Toward a Web-based Speech Corpus for Algerian Arabic Dialectal Varieties

Soumia Bougrine[1] **Aicha Chorana**[1] **Abdallah Lakhdari**[1] **Hadda Cherroun**[1]
[1]Laboratoire d'informatique et Mathématiques
Université Amar Telidji Laghouat, Algérie
{sm.bougrine,a.chorana,a.lakhdari,hadda_cherroun}@lagh-univ.dz

Abstract

The success of machine learning for automatic speech processing has raised the need for large scale datasets. However, collecting such data is often a challenging task as it implies significant investment involving time and money cost. In this paper, we devise a recipe for building large-scale Speech Corpora by harnessing Web resources namely YouTube, other Social Media, Online Radio and TV. We illustrate our methodology by building KALAM'DZ, An Arabic Spoken corpus dedicated to Algerian dialectal varieties. The preliminary version of our dataset covers all major Algerian dialects. In addition, we make sure that this material takes into account numerous aspects that foster its richness. In fact, we have targeted various speech topics. Some automatic and manual annotations are provided. They gather useful information related to the speakers and sub-dialect information at the utterance level. Our corpus encompasses the 8 major Algerian Arabic sub-dialects with 4881 speakers and more than 104.4 hours segmented in utterances of at least 6 s.

1 Introduction

Speech datasets and corpora are crucial for both developing and evaluating Natural Language Processing (NLP) systems. Moreover, such corpora have to be large to achieve NLP communities expectations. In fact, the notion of "More data is better data" was born with the success of modeling based on machine learning and statistical methods.

The applications that use speech corpora can be grouped into four major categories: speech recognition, speech synthesis, speaker recognition/verification and spoken language systems. The need for such systems becomes inevitable. These systems include real life wingspan applications such as speech searching engines and recently *Conversational Agents*, conversation is becoming a key mode of human-computer interaction.

The crucial points to be taken into consideration when designing and developing relevant speech corpus are numerous. The necessity that a corpus takes the within-language variability (Li et al., 2013). We can mention some of them: The corpus size and scope, richness of speech topics and content, number of speakers, gender, regional dialects, recording environment and materials. We have attempted to cover a maximum of these considerations. We will underline each considered point in what follows.

For many languages, the state of the art of designing and developing speech corpora has achieved a mature situation. On the other extreme, there are few corpora for Arabic (Mansour, 2013). In spite that geographically, Arabic is one of the most widespread languages of the world (Behnstedt and Woidich, 2013). It is spoken by more than 420 million people in 60 countries of the world (Lewis et al., 2015). Actually, it has two major variants: Modern Standard Arabic (MSA), and Dialectal Arabic. MSA is the official language of all Arab countries. It is used in administrations, schools, official radios, and press. However, DA is the language of informal daily communication. Recently, it became also the medium of communication on the Web, in chat rooms, social media etc. This fact, amplifies the need for language resources and language related NLP systems for dialects.

For some dialects, especially Egyptian and Levantine, there are some investigations in terms of building corpora and designing NLP tools. While,

Proceedings of The Third Arabic Natural Language Processing Workshop (WANLP), pages 138–146,
Valencia, Spain, April 3, 2017.

very few attempts have considered Algerian Arabic dialect. Which, make us affirm that the Algerian dialect and its varieties are considered as under-resourced language. In this paper, we tend to fill this gap by giving a complete recipe to build a large-size speech corpus. This recipe can be adopted for any under-resourced language. It eases the challenging task of building large datasets by means of traditional direct recording. Which is known as time and cost consuming. Our idea relies on Web resources, an essential milestone of our era. In fact, the Web 2.0, becomes a global platform for information access and sharing that allows collecting any type of data at scales hardly conceivable in near past.

The proposed recipe is to build a speech corpus for Algerian Arabic dialect varieties. For this preliminary version, the corpus is annotated for mainly supporting research in dialect and speaker identification.

The rest of this paper is organized as follows. In the next section, we review some related work that have built DA corpora. In Section 3 we give a brief overview of Algerian sub-dialects features. Section 4 is dedicated to describe the complete proposed recipe of building a Web-based speech dataset. In Section 5, we show how this recipe is narrated to construct a speech corpus for Algerian dialectal varieties. The resulted corpus is described in Section 6. We enumerate its potential uses in Section 7

2 Related Work

In this section, we restricted our corpora review to speech corpora dealing with Arabic dialects. We classify them according to two criteria: *collecting method* and *Intra/Inter country dialect collection context*. They can be classified into five categories according to the collecting method. Indeed, it can be done by recording broadcast, spontaneous telephone conversations, telephone responses of questionnaires, direct recording and Web-based resourcing. The second criterion distinguishes the origin of targeted dialects in either Intra-country/region or Inter-country, which means that the targeted dialects are from the same country/region or dialects from different countries. This criterion is chosen because it is harder to perform fine collection of Arabic dialects belonging to close geographical areas that share many historic, social and cultural aspects.

In contrast of relative abundance of speech corpora for Modern Standard Arabic, very few attempts have considered building Arabic speech corpora for dialects. Table 1 reports some features of the studied *DA* corpora. The first set of corpora has exploited the limited solution of telephony conversation recording. In fact, as far as we know, development of the pioneer DA corpus began in the middle of the nineties and it is *CALLFRIEND Egyptian* (Canavan and Zipperlen, 1996). Another part of *OrienTel* project, cited below, has been dedicated to collect speech corpora for Arabic dialects of Egypt, Jordan, Morocco, Tunisia, and United Arab Emirates countries. In these corpora, the same telephone response to questionnaire method is used. These corpora are available via the ELRA catalogue [1].

The *DARPA Babylon Levantine* [2] Arabic speech corpus gathers four Levantine dialects spoken by speakers from Jordan, Syria, Lebanon, and Palestine (Makhoul et al., 2005).

Appen company has collected three Arabic dialects corpora by means of spontaneous telephone conversations method. These corpora [3] uttered by speakers from Gulf, Iraqi and Levantine. With a more guided telephone conversation recording protocol, *Fisher Levantine Arabic* corpus is available via LDC catalogue [4]. The speakers are selected from Jordan, Lebanon, Palestine, Lebanon, Syria and other Levantine countries.

TuDiCoI (Graja et al., 2010) is a spontaneous dialogue speech corpus dedicated to Tunisian dialect, which contains recorded dialogues between staff and clients in the railway of Sfax town, Tunisia.

Concerning corpora that gather MSA and Arabic dialects, we have studied some of them. *SAAVB* corpus is dedicated to speakers from all the cities of Saudi Arabia country using telephone response of questionnaire method (Alghamdi et al., 2008). The main characteristic of this corpus is that, before recording, a preliminary choice of speakers and environment are performed. The selection aims to control speaker age and gender and telephone type.

Multi-Dialect Parallel (MDP) corpus, a free

[1] Respective code product are ELRA-S0221, ELRA-S0289, ELRA-S0183, ELRA-S0186 and ELRA-S0258.

[2] Code product is LDC2005S08.

[3] The LDC catalogue's respective code product are LDC2006S43, LDC2006S45 and LDC2007S01.

[4] Code product is LDC2007S02.

Corpus	Type	Collecting Method	Corpus Details
Al Jazeera multi-dialectal	Inter	Broadcast news	57 hours, 4 major Arabic dialect groups annotated using crowdsourcing
ALG-DARIDJAH	Intra	Direct Recording	109 speakers from 17 Algerian departments, 4.5 hours
AMCASC	Intra	Telephone conversations	3 Algerian dialect groups, 735 speakers, more than 72 hours.
KSU Rich Arabic	Inter	Guided telephone conversations and Direct recording.	201 speakers from nine Arab countries, 9 dialects + MSA.
MDP	Inter	Direct Recording	52 speakers, 23% MSA utterances, 77% DA utterances, 32 hours, 3 dialects + MSA.
SAAVB	Inter	Selected speaker before telephone response of questionnaire	1033 speakers; 83% MSA utterances, 17% DA utterances, Size: 2.59 GB, 1 dialect + MSA
TuDiCoI	Inter	Spontaneous dialogue	127 Dialogues, 893 utterances, 1 dialect.
Fisher Levantine	Inter	Guided telephone conversations	279 conversations, 45 hours, 5 dialects.
Appen's corpora	Inter	Spontaneous telephone conversations	3 dialects, Gulf: 975 conver, $\backsim$ 93 hours; Iraqi: 474 conver, $\backsim$ 24 hours; Levantine: 982 conver, $\backsim$ 90 hours.
DARPA Babylon Levantine	Inter	Direct recording of spontaneous speech	164 speakers, 75900 Utterances, Size: 6.5 GB, 45 hours, 4 dialects.
OrienTel MCA	Inter	Telephone response of questionnaire	5 dialects, # speakers: 750 Egyptian, 757 Jordanian, 772 Moroccan, 792 Tunisian and 880 Emirates.
CALLFRIEND	Inter	Spontaneous telephone conversations	60 conversations, lasting between 5-30 minutes, 1 dialect.

Table 1: Speech Corpora for Arabic dialects.

corpus, which gathers MSA and three Arabic dialects (Almeman et al., 2013). Namely, the dialects are from Gulf, Egypt and Levantine. The speech data is collected by direct recording method.

KSU Rich Arabic corpus encompasses speakers by different ethnic groups, Arabs and non-Arabs (Africa and Asia). Concerning Arab speakers in this corpus, they are selected from nine Arab countries: Saudi, Yemen, Egypt, Syria, Tunisia, Algeria, Sudan, Lebanon and Palestine. This corpus is rich in many aspects. Among them, the richness of the recording text. In addition, different recording sessions, environments and systems are taken into account (Alsulaiman et al., 2013).

Al Jazeera multi-dialectal speech corpus, a larger scale, based on Broadcast News of Al Jazeera (Wray and Ali, 2015). Its annotation is performed by crowd sourcing technology. It encompasses the four major Arabic dialectal categories.

In an intra country context, there are two corpora dedicated to Algerian Arabic dialect varieties: *AMCASC* (Djellab et al., 2016) and *A*LG-*D*ARIDJAH (Bougrine et al., 2016). *AMCASC* corpus, based on telephone conversations collecting method, is a large corpus that takes three regional dialectal varieties. While *A*LG-*D*ARIDJAH corpus is a parallel corpus that encompasses Algerian Arabic sub-dialects. It is based on direct recording method. Thus, many considerations are controlled while building this corpus. Compared to *AMCASC* corpus, the size of *A*LG-*D*ARIDJAH corpus is restricted.

According to our study of these major Arabic dialects corpora, we underline some points. First, these corpora are mainly fee-based and the free ones are extremely rare. Second, almost existing corpora are dedicated to inter-country dialects. Third, to the best of our knowledge, there is no Web-based speech dataset/corpus that deals with Arabic speech data neither for MSA nor for dialects. While for other languages, there are some investigations. We can cite the large recent col-

lection *Kalaka*-3 (Rodríguez-Fuentes et al., 2016). This is a speech database specifically designed for Spoken Language Recognition. The dataset provides TV broadcast speech for training, and audio data extracted from YouTube videos for testing. It deals with European languages.

3 Algerian Dialects: Brief Overview

Algeria is a large country, administratively divided into 48 departments. Its first official language is Modern Standard Arabic. However, Algerian dialects are widely the predominant means of communication. In Figure 1, we depict the main Algerian dialect varieties. In this work, we focus on Algerian Arabic sub-dialects as they are spoken by 75% to 80% of the population. The Algerian dialect is known as Daridjah to its speakers.

Algerian Arabic dialects resulted from two Arabization processes due to the expansion of Islam in the 7^{th} and 11^{th} centuries, which lead to the appropriation of the Arabic language by the Berber population.

According to both Arabization processes, dialectologists (Palva, 2006), (Pereira, 2011) show that Algerian Arabic dialects can be divided into two major groups: Pre-Hilālī and Bedouin dialects. Both dialects are different by many linguistic features (Marçais, 1986) (Caubet, 2000).

Firstly, Pre-Hilālī dialect is called a sedentary dialect. It is spoken in areas that are affected by the expansion of Islam in the 7^{th} century. At this time, the partially affected cities are: Tlemcen, Constantine and their rural surroundings. The other cities have preserved their mother tongue language (Berber).

Secondly, Bedouin dialect is spoken in areas which are influenced by the Arab immigration in the 11^{th} century (Palva, 2006) (Pereira, 2011). Marçais (1986) has divided Bedouin dialect into four distinct dialects: i) *Sulaymite* dialect which is connected with Tunisian Bedouin dialects, ii) *Ma'qilian* dialect which is connected with Moroccan Bedouin dialects, iii) *Hilālī* dialect contains three nomadic sub-dialects. *Hilālī-Saharan* that covers the totality of the Sahara of Algeria, the *Hilālī-Tellian* dialect which its speakers occupy a large part of the Tell of Algeria, and the *High-plains of Constantine*, which covers the north of Hodna region to Seybouse river. iv) *Completely-bedouin dialect* that covers Algiers' Blanks, and some of its near sea coast cities. Regarding to some linguistic differences, we have divided this last dialect into two sub-dialects, namely *Algiers-Blanks* and *Sahel-Tell*.

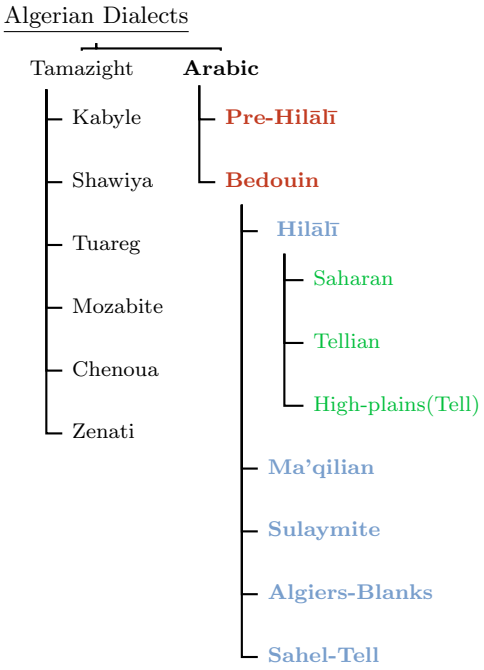

Figure 1: Hierarchical Structure of Algerian Dialects.

Arabic Algerian dialects present complex linguistic features and many linguistic phenomena can be observed. Indeed, there is many borrowed words due to the deep colonization. In fact, Arabic Algerian dialects are affected by other languages such as Turkish, French, Italian, and Spanish (Leclerc, 30 avril 2012). In addition, code switching is omnipresent especially from French.

Versteegh et al. (2006) used four consonants (the dentals fricative */t, d̲, ḍ̲/* and a voiceless uvular stop */q/*) to discriminate the two major groups: Pre-Hilālī and Bedouin dialect. In fact, he shows that Pre-Hilālī dialect are characterized by: */q/* is pronounced */k/* and the loss of inter-dentals and pass into the dentals */t, d, ḍ/*. For Algerian Bedouin dialect, the four discriminative consonants are characterized by: /q/ is pronounced /g/ and the inter-dentals are fairly preserved. For more details on Algerian linguistic features refer to (Embarki, 2008) (Versteegh et al., 2006) (Harrat et al., 2016).

4 Methodology

In this section, we first describe, in general way, the complete recipe to collect and annotate a Web-based spoken dataset for an under-resourced language/dialect. Then, we illustrate this recipe to build our Algerian Arabic dialectal speech corpus mainly dedicated to dialect and speaker identification.

Global View of the Recipe

The recipe described in the following can be easily tailored according to potential uses of the corpus and on the specificities of the targeted language resources and its spoken community.

1. *Inventorying Potential Web sources:* First, we have to identify sources that are the most targeted by the communities of the languages/dialects in concerns. Indeed, depending on their culture and preferences, some communities show preference for dealing with some Web media over others. For example, Algerian people are less used to use *Instagram* or *Snapchat* compared with Middle Est and Gulf ones. Moreover, each country has its own most used communication media. For instance, some societies (Arabs ones) are more productive on TVs and Radios, compared with west communities that are more present and productive on social media.

2. *Extraction Process:* In order to avoid crawling useless data, this steps is achieved by three stages

   (a) *Preliminary Validation Lists*: For each chosen Web source, we define the main keywords that can help automatically search video/audio lists. When such lists are established, a first cleaning is performed keeping only the potential suitable data. Sizing such lists depends on the sought scale.

   (b) *Providing the collection script*: For each resource, we fix and implement the suitable way to collect data automatically. Open Source tools are the most suitable. In fact, downloading a speech from a streaming or from YouTube or even from online Tv needs different scripts. The same fact has to be taken into account concerning their related metadata [5] which are very useful for annotation.

   (c) *Downloading*: This is a time consuming task. Thus, it is important to consider many facts such as preparing storage and downloading the related metadata, ...

   (d) *Cleaning*: Now, the videos/audios are locally available, a first scan is performed in order to keep the most appropriate data to the corpus concerns. This can be achieved by establishing a strategy depending on the corpus future use.

3. *Annotation and Pre-processing*: For a targeted NLP task, pre-processing the collected speech/video can include segmentation, White noise removing.... Some annotations can simply be provided from the related metadata of the Web-source when they exist. However, this task makes use of other annotation techniques like crowdsourcing where crowd are called to identify the targeted dialect/speaker or/and perform translations.

The method can be generalized to other languages/dialects without linguistic and cultural knowledge of the regional language or dialect by using video/audio search query based on the area (location) of targeted dialect/language. Then use the power of crowdsourcing to annotate corpus.

5 Corpus Building

For the context of the Algerian dialects, in order to build a speech corpus that is mainly dedicated to dialect/speaker identification using machine learning techniques, we have chosen several resources.

5.1 Web Sources Inventory

The main aim is to allow the richness of the corpus. In fact, it is well known that modeling a spoken language needs a set of speech data counting the within-language/intersession variability, such as speaker, content, recording device, communication channel, and background noise (Li et al., 2013). It is desirable to have sufficient data that include the intended intersession effects.

Table 2 reports the main Web sources that feed our corpus. Let us observe that there are several

[5] YouTube video Metadata such as *published_date, duration, description, category...*

speech Topics which allows capturing more linguistics varieties. In fact, this inventory contains "Local radio channels" resources. Fortunately, each Algerian province has at least one local radio (a governmental one). It deals with local community concerns and exhibits main local events. Some of their reports often use their own local dialect. It is the same case for amateur radios. Both, these radio channels and Tvs are Web-streamed live.

In addition, we have chosen some Algerian TVs for which most programs are addressed to large community. So, they use dialects. Finally, we have targeted some YouTube resources such as Algerian PodCasts, Algerian Tags, and channels of Algerian YouTubers.

Source	Sample	Topics
Algerian Tv	Ennahar	News
	El chorouk	News, General
	Samira, Bina	Cook
Local Radios	48 departments	Social, local, General
	On YouTube	
Algerian PodCast	Anes Tina	Politic, Culture, Social
Algerian YouTubers	Khaled Fkir	Blogs, Cook
	CCNA DZ	Tips, Fun
	Mesolyte	Advices, Beauty Technology, Vlog
Algerian TAG	–	Advices, Tips social discussions

Table 2: Main Sources of Videos

5.2 Extraction Process

Now having these Web sources, and as they are numerous, we process in two steps in order to acquire video/audio speech data. First, we drawn up lists by crawling information mainly meta data about existing data related to potential videos/audios that potentially contain Algerian dialect speech. The deployed procedure relies on mainly two different scripts according to the Web resource type.

In order to collect speech data from local radio channels, we refer back manually to the programs of radio to select report and emission lists that are susceptible to contain dialectal speech.

For data from YouTube, the lists are fed by using YouTube search engine through its Data API [6]. In addition, the extraction of related metadata are performed using Python package BeautifulSoup $V4$ dedicated to Web crawling (Nair, 2014).

In order to draw up search queries, we have used three lists. The first one Dep, contains the names of the 48 Algerian provinces, spelled in MSA, dialect and French. While the second list Cat contains the selected YouTube categories. Among the actual YouTube categories, we have targeted: *People & Blogs, Education, Entertainment, How-to & Style, News & Politics, Non-profits & Activism categories, and Humor*. The third list $Comm\_Word$ contains a set of common Algerian dialect word (called White Algerian terms) that are used by all Algerians. These chosen set is reported in Table 3. Then, we iterate searching using a search query model that has four keywords. The first and the second ones are from Dep and Cat lists respectively. The remaining two keywords are selected arbitrary from $Comm\_Word$. This query formulation can guarantee that speakers are from the fixed province and the content topics are varied thanks to YouTube topic classification.

Concerning Algerian TVs source, the search queries are drawn up using mainly two keywords. The first one is the name the channel and the second word refers to the report name. In fact, a prior list is filled manually with emission and report names that uses dialects. Easily, videos from Algerian YouTubers/Podcasts channels are searched using the name of the corresponding author.

واش	What	مشاكل	Problems	نقاش	Discussion
جرنان	Journal	الحقرة	Injure	بلاك	Maybe
تريسيتي	Electricity	يصرا	Happen	صاري	Happened
دارجة	Colloquial	دزيرية	Algerian	الدشرة	Village
النشرة	News	شوف	See	جزايري	Algerian
دزاير	Algeria	لولاية	Department	الباك	Baccalaureat
بروبليم	Problem	ماكلة	food	وشرايك	Your Opinion

Table 3: Common Algerian terms used as Keywords

For all YouTube search queries, we selected the first 100 videos (when available). When lists are drawn up, we start the cleaning process that discards the irrelevant video entries. In fact, we remove all list entries whose duration is less than 5s. The videos whose topic shows that it doesn't deal with dialects are also discarded. This is done by analyzing manually the video title then its description and keywords.

In addition, to be in compliance with YouTube Terms of Services, first, we take into account the existence of any Creative Commons license asso-

[6] YouTube Data API, `https://developers.google.com/YouTube/v3/`

ciated to the video.

The whole cleaning process leads us to keep 1182 Audios/videos among 4 000 retrieved ones. Video's length varies between 6s and 1 hour. The cleaning task was carried out by the authors of this paper.

Now, the download process is launched using the resulting cleaned lists. Concerning the data from radio channels, we script a planned recording of the reports from the stream of the radio. Here also the recording amount depends on the desired scale.

Concerning the data from Radios, we deploy, in the download script, mainly the *VLC* Media Player tool [7] with the *cron* Linux command. In order to download videos from YouTube, we have deployed YouTube-dl [8] a command-line program.

5.3 Preprocessing and Annotation

A collection of Web speech data is not immediately suitable for exploration in the same way a traditional corpus is. It needs more cleaning and preprocessing. Let us recall, that our illustrative corpus will serve for dialect/speaker identification using machine learning techniques. For that purpose, for each downloaded video, we have applied the following processing:

1. *Audio extraction*: FFmpeg [9] tool is used to extract the audio layer. In addition, the SoX [10] tool, a sound processing program, is applied to get single-channel 16 kHz WAV audio files.

2. *Non-speech segments removal*: such as music or white noise by running a VAD (Voice Activation Detection) tool to remove as many as possible.

3. *Speaker Diarization*: is performed to determine who speaks when, and to assign for each utterance a speaker ID. It is achieved using VoiceID Python library based on the LIUM Speaker Diarization framework (Meignier and Merlin, 2010). The output from VoiceID segmentation is a set of audio files with information about speaker ID, and utterance duration.

[7] VLC media player V 2.2.4 `https://www.videolan.org/vlc/`

[8] YouTube-dl V3 `http://YouTube-dl.org/`

[9] FFmpeg http://www.ffmpeg.org/ V3.2

[10] SoX `http://sox.sourceforge.net/`

Number of Dialects	8
Total Duration	104.4 hours
Clean Speech Ratio	39, 15 %
Number of speakers	4881
Speech duration by Speaker	6s – 9hours

Table 4: Corpus Global Statistics.

For this preliminary version, most manual annotations are made thanks to authors themselves with the help of 8 volunteers. These In Lab annotations concern assigning for each utterance the spoken dialect, validation of speaker gender (previously detected automatically by VoiceID). During these manual annotations, we check that utterances deal with dialect. Otherwise, they are discarded.

6 Corpus Main Features

First of all, we note that this preliminary version of our corpus is collected and annotated in less than two months, and the building is still an ongoing process. A sample of our corpus is available online [11]. KALAM'DZ covers most major Arabic sub-dialects of Algeria. Table 4 reports some global statistics of the resulted corpus. *Clean Speech Ratio* row gives the ratio of Speeches that have good sound quality. The remaning portion of speeches present some noise background mainly music or they are recorded in outdoor. However, they can be used to build dialect models in order to represent the within-language variability.

More details on KALAM'DZ corpus are reported in Table 5. Let us observe that some dialects are more represented. This is due to people distribution and Web culture. For instance, Algiers and Oran are metropolitan cities. So their productivity on the Web is more abundant.

In order to facilitate the deployment of KALAM'DZ corpus, we have adopted the same packaging method as TIMIT Acoustic-Phonetic Continuous Speech Corpus (Garofolo et al., 1993).

7 Potential Uses

KALAM'DZ built corpus can be considered as the first of its kind in term a rich speech corpus for Algerian Arabic dialects using Web resources. It can be useful for many purposes both for NLP and computational linguistic communities. In fact, it

[11] `https://github.com/LIM-MoDos/KalamDZ`

Sub-Dialect	Departments/Village	# Speakers	Web-sources (h)				
			Algerian Tv	Local Radios	On YouTube	Total (h)	Good Quality (%)
Hilali-Saharan	Laghouat, Djelfa, Ghardaia, Adrar, Bechar, Naâma' South	1338	12.7	16.5	-	29.4	38.5
Hilali-Tellian	Setif, Batna, Bordj-Bou-Arreridj	605	3.6	-	-	3.6	32.2
High-plains	Constantine, Mila, Skikda	297	2.0	-	-	2.0	42.7
Ma'qilian	Sidi-Bel Abbas, Saïda, Mascara, Relizane, Oran, Aïn Timouchent, Tiaret, Mostaganem, Naâma' North	421	4.1	-	20.7	24.8	38.3
Sulaymite	Annaba, El-Oued, Souk-Ahras, Tebessa, Biskra, Khanchela, Oum El Bouagui, Guelma, El Taref	914	6.7	-	6.9	13.6	39
Algiers Blanks	Algiers, Blida, Boumerdes, Tipaza	723	5.1	-	16.1	21.2	42.3
Sahel-Tell	Médea, Chlef, Tissemsilt, Ain Defla	447	3.1	6.0	-	9.1	45.5
Pre-Hilālī	Tlemcen Nadrouma, Dellys, Jijel, Collo, Cherchell	136	0.7	-	-	0.7	34.7
Global		4881	38.2	22.5	43.7	104.4	39.1

Table 5: Distribution of speakers and Web-sources per sub-dialect in KALAM'DZ corpus.

Sub-Dialect	N&P	Edu.	Ent.	H&S	P&B	Hum.
Hilali-Saharan	29.4	-	-	-	-	-
Hilali-Tellian	3.6	-	-	-	-	-
High-plains	2.0	-	-	-	-	-
Ma'qilian	4.1	1.5	-	7.4	10.2	1.6
Sulaymite	6.7	-	-	-	6.9	-
Algiers Blanks	5.1	2.2	8.7	-	1.4	2.8
Sahel-Tell	9.1	-	-	-	-	-
Pre-Hilālī	0.7	-	-	-	-	-
Total	60.7	3.7	8.7	7.4	19.5	4.4

Note: News & Politics (N&P), Educations (Edu.), Entertainment (Ent.), How-to & Style (H&S), People & Blogs (P&B), and Humor (Hum.).

Table 6: Distribution of categories per dialect (in hours).

can be used for building models for both speaker and dialect identification systems for the Algerian dialects. For linguistic and sociolinguistics communities, it can serve as base for capturing dialects characteristic.

All videos related to extracted audio data are also available. This can be deployed to build another corpus version to serve any image/video processing based applications.

8 Conclusion

In this paper, we have devised a recipe in order to facilitate building large-scale Speech corpus which harnesses Web resources. In fact, the used methodology makes building a Web-based corpus that shows the within-language variability. In addition, we have narrated this procedure for building KALAM'DZ a speech corpus dedicated to the whole Algerian Arabic sub-dialects. We have been ensured that this material takes into account numerous speech aspects that foster its richness and provides a representation of linguistic varieties. In fact, we have targeted various speech topics. Some automatic and manual annotations are provided. They gather useful information related to the speakers and sub-dialect information at the utterance level. This preliminary KALAM'DZ version encompasses the 8 major Algerian Arabic sub-dialects with 4881 speakers and more than 104.4 hours.

Mainly developed to be used in dialect identification, KALAM'DZ can serve as a testbed supporting evaluation of wide spectrum of NLP systems.

In future work, we will extend the corpus by collecting Algerian sub-dialects uttered by Berber native speakers. As the corpus building is still an ongoing work, its evaluation is left to a future work. In fact, we plan to evaluate the corpus on dialects identification in intra-country context.

References

Mansour Alghamdi, Fayez Alhargan, Mohammed Alkanhal, Ashraf Alkhairy, Munir Eldesouki, and Ammar Alenazi. 2008. Saudi Accented Arabic Voice Bank. *Journal of King Saud University-Computer and Information Sciences*, 20:45–64.

K. Almeman, M. Lee, and A. A. Almiman. 2013. Multi Dialect Arabic Speech Parallel Corpora. In *Communications, Signal Processing, and their Applications (ICCSPA)*, pages 1–6, Feb.

Mansour Alsulaiman, Ghulam Muhammad, Mohamed A Bencherif, Awais Mahmood, and Zulfiqar Ali. 2013. KSU Rich Arabic Speech Database. *Journal of Information*, 16(6).

Peter Behnstedt and Manfred Woidich. 2013. Dialectology.

S. Bougrine, H. Cherroun, D. Ziadi, A. Lakhdari, and A. Chorana. 2016. Toward a Rich Arabic Speech Parallel Corpus for Algerian sub-Dialects. In *LREC'16 Workshop on Free/Open-Source Arabic Corpora and Corpora Processing Tools (OSACT)*, pages 2–10.

Alexandra Canavan and George Zipperlen. 1996. CALLFRIEND Egyptian Arabic LDC96S49. Philadelphia: Linguistic Data Consortium.

Dominique Caubet. 2000. Questionnaire de dialectologie du Maghreb (d'après les travaux de W. Marçais, M. Cohen, GS Colin, J. Cantineau, D. Cohen, Ph. Marçais, S. Lévy, etc.). *Estudios de dialectología norteafricana y andalusí, EDNA*, 5(2000-2001):73–90.

Mourad Djellab, Abderrahmane Amrouche, Ahmed Bouridane, and Noureddine Mehallegue. 2016. Algerian Modern Colloquial Arabic Speech Corpus (AMCASC): regional accents recognition within complex socio-linguistic environments. *Language Resources and Evaluation*, pages 1–29.

Mohamed Embarki. 2008. Les dialectes arabes modernes: état et nouvelles perspectives pour la classification géo-sociologique. *Arabica*, 55(5):583–604.

J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, and N. L. Dahlgren. 1993. DARPA TIMIT Acoustic Phonetic Continuous Speech Corpus CDROM.

Marwa Graja, Maher Jaoua, and L Hadrich-Belguith. 2010. Lexical Study of A Spoken Dialogue Corpus in Tunisian Dialect. In *The International Arab Conference on Information Technology (ACIT), Benghazi, Libya*.

Salima Harrat, Karima Meftouh, Mourad Abbas, Khaled-Walid Hidouci, and Kamel Smaili. 2016. An algerian dialect: Study and resources. *International Journal of Advanced Computer Science and Applications-IJACSA*, 7(3).

Jacques Leclerc. 30 avril 2012. Algérie dans l'aménagement linguistique dans le monde.

M. Paul Lewis, F. Simons Gary, and D. Fenning Charles. 2015. Ethnologue: Languages of the World, Eighteenth edition. Web.

H. Li, B. Ma, and K. A. Lee. 2013. Spoken language recognition: From fundamentals to practice. *Proceedings of the IEEE*, 101(5):1136–1159, May.

J. Makhoul, B. Zawaydeh, F. Choi, and D. Stallard. 2005. BBN/AUB DARPA Babylon Levantine Arabic Speech and Transcripts. Linguistic Data Consortium (LDC). LDC Catalog Number LDC2005S08.

Mohamed Abdelmageed Mansour. 2013. The Absence of Arabic Corpus Linguistics: A Call for Creating an Arabic National Corpus. *International Journal of Humanities and Social Science*, 3(12):81–90.

Philippe Marçais. 1986. *Algeria*. Leiden: E.J. Brill.

Sylvain Meignier and Teva Merlin. 2010. Lium spkdiarization: an open source toolkit for diarization. In *in CMU SPUD Workshop*.

Vineeth G. Nair. 2014. *Getting Started with Beautiful Soup*. Packt Publishing.

Heikki Palva. 2006. Dialects: classification. *Encyclopedia of Arabic Language and Linguistics*, 1:604–613.

C. Pereira. 2011. Arabic in the North African Region. In S. Weniger, G. Khan, M. P. Streck, and J. C. E. Watson, editors, *Semitic Languages. An International Handbook*, pages 944–959. Berlin.

Luis Javier Rodríguez-Fuentes, Mikel Penagarikano, Amparo Varona, Mireia Diez, and Germán Bordel. 2016. Kalaka-3: a database for the assessment of spoken language recognition technology on youtube audios. *Language Resources and Evaluation*, 50(2):221–243.

Kees Versteegh, Mushira Eid, Alaa Elgibali, Manfred Woidich, and Andrzej Zaborski. 2006. Encyclopedia of Arabic Language and Linguistics. *African Studies*, 8.

Samantha Wray and Ahmed Ali, 2015. *Crowdsource a little to label a lot: Labeling a speech corpus of dialectal Arabic*, volume 2015-January, pages 2824–2828. International Speech and Communication Association.