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

Małgorzata E. Ćavar, Damir Cavar, Hilaria Cruz
Indiana University, University of Kentucky
Bloomington, IN, Lexington, KY
{mcavar,dcavar}@indiana.edu, hcr224@g.uky.edu

Abstract

This project approaches the problem of language documentation and revitalization from a rather untraditional angle. To improve and facilitate language documentation of endangered languages, we attempt to use corpus linguistic methods and speech and language technologies to reduce the time needed for transcription and annotation of audio and video language recordings. The paper demonstrates this approach on the example of the endangered and seriously under-resourced variety of Eastern Chatino (CTP). We show how initial speech corpora can be created that can facilitate the development of speech and language technologies for under-resourced languages by utilizing Forced Alignment tools to time align transcriptions. Time-aligned transcriptions can be used to train speech corpora and utilize automatic speech recognition tools for the transcription and annotation of untranscribed data. Speech technologies can be used to reduce the time and effort necessary for transcription and annotation of large collections of audio and video recordings in digital language archives, addressing the transcription bottleneck problem that most language archives and many under-documented languages are confronted with. This approach can increase the availability of language resources from low-resourced and endangered languages to speech and language technology research and development.

Keywords: Chatino, Speech and Language Corpus, Speech Technology

1. Introduction

Endangered language documentation includes, among others, the creation of language resources in form of word lists, audio and video recordings, notes, or grammar fragments. Once the data is collected, various institutions offer archiving services for these language resources, however, irrespective of the availability of archives, another big issue needs to be addressed. Large collections of recordings and data sets face a problem that is commonly referred to as the Transcription Bottleneck.

The Transcription Bottleneck is a metaphor that is often used to describe the problem that large amounts of digital audio and video language recordings from language documentation projects are only available in the original audio or video format. In other words, the recordings in digital language archives are often accompanied by some meta-information that includes, for example, information about authors, speakers, origin, or source of the material but, crucially, transcriptions of the audio or video material are not available or existent for most of the archived recordings, neither linguistic annotations or even translations.

To transcribe the resources collected and recorded in the field or in interactions with native speakers it is necessary but not sufficient to just keep track of the meta information related to the source, time, location, or purpose of the recording, and the technology that was used. If they are not transcribed, annotated and translated the collected language resources are only accessible to the native speakers of the language or experts. If the recordings can be deciphered only by native speakers or experts, this in itself presents a problem for the low-resourced or endangered languages we work with as the numbers of the potential users of the resources steadily diminish.

To transcribe the audio and video data using traditional methods – i.e. manually – is, however, not a trivial task. This work is often estimated at some 50 to 100 hours of transcription and annotation by experts per 1 hour of recording, thus, it is extremely cost- and labor-intensive. The fact that only experts/native speakers can prepare transcriptions raises the costs of the transcription further, all this resulting in the Transcription Bottleneck.

Digital language archives host numerous collections of recordings from languages that otherwise have not been documented using corpora or digital annotations based on standards from disciplines like corpus linguistics or Natural Language Processing (NLP). The material has been recorded in different ways using a variation of strategies that is common in documentary linguistic work, but not ideal for corpus linguistic processes. Audio recordings – for example – have not been collected using strategies that minimize noise and maximize the quality of the speaker's voice. The fact that in many of the recordings that we looked at we find free interactions with native speakers in dialogues with overlaps and background noise, inseparable speaker voices, or variation in sound quality, make processing and analysis of the recorded material very difficult, not just for human analysts, but also for speech and language technologies. Many extremely valuable resources remain in archives without transcription or annotation.

For example, the Archive of the Indigenous Languages of Latin America (AILLA) at UT Austin,[1] or Archive of Traditional Music (ATM) at Indiana University host collections of recorded audio and video material from many extinct, endangered, or extremely under-resourced languages, and in particular of Eastern Chatino. The archived audio and video language recordings are mostly not transcribed. For

[1] See `http://www.ailla.utexas.org` for more details.

most of the material the archives do not provide translations or any kind of linguistic analysis in form of annotations or other forms of analysis.

To make the archive language resources content accessible we might have to provide:
– transcription
– linguistic annotation (lexical, morphological, syntactic)
– translation to some major language
We are not aware of any systematic study that makes clear predictions about the necessary workload per task for the transcription of audio and video recordings from language documentation projects for any given language. Intuitively the estimate of 50 to 100 hours seems appropriate, if we assume that the transcription is accompanied by additional linguistic annotation and translation.

Newman (2013) named such digital language archives *language graveyards*, using a notion introduced earlier by Lehmann (2001, 85) and Himmelmann (2006, 4) that emphasizes the exclusion of the archived material from active research or academic discourse due to a lack of analysis. The extremely valuable and interesting language material from low-resourced and endangered languages in these archives is not accessible to non-experts, and real experts would have to invest significant time and effort to study the collections. Besides, there are usually only a few experts for most endangered languages.

There are various other issues with the archived language data that range from complicated licensing and copyright restrictions, to a limited quality and value for non-documentary purposes, e.g. in other types of language related research projects that are mainly related to computational linguistics or speech and language technology.

We will focus in this paper on Eastern Chatino (CTP), but the problem affects numerous languages and data in the respective digital archives. Our approach should also be seen as a study of possibilities to reduce the transcription and annotation effort for all low-resourced languages. In addition to Chatino, we also experimented with the creation of a Burmese and Croatian speech corpus, for which we assume that speech corpora with part-of-speech annotation and translation are not freely available, neither phonotactic (acoustic) or language models, as appropriate for common speech technologies.

In this paper we present our approach to the problem of the Transcription Bottleneck by using corpus and computational linguistic methods to create speech and language technology resources that can reduce the transcription time and effort significantly. We propose a methodology to rapidly develop spoken word corpora for endangered and low-resourced languages, which can potentially facilitate the annotation of already existing audio and video recordings in the different digital language archives. This approach may have a positive impact on linguistic science and language documentation in general. It can reduce the time for transcription by at least 50%, bringing together the world of language documentation, linguistic fieldwork with speech and language technologies.

2. Eastern Chatino of San Juan Quiahije

Eastern Chatino of San Juan Quiahije (SJQ)[2] is a Zapotecan language spoken in the municipality of San Juan Quiahije (16° 18' N 97° W), Oaxaca, Mexico by some 3,000 speakers. Widespread poverty, unemployment, deficient infrastructure in terms of schools, hospitals, and lack of arable land, all lead to massive migration and further diminishing of the number of speakers. The compulsory use of Spanish in the public context, such as in health care institutions, courts, and public education further negatively impact the vitality of the Chatino language.
Members of the Chatino Language Documentation Group (CLDP) developed – and continue to work on – a practical orthography for the SJQ and other Chatino varieties from 2004 on (Cruz and Woodbury, 2014). The great majority of speakers do not write or read in their language; however, there is a small group of young Chatinos who are beginning to write Chatino on social media (e.g. on Facebook and Twitter) using the system developed by the CLDP.
The linguistic description of Chatino languages in general is limited. Previous approaches to document the language involved the collection of audio and video recording of speakers in formal and informal settings including ceremonial language, everyday conversation, and grammatical elicitation. This documentary collection is kept at the Archive of the Indigenous Languages of Latin America (AILLA) at the University of Texas at Austin[3] and the Endangered Languages Archive (ELAR) at SOAS, University of London.[4] However, only a very small number of these recordings are transcribed or annotated.
To enable researchers, students, and the broader public the full access to language recordings for research and other purposes, it is necessary that these materials are transcribed and annotated. Ideally, the transcription and annotation has to be time-aligned. Unfortunately, as already mentioned, this effort is estimated to consume 50 to 100 times real time and it is thus difficult to identify and allocate the necessary human and financial resources to transcribe and annotate manually all the already collected Chatino recordings. The same holds for the recordings of many other endangered, low-resourced or under-documented languages with thousands of hours of audio and video resources stored as dormant treasures in many digital language archives.

2.1. Linguistic challenges

Chatino is a group of three language varieties (Zenzontepec, Tataltepec, and Eastern Chatino) forming an independent genetic unit within the Zapotecan language family (Cruz and Woodbury, 2013). Zenzontepec Chatino is spoken in about twelve towns in the Municipality San Cruz Zenzontepec (Cruz, 2011). Tataltepec Chatino is spoken in one town, Tataltepec de Valdez, and Eastern Chatino is spoken in about eighteen towns (Cruz, 2011). SJQ Chatino belongs to the Eastern Chatino variety. All Chatino languages

[2] The ISO 639-3 code for Eastern Chatino is CTP. The commonly used abbreviation for San Juan Quiahije is SJQ.

[3] See `http://www.ailla.utexas.org` for more details.

[4] See `http://elar.soas.ac.uk` for more details.

present a number of potential challenges and research questions that need to be addressed when developing language models and NLP tools.

2.2. Mutual intelligibility

There is no mutual intelligibility between the three Chatino languages. Additionally, the intelligibility among the speakers in different communities of Eastern varieties is frequently difficult due to lexical and syntactic, but primarily due to phonetic tone differences (Campbell, 2011). The current assumptions are that individual villages within the Eastern Chatino grouping have different phonetic tone registers. For example, if a low level tone that a female speaker would produce in SJQ is approximately 200 Hz, the same tone by a comparable speaker in a different town may be realized as 220 Hz. The same can be observed for the person and number marking on verbs and nouns. For example, if a third singular (3SG) person and number inflection is marked by an ascending tone in SJQ, here annotated as 42 (see table 1), the same 3SG will fit into the range of tone 32 in another Eastern Chatino dialect (as noted in section 2.3. below, table 1 shows the tone registers with 1 as the label for the relatively highest tone). Consequently, phonetic tonal differences may lead to contradictory interpretation of meaning across different dialects.

2.3. Challenge: Tones

It is often assumed that SJQ belongs among languages with most likely the richest tonal systems in the world. It has four tone levels and altogether thirteen different phonemic tonal contrasts, including level and contour tones. Tones are associated with syllables and in the CLDP system are expressed using numbers:

0: floating, SH	04: SH to low	40: L to SH
1: H	14: H to L	20: M to SH
2: M H	24: M to L	10: H to SH
3: L M	42: L to H	140: H to L to SH
4: L	32: M to H	

Table 1: H is *high*, SH is *super high*, L is *low*, M is *mid tone*.

Further complication is added by the fact that SJQ Chatino has also tone sandhi effects, that is, the tones for lexical items is modified depending on the context in which they are found. For example, the words *lo4* ’ón’ and *ke4* ’rock’ in isolation have both low tones, but - when these words come together in a sentence – the tone of the noun *ke4* ’rock ’becomes an ascending tone: *lo4 ke32* ’ón the rock’. 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 to us whether the tonal properties pose a potential difficulty for speech technologies and common Automatic Speech Recognition (ASR) system at all. One purpose of this project is to study the effect of particular feature extraction algorithms for Forced Aligners and ASRs on the recognition accuracy using speech and language data of this kind.

2.4. Previous Work on Chatino

2.4.1. Linguistic description

Linguistic research in SJQ Chatino began in earnest in 2003 with work by Emiliana Cruz and later Hilaria Cruz, both native speakers of SJQ Chatino, then graduate students, and their mentor, Anthony Woodbury.

Given that Chatino is an extremely tonal language, no research could be done without a good understanding of these tone patterns (Cruz and Woodbury, 2014). Consequently, a large part of the research on SJQ Chatino has been on phonological patterns of tones. E. Cruz (Cruz, 2011) dissertation focuses on the phonology, morphology, and functional aspects of SJQ tones. Tonal sandhi has been researched in E. Cruz and Woodbury (2006).

Some initial research on morphology has been conducted, among others, Emiliana Cruz, Hilaria Cruz, Thom Smith-Stark (Smith-Stark et al., 2008) reported on SJQ complementation, Woodbury (Woodbury, 2008) has investigated tones in the inflection of person and number. There is also some preliminary work on compounding and tones (Cruz and Woodbury, 2013). Emiliana Cruz and Ryan Sullivant (Cruz and Sullivant, 2012) have reported on demonstratives. Poetic patterns in SJQ discourse, is another aspect of the language that been researched in SJQ (Cruz, 2014). Currently Lynn Hou and Kate Mesh are investigating child acquisition of sign language (Hou forthcoming). Similarly, Kate Mesh compares speech gestures among deaf and speaking citizens in SJQ.

2.5. Existing corpora

Chatino language data and collections can be found in different digital language archives, e.g. the Archive of Traditional Music (ATM) at Indiana University or at AILLA. As of March 2016, AILLA houses roughly one hundred and seven (107) hours of audio and ten hours of video recordings of SJQ Chatino in total. Some of the materials have restricted access because sensitive information about living individuals in the community is present within certain narratives. Roughly ten hours of these audio and video recordings are transcribed. The transcribed speech is primarily based on recordings of grammatical elicitation and does not include ceremonial or continuous everyday speech.

Table 2 offers a summary of resources in the Collection “Chatino Documentation of Hilaria Cruz” archived at AILLA. This collection includes conversations, field notes, commentary, ceremonial dialogue, and many more genres. Approximately 2% of the recordings in this collection include associated transcriptions.

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%

Table 2: Collection “Chatino Documentation of Hilaria Cruz” in AILLA

To make these resources more accessible and searchable for data analysis or for the creation of a text corpus, there is an urgent need to completely transcribe the already existing resources. Linguistically annotated text corpora that would allow for the creation of qualitative education material for Chatino, or any kind of qualitative and quantitative study of language use, do not seem to exist. Ongoing efforts to document, study, or revitalize and teach Chatino would – however – significantly benefit from such a corpus.

3. Procedures

As mentioned earlier, manual transcription of documentary linguistic recordings consumes an excessive amount of time. We expect that current speech processing technologies can offer a significant reduction of this effort.

On the one hand, Forced Alignment as a technology to automatically time-align text and audio signal can potentially significantly reduce the time for the creation of speech corpora by automatically providing time-alignment information. This process expects a transcription and the corresponding audio as input and it generates time boundaries for words in the transcription. Forced Aligners do not provide any transcription for speech recordings, just the time-alignment for existing transcriptions. This time-information is an essential part of a speech corpus that is used to train automatic speech recognizers (i.e. tools that provide a full word-level transcription with the appropriate time alignment for some audio recordings).

Automatic speech recognition (ASR) technologies could be trained using time-aligned speech corpora and then used for automatic transcription of hundreds of hours of recorded speech. This two-step strategy, first, creating manually a speech corpus for the training of a Forced Alignment tool and, second, using manually transcribed, and manually and automatically aligned speech data for training of ASRs, can significantly reduce the workload for the transcription of large collections of speech recordings.

There are various implementations of Forced Aligners that are based on Hidden Markov Models (HMM) and in particular the Hidden Markov Model Toolkit (HTK)[5] The Prosodylab Aligner (Gorman et al., 2011)[6], for example, requires about two hours of time-aligned speech and accompanying acoustic models. An alternative approach to generate a model for a particular language for forced alignment is to use Praat (Boersma, 2001; Boersma and Weenink, 2016) and its integrated *analysis by synthesis* approach for time-alignment. Praat makes use of a speech synthesis or *text to speech* algorithm to automatically generate an audio signal from text (i.e. from the transcription). It time-aligns the audio signal itself in a recorded audio signal to generate hypotheses about the time-alignment of some text in a corresponding audio recording. This approach does not require any speech corpus, but instead relies on a language model for *text-to-speech* using the Espeak[7] algorithm or system that Praat makes use of.

Our approach is to transcribe and time-align at least two hours of recorded speech to be able to train an HTK-based Forced Aligner, and to also develop a *text-to-speech* language model for the use in Espeak and Praat. This way we can compare two methods of forced alignment. At the same time we develop a human annotated and validated test corpus for evaluation and comparison purposes.

3.1. Initial transcriptions and recordings

To generate an initial speech corpus for Chatino, we started with the already existing texts and transcripts from field work and language documentation projects. We used ritual texts collected and transcribed by Hilaria Cruz (2014), a Chatino researcher and native speaker of the language herself. The texts are written in the CLDP transcription schema, which is an ASCII-based phonemic transcription system that uses symbols for phonemes and tones.

The CLDP transcription schema represents a detailed phonemic transcription that might turn out to be too complex to serve as an ultimate orthography, but it serves as a good starting point for a potential orthographic standard. Given the convention to use digits to mark tone, it also enables us to experiment with alternative models, e.g. one can easily remove transcriptions of tone and reduce the pronunciation dictionary in HTK-based Forced Aligners or ASRs.

To reduce the transcription workload and to generate sufficient audio material for a speech corpus that can serve as a quantitative and qualitative base for models necessary for Forced Alignment and ASR training, we recorded spoken language under near studio conditions using a native speaker who read the pre-collected transcripts and texts. The initial recordings were created using high quality audio equipment and 96 kHz with 24 Bit sampling. The recordings were stored in an uncompressed WAVE-form audio file format.

It is estimated that an initial corpus of manually time aligned transcriptions of at least two hours of speech would be sufficient to train a Forced Aligner. In the first phase we recorded approximately three hours of speech using careful reading of the prepared and pre-existing texts. This way the manual transcription and time alignment effort is minimized to copy and paste and setting of time boundaries in the speech signal using common annotation tools.

3.2. Initial time alignment and annotation

The initial time alignment has been created manually using ELAN (Wittenburg et al., 2006). The process of time alignment is time-optimized, since in this case the transcription already exists, consequently, only copy-and-paste and the time alignment using common transcription tools like ELAN is necessary. ELAN was used to initially time-align the transcription along the utterance level. Its design allows for rapid setting of time-boundaries and multi-tier annotations with tier dependencies. On the other hand, since it does not offer a fine grained picture of the spectrogram, it does not provide a good environment to place exact time-boundaries beneath the utterance level.

The transcription of Chatino contains multiple tiers. Besides a transcription tier for the CLDP transcription, we added a tier for part-of-speech tags and an additional translation tier for the utterance-level translation to English. Since the CLDP transcription is a simplified phonetic, or

[5] See `http://htk.eng.cam.ac.uk/` for more details.

[6] See `https://github.com/prosodylab/Prosodylab-Aligner` for more details.

[7] See (`http://espeak.sourceforge.net` for more details.

rather, phonemic transcription schema, we do not add any other phonemic or phonetic transcription tier.

For the transcription and time-alignment of the initial three hours of the speech corpus, we estimate that the process to create a time-aligned speech corpus was reduced to max. 5 times the duration of the recording. To train a forced aligner for a larger corpus of audio recordings and corresponding textual transcriptions requires additional data structures and corpus work.

In addition to the manually time-aligned speech corpus, a pronunciation dictionary and a language model are necessary, as well as a list of utterance transcriptions and audio sequences. The pronunciation dictionary consists of a transcription for a word and the corresponding tokenized sequence of phonemes (or phones). Given the CLDP transcription schema that is used in the speech corpus, there is no mismatch between the *orthography* and the phonemic representation. The only additional information provided in it in this specific case is the tokenization of the phonemic symbols.

sten24en	s t en 24 en
tsan4	ts an 4
keq3	k e q 3
kang4	k a ng 4

Table 3: Sample of a Chatino pronunciation dictionary

The pronunciation dictionary in table 3 contains the digits for the tones that are assumed in the cited literature to be phonemic. We generate a reduced second type of pronunciation dictionary from this one that does not contain transcriptions of tones. This way we simplify the output mapping by reducing the number of types and increasing the token frequencies.[8] These pronunciation dictionaries are also essential for training ASRs that are based on HMMs, e.g. using HTK or Sphinx.[9]

In addition to this pronunciation dictionary, we needed for the HTK-based Forced Aligners a corpus of individual audio files and their corresponding transcriptions in text-files. These utterances correspond to our initial time-alignment in ELAN. To extract the transcription tier and cut the corresponding audio sequences from the ELAN annotation files, we created the software tool ELAN2split[10] that makes use of the Sound eXchange[11] library and tools for cutting audio sequences in chunks. ELAN2split generates a corpus of audio files and text-files from one ELAN annotation file using a specific tier.

It is worth emphasizing here that in order to create useful corpora and data for speech and language technologies, we deviate here from classical or traditional methods. Our approach is different from traditional language documentation and fieldwork methods since, instead of collecting word lists and recordings of elicited speech from speakers, we create initially a speech corpus using carefully spoken speech of near studio quality. One problem with the traditional 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 cost- and time-consuming. In contrast, we record a high-quality speech corpus with minimized noise and optimized speech signal from prepared text to create an initial corpus with little time investment. This initial low-investment speech corpus allows us to train speech and language technologies for a more rapid extension of the volume of automatically annotated recordings, as well as further bootstrapping of speech and language technologies for the particular languages.

Using a limited amount of manually time-aligned annotations, approximately two hours of a speech corpus, we train a Forced Aligner that allows us to facilitate the time-alignment of a larger data set with existing raw textual transcription. The time aligner generates automatic alignments of audio or video recordings and raw textual transcriptions. The output of the Prosodylab Aligner is a Praat-compatible TextGrid file. This automatic time-alignment can be manually validated and corrected using Praat. Since ELAN can import Praat-based TextGrid files, we can perform a full cycle of retraining of the HTK-based Forced Aligner by regenerating the training corpus with the improved and corrected alignment over a larger corpus.

Given acoustic and language models we are able to improve the performance of the Forced Aligner and generate an initial Automatic Speech Recognition system.

4. Tools and technologies

In our experiments and in the development environment we make use of the following speech and language technologies:

- Prosodylab Aligner (Gorman et al., 2011),[12]
- ELAN (Wittenburg et al., 2006) for initial multi-tier annotation and time-alignment,
- ELAN2split[13] for corpus creation for HTK-based Forced Aligners using ELAN Annotation Files,
- Espeak and Text-to-Speech-models[14] for the Praat-based forced alignment feature,
- Praat (Boersma, 2001; Boersma and Weenink, 2016) for detailed utterance and word-level time-alignment.

[8] There is a certain amount of tokens that differs only with respect to tones and their distribution. By eliminating the tone annotation, we create more homographs and simplify the acoustic model.

[9] For Sphinx see `http://cmusphinx.sourceforge.net/` for more details.

[10] ELAN2split is a freely available open-source tool that is implemented in C++11 and known to compile on common operating systems. See `https://bitbucket.org/dcavar/elan2split` for more details.

[11] See `http://sox.sourceforge.net/` for more details.

[12] See `https://github.com/prosodylab/Prosodylab-Aligner` for more details.

[13] See `https://bitbucket.org/dcavar/elan2split` for more details.

[14] See (`http://espeak.sourceforge.net` for more details.

We use ELAN (Wittenburg et al., 2006) for the multi-tier transcription and annotation of speech recordings. Our own software ELAN2split (`https://bitbucket.org/dcavar/elan2split`) extracts tuples of time-aligned transcription and audio sequences from ELAN Annotation Files and corresponding WAVE-media files. This ELAN2split software generates a training corpus for e.g. the Prosodylab Aligner from existing ELAN transcriptions. This corpus consists of audio and text-transcription pairs for individual time intervals in the ELAN transcription and time alignment.

The difference between Praat-based speech corpus annotation and ELAN-based time-alignment is that Praat allows for much more detailed alignment based on speech signal spectogram views. ELAN does not allow for such a fine-grained annotation. In addition to that, Praat provides a Forced Alignment functionality that is based on Espeak, an open-source text-to-speech engine. Given a speech signal and a corresponding transcription text, Praat generates a speech signal using the text-to-speech engine and, then, hypotheses about the alignment of the generated speech signal and the recorded one, i.e. it utilizes a method that could be described as analysis by synthesis. This method requires the development of models that are mappings from orthographic to phonemic or sound representations that are specific to Espeak.

Espeak and *text-to-speech* models as utilized in Praat not only enable the alignment functionality of Praat for a specific language, but they are also a very interesting documentation tool. The models contain mappings of orthographic representations to the Espeak internal pronunciation symbols and additional exception lexicons. The notation allows for the markup of stress or tone and regularities of stress distribution. A documentation of the acoustic properties of lexical items using this system extends the classical phonetic transcription of lexical lemmata with potential exception lists.

The Prosodylab Aligner is a Hidden-Markov-Model-Toolkit-(HTK)-based Forced Aligner that requires a pronunciation dictionary and the audio transcription pairs on the utterance level to train a model.[15] It generates TextGrid files that are the underlying Praat annotation and time-alignment information file format. These TextGrid files can be opened and edited in Praat.

5. Outcomes

The project has generated and will generate further data and outcomes that should improve the documentation process of endangered and low-resourced languages. The results are also interesting from the point of view of speech and language technology engineering. First, we generate language resources and 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 and modeling challenges due to the specific acoustic and linguistic properties of Chatino compared to most well-resourced languages.

Second, the initial speech and language 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. As mentioned, the ATM at Indiana University contains some Chatino resources, but most of the recordings can be found in AILLA at the University of Texas at Austin. The transcription of these resources using our bootstrapping corpora and technologies can increase the volume of the analyzable corpora for Chatino (or other closely related variants) and give 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 could be used for teaching Chatino either as a second language or to be introduced to schools along with Spanish.

Further, 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 ASR system, including frequent reductions in fast speech, the multitude of distinct tonal patterns and complex “tonal sandhi”. 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.

Given larger volume of the corpora we also expect that transcriptions of the recordings and the recordings themselves, can be analyzed now in a very different way, namely, from the point of linguistic patterns, lexically, phonologically, or even at the level of syntax or semantics. Creating larger volume of Chatino transcriptions will open up new research possibility in the field of history, anthropology or sociology. All outcomes will ultimately benefit the speaker community.

5.1. Rapid expansion of the training corpus

To create a larger volume of recordings we intend to use the same texts as for the initial corpus created with one native speaker only, and record these utterances as spoken by multiple speakers. Since native speakers of Chatino are educated in Spanish, they are not familiar with the academically-motivated transcription schema developed for their native language by CLDP. 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 that is more simplified than an academically motivated phonemic transcription schema, and to circumvent the fact that native speakers are not familiar with any written Chatino texts, we decided to introduce a new method for corpus elicitation. In our second setting of recording sessions we use the recorded material from the initial speech corpus as acoustic stimuli and present it to the native speak-

[15] See for the Hidden Markov Model Toolkit (HTK) `http://htk.eng.cam.ac.uk` and Young et al. (2006).

ers whom we record. Native speakers are asked to listen to the original recording over headphones and repeat it. This way we also avoid fluency fluctuations associated with read language. Repeated spoken language tends to be more fluent and natural than 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 ten hours of spoken language. From the ten hours or more of speaker 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 ten or more hours of speech recordings, we will use the existing transcriptions and the two different automatic forced-aligning methods to generate a fully time-aligned corpus. As described, the HTK-based 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 spoken-language repetition task. We do know from previous experiments on Croatian that subjects tend to subconsciously “auto-correct” the acoustically perceived language by rendering the most unmarked word sequence for their personal speech variety. Similarly, since some of the texts that are used in the Chatino corpus are of poetic or ritual nature, the repetition task might render some significant variation. For this kind of variation, the transcriptions will need to be corrected manually for each file. The performance of the forced aligner is not assumed to be affected by this kind of variation, though.

6. 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/` (Cavar et al., 2016), this volume. All resources are freely accessible under the Creative Commons Attribution and Share Alike (CC BY-SA) license. The GORILLA platform is currently being developed and interested users can be shared on a Dropbox-folder[16] or using a Bitbucket[17] repository to get access to the corpora, as long as the GORILLA website is not fully accessible for downloads, commenting, and submission.

The software that we developed for data processing and corpus creation for the Forced Aligners is linked on the GORILLA page and made available under the Apache 2.0 license.[18]

7. Conclusions

The described process documents the initial phases of the project on Chatino. It also describes to some extend the processes that we tested for other low-resourced languages where we ran similar tests and experiments with speech corpus development, training of Forced Aligners and basic ASRs.

The main goal of this project is to identify new methods and specify new processes that can help us to significantly reduce the time necessary for transcription and annotation of audio and video recordings from endangered and low-resourced languages, i.e. we look for ways to widen the Transcription Bottleneck. Ultimately we intend to generate speech and language corpora and technologies for Eastern Chatino to facilitate quantitative and qualitative studies and the development of resources for education.

The initial rationale to experiment with the issues related to corpus creation for endangered and low-resourced languages was also that these speech and language resources are not accessible for speech and language technology related research.

Our goal is to define processes and to develop tools and an environment for low time and effort solutions which can maximize the resource generation, potentially rapidly expanding on the volume of transcriptions of texts for many endangered and low-resourced languages, and in particular for SJQ Chatino. The project has resulted – among others – in the development of strategies for the rapid expansion of the transcribed speech corpora for languages without writing tradition.

Apart from Chatino, we have tested the methods on other languages, in particular, we created a Burmese and a Croatian speech corpus in a similar way. The difference between Chatino, on the one hand, and Croatian and Burmese, on the other hand, is that there are large amounts of texts available for the latter, but no qualitative speech corpus that is published under a free CC BY-SA license. In order to generate rapidly a time-aligned speech corpus for these languages we used freely available Wikipedia texts in these languages. Native speakers have read this text and time-aligned them in ELAN. Using the same technologies and cycles described for Chatino, not only have we developed the resources to train a basic ASR system, but we also provide various models, as for example *text to speech* models for Espeak for Burmese, Finite State Transducer based tokenizers and morphological analyzers.

This approach reduces the time and effort invested in speech corpus creation for under-resourced or endangered languages significantly. The resulting technologies can facilitate the transcription of larger amounts of already existing recordings. The corpora are also crucial for fast and efficient creation of language material for education, revitalization, but also for the documentation and research, or the development of speech and language technologies.

8. Acknowledgments

This project emerged as a spin-off from the NSF-funded AARDVARC project (NSF number 1244713). We are grateful to a large group of colleagues who participated in workshops and conferences organized by the AARDVARC team. We are most grateful for ideas, suggestions, help, and comments from Helen Aristar-Dry, Anthony Aristar, and Douglas H. Whalen, the main team members and organizers of AARDVARC. This particular project on Chatino

[16] See `https://www.dropbox.com/` for more details.

[17] This is a git-based repository. For more details see `https://bitbucket.org`.

[18] The Apache 2.0 license can be found here `http://www.apache.org/licenses/LICENSE-2.0`.

– however – was not funded by any funding agency or with third party funds.

We are grateful to the interns and visiting students at LINGUIST List during the summer of 2015 for their help with the corpus creation and annotation process using ELAN, the Prosodylab Aligner, and Praat and Espeak based *text to speech* model creation and experimenting.

Many graduate students from the Department of Linguistics at Indiana University have participated and contributed to the resources. In particular, we are grateful to Aaron Albin, Zac Branson, Levi King, and Alec Wolyniec for their help and support in the early phases. Michael Abramov helped us transcribe and time-align Chatino in ELAN. All editors and graduate assistants of LINGUIST List – Sara Couture, Andrew Lamont, Lwin Moe, Anna White, Ashley Parker, Amanda Foster – were helpful with their suggestions, comments and other contributions.

9. Bibliographical References

Boersma, P. and Weenink, D. (2016). Praat: doing phonetics by computer [computer program]. Version 6.0.14, retrieved 11 February 2016.

Boersma, P. (2001). Praat, a system for doing phonetics by computer. *Glot International*, 5(9/10):341–345.

Campbell, E. W. (2011). Del proto-zapotecano al proto-chatino. El Quinto Congreso de Idiomas Indígenas de Latinoaméica. 7 octubre, 2011. Volante.

Cavar, D., Cavar, M., and Moe, L. (2016). Global Open Resources and Information for Language and Linguistic Analysis (GORILLA). In *Proceedings of LREC 2016*. ELRA.

Cruz, E. and Sullivant, R. (2012). Demonstrativos próximos y distales: un estudio comparativo de dos lenguas chatinas. In *Paper presented at the V Coloquio sobre Lenguas Otomangues y Vecinas in Oaxaca, Mexico*.

Cruz, E. and Woodbury, A. C. (2006). El sandhi de los tonos en el chatino de quiahije. In Las memorias del Congreso de Idiomas Indígenas de Latinoamérica-II. Archive of the Indigenous Languages of Latin America.

Cruz, E. and Woodbury, A. C. (2013). Tonal complexity in san juan quiahije eastern chatino compound verb inflection. SSILA Special Session: Inflectional Classes in the Languages of the Americas. Boston, MA, January 6, 2013.

Cruz, E. and Woodbury, A. C. (2014). Collaboration in the context of teaching, scholarship, and language revitalization: Experience from the chatino language documentation project. *Language Documentation and conservation*, 8:262–286. Special issue: Community Collaboration in the Americas.

Cruz, E. (2011). *Phonology, tone, and the functions of tone in San Juan Quiahije Chatino*. Ph.D. thesis, University of Texas at Austin. Doctoral dissertation.

Cruz, H. (2014). *Linguistic Poetics and Rhetoric of Eastern Chatino of San Juan Quiahije*. Ph.D. thesis, University of Texas at Austin. Ph.D. dissertation.

Gorman, K., Howell, J., and Wagner, M. (2011). Prosodylab-aligner: A tool for forced alignment of laboratory speech. *Canadian Acoustics*, 39(3):192–193.

Himmelmann, N. P. (2006). Language documentation: What is it and what is it good for? In J. Gippert, et al., editors, *Essentials of Language Documentation*, pages 1–30. Mouton de Gruyter, Berlin.

Lehmann, C. (2001). Language documentation: A program. In Walter Bisang, editor, *Aspects of Typology and Universals*, pages 83–97. Academie Verlag, Berlin.

Newman, P. (2013). The law of unintended consequences: How the endangered languages movement undermines field linguistics as a scientific enterprise. Paper presented at the Linguistics Departmental Seminar Series, SOAS, University of London.

Smith-Stark, T., Cruz, H., and Cruz, E. (2008). Complementación en el chatino de san juan quiahije. Proceedings of the Conference on Indigenous Languages of Latin America-III. Organized by the Center for Indigenous Languages of Latin America (CILLA), Teresa Lozano Long Institute of Latin American Studies at the University of Texas at Austin.

Wittenburg, P., Brugman, H., Russel, A., Klassmann, A., and Sloetjes, H. (2006). Elan: a professional framework for multimodality research. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC 2006)*, pages 1556–1559.

Woodbury, A. C. (2008). Guide to data usage in toolbox sjq project1.0. Manuscript. University of Texas at Austin.

Young, S. J., Evermann, G., Gales, M. J. F., Hain, T., Kershaw, D., Moore, G., Odell, J., Ollason, D., Povey, D., Valtchev, V., and Woodland, P. C. (2006). *The HTK Book, version 3.4*. Cambridge University Engineering Department, Cambridge, UK.