Towards a Unified ASR System for the Armenian Standards
Samuel Chakmakjian, Ilaine Wang

To cite this version:
Samuel Chakmakjian, Ilaine Wang. Towards a Unified ASR System for the Armenian Standards. DigitArm Workshop at the International Conference on the 13th Language Resources and Evaluation (LREC 2022), Jun 2022, Marseille, France. pp.38-42. hal-04059100

HAL Id: hal-04059100
https://hal.science/hal-04059100
Submitted on 6 Apr 2023

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Distributed under a Creative Commons Attribution - NonCommercial| 4.0 International License
Towards a Unified ASR System for the Armenian Standards

Samuel Chakmakjian¹,², Ilaine Wang²
¹SeDyL (CNRS-Inalco),²ERTIM (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 (SW A) inventory being comprised of 30 phonemes (24 consonants, 6 vowels).

The vocalic systems of SEA and SW A 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, ɾ, h, j, l/). SEA distinguishes between two rhotics, a tap /ɾ/ and a trill /r/, whereas SW A 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 SW A and SEA. SEA’s plosive and affricate phonemes have a three-way voicing distinction: voiced, voiceless, and voiceless aspirated. Modern SW A has a two-way voicing system of voiced and voiceless aspirated. The plosive and affricate phonemes of SEA are therefore the following: /b, p, pʰ, d, t, tʰ, g, k, kʰ, dz, ts, tsʰ, dʒ, tʃ, tʃʰ/, and the plosive and affricate phonemes of SW A are as follows: /b, pʰ, d, tʰ, g, kʰ, dz, tsʰ, dʒ, tʃʰ/.

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

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. Furthermore, 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 SW A (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 SW A-speaking diasporan communities, and an increasing presence of SW A 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 SW A-speaking interviewee, or a discussion between a SW A-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.

<table>
<thead>
<tr>
<th>Item</th>
<th>SEA</th>
<th>SW A</th>
<th>Translation</th>
</tr>
</thead>
<tbody>
<tr>
<td>⟨ pun ⟩</td>
<td>[pún]</td>
<td>[pʰ(un)]</td>
<td>‘word’</td>
</tr>
<tr>
<td>⟨ upun ⟩</td>
<td>[púren]</td>
<td>[b̥ar]</td>
<td>‘dance’</td>
</tr>
<tr>
<td>⟨ qun ⟩</td>
<td>[pʰun]</td>
<td>[b̥ar]</td>
<td>‘placenta’</td>
</tr>
</tbody>
</table>

Table 1: Three words and their pronunciations in SEA vs. SW A
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 model 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 Scribe①, a company based in Barcelona, Spain, proposes an automatic transcription and automatic subtitling service for 63 languages, including Armenian. Another such example is VocalMatic② 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 present③. 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 specified④.

#### 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 CMUSphinx project⑤. Another important Armenian-specific project is that of the National Center of Communication and Artificial Intelligence Technologies (NCCAIT⑥), 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 project⑦ 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[https://www.happyscribe.com/](https://www.happyscribe.com/)
②[https://vocalmatic.com/languages/](https://vocalmatic.com/languages/)
③[https://cmusphinx.github.io](https://cmusphinx.github.io)

⑥Vardanyan (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.


1Especially in languages other than English.
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 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 Amari 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 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)[9] (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[10]

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[11]. Such conversion is useful as we are using 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, transcripts meant to be analysed by linguists also contain extralinguistic information such as the speaker’s attitude, and Daniel, 2009[9] (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[10]. Such conversion is useful as we are using 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, transcripts meant to be analysed by linguists also contain extralinguistic information such as the speaker’s attitude,
laughs, pauses or overlapping sequences, as shown in Figure 4. 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 ( врач doctor, служащий employee etc.) or an identification code (S1/S2, V1/V2 etc.).

Figure 1: Excerpt of a transcription of a dialogue from EANC [dialogue_in_the_shop1]

This decision was mainly led by the fact that it wraps aeneas (Petarin, 2023) as a segmentation mark because such units would be too small.

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 its efficacy and accuracy.

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.
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
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 deutsches sprache der gegenwart, 18:13–40.