A Crowdsourcing-based Approach for Speech Corpus Transcription Case of Arabic Algerian Dialects

Ilyes Zine, Mohamed Cherif Zeghad, Soumia Bougrine and Hadda Cherroun
Laboratoire d'Informatique et Mathématique (LIM)
Université Amar Telidji Laghouat, Algérie
{i.zine,m.zeghad,sm.bougrine,hadda_cherroun}@lagh-univ.dz

Abstract

In this paper we describe a corpus annotation project based on crowdsourcing technique that performs orthographic transcription of KALAM'DZ corpus (Bougrine et al., 2017c). This latter is a speech corpus dedicated to Arabic Algerian dialectal varieties. The recourse to crowdsourcing solution is deployed to avoid time and cost consuming solutions that involves experts. Since Arabic dialects have no standard orthographic, we have fixed some guidelines that helps crowd to get more normalized transcriptions. We have performed experiments on a sample of 10% of KALAM'DZ corpus, totaling 8.75 hours. The quality control of the output transcription is ensured within three stages: Pre-qualification of crowd, online filtering and in lab validation and revision. A baseline resource is used to evaluate both first stages. It consists on 5% of the targeted dataset transcribed by well trained transcribers. Our results confirm that the crowdsourcing solution is an effective approach for speech dialect transcription when we deal with under-resourced dialects. Before the validation of the well trained transcribers the accuracy of transcriptions reached 74.38. In addition, we present a set of best practices for crowdsourcing speech corpus transcription.

1 Introduction

The transcription task is the process of language representation in written form. The source can either be speech or a text in another writing system. Transcribed Speech Corpora are crucial for both developing and evaluating NLP systems such speech recognition. Such corpora have to respond to NLP communities expectations and allow to be exploited in machine learning based solutions.

For many languages, the state of the art of NLP systems have achieved accurate mature situation thanks to large and well designed corpora. On the other extreme, there are few corpora for Arabic (Surowiecki, 2004). Moreover, very few attempts have been considered for Algerian Arabic dialect (Mansour, 2013). Recently, KALAM'DZ corpus (Bougrine et al., 2017c) has been developed to cover the Arabic dialectal varieties of Algeria. This corpus is collected using web-based sources. Despite its important size, about more than 104 hours, very few annotations are available. In fact, only dialect and speaker annotations are provided. In this paper, we investigated a crowdsourcing-based approach to transcribe its speeches. Transcribing dialectal speeches is a very challenging task as dialects have no linguistic rules and a recourse to experts transcription is time and cost consuming.

The rest of this paper is organized as follows. In the next section, we review some related work that have dealt with speech corpus transcription for Arabic. In Section 3, we give brief glance to Algerian dialects linguistic properties. In Section 4 we describe the target corpus KALAM'DZ. Section 5 is dedicated to our crowdsourcing solution, in which we explain the designed crowdsourcing project and the deployed quality control strategy. A list of best practices based on these crowdsourcing experiments is compiled in Section 6.

2 Related Work

The existing speech corpora annotated by orthographic transcripts, could be classified into two major groups: Pre-transcribed and Post-transcribed speech corpus. In fact, pre-transcribed speech datasets are mostly collected by recording audio files directly from a set of text files prepared to be uttered by various speakers. While, post-transcribed corpora represent speech datasets collected from Internet or by recording sponta-

Corpus	Transcription Type	Language	Details
***A-SpeechDB* (2005)**	Automatic + Manual Revision	MSA	20 hours of continuous speech, 30% of females and 70% of males
***NetDC* (2004)**	Manual transcription by experts	MSA	Using Transcriber tool (1998),22 hours of broadcast news speech
Fisher (2004)	Manual transcription by experts	Levantine Arabic Dialect	250 hours of telephone conversations, Using AMADAT tool
***CallHome* (1997)**	Manual transcription by experts	Egyptian Arabic Dialect	120 telephone conversations
***SAAVB* (2008)**	Manual transcription by experts	Saudi Dialect	96 hours distributed among 60 947 files
***STAC* (2015)**	Manual transcription by experts	Tunisian Dialect	5 hours, Using Praat tool (2001)
***MD-ASPC* (2013)**	Pre-transcribed	MSA, Gulf, Egypt, Levantine	32 hours
***Aljazeeras Corpus* (2015)**	Manual transcription using crowdsourcing	Egyptian, Levantine, Gulf, Maghrebi	Using CrowdFlower
***Alg-Daridjah* (2016)**	Manually transcribed	Arabic Algerian dialects	4h30mn, 6213 utterances
***MGB-2* (2016)**	Manually transcribed	MSA, Egyptian, Levantine, Gulf, Maghrebi	1200 hours, 70% of the speech is MSA, and the rest is in different Dialectal Arabic
***MGB-3* (2017)**	Manually transcribed	Egyptian dialectal Arabic	16 hours extracted from 80 YouTube videos

Table 1: Details on Corpora Transcription Approaches

neous/random conversations. Thus, the second category requires a transcription process.

Regarding transcribing approaches, we can classify them according to the used method into two categories: manual and semi-automatic transcription. This latter way is usually used to transcribe a non-colloquial language such as English, French or Modern Standard Arabic (MSA). The transcription process is achieved into two passes. By the first pass, an Automatic Speech Recognition (ASR) is used in order to generate a rough transcription that is manually reviewed in the second pass. On the other hand, manual transcription, is divided according to the transcriber level into two classes: experts or non-expert (crowd).

In this literature review, we focus on transcribed Arabic Speech corpora and their related transcription process. Let us note that the major Arabic dialects corpora are available through the Linguistic Data Consortium (LDC) as well as European Language Resources Association (ELRA) catalogues. Table 1 summarizes the reviewed transcribed speech corpora.

A-SpeechDB[1] is an MSA speech database suited for training acoustic models. The transcriptions are automatically generated. In addition, each transcribed sentence is augmented by a manually revised version (2005). NetDC[2] (Network of Data Centers) (Choukri et al., 2004), is an Arabic broadcast news speech corpus. It is dedicated to the Modern Standard Arabic from the Middle East region. The corpus is transcribed manually using *Transcriber*[3] software (Barras et al., 1998).

As regards LDC Catalogue, we can review Fisher Levantine Arabic[4] and CallHome[5] Egyptian Arabic projects. Fisher Levantine Arabic corpus contains a collection of 2000 telephone calls of 9400 speakers from the Northern, Southern and Bedwi dialects of Levantine Arabic (Maamouri et al., 2004). The transcription was done by experts using Arabic Multi-Dialectal Transcription Tool (AMADAT). Besides, the colloquial corpus called CallHome Egyptian Arabic is transcribed manually by Gadalla et al. (1997).

Saudi Accented Arabic Voice Bank (SAAVB) is dedicated to Saudi Arabic dialect. It is a very rich corpus in terms of its speech sound content and speaker diversity within the Saudi Arabia (Alghamdi et al., 2008). The transcription was done manually by experts using their own transcription interface.

Zribi et al. (2015) have built a Spoken Tunisian Arabic Corpus (STAC). It is transcribed manually by experts using Praat[6] tool (Boersma and Van Heuven, 2001). The transcription was done respect to OTTA an Orthographic Transcription of Tunisan dialect (Zribi et al., 2013).

Almeman et al. (2013) have built a Multi-Dialect Arabic Speech Parallel Corpus (MD-

[1]Code product: ELRA catalogue ELRA-S0315.
[2]Code product: ELRA catalogue ELRA-S0157

[3] www.transcriber.com
[4]LDC Catalogue No. LDC2007T04
[5]LDC Catalogue No. LDC97T19
[6]www.praat.org

ASPC). It contains written MSA prompts translated to dialects and then recorded. This one is an illustration of pre-transcribed speech corpora.

Wray et al. (2015) have transcribed a speech dataset collected from programs uploaded to Aljazeeras website. The transcription is performed by a crowdsourcing technique through the CrowdFlower platform.

Bougrine et al. (2016) have build an Arabic speech corpus for Algerian dialects, by recording 109 native speakers from 17 different provinces. The transcription was done manually by authors.

The Arabic Multi-Genre Broadcast (MGB-2) Challenge used recorded programs from 10 years of Aljazeera Arabic TV channel (Ali et al., 2016; Khurana and Ali, 2016). These programs were manually captioned on their Arabic website[7] with no timing information (Ali et al., 2016). Thus, an alignment was required for the manual captioning in order to produce speech segments for training speech recognition (Khurana and Ali, 2016). Furthermore, the Arabic MGB-3 Challenge (Ali et al., 2017), unlike Arabic MGB-2 Challenge, emphasizes dialectal Arabic using a multi-genre collection of Egyptian YouTube videos. The speech transcription was done manually using Transcriber tool, without a strict guidelines for standardizing DA orthography.

We observed that most reviewed transcribed corpora did not use crowdsourcing for speech transcription. Plus, Algerian Dialect has not received any attention.

3 Algerian Dialects

Algeria is a large country, administratively divided into 48 provinces. Its first official language is Modern Standard Arabic (MSA). However, Algerian dialects are widely the predominant means of communication.

Algerian Arabic dialects resulted from two Arabization processes due to the expansion of Islam in the 7th and 11th centuries, which lead to the appropriation of the Arabic language by the Berber population. According to both Arabization processes, Algerian Arabic dialects can be divided into two major groups: Pre-Hilālī and Bedouin dialect. Both dialects are different by many linguistic features (Gibb et al., 1986; Caubet, 2000). Bougrine et al. (2017b) give a preliminary version of an hierarchy structure for Arabic Algerian dialects (Figure 1).

Algerian dialect is considered among the most complex Arabic dialects with a lot of linguistic phenomena. For the current purpose, let us focus on some lexical, morphological and syntactic properties. Algerian DA vocabulary is mostly issued from MSA with many phonological alteration and many borrowed words from other languages, such as Turkish, French, Italian, and Spanish due to the deep colonization. In addition, code switching is omnipresent especially from French (Harrat et al., 2016; Saadane and Habash, 2015; Bougrine et al., 2017c).

Algerian DA morphology is similar to MSA excepts for some features. Some variations make Algerian DA morphology simpler than MSA. Essentially in some aspects of inflection and inclusion system, by eliminating several clitics and rules. Whereas negation in Algerian DA, including other Arabic dialects, is more complex than MSA. It is expressed by the circum-clitic negation ما and ش surrounding the verb with all its clitics or the indirect object pronouns (Harrat et al., 2016; Saadane and Habash, 2015).

As regards Algerian DA syntax, the words order of a declarative sentence is relatively flexible and all orders are allowed. The speaker begins the phrase with what he wants to highlight (Harrat et al., 2016). But the most commonly used order is the SVO order (Subject-Verb-Object) (Souag, 2006).

For more details on Algerian linguistic features refer to Embarki (2008); Saadane and Habash (2015); Harrat et al. (2016).

4 Targeted Corpus

Few speech corpora for Algerian Dialectal varieties are available (Bougrine et al., 2016, 2017c). For this study purpose, we have chosen KALAM'DZ corpus (Bougrine et al., 2017c). KALAM'DZ is a large speech corpus dedicated to Algerian Arabic dialectal varieties (Bougrine et al., 2017c). It covers eight major Arabic dialects spoken in Algeria. This corpus is collected from web sources namely YouTube, Online Radio stations, and TV channels. The size of the corpus is about 104 hours with 4881 speakers. All annotations are extracted from the related web sources metadata which are namely the ti-

[7] www.aljazeera.net

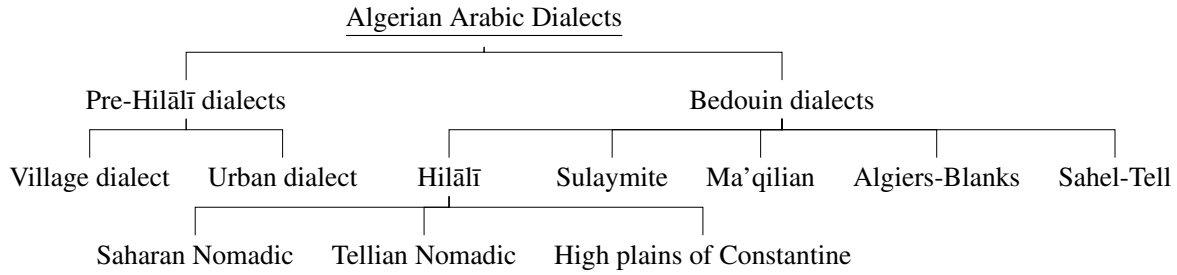

Figure 1: Hierarchy Structure for Algerian Dialects.

tle, category, location from where the source is posted, and the identity of the publisher. In addition, speaker gender is detected automatically by VoiceID tool. Concerning the dialect annotation, they are performed thanks to a crowdsourcing solution (Bougrine et al., 2017a).

In the current crowdsourcing task, we consider more than 8.75h hours to be transcribed. It contains 5122 speech segments with an average size of 6.2 seconds. Table 2 gives the distribution of speeches per Algerian dialect.

Sub-Dialect	# Segments	Duration (hour)
Hilālī-Saharan	1495	2.00
Sulaymite	1268	2.25
Algiers-blanks	1445	2.50
Ma'qilian	914	2.00
Total	**5122**	**8.75**

Table 2: Distribution of the Targeted Sample per Dialect.

5 Transcription Project

In order to transcribe the part of KALAM'DZ corpus, we have relied on crowdsourcing solution. To make these annotations scalable and of high quality, we have followed the crowdsourcing engineering process defined by Sabou et al. (2014). It suggests designing the system in four stages: project definition, data preparation, project execution, and data aggregation & evaluation. The project is baptized SPEECH2TEXT'DZ.

5.1 Project Definition

In this stage, we define the crowdsourcing task as well as the choice of crowdsourcing genre. As a basic task:" The contributor will be asked to listen to a short audio segment then write what they have heard exactly using Arabic letters and some shortcuts". The latter are deployed to facilitate the task and avoid contributor workload.

In order to make more interaction, users will be paid. Funding crowdsourcing projects is still not a common practice within the Algerian research community. Thus, we decided to go with a modest paid-for crowdsourcing. Where a user can collect points with a variable rate per task. These points can be used for mobile phones recharging.

5.2 Data Preparation

In this second stage, we build the project user and management interfaces. In order to collect crowdsourced transcripts, we have developed our own crowdsourcing platform[8] due to many constraints. Indeed, our targeted communities presence in crowdsourcing platforms as client is very modest. In addition to the administration profile, two roles are allowed: Transcriber and Well-Trained Transcriber (WTT). The *transcribers* are the crowd that can submit transcriptions. While WTT are users with more privileges. They are allowed to control transcribers' submissions. They are mainly lab members.

Concerning the transcriber interface, we have designed a form containing a text editor frame where the crowd transcribes the given speech segment, a set of shortcuts to help the crowd, and a link to a video that demonstrates the transcription guidelines. Our task is restricted mainly to Algerian users for that the form is written in Arabic.The management interface allows WTT validating and revising transcribers' output.

5.3 Project Execution

This is the main phase of any crowdsourcing project. In this step we performed three jobs: recruit contributors, train/retain contributors and manage/monitor crowdsourcing tasks.

[8] www.speech2text-dz.com

Publishing and advertising for attracting and retaining a large number of contributors is a key of success of any crowdsourcing system. We have decided to follow a simple strategy to advertise our platform. Social networks are always a good choice; we have gone with Facebook as preferable way for our targeted community.

Given that dialectal Arabic lacks a standardized orthography, we have defined an Orthographic Transcription Guideline that help to deliver a normalized transcription as much as possible. Our designed guideline is inspired from Saadane and Habash (2015) and Wray et al. (2015). In fact, we have designed some rules based on the Conventional Orthography for Dialectal Arabic (CODA) due to Habash et al. (2012) and adapted for Algerian dialect by Saadane and Habash (2015). Some other rules are added following the recommendations for crowdsourcing Arabic speech transcription due to Wray et al. (2015). This guideline is delivered through a video demonstration. Among these rules:

- The transcription is done in Arabic Script.
- To have a normalized spelling, the crowd has to transcribe colloquial words as close as possible to appropriate MSA spelling.
- In order to facilitate future potential Part-Of-Speech (POS) tagging task; foreign words, named entities, places and proper names should be transliterated in Arabic and guarded by some predefined tags. For example: [علم : مكان الجزائر] ([Named Entity: Place Algeria]) is used to tag that [الجزائر] is a proper noun indicating Algeria country.
- For more uniform transcription, a given spoken form is always written the same way.
- To be more faithful when transcribing; all non-speech sounds should be transcribed. For instance music, noise, breathing, laughs, they have to use respectively the predefined tags [موسيقى] [ضجيج] [تنفس] [ضحك].

Quality Control

A front-end verification process makes sure that transcribers respect the given guideline. In fact, two JavaScript functions are deployed, one function forces transcribers to type using only Arabic letters, and the second function to make sure that no spamming data are collected by disabling Copy/Paste functionality.

In order to ensure the quality control of the output transcriptions, we have acted in three stages: *Transcriber Pre-qualification*, *Online Filtering*, and *WTT revision*.

For the two first stages, we use an in lab transcripts as a *Baseline Resource* (BR) coupled with a mechanism of *Transcriber Trusting*. BR resource contains 256 transcribed utterances which represents 5% of the targeted sample. In brief, the mechanism works as follows. Initially, an arbitrary score of 50% is assigned to any new transcriber. This score changes every time that the transcriber has to pass a trusting control by means of transcribing a speech segment belonging to BR. In fact, his transcript is confronted to the corresponding BR one. The comparison is done by means of Levenshtein distance and similar tests.

Now, let us explain how the control quality is performed:

- Within the *Pre-qualification* stage, the transcriber should go through a trust test. In fact, they have to perform 5 successful transcripts. Then, he will be allowed working. Otherwise the transcriber is invited to check the guideline once again, and every transcriber has 3 attempts before suspending their account.
- Once trusted, this is not for ever, the *Online Filtering* stage is activated. In fact, a verification process is launched after every 5 submitted transcripts. Where the system ask the transcriber to transcribe one speech among BR. Here also users are invited to check the guideline once again, if their scores are lowered. Users with score higher than or equal to 70% will be considered as a trusted transcriber so he will be tested every 10 transcriptions instead of 5.
- In parallel, the *WTT revision* step is launched. It is added to get more accurate transcriptions. In fact, the well trained transcribers, mainly the authors and lab members, reviewed the transcriptions submitted by users with score less than 70%. If the task needs a bit revision they performed it. Otherwise, they list the task again. Figure 2 shows WTT interface to validate/revise transcriptions.

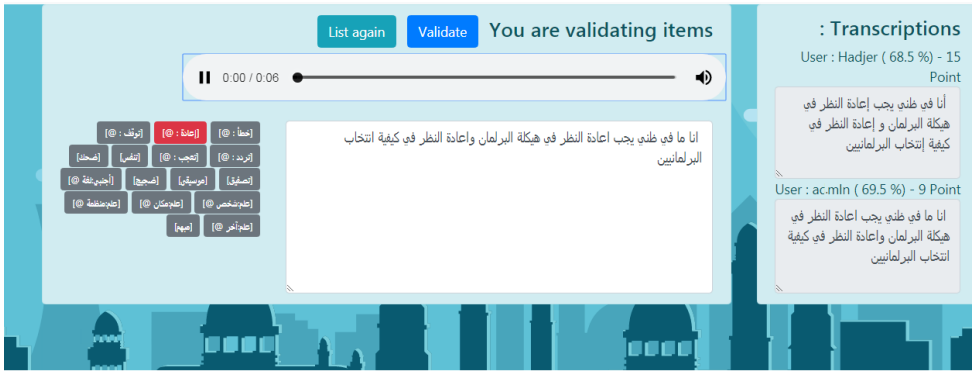

Figure 2: WTT Review and Validate Transcriptions Page

5.4 Project Data Evaluation and Aggregation

SPEECH2TEXT'DZ project was launched on May 2018. Contributors were invited to participate.

Total duration	51 Days
Number of crowd	208
Number of transcription	5335
Number of audio transcribed twice	277
Number of audio transcribed more than twice	312
Average Transcriptions per user	25.65
Guideline video views per day	33
Average Transcription time	3min 21s

Table 3: Global Statistics about the Project Execution.

After 51 days of web application hosting, more than 208 users registered. According to Google Analytics tool and our platform administration page we have got some statistics and details regarding user participation and behaviors. Table 3 gives global statistics about the project execution. In average a time of 3min 21s is needed for one transcription. This fact, shows that the transcription task is very challenging despite that utterance size is about 6.2s in average. This is also confirmed by the fact that in average a user transcribes less than 26 speeches. .

In order to ensure transcription quality, all works took less than 20 seconds are treated as malicious work and been consequently eliminated. Moreover and as explained, WTT can validate and review users transcriptions and list a task again if it is needed.

For evaluating the crowdsourcing solution, we consider the transcription quality by the crowd transcribers before the WTT revision stage. Table 4 shows the distribution of users according to their achieved scores and the related number of transcribed utterances.

Scores show that the well transcribed utterances were performed by less than 21 crowds. While the 73 transcribers reached a score between 60% and 80%.

The overall precision Pr achieved is computed using the following formula:

$$Pr = \frac{\sum_{i=1}^{NT} \#Utt_i * Score_i}{N}$$

Where N is the total number of transcribed utterances, NT the number of transcribers, $\#Utt_i$ and $Score_i$ are respectively, the number of transcribed utterances and the average score of a user i. Accordingly, we have got a precision about 74.38%. which can be considered as an acceptable result according to the challenging dialect transcription task.

Let us mention that after the WTT revision step all the transcription are considered as well transcribed according to the defined guideline.

Figure 3 illustrates a sample of transcriptions confronted to the well-trained transcribers' ones. We have observed that the most common mistakes and errors are due to the misunderstanding of guideline or also from the fact that users ignore watching the video tutorial that demonstrates how to transcribe and use the platform. Also some users misuse the defined tags, for example instead of using the tag [تردد] they used [تعجب].

%	55 <	55–60	60–70	70–80	> 80
# Users	33	81	39	34	21
# Transcribed Utterances	1136	1011	1186	936	1086

Table 4: Users Score Quality Rates and Transcriptions Distribution by Score.

Expert	بينما الواقع نتاع الرياضة [تردد] واش نقلك ؟ شباب طموح جدا جدا داير دورة رياضية
Crowd	بينما لواقع تاع رياض وشن نقلك شباب طموح جدا [عادة : جدا] دار دورة رياضا
Crowd	بينما لواقع نتاع رياضا واشن نقلك شباب طموح جدا [إعادة] دير دورة رياضا
Expert	كاين بزاف مواطنين مازالو ينيروا على الشموع [ضحيج]
Crowd	كاينة بزاف مواطنين مزالو ينيرو على شموع
Crowd	كاينا بزاف مواطنين مزلو ينرو على شمع
Expert	[تردد : ال] الموال راه عندو مصاريف يصرفها على هذي الشاه ما يحي يلحقها لوقت العيد حتان !
Crowd	الموال راه عندو مصاريف يصرفها على هذي الشاه ما يحي يلحقها لوقت العيد حتى
Crowd	[تردد : ال] الموال راه عندو مصاريف يصرفها على هذي الشاه ما يحي يلحقها لوقت العيد [توقف : حتان]

Figure 3: A Sample of Expert vs. Crowd Transcriptions.

6 Best Practices

Based on the experiments of this crowdsourcing-based solution and the resented results, we have dedicated some rules for a good validation of dialect transcription :

- Dialect speech transcription is a hard task, for that the size of the speech segments must be managed.
- Daily observation must be done to check the progress of completed tasks to recall new users when it needed.
- A part of quality control must be implemented on the project to avoid malicious work and get accurate result.
- The online filtering stage is very important to ensure quality control and avoid useless workload.
- The time of launching calls must be considered to get a large participation.

7 Conclusion and Future Work

For many researchers and institutions, crowdsourcing has become a popular method in NLP for lowering time and cost comparing to expert requirements. In this paper, we have investigated a paid crowdsourcing solution in order to transcribe a part of the speech utterances of KALAM'DZ corpus. We have followed two strategies to ensure the control quality of users transcriptions. First a predefined guideline is provided in order to help and train the crowd to deliver as normalized transcriptions as possible. The second control quality strategy is ensured using three control stages: Prequalification of transcribers, online filtering and revision step.

The results show that using crowdsourcing with a well tuned quality control mechanisms is an effective way for speech dialect transcription. In fact, the reached transcription results shows that the precision of the transcripts is more than 74.38% according to a baseline resource.

In addition, we have determined a list of best practices for crowdsourcing-based solutions for corpus transcription.

This crowdsourcing-based solution has proved its accuracy, in an ongoing work we are enlarging the dataset to be transcribed by improving the crowd recruitment strategy.

As future work, we plan to extend the usage of crowdsourcing in order to cover further annotation and validation to KALAM'DZ corpus such start POS tagging the sentences to build a treebank-like resource.

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.

Ahmed Ali, Peter Bell, James Glass, Yacine Messaoui, Hamdy Mubarak, Steve Renals, and Yifan Zhang. 2016. The MGB-2 Challenge: Arabic Multi-Dialect Broadcast Media Recognition. In *2016 IEEE Spoken Language Technology Workshop (SLT)*, pages 279–284. IEEE.

Ahmed Ali, Stephan Vogel, and Steve Renals. 2017. Speech Recognition Challenge in the Wild: Arabic

MGB-3. In *2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 316–322. IEEE.

Khalid Almeman, Mark Lee, and Ali Abdulrahman Almiman. 2013. Multi Dialect Arabic Speech Parallel Corpora. In *2013 1st International Conference on Communications, Signal Processing, and their Applications (ICCSPA)*, pages 1–6. IEEE.

Claude Barras, Edouard Geoffrois, Zhibiao Wu, and Mark Liberman. 1998. Transcriber: a Free Tool for Segmenting, Labeling and Transcribing Speech. In *First international conference on language resources and evaluation (LREC)*, pages 1373–1376.

Paul Boersma and Vincent Van Heuven. 2001. Speak and unSpeak with PRAAT. *Glot International*, 5(9/10):341–347.

Soumia Bougrine, Hadda Cherroun, and Ahmed Abdelali. 2017a. Altruistic Crowdsourcing for Arabic Speech Corpus Annotation. *Procedia Computer Science*, 117:137 – 144.

Soumia Bougrine, Hadda Cherroun, and Djelloul Ziadi. 2017b. Hierarchical Classification for Spoken Arabic Dialect Identification using Prosody: Case of Algerian Dialects. *CoRR*, abs/1703.10065.

Soumia Bougrine, Hadda Cherroun, Djelloul Ziadi, Abdallah Lakhdari, and Aicha Chorana. 2016. Toward a Rich Arabic Speech Parallel Corpus for Algerian sub-Dialects. In *The 2nd Workshop on Arabic Corpora and Processing Tools 2016 Theme: Social Media*, pages 2–10. European Language Resources Association (ELRA).

Soumia Bougrine, Aicha Chorana, Abdallah Lakhdari, and Hadda Cherroun. 2017c. Toward a Web-based Speech Corpus for Algerian Arabic Dialectal Varieties. In *Proceedings of the Third Arabic Natural Language Processing Workshop*, pages 138–146. Association for Computational Linguistics.

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)*, 5:73–92.

Khalid Choukri, Mahtab Nikkhou, and Niklas Paulsson. 2004. Network of data centres (NetDC): BNSC - an Arabic broadcast news speech corpus. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, pages 889–892. European Language Resources Association (ELRA).

European Language Resources Association (ELRA). 2005. A-SpeechDB ID: ELRA-S0315. `http://catalog.elra.info/en-us/repository/browse/ELRA-S0315/`. Accessed: 2018-10-30.

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

Hassan Gadalla, Hanaa Kilany, Howaida Arram, Ashraf Yacoub, Alaa El-Habashi, Amr Shalaby, Krisjanis Karins, Everett Rowson, Robert MacIntyre, Paul Kingsbury, David Graff, and Cynthia McLemore. 1997. CALLHOME Egyptian Arabic Transcripts. *Linguistic Data Consortium, Philadelphia*.

Hamilton Alexander Rosskeen Gibb, Johannes Hendrik Kramers, Évariste Lévi-Provençal, Bernard Lewis, Charles Pellat, Joseph Schacht, et al. 1986. *The Encyclopaedia of Islam*, new edition, volume 1, chapter Algeria. E. J. Brill, Leiden.

Nizar Habash, Mona Diab, and Owen Rambow. 2012. Conventional Orthography for Dialectal Arabic. In *Proceedings of the Language Resources and Evaluation Conference (LREC), Istanbul*, pages 711–718. European Language Resources Association (ELRA).

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):384–396.

Sameer Khurana and Ahmed Ali. 2016. QCRI advanced transcription system (QATS) for the Arabic Multi-Dialect Broadcast media recognition: MGB-2 challenge. In *2016 IEEE Spoken Language Technology Workshop (SLT)*, pages 292–298. IEEE.

Mohamed Maamouri, Tim Buckwalter, and Christopher Cieri. 2004. Dialectal Arabic Telephone Speech Corpus: Principles, Tool Design, and Transcription Conventions. In *NEMLAR International Conference on Arabic Language Resources and Tools, Cairo*, pages 22–23. Linguistic Data Consortium (LDC).

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.

Houda Saadane and Nizar Habash. 2015. A Conventional Orthography for Algerian Arabic. In *Proceedings of the Second Workshop on Arabic Natural Language Processing*, pages 69–79. Association for Computational Linguistics.

Marta Sabou, Kalina Bontcheva, Leon Derczynski, and Arno Scharl. 2014. Corpus Annotation through Crowdsourcing: Towards Best Practice Guidelines. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 859–866. European Language Resources Association (ELRA).

Mostafa Lameen Souag. 2006. *Explorations in the Syntactic Cartography of Algerian Arabic*. Ph.D. thesis, University of London, School of Oriental and African Studies, London.

James Surowiecki. 2004. *The wisdom of crowds*, first edition. Anchor Books, New York.

Samantha Wray, Hamdy Mubarak, and Ahmed Ali. 2015. Best Practices for Crowdsourcing Dialectal Arabic Speech Transcription. In *Proceedings of the Second Workshop on Arabic Natural Language Processing*, pages 99–107. Association for Computational Linguistics.

Inès Zribi, Mariem Ellouze, Lamia Hadrich Belguith, and Philippe Blache. 2015. Spoken Tunisian Arabic Corpus "STAC": Transcription and Annotation. *Research in computing science*, 90:123–135.

Inès Zribi, Marwa Graja, Mariem Ellouze Khmekhem, Maher Jaoua, and Lamia Hadrich Belguith. 2013. Orthographic Transcription for Spoken Tunisian Arabic. In *International Conference on Intelligent Text Processing and Computational Linguistics*, pages 153–163. Springer-Verlag Berlin Heidelberg.