X ne-c ti-h xka-l yaq-a, wa-c qne-e ne-c ntyqo-h chaq-f k...	00:01:47.360	00:01:51.330	00:00:03.970
	11	wa-c ntqan-e qan-e ntqan-e ten-j qo-e wa-f ne-c jin-h in-h	00:01:52.100	00:01:54.895	00:00:02.795
	12	Jan-a jan-f	00:02:13.080	00:02:13.930	00:00:00.850

00:13:16.635   Selection: 00:13:16.635 - 00:13:19.350   2715

|◄◄ |◄ 1◄ F◄ -|◄ ► ►|+ ►F ►1 ►| ►|| ►S S̸ ←   ← → ↓ ↑   ☐ Selection Mode   ☐ Loop Mode   🔊

2015-03-21...

| | 00:13:57.000 | 00:13:58.000 | 00:13:59.000 | 00:14:00.000 | 00:14:01.000 | 00:14:02.000 | 00:14:0 |

Maestra-Epifania [3]

Maestra-Epifania [177]

Elena_Chatino [208]   kwan-h nya-j ska-a janq-g ja-a nkwa-c tykwiq-e qo-e ska-a janq-a nkwa-c tyqan-e. Ja-a nkwa-c tykwiq-e janq-g janq-h in-h qo-e ja-a nkwa-c tykwiq-e janq-a janq-h. Qo xka-i janq-a in-j qan-a jan...

Elena_Spanish [2]

Elena_English [204]   "there were two of them. One of them could not speak and could not walk and the other one was able to walk, but could not speak

Elena-relative_Cha [43]

Elena_relative_Eng [40]

Chatino

- Recording more voices:
  - Native-speakers cannot read the Chatino transcription schema
  - Presenting stimuli to speakers via headphones
  - Recording repetition of audio stimuli (and possible errors)
- Goals:
  - Variation of voices by gender and age
  - Collection >10 hours of audio (affecting signal feature extraction, not vocabulary or language model)
  - Transcription: copy and paste + time alignment using Forced Aligner
  - Subsequent transcription and FA of more data

Previous experience with audio repetition

- Stimuli containing variation in word order to test gradual grammaticality judgements, and acoustic perception/articulation:
  - Neo-Shtokavian
  - Chakavian
- Error rate could be used to rank grammaticality or complexity at respective linguistic level.

Forced Alignment

- Speech recognition without Language Model
  - Im-/Explicit in the transcription to be aligned

- Simplified process using:
  - Prosodylab Aligner
  - ELAN2split
  - Tokenizer script for unwritten languages or lexicon for written ones

Prosodylab Aligner

- Requirements
  - HTK
  - SoX
  - Python
- Useful additions:
  - ELAN2split (C++11 tool, using Xerces-C++ lib and SoX)
    - https://bitbucket.org/dcavar/elan2split
- Platforms:
  - Mac or Linux (in principle doable on Windows, too)

Prosodylab Aligner

- Comes with English or French, some dictionaries available (e.g. German, Dutch)
- For new acoustic model:
  - Min. one hour of audio with corresponding word-level transcription
  - Pronunciation dictionary
  - Alignments are learned during training phase
  - That is:
    - One only needs utterance, sentence, or breath group segmentation (length not really relevant)

Pronunciation dictionary

- ASCII (or Unicode) text (see HTK tutorial)

```
AABERG AA1 B ER0 G
AACHEN AA1 K AH0 N
AACHENER AA1 K AH0 N ER0
AAKER AA1 K ER0
AALSETH AA1 L S EH0 TH
AAMODT AA1 M AH0 T
AANCOR AA1 N K AO2 R
```

Pronunciation dictionary

- For unwritten languages (e.g. Chatino):
    - Phonemic transcription of word followed by tokenized sequence of phonemes
- For written languages:
    - Use phonemic/phonetic transcription from dictionaries or manually transcribe
        - Raphael Finkel's Yiddish dictionary for our Yiddish corpus
        - Manual for Croatian corpus
        - etc.

ELAN2split output

- Corpus of utterance tuples:
  - WAV-section that is marked up in ELAN
  - Corresponding transcript in a particular tier

- Deep parsing of ELAN EAF-XML

- SoX-based cutting of WAV-file and segment generation

Praat aligner

- Analysis via Synthesis
  - Espeak based text-to-speech module (in Annotate, see Interval)
  - Declarative rule-based text-to-speech language model
    - Phone inventory
    - Exception dictionary
    - Stress, tone, contours, breath, etc.
  - Generation of Wave from transcription-token and fitting into the recorded Wave
  - Rich collection of language models
- Our models:
  - Yiddish, Burmese, ...

Burmese TXT2Speech

- Example: Burmese (Lwin Moe)
  - How are you? (Espeak): https://www.youtube.com/watch?v=Cr7JbO3SlIM
  - How are you? (Lwin Moe): https://www.youtube.com/watch?v=THu6957qebU
  - I am fine (Espeak): https://www.youtube.com/watch?v=ryupSFaJdUk
  - I am fine (Lwin Moe): https://www.youtube.com/watch?v=i6Vm-SaPDkw
  - How are you? (In English): https://www.youtube.com/watch?v=I5tl3l63b3g