Endangered Languages Repositories Report

Kristen Howell

August 26, 2016

1 Introduction

As part of an important effort to document endangered languages before they become extinct, a variety of endangered language repositories have emerged to provided shared locations for field linguists to store data. These repositories vary greatly in the way their collections are organized and in the metadata they collect from depositors. The availability of collections of low resource language data has the potential for some interesting tasks in Natural Language and Speech Processing. However, the variety of formats in these repositories makes it difficult to know how much of the available data is suitable for such a task.

I investigated a number endangered language repositories including the Alaskan Native Languages Archive (ANLA)[1], the Pacific and Regional Archive for Digital Sources (PARADISEC)[2], the Endangered Languages Archive (ELAR)[3], the Archive of Indigenous Languages in Latin America (AILLA)[4] and Kaipuleohone[5]. Over the course of this investigation, I looked through roughly 300 collections from these repositories, creating a record of possible candidates for an automatic language processing task. For possible candidates I recorded the repository, language, contributor, type of information in the collection, type of data, recording conditions and quality, type of annotation, glossing and quantity of data. I also noted whether the collection contained a phonological inventory or any speaker and language metadata.

2 Summary of Findings

While the collections surveyed would be useful for linguistic research in the particular phenomena they represent, many are not well suited to automatic language processing tasks. A number of collections restrict access or are very

[1] https://www.uaf.edu/anla/

[2] http://www.paradisec.org.au/

[3] http://elar.soas.ac.uk/

[4] http://www.ailla.utexas.org/site/welcome.html

[5] https://scholarspace.manoa.hawaii.edu/handle/10125/4250

small. Of those that are accessible and sufficiently large, many collections include scans of handwritten field notes, audio recordings without transcription, audio recordings with excessive background noise or transcribed data without any gloss. Nevertheless, a few strong candidates for speech processing and Natural Language Processing tasks did emerge from the repositories surveyed.

Although I looked into a number of archives, most of the strong candidates that I have identified for automatic language processing tasks come from the ELAR archive. This is likely to be the case, not only because of the presence of collections with clear recordings, time-aligned transcription and Interlinear Glossed Text (IGT), but also because the database provides a great deal of metadata and filters, making it much easier to identify collections with a particular data format.

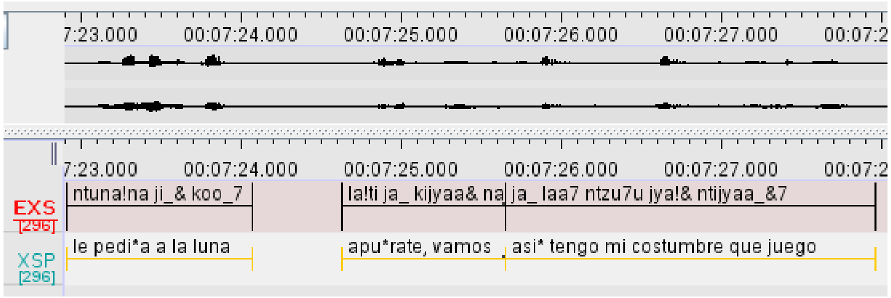
By sifting through the repositories and seeking recommendations from repository managers, I identified about 25 possible candidates for either a speech processing or text processing task. Some of these candidates were stronger than others and the following is a summary of the 12 most suitable collections and their merits. These collections each include a deposit summary and list the group represented, language information, location and the name of the depositor. Only one unrestricted collection that I found includes speaker data (see Glavda below). However, some collections are only partially restricted and it is possible that speaker metadata is among the restricted files and the appropriate permissions may be acquired for a specific research project. All of the collections summarized below include language metadata such as dialect and region. No collections among those I explored explicitly give the language's phonology.

2.1 Collections of Audio with Transcription

The first set of data includes audio files with time-aligned transcription in .eaf or .trs format. These collections can all be found in the ELAR database with the exception of Palu'e, which can be found in the Kaipuleohone repository.

Zenzontepec Chatino (Oto-Manguean). This collection, contributed to ELAR by Eric Campbell, contains audio in .wav format with time-aligned transcription in .eaf format. Transcription is time-aligned by breath group with a Spanish translation, as shown in Figure 1. The amount of data in this collection is on the order of tens of hours and contains some high quality recordings, although some of the recordings have background noise.

Figure 1[6]

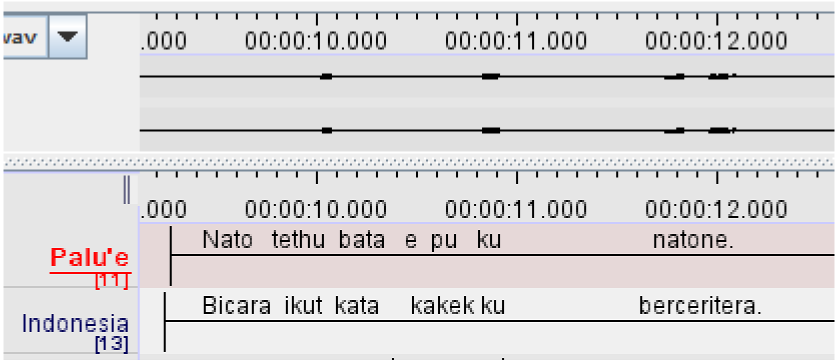


ntuna!na ji_& koo_7 | la!ti ja_ kijyaa& na | ja_ laa7 ntzu7u jya!& ntijyaa_&7

le pedi*a a la luna | apu*rate, vamos | asi* tengo mi costumbre que juego

Palu'e (Austronesian). This collection, contributed to Kaipuleohone by Stefan Danerek, contains recordings of narratives or speakers describing topics in .wav format. Time-aligned transcription is included for each recording in .eaf format. The transcription appears is time-aligned by the breath group with an Indonesian translation as shown in Figure 2. The recordings are good quality and data in the collection is on the order of tens of hours.

Figure 2[7]



Nato tethu bata e pu ku natone.

Bicara ikut kata kakek ku berceritera.

Sakun (Afro-Asiatic). This collection, contributed to ELAR by Michael Thomas, includes recordings of a speaker discussing various topics in .wav with time-aligned transcription in .eaf. Time -alignment varies between the sentence level and breath group. Some recordings are quite good, while others have a lot of background noise. The quantity is on the order of tens of hours. There is probably quite a bit of usable data in this collection, but the quantity of recordings with background noise will require some sifting to find the usable data.

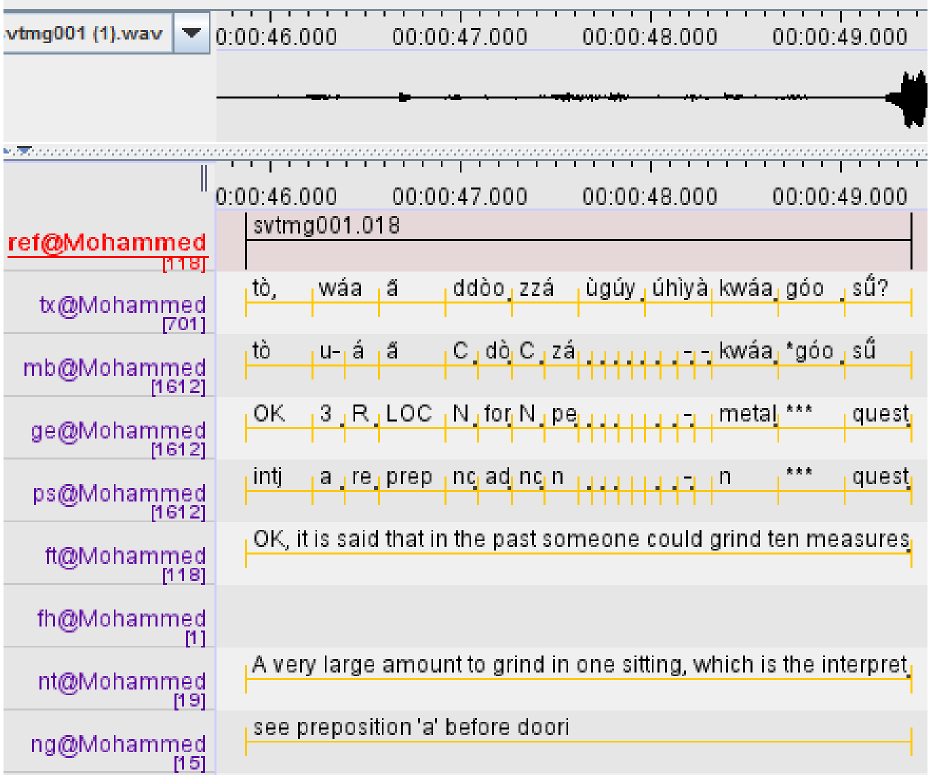
[6] The full audio and transcription files (tcr049.wav and tcr049.eaf) can be found here http://elar.soas.ac.uk/deposit/0185 under the title "Cotita".

[7] The full audio and transcription files (SD1-098.wav and SD1-098.eaf) can be found here http://hdl.handle.net/10125/39001.

Goemai (Afro-Asiatic). This collection, contributed to ELAR by Birgit Hellwig, includes recordings of a speaker discussing various topics in .wav format. Transcription is time-aligned on the sentence level with English translations in .eaf format. There is also IGT transcription in .txt format (discussed in the next section). Data quantity is on the order of hours, but recordings are good quality.

Cicipu (Niger-Congo). This collection, contributed to ELAR by Stuart McGill, includes recordings in .wav of a speaker talking about history and related topics as well as some interviews. Time-aligned transcription is in .trs and .eaf format. Transcription is time-aligned on the word level as well as the morpheme level. Morpheme gloss, English translation and additional comments are included as tiers in the transcription as well, as shown in Figure 3. Data quantity is on the order of hours, but recordings are good quality.

Figure 3[8]



ref@Mohammed [118] svtmg001.018

tx@Mohammed [701] tò, wáa ã̋ ddòo zzá ùgúy úhìyà kwáa góo sű?

mb@Mohammed [1612] tò u- á ã̋ C dò C zá - - kwáa *góo sű

ge@Mohammed [1612] OK 3 R LOC N for N pe - metal *** quest

ps@Mohammed [1612] intj a re prep nc ad nc n - n *** quest

ft@Mohammed [118] OK, it is said that in the past someone could grind ten measures

fh@Mohammed [1]

nt@Mohammed [19] A very large amount to grind in one sitting, which is the interpret

ng@Mohammed [15] see preposition 'a' before doori

Uncunwee (Nilo-Saharan). This collection, contributed to ELAR by Robert Williams, includes a mixture of multi-speaker discussions and single speaker narratives. Recordings are in .wav format and time-aligned transcrip-

[8] The full audio and transcription files (svtmg001.wav and svtmg001.eaf) can be found here http://elar.soas.ac.uk/deposit/0052 under the title ”Interview about the olden days”.

tions are in .eaf. Transcriptions are time-aligned on the sentence level and include English translations. The collection also contains some transcribed proverbs with IGT in .doc, the there is very little data in this format. Audio data quantity is on the order of hours, but recordings are good quality.

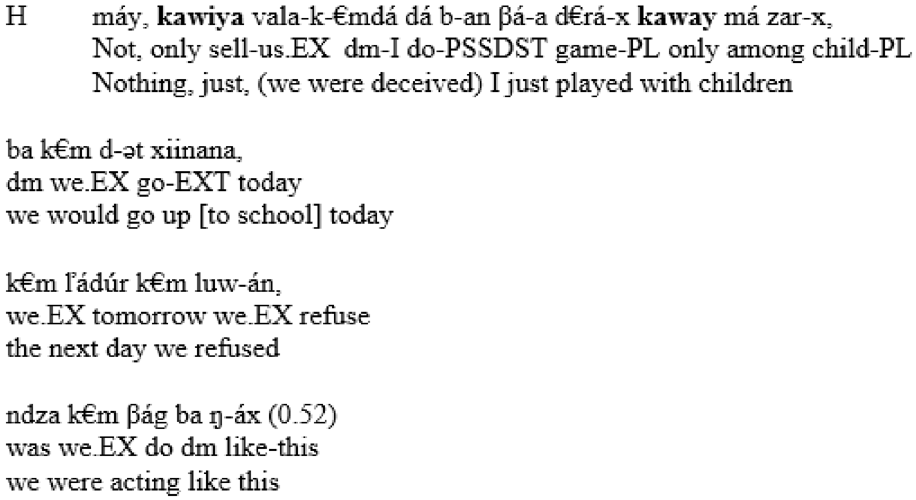
Mavea (Austronesian). This collection, contributed by to ELAR Valérie Guérin, includes recorded narratives in .wav with time-aligned transcriptions on the sentence level in .trs. Some of the transcribed audio files have English translations, but not all. Data quantity is on the order of hours, but the recordings are good quality.

Ikaan (Niger-Congo) Salffner [2010]. This collection, contributed to ELAR by Sophie Salffner contains dialogues and elicitations in .wav and sentence level time-aligned transcriptions with English translations in .eaf. The quality of recordings varies, but the quantity is on the order of tens of hours. This collection likely has a lot of usable data, but identifying that data will take some work, because many recordings are very quiet or have a lot of background noise.

Great Andamanese (Andamanese). This collection, contributed to ELAR by Anvita Abbi, contains 1,200 recordings of individual sentences in .wav with transcriptions in .doc and .pdf. These transcription include English and Hindi translations and Hindi IGT. The quality of the recordings is very good. The collection also contains a couple of narratives with IGT transcriptions.

Glavda (Afro-Asiatic). This collection, contributed to ELAR by Jonathan Owens, contains audio data in .wav format in which a speaker talks about daily life. Some recordings are monologues and some are interviews. The data is transcribed in .pdf format and includes some IGT and English translations, as shown in Figure 4. The recordings are good quality and the quantity is on the order of hours. Some metadata is also included for the speakers, including education, sex, other languages they speak and some other information.

Figure 4[9]



H máy, **kawiya** vala-k-€mdá dá b-an βá-a d€rá-x **kaway** má zar-x,
Not, only sell-us.EX dm-I do-PSSDST game-PL only among child-PL
Nothing, just, (we were deceived) I just played with children

ba k€m d-ət xiinana,
dm we.EX go-EXT today
we would go up [to school] today

k€m ľádúr k€m luw-án,
we.EX tomorrow we.EX refuse
the next day we refused

ndza k€m βág ba ŋ-áx (0.52)
was we.EX do dm like-this
we were acting like this

Arapaho (Algic). This collection, contributed to ELAR by Lisa Conathan, contains recordings of narratives in .wav format with transcription and English translation ins .txt. The quality of recordings is good and quantity is on the order of hours.

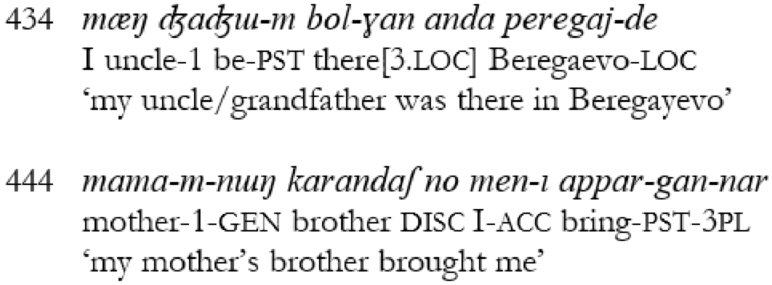
2.2 Collections of Written Data with IGT

There are a few text collections with IGT glossing. The format of this data is somewhat undesirable (.doc and .pdf) and would require some curating. These collections can all be found in the ELAR database.

Middle Chulym (Turkic). This collection, contributed to ELAR by David Harrison, contains .doc and .pdf files containing sentences with IGT, as shown in Figure 5, on the order of tens of sentences.

[9] The full audio and transcription files (12H.wav and 12H.pdf) can be found here http://elar.soas.ac.uk/deposit/0109 under the title "Hawa John on Glavda daily life".

Figure 5[10]



434 *mæŋ ʤaʤɯ-m bol-ɣan anda peregaj-de*
I uncle-1 be-PST there[3.LOC] Beregaevo-LOC
‘my uncle/grandfather was there in Beregayevo’

444 *mama-m-nɯŋ karandaʃ no men-ı appar-gan-nar*
mother-1-GEN brother DISC I-ACC bring-PST-3PL
‘my mother’s brother brought me’

Great Andamanese (Andamanese). This collection, contributed to ELAR by Anvita Abbi, was also mentioned in the Transcribed Audio section and contains recordings in .wav format. In addition 800 sentences are transcribed with English and Hindi translations and Hindi IGT in .txt and .doc format.

Goemai (Afro-Asiatic). This collection, contributed to ELAR by Birgit Hellwig, was also described in the previous section. It contains hours of transcribed recordings with IGT in .eaf and .txt format.

3 Conclusion

This survey explored existing endangered languages repositories to identify collections that could provide data suitable for shared tasks in Natural Language Processing and speech processing. A total of 12 collections from a variety of language families (including Afro-Asiatic, Algic, Andamanese, Austronesian, Niger-Congo, Nilo-Saharan, Oto-Manguean and Turkic) were identified for use in such tasks- 11 for speech processing and 3 for NLP, as two of the collections include useful data for both types of shared tasks. These collections provide transcribed audio on the order of tens of hours and IGT on the order of hundreds of sentences.

4 Acknowledgments

This material is based upon work supported by the National Science Foundation under Grant No. "EL-STEC: Shared Task Evaluation Campaigns for Endangered Languages" (NSF #: 1500157).

Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

[10] The full document with IGT (`Chulym_Bull_attack`.doc) can be found here http://elar.soas.ac.uk/deposit/0022 under the title "Chulym Bull Attack".

References

Sophie Salffner. *Ikaan and related dialects of Ukaan: an archive of language and cultural material from the Akaan people of Ikakumo (Ondo State, Nigeria)*. SOAS, Endangered Languages Archive, 2010. URL `http://elar.soas.ac.uk/deposit/0034`.