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

Malgorzata E. Cavar, Damir Cavar, Hilaria Cruz

**Goals.** This project aims to improve and facilitate language documentation of endangered languages using corpus and computational linguistic methods and technologies, i.e. approaching the problem of documentation and revitalization from a rather untraditional angle. It demonstrates this approach on the basis of the endangered and seriously under-resourced variety of Eastern Chatino.

**Background.** Eastern Chatino of San Juan Quiahije (CTP) is a Zapotecan language spoken in the municipality of San Juan Quiahije, Oaxaca, Mexico by some 3,000 speakers. Widespread poverty, unemployment, deficient infrastructure in terms of schools, hospitals, and roads and lack of arable land, all lead to massive migration and further diminishing of the number of speakers. The other source of the language attrition is the dominance of Spanish and the absence of Chatino from education. There exists no standard orthography for Chatino and only scarce incomplete linguistic description. Previous approaches to document the language involved the collection of audio and video recording of speakers in free conversations or using ceremonial language. The large amounts of data are kept in archives and only a small portion of the recordings is transcribed or otherwise annotated. To enable such recordings, also recordings from other languages for research and other purposes, they need to be transcribed and annotated. Ideally, the transcription and annotation has to be time-aligned. This effort is estimated to consume 50 to 100 times real time. It is basically impossible to transcribe and annotate all the already collected Chatino recordings manually, not to mention the recordings of all other languages with thousands of hours of audio and video resources.

**Procedures.** Our approach is different from the traditional language documentation and fieldwork methods. Instead of collecting word lists and recording elicited speech from speakers we create first speech corpora. One problem with this approach is that the recordings contain speech that is unique and as such needs to be transcribed for every single recording: recordings from elicited speech require subsequent manual transcription and annotation. This transcription process is very costly and time consuming. In contrast, we first build a speech corpus that allows us to train speech and language technologies for a more rapid extension of the volume of annotated recordings, as well as further bootstrapping of automatic tools for the particular languages.

To generate a speech corpus for Chatino, we prepared existing texts and transcripts from field work and language documentation projects. We used ritual texts collected and transcribed by Cruz (2014), a Chatino researcher and native speaker of the language herself.

The texts are in fact using an ASCII-based phonemic transcription schema, because the language has no standard orthography yet. A native speaker read the text under near-studio conditions using high quality audio equipment (96kHz, 24 Bit, uncompressed WAVE-form audio file format).

This way approximately 3 hours of recordings were generated in the first phase - all texts were read by a trained native speaker.  The initial time alignment in ELAN has been created manually. The process of time alignment is optimized, since in this case the transcription does already exist, consequently, only the time alignment using common transcription tools is necessary. We used ELAN to initially time-align the transcription, part-of-speech tag the words, and translate the utterances to English.We estimate that the process to create a time-aligned speech corpus was reduced to 5 times real time for the initial 3 hours of recordings. Using the recordings and manually time-aligned annotation we could train a forced aligner to be used for a larger amount of recordings.

We record a speech corpus from prepared text to create an initial corpus with little time investment. Using a limited amount of manually time-aligned annotations, approximately 2 hours of a speech corpus, we train a Forced Aligner that allows us to facilitate the time-alignment of a larger data set with raw textual transcription. The time aligner generates automatic alignments of audio or video recordings and raw textual transcriptions. We correct this automatically generated alignment and create a larger speech corpus in the second phase. Given acoustic and language models we are able to improve the Forced Aligner and generate an initial Automatic Speech Recognition system. This approach reduces the time and effort invested in speech corpus creation for an under-resourced or endangered languages significantly. The resulting technologies can facilitate the transcription of large amounts of recordings and significantly reduce the time necessary for corpus creation. The corpora are also crucial for fast and efficient creation of language material for education, revitalization, but also documentation and research, or the development of speech and language technologies.

To create a larger volume of recordings we used the same texts as for the initial corpus created with one native speaker only and record the same utterances from multiple speakers. Since native speakers of Chatino are educated in Spanish, they are not familiar with the academically motivated transcription schema used for their native language. It would be too time-consuming to train them to read aloud the existing transcriptions. To avoid the problem related to the lack of a standard orthography and native speakers lack of familiarity with written Chatino texts, we decided to introduce a new method. In our second recording setting we use the recorded material from the initial speech corpus as acoustic stimuli and present it to the native speakers who we record. We ask them to listen to the original recording over headphones and repeat after the recording. This way we also avoid fluency fluctuations associated with read language. Repeated spoken language tends to be more fluent and natural than purely read language. In the second corpus creation phase we aim at a corpus including both male and female voices in approximately equal proportions, ultimately extended to above 10 hours. From the 10 hours or more of recordings we are able to create a fully time aligned transcription, i.e. a part-of-speech tagged, and translated speech corpus using the already existing transcriptions and annotations.

For the resulting 10+ hours of speech recordings, we use the existing transcription and the force-aligner to generate a fully time aligned corpus. As mentioned, the Forced Aligner is trained on the manually aligned initial portion of the corpus. We expect a certain dose of variation in the recordings that are created using the repetition task, though.  From previous experience we know that subjects tend to subconsciously “correct” the heared language by rendering the most unmarked word sequence for the speech variety. Since some of the texts used are of poetic, ritual nature, the repetition task might render some variation, which still needs to be corrected manually. The use of the forced aligner is not affected by this kind of variation, though.

**Outcomes.**The project has generated and will generate further data and outcomes interesting from the point of view of speech and language technology engineering but also linguistic research. Both outcomes will ultimately benefit the speaker community. On the one hand, we generate language models for a language that is typologically very different from the languages that most of the ASR resources have been developed for: e.g. various Indo-European languages like English, German, Spanish, and some semitic and Asian languages. The ASR for Chatino is likely to encounter specific technical challenges due to the specific acoustic properties of Chatino compared to most well-resourced languages. Given larger corpora we also expect that transcriptions of the recordings and the recordings themselves, can be analyzed now in a very different way at the level of speech patterns, but also lexically, phonologically, or even at the level of syntax or semantics.

The initial speech models and technologies that we generate can be tested against a broader volume of recordings, for example recordings collected in the past and stored without transcription in various digital language archives. The transcription of these resources using our bootstrapping corpora and technologies thus can increase the volume of the analyzable corpora for Chatino (or other closely related variants) and giving it, additionally, a diachronic depth. In this case, even a relatively inaccurate speech recognition system can assist a researcher with the command of the language in fast annotation of the heritage resources. More accurate language models can be utilized to subsequently develop apps and tools for Chatino , for language education and revitalization. Further, the transcriptions themselves can be used for the generation of word lists, dictionaries, grammars and other materials that can be used for teaching Chatino either as a second language or to be introduced to schools along with Spanish.

We experiment with the amount of data necessary to bootstrap ASR language models. Chatino has a number of linguistic characteristics that might present a challenge to the SR system. Apart from frequent reductions in fast speech, Chatino has no less than fourteen distinct tonal patterns and complex “tonal sandhi”. Some of these tonal patterns are complex, consisting of a tone that is realized on the host word, plus a second “floating” tone that is realized only when the host word is followed by another word. It is not clear yet from our experiments and settings, how these acoustic properties of Chatino impact existing speech recognition algorithms. We will report more results from these studies at the LREC meeting.

**Tools and technologies.** In our environment we make use of the following technologies:

1. Prosodylab Aligner (Goman et al., 2011) (https://github.com/prosodylab/Prosodylab-Aligner): to train an HTK-based Forced Aligner that requires a pronunciation dictionary and the audio transcription pairs on the utterance level.

2. ELAN and our own software ELAN2split (https://bitbucket.org/dcavar/elan2split): for the generation of a training corpus for the Prosodylab Aligner, i.e. the generation of audio and transcription pairs for individual time intervals in the ELAN transcription and time alginment.
3. Espeak and Text-to-Speech-models for Espeak (http://espeak.sourceforge.net/): to generate spoken language audio from existing transcriptions. This text-to-speech generated audio signal is used in Praat to align the transcription of a recording. Since a basic approximation of the Espeak output seems to be enough to generate qualitatively acceptable results in Praat, this is a feasible approach for language documentation and speech corpus generation.
4. Praat and the Praat-based forced aligner: that is, Praat provides a Forced Aligner that is based on Espeak and the language specific text-to-speech model.

**Collaborators and the philosophy.** The corpora, all models and the documentation of the annotation schema are hosted at GORILLA, an archive and language resource platform at The LINGUIST List and The Archive of Traditional Music, at Indiana University: http://gorilla.linguistlist.org/. All resources are freely accessible under the Creative Commons Attribution and Share Alike (CC BY-SA) license. (As long as the GORILLA platform is being developed, interested users can be shared on a Dropbox-folder to get access to the corpora; the GORILLA website will be fully accessible for downloads, commenting, and submission before the LREC 2016 conference.)

## References

Boersma, Paul & Weenink, David (2015). Praat: doing phonetics by computer [Computer program]. Version 5.4.22, retrieved 8 October 2015 from http://www.praat.org/

Cruz, Hilaria. 2014. Linguistic Poetics and Rhetoric of Eastern Chatino of San Juan Quiahije. Ph.D. dissertation. University of Texas at Austin.

ELAN: http://tla.mpi.nl/tools/tla-tools/elan/ Max Planck Institute for Psycholinguistics, The Language Archive, Nijmegen, The Netherlands.

Gorman, Kyle, Jonathan Howell and Michael Wagner. 2011. Prosodylab-Aligner: A Tool for Forced Alignment of Laboratory Speech. Canadian Acoustics. 39.3. 192–193.

Wittenburg, P., Brugman, H., Russel, A., Klassmann, A., Sloetjes, H. (2006). ELAN: a Professional Framework for Multimodality Research. In: Proceedings of LREC 2006, Fifth International Conference on Language Resources and Evaluation.