Proceedings of the Workshop on Processing Language Variation: Digital Armenian (DigitAm 2022) @LREC2022, pages 38–42
Marseille, 20 June 2022

Towards a Unified ASR System for the Armenian Standards

Samuel Chakmakjian[1,2], Ilaine Wang[2]
[1]SeDyL (CNRS-Inalco), [2]ERTIM (Inalco)
7 rue Guy Môquet 94800 Villejuif, 2 rue de Lille 75007 Paris
{samuel.chakmakjian, ilaine.wang}@inalco.fr

Abstract

Armenian is a traditionally under-resourced language, which has seen a recent uptick in interest in the development of its tools and presence in the digital domain. Some of this recent interest has centred around the development of Automatic Speech Recognition (ASR) technologies. However, the language boasts two standard variants which diverge on multiple typological and structural levels. In this work, we examine some of the available bodies of data for ASR construction, present the challenges in the processing of these data and propose a methodology going forward.

Keywords: speech corpus, ASR, forced alignment

1. The Problem

Armenian is a traditionally under-resourced language, which has seen a recent uptick in interest in the development of its tools and presence in the digital domain. Some of this recent interest has centred around the development of Automatic Speech Recognition (ASR) technologies. However, the language boasts two standard variants which diverge on multiple typological and structural levels.

1.1. A Tale of Two Phonologies

This structural divide is the most salient at a phonetic-phonological level, with Standard Eastern Armenian's (SEA) phonemic inventory containing 36 phonemes (30 consonants, 6 vowels), and Standard Western Armenian's (SWA) inventory being comprised of 30 phonemes (24 consonants, 6 vowels).

The vocalic systems of SEA and SWA are largely the similar, with the five cardinal vowels /i, e, a, o, u/ and a mid-central vowel /ə/. The consonant systems share the same nasals, fricatives, and approximants (/m, n, f, v, s, z, ʃ, ʒ, χ, ʁ, h, j, l/). SEA distinguishes between two rhotics, a tap /r/ and a trill /r/, whereas SWA does not make such a distinction. The most problematic feature of the divergence in phonologies however, is that of the plosive and affricate series in SWA and SEA. SEA's plosive and affricate phonemes have a three-way voicing distinction: voiced, voiceless, and voiceless aspirated. Modern SWA has a two-way voicing system of voiced and voiceless aspirated. The plosive and affricates phonemes of SEA are therefore the following: /b, p, p^{h}, d, t, t^{h}, g, k, k^{h}, dz, ts, ts^{h}, dʒ, tʃ, tʃ^{h}/, and the plosive and affricate phonemes of SWA are as follows: /b, p^{h}, d, t^{h}, g, k^{h}, dz, ts^{h}, dʒ, tʃ^{h}/.

Table 1 provides an example of the diverging phonetic realisations of three similar items.

Item	SEA	SWA	Translation
⟨ բառ ⟩	[**b**ɑr]	[**p**$^{\mathbf{h}}$ɑɾ]	'word'
⟨ պար ⟩	[**p**ɑɾ]	[**b**ɑɾ]	'dance'
⟨ փառ ⟩	[**p**$^{\mathbf{h}}$ɑr]	[**p**$^{\mathbf{h}}$ɑɾ]	'placenta'

Table 1: Three words and their pronunciations in SEA *vs.* SWA

1.2. Towards a Multivariant Culture

Despite this divergence in phonemic inventories, many factors render a unified system preferable. The two variants share a writing system and base lexicon, and while the two variants may be clearly distinct from one another, their speech communities are not. Amongst proficient speakers, there is a high level of mutual intelligibility. Furthurmore, the social realities of increased contact between speakers of SEA (traditionally found in the Republic of Armenia, Iran and countries of the post-Soviet zone) and speakers of SWA (traditionally found in post-Ottoman diasporan communities founded in the Middle East, Europe and the Americas) manifest in multivariant households, and sometimes multivariant speakers.

An increasing presence of SEA speakers in traditionally SWA-speaking diasporan communities, and an increasing presence of SWA in the Republic of Armenia (a traditionally SEA-speaking zone) pose more of a technical problem than a social one. While speakers frequently overcome these barriers, it would be very challenging for a single-variant ASR system to generate automatic subtitles for a video of a SEA-speaking journalist and a SWA-speaking interviewee, or a discussion between a SWA-speaking educator and a SEA-speaking student. If single-variant ASR were employed for the purposes of home-assistant technologies, a device would risk understanding one spouse in a multivariant household, and not the other.

Armenian's orthography (in both variants) is largely phonemic (Vaux, 1998), and maintains a representation of three graphemes for each of the plosive/affricate voicing sequences, making rule-based speech synthesis of either pronunciation feasible from the same text. However, producing text from speech input poses a challenge when some acoustically identical inputs correspond to the same grapheme, while other sets of identical input are to be recognised as different graphemes.

Armenian can be described as a pluricentric language (Cowe, 1992; Muhr, 2016). We can draw inspiration from attempts that have been made to construct ASR systems for other pluricentric languages. Many attempts rely at their core on a Grapheme to Phoneme approach (G2P) (Bisani and Ney, 2008). For example for Spanish, Caballero et al. (2009) define a "...multidialectal phone set [which] leads to a full dialect-independent recognizer." Another approach builds off of the process of discriminating between similar languages (DSL) (Zampieri et al., 2017) in creating a mechanism to determine which variant of a multivariant language is being spoken, such as the case of Arabic (Ali, 2018). Attempts at solving this issue for Armenian will rely upon a combination of these two approaches, due to the complication of Armenian's phone sets including an inversion and a merger.

Recent literature acknowledges a slight performance gap, with end-to-end (E2E) ASR systems slightly under-performing when compared to hybrid ASR models[1], but also, that recent innovations are closing that gap (Perero-Codosero et al., 2022). We will present our preliminary study of the main phonemic considerations which are a challenge for an ASR system to address the SEA:SWA variation issue. Our work to construct an ASR model for Armenian is conducted in the framework of the DALiH project, within which we expect to take advantage of the two major transcribed audio corpora, described in Section 3. Those will be used to implement E2E and hybrid models which, in turn, will be used in comparative/contrastive studies to have a more informed view of how SEA:SWA variations can be efficiently taken into account by a unified ASR system.

2. The State of Armenian ASR

The budding presence of ASR technologies for Armenian is underway, however there often exist many roadblocks in terms of access of information, material and data for the scientific and research communities. We can group the attempts to approach Armenian ASR into two categories: (1) multilingual approaches which include Armenian, and (2) Armenian-specific approaches.

2.1. Multilingual Models

In the case of (1) one can site companies who create models adapted to multiple languages. For example, Happy Scribe[2], a company based in Barcelona, Spain, proposes an automatic transcription and automatic subtitling service for 63 languages, including Armenian. Another such example is VocalMatic[3], based in Toronto, Canada. Similarly to Happy Scribe, VocalMatic boasts speech-to-text models for more than 100 languages (including Armenian). Lastly, amongst the three corporations often credited with bringing ASR technology into private homes via personal assistants (Google, Amazon, and Apple), only Google has a voice recognition option for Armenian at present[4]. In none of the aforementioned instances is the variant of Armenian specified, but when this is the case, the underlying assumption is that "Armenian" refers only to SEA. Otherwise, the variant or dialect would be specified[5].

2.2. Armenian-specific Models

In regards to case (2), Armenian-specific approaches date back at least to 2016, such as the system of Vardanyan (2016), an ASR system constructed based on tools from the open-source CMUSphynx project[6]. Another important Armenian-specific project is that of the National Center of Communication and Artificial Intelligence Technologies (NCCAIT[7]), which builds its corpus progressively through audio submissions provided by volunteers who read pre-selected texts. These two projects work on SEA primarily, but recently, the NCCAIT introduced a new analogous, but seemingly separate project[8] which operates in a similar manner for SWA.

Both the multilingual approaches and Armenian-specific approaches are promising in that they show evidence of the advancement of the technology, however the multilingual approaches are all explicitly private, and it remains unclear whether the NCCAIT resources will ultimately be open-source. The broader scientific community therefore lacks access to their information,

[1] Especially in langauges other than English.

[2] https://www.happyscribe.com/transcribe-armenian

[3] https://vocalmatic.com/languages/transcribe-armenian-armenian-to-text

[4] Google Translation has speech-to-text capacities for Armenian, indicated by the microphone button in the input box https://translate.google.com/?hl=fr&sl=hy&tl=en&op=translate

[5] For example, Vardanyan (2016) wrote an entire master's thesis on the creation of an "Armenian" ASR system, in which the variant is never specified, all of the data and analyses pertain exclusively to SEA

[6] https://cmusphinx.github.io

[7] http://3.144.127.191/mt/#

[8] https://aws.ican24.net/hywrec/index.php

training corpora, and above all, the methodologies behind the creation of their systems. Furthermore, none of the programmes mentioned above have the explicit objective of functioning on a bi-variant basis; they either ignore this complication (by referring only to "Armenian", understood to mean SEA) or in the case of NCCAIT, they isolate the variants from each other in constructing separate models.

3. Resources

While Armenian has traditionally been considered an under-resourced language when compared to languages of wider-spread speakerships, the language benefits from a developed literary history and extensive textual corpora. In recent years, significant advances have been made in the digitisation of Armenian texts, and the compilation of oral corpora as well. Any further research into the development and refining of Armenian ASR technologies will depend on bare audio data for processing, as well as transcribed and aligned audio data for verification and training. Our research within the DALiH framework will benefit from two major available oral corpora, one of each of the standard variants.

3.1. Available Speech Corpora

3.1.1. Western Armenian

A major source of audio data for standard Western Armenian is the Rerooted[9] archive, an archive of interviews carried out starting in 2017 with Western Armenian speakers from Syria, who relocated to the Republic of Armenia as a result of the war in their birth country. Each interview generally last between 45 minutes and 1.5 hours, in which an interviewer poses question (often in Western Armenian, but sometimes in English) and the interviewee responds at length in Western Armenian. The vast majority of the audio documents available are not only transcribed in SWA, but also translated into English, as the project's primary goal concerns the transmission of memory of a displaced community. The full length interviews are available through the Rerooted website and housed on YouTube, where the transcriptions and translations serve as subtitles (and are therefore aligned by phrase). These aligned transcriptions were produced using the online subtitling platform Amara[10], from where we have been granted access to the aligned transcriptions in SRT (standard subtitling) format. In the framework of the DALiH project, we aim to make these resources publicly available as well. In total the exploitable aligned audio data from the Rerooted archive amounts to 90 documents, or 81 hours and forty minutes.

3.1.2. Eastern Armenian

A primordial source of Eastern Armenian audio data is the Eastern Armenian National Corpus (Khurshudian and Daniel, 2009)[11] (EANC), an online written and speech corpus compiled by an international team of linguists, scholars and software professionals, in the framework of an eponymous project launched in 2006. Amongst EANC's collected and processed materials are audio data of diverse genres: spontaneous speech, public discourse, online communications and task-oriented discourse. All together the aforementioned materials amount to 774 transcribed audio documents, or 3.5 million tokens.[12]

	Interviews	Hours
Translated (ENG)	100	87:46:03
Transcribed (ARM) + aligned	90	81:39:41
Total available	102	89:50:04

Table 2: Rerooted Archives' database

Rerooted and EANC both provide a healthy base of semi-processed audio data, originating from speakers of diverse ages and backgrounds, upon which further research and testing of ASR models will depend.

3.2. Data Preprocessing

None of the two corpora described in this section were built to train an ASR model. The use of such resources therefore requires preprocessing.
As mentioned in the previous section, most of Rerooted videos already have subtitles in Armenian. No further data processing is needed other than a trivial format conversion, from SRT to TextGrid[13]. Such conversion is useful as we are using Praat (Boersma and Weenink, 2022) to visualise the data and running Praat scripts to study variant-related phenomena.

On the other hand, transcriptions for EANC have to be aligned to be used. Considering the amount of data to be processed, we developed a simple automatic processing chain:

1. Extraction of the transcription from Word files
2. Automatic segmentation into utterances, units that are broadly equivalent to sentences in written texts
3. Forced alignment of those units with the sound

Extraction of the transcription While subtitles only transcribe what was pronounced by speakers, transcriptions meant to be analysed by linguists also contain extralinguistic information such as the speaker's attitude,

[9] https://www.rerooted.org

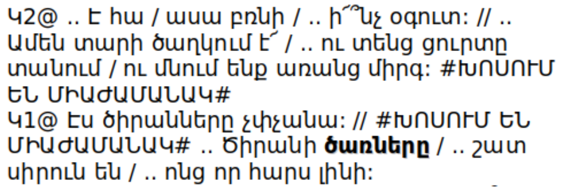
[10] https://amara.org/fr/

[11] http://www.eanc.net

[12] Unfortunately we cannot report the quantity of data in hours, because that information is unavailable to us.

[13] The conversion was successfully tested on a sample using a slightly modified version of the script available on https://github.com/tanmaysurana/srt2textgrid.

laughs, pauses or overlapping sequences, as shown in Figure 1. In this example, we can see that the annotator explicitly indicated that the two speakers were "talking at the same time", using a specific marker, #, to signal that this is not a transcription but an annotation. The first step of our processing chain consists of removing this extralinguistic information along with the speaker's identification which can be either their name, their status (Բժիշկ *doctor*, Աշխատող *employee* etc.) or an identification code (S1/S2, Կ1/Կ2 etc.).



Կ2@ .. Է հա / ասա բռնի / .. ի՞՞նչ օգուտ։ // .. Ամեն տարի ծաղկում է՛ / .. ու տենց ցուրտը տանում / ու մնում ենք առանց միրգ։ #ԽՈՍՈՒՄ ԵՆ ՄԻԱԺԱՄԱՆԱԿ#
Կ1@ Էս ծիրանները չփչանա։ // #ԽՈՍՈՒՄ ԵՆ ՄԻԱԺԱՄԱՆԱԿ# .. Ծիրանի **ծառները** / .. շատ սիրուն են / .. ոնց որ հարս լինի։

Figure 1: Excerpt of the transcription of a dialogue from EANC [dialogue_in_the_shop1]
translation:
K1@ Well yeah / I said hold / .. what's the use.// .. Every year it blossoms / .. and that kind of cold in the house / and we remain without fruit. #TALKING AT THE SAME TIME#
K2@ don't let these apricots spoil. // #TALKING AT THE SAME TIME# the apricot ***trees*** *are very pretty/ .. like a bride would be.*

Automatic segmentation Text segmentation is necessary for alignment. Speech data typically does not have punctuation and automatic speech segmentation may therefore rely on prosodic cues (such as lengthening of vowels or contours) or the length of pauses between words. However, EANC's transcription guidelines seem to include punctuation marks as well as segmentation marks in some cases such as in Figure 1 where / and // seem to be used to segment the utterances into smaller units. The second step of our processing chain made use of punctuation marks (namely, the comma and the :[14] (*verjaket*) used as a full stop) and :// in dialogues[15].

Forced alignment Aligning orthographic transcriptions with their corresponding speech is a costly task in terms of time. For the last part of our processing chain, we use a well-documented Python package for forced alignment called aeneas (Pettarin, 2022)[16]. This decision was mainly led by the fact that it wraps eSpeak[17], an open-source speech synthesizer, allowing support for both standards of Armenian. Even if this support has been implemented naively with no feedback neither from Eastern nor Western Armenian native speakers yet, preliminary results are quite good for high quality recordings.

The alignment was manually evaluated by a native speaker on a small sample of different types of speech from EANC:

- monologues including TV speech recordings and interviews in which the interviewer only asks a question at the very beginning of the recordings;
- dialogues : conversations with two participants (on the phone, at the office, when shopping etc.).
- polylogues : conversations with more than two participants, such as friends having a meal together.

It is noteworthy that there is a discrepancy in the quality of those samples, some being recorded in a quiet room, while others were recorded in the street where cars or construction work can be heard in the background. Unsurprisingly, the results are unequal: very good on monologues (especially for the interviews) but quite bad on polylogues, especially with noise in the background and/or when speakers' speech overlaps frequently. While the use of our aligner is promising on monologues, we now have to do a formal evaluation to assess whether or not providing our annotators with automatically aligned recordings of polylogues will help them or if segmenting from scratch takes less time than the manual correction of segments' boundaries.

4. A Unified System

As explained in Section 2, while advances in Armenian ASR are well underway, there remain large issues in terms of availability to the academic community. Additionally, none of the existing projects propose a model which addresses the community's need for a unified, or bi-variant system. In order to proceed forward in this research, and keeping in mind the limitations of resources, we propose that a hybrid method is more appropriate in the immediate future than an End-to-end (E2E) ASR system. In following other recent approaches to automatic transcription for lesser-endowed languages (such as (Guillaume et al., 2022)), we suggest that a hybrid system would enable us to employ neuronal systems such as wav2vec for feature extraction, informing our acoustic model, which we would fine-tune manually. We would then pass to a sole traditional lexicon model, and finally to a language model.

In employing this strategy, the pre-processing (i.e. alignment) and processing of audio data becomes all the more crucial in order to train our model, and also to measure it's efficacy and accuracy.

[14] The *verjaket* looks like a Latin colon but is part of the Armenian script and is encoded U+0589 in Unicode, so this punctuation mark is quite reliable as a segmenter.

[15] We decided not to use / as a segmentation mark because such units would be too small.

[16] Freely available on https://github.com/readbeyond/aeneas/.

[17] http://espeak.sourceforge.net/

5. Conclusion

We have outlined the major challenges in the development of Armenian ASR, especially as it pertains to a system which would understand both of the language's standard variants. Despite major advancements in Armenian ASR, this central issue remains largely unaddressed. We present the available oral corpora, and with the data available to us we ran a preliminary forced-alignment test, which showed varying results, confirming the need for the development of tools and resources. Lastly we proposed a basic methodology for moving forward.

6. Acknowledgements

DALiH (*Digitizing Armenian Linguistic Heritage*) is a project funded by the ANR, the French National Research Agency (ANR-21-CE38-0006).

7. Bibliographical References

Ali, A. M. A. M. (2018). *Multi-dialect Arabic broadcast speech recognition.* Ph.D. thesis.

Bisani, M. and Ney, H. (2008). Joint-sequence models for grapheme-to-phoneme conversion. *Speech communication*, 50(5):434–451.

Boersma, P. and Weenink, D. (2022). Praat: doing phonetics by computer [Computer program]. Version 6.2.12, retrieved 26 April 2022 from http://www.praat.org/.

Caballero, M., Moreno, A., and Nogueiras, A. (2009). Multidialectal spanish acoustic modeling for speech recognition. *Speech Communication*, 51(3):217–229.

Cowe, P. (1992). Amn tel hay kay: Armenian as a pluricentric language. *M. Clyne*.

Guillaume, S., Wisniewski, G., Galliot, B., Nguyễn, M.-C., Fily, M., Jacques, G., and Michaud, A. (2022). Plugging a neural phoneme recognizer into a simple language model: a workflow for low-resource settings. working paper or preprint, March.

Khurshudian, V. and Daniel, M. (2009). Eastern Armenian National Corpus. In *Computational Linguistics and Intellectual Technologies: Papers from the Annual International Conference "Dialogue'2009"*, pages 509–518.

Muhr, R. (2016). The state of the art of research on pluricentric languages: Where we were and where we are now. *Pluricentric Languages and Non-Dominant Varieties Worldwide, Österreichisches deutsch sprache der gegenwart*, 18:13–40.

Perero-Codosero, J. M., Espinoza-Cuadros, F. M., and Hernández-Gómez, L. A. (2022). A comparison of hybrid and end-to-end asr systems for the iberspeech-rtve 2020 speech-to-text transcription challenge. *Applied Sciences*, 12(2):903.

Pettarin, A. (2022). aeneas [Computer program]. Version 1.7.3, retrieved 26 April 2022 from https://www.readbeyond.it/aeneas/.

Vardanyan, A. (2016). Noise-robust speech recognition system for armenian language. Master's thesis, American University of Armenia.

Vaux, B. (1998). *The phonology of Armenian.* Oxford University Press.

Zampieri, M., Malmasi, S., Ljubešić, N., Nakov, P., Ali, A., Tiedemann, J., Scherrer, Y., and Aepli, N. (2017). Findings of the vardial evaluation campaign 2017. In *Proceedings of the fourth workshop on NLP for similar languages, varieties and dialects*.