Best Practices for Crowdsourcing Dialectal Arabic Speech Transcription

Samantha Wray, Hamdy Mubarak, Ahmed Ali
Qatar Computing Research Institute
Hamad Bin Khalifa University
Doha, Qatar
{swray,hmubarak,amali}@qf.org.qa

Abstract

In this paper, we investigate different approaches in crowdsourcing transcriptions of Dialectal Arabic speech with automatic quality control to ensure good transcription at the source. Since Dialectal Arabic has no standard orthographic representation, it is very challenging to perform quality control. We propose a complete recipe for speech transcription quality control that includes using output of an Automatic Speech Recognition system. We evaluated the quality of the transcribed speech and through this recipe, we achieved a reduction in transcription error of 1.0% compared with 13.2% baseline with no quality control for Egyptian data, and down to 4% compared with 7.8% for the North African dialect.

1 Introduction

Crowdsourcing is the process of segmenting a complex task into smaller units of work and distributing them among a large number of non-expert workers at a lower cost and for less time than professional companies.

The usage of popular crowdsource platforms such as Amazon Mechanical Turk (MTurk) and CrowdFlower (CF) for the acquisition, transcription, and annotation of speech data has been well demonstrated (Evanini et al., 2010; Parent and Eskenazi, 2010; Zaidan and Callison-Burch, 2011; Novotney and Callison-Burch, 2010; Marge et al., 2010b), among others. However, using crowdsourcing for the transcription of speech for languages with nonstandard orthographies is less explored, especially with regards the development of quality control protocols in the absence of established writing standards.

Although the writing system of Modern Standard Arabic (MSA) is standardized, the varieties of Dialectal Arabic (DA) are written without standard orthography, typically by utilizing the writing system of MSA. In this paper, we present best practices for crowdsourcing transcriptions of report and conversational DA and present results of experiments varying automatic quality control parameters that led to the creation of these best practices. We show that comparing output from an MSA-based Automatic Speech Recognition (ASR) system trained on a minimal amount of DA to output from a human transcriber outperforms other methods of quality control and results in low rates of data attrition. We show that utilizing a forgiving edit distance algorithm to compare ASR and user transcripts retains natural variation in orthographic usage without sacrificing quality.

This paper is organized as follows: in Section 2 we discuss issues in crowdsourcing written DA, with particular reference to the usage of nonstandard orthography. Section 3 outlines the utilization of professional transcription for DA and compares it in general terms to the usage of crowdsourcing for the same task. The DA audio data used in this study is described in detail in Section 4. Crowdsourcing experiments are detailed in Section 5, and best practices based on the results of these experiments are presented in Section 6. Section 7 summarizes our findings.

2 Challenges in Crowdsourcing DA

The nonstandard features of the written form of DA complicate efforts for designing effective quality control in crowdsourcing because many typical methods are not effective, as we outline here.

2.1 Overview of DA

MSA is the variety used for formal communication, and written materials such as books, newspapers, etc. while DA varieties are used for daily communication between people in the Arab world.

Proceedings of the Second Workshop on Arabic Natural Language Processing, pages 99–107,
Beijing, China, July 26-31, 2015. ©2014 Association for Computational Linguistics

Nowadays, there are many available resources for MSA such as corpora, morphological analyzers, Part Of Speech taggers, parsers, and so forth. However, there is still a need to build such resources for DA.

MSA resources do not typically perform well for handling DA. Darwish et al. (2014) showed that there are differences between MSA and the Egyptian dialect of DA at almost all levels: lexical, morphological, phonological, and syntactic.

Another challenge for DA is the nonstandard orthography, and words may be written in many different ways. For example, the future marker in Egyptian DA can be spelled with two different MSA characters: ه or ح . (For a complete overview of these issues, see Eskander et al. (2013)). There are some proposed rules for standardizing DA such as the Conventional Orthography for Dialectal Arabic (CODA) (Habash et al., 2012) which is very useful for many applications like ASR, and natural language processing (NLP). Although these effective tools and others (such as Zribi et al. (2014)) exist for training annotators to write DA in a particular way and for automatic normalization of text after the fact, our aims are to obtain a transcribed speech corpus which exhibits natural orthographic variation among speakers, so normalization tools would not be appropriate for this task.

2.2 Quality Control in Crowdsourcing

Crowdsourcing is now considered a promising alternative to the employment of transcription experts to create large corpora of transcribed speech in languages such as English (Lee and Glass, 2011; Marge et al., 2010b; Marge et al., 2010a; Hämäläinen et al., 2013), Spanish (Audhkhasi et al., 2011), Swahili, Amharic (Gelas et al., 2011), Korean, Hindi, and Tamil (Novotney and Callison-Burch, 2010).

One of the main challenges in crowdsourcing is quality control. There is great incentive to performing automatic quality control as opposed to leaving the cleaning of data to post-processing. Automatic quality control which issues warning messages to a user or rejects submission of spammy data reduces overall data attrition.

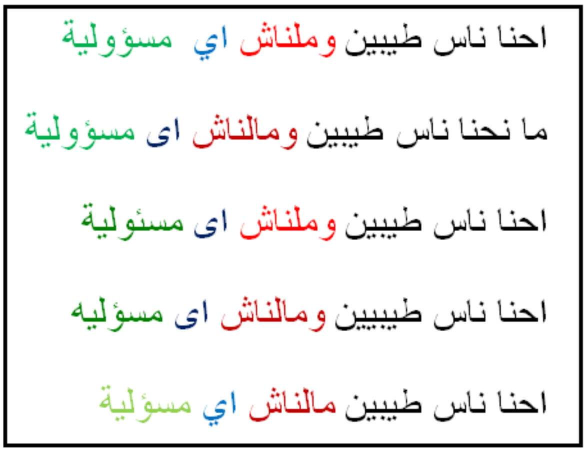
A typical way of performing automatic quality control is the usage of a gold standard to be used as test questions. Users having low quality with respect to these questions will be excluded and their work will be rejected.

Sprugnoli et al. (2013) compared different automatic quality control methods for crowdsourcing speech transcription for Italian and German:

- **The iterative dual pathway method**
  In this method, the speech segment is randomly assigned to four annotators in two independent pathways. When four transcriptions, two from each pathway, match each other, the segment is considered as transcribed correctly. The key advantage of this method is to have accurate transcriptions without having explicit quality control or preparing test questions.

- **The gold standard method**
  In this method, at least 10% of the segments are transcribed by experts and this is used to distinguish between trusted and untrusted transcribers.

These quality control methods cannot be applied to DA because there is no standard orthography and it may happen in many cases that there will not be exact match between annotators (first method) nor with the gold standard (second method). Figure 1 shows real transcription outputs for the same speech segment in which there is no single match between the whole transcription among transcribers because words in colors are written differently and all are correct.



احنا ناس طيبين وملناش اي مسؤولية

ما نحنا ناس طيبين ومالناش اى مسؤولية

احنا ناس طيبين وملناش اى مسئولية

احنا ناس طيبيين ومالناش اى مسؤليه

احنا ناس طيبين مالناش اي مسؤلية

Figure 1: Non Standard Orthography for transcribing DA

For the current study, we utilize CF which draws users from worker channels including microworking and rewards sites. CF has a robust userbase in the Arab world, and users can be selected by country of origin, which is an attractive option for studies which focus on regional DA varieties. CF also allows users to obtain a High Quality status, which allows task designers to target only High Quality users for a task. The opposing setting is High Speed which allows any user in the targeted country to complete the task. CF also has a built-in gold standard system which performs quality control. However, options for fuzzy text matching using the built-in system are extremely limited, and as outlined earlier, exact matching will not suffice for DA's nonstandard orthography.

Automatic quality control for translation and transcription tasks which do not rely on typical gold standard or multi-pass quality control methods include utilizing a series of checks which prevent submissions with text similar to the instructions of the task (Gelas et al., 2011) or which violate set word minimum/maximum sizes (Gelas et al., 2011), using a support vector machine (SVM) classifier to determine if a transcript is of good or poor quality (Lee and Glass, 2011), comparing to a language model built from an existing text corpus (Gelas et al., 2011; Zaidan and Callison-Burch, 2011) and utilizing vocabulary size of transcript (Lee and Glass, 2011).

For the current study, we employ typical gold standard question quality control, as well as two other methods: one which does not rely on the audio content, and one which does. The former relies on comparing a user's transcript to expected norms for legal Arabic text (following Gelas et al. (2011) for Swahili and Amharic.) The latter utilizes ASR output of each audio segment. This was explored by Lee and Glass (2011) in the form of integration of auxiliary features from ASR including the Word Error Rate (WER) for the top N best, as well as Phoneme Error Rate (PER) for the same number of hypotheses. They used these features as a form of automatic quality control to reject transcripts which deviated from expected input. One reason for such a system to perform reasonably is access to mature ASR such as for English in this case. However, for the current study Arabic ASR is still facing major challenges, and performance for DA ASR systems is behind even compared to MSA ASR, to say nothing of more mature ASR systems such as English and French. WER for Arabic ASR are appreciably higher than WER for these mature systems. (See Section 4.2 for a detailed look at Arabic ASR). Thus, in the current study, we quantify transcription quality based on edit distance from the expected string itself instead of relying on WER.

3 Crowdsourcing versus Transcription House

To determine the potential advantages in cost and speed of transcribing using CF, we submitted several tasks of Egyptian DA audio totaling approximately 10 min of speech and selected the High Speed option, which allows every user in the selected country to participate. We collected 5 transcripts for each item. On average, the task was completed after 3 hours from its launch and the total cost was USD 7. Our calculated cost of transcribing 1 hour of speech is USD 42 taking 18 hours.

When we transcribed the same audio using a professional company, the cost was USD 300 and it took 4 days. Furthermore, only one transcript per audio segment was provided.

It's clear from this comparison that using crowdsourcing reduces the cost and time significantly. Another important benefit is having multiple transcriptions and different writings which is very useful for building resources for DA such as corpora and morphological analyzers.

The high quality of data from crowdsourcing when compared to professional experts has been well explored not only here but also by Zaidan and Callison-Burch (2011) and Williams et al. (2011), and many others.

4 Data

The data for this study took two forms: the speech audio to submit for transcription and automatic transcription of that audio from an ASR system.

4.1 DA Speech Data

Audio for the transcription task was taken from debate and news programs uploaded to Al Jazeera's website between June 2014 and January 2015.

The audio underwent a series of preprocessing steps before being submitted to CF for transcription. Voice Activation Detection was performed to remove non-speech audio such as music or white noise, followed by processing using the LIUM SpkDiarization toolkit, which is a software package dedicated to speaker segmentation and clustering, diarization, and speaker linking within the same episode (Meignier and Merlin, 2010). The output from LIUM segmentation is typically small chunks of audio files containing information about speaker ID, and duration of utterance.

In addition, a crucial preprocessing step took place: classification of dialect group. This was performed using human computation, which also occurred via CF. Utterances underwent dialect classification by 3-9 annotators per audio file into five broad Arabic dialect groups: Modern Standard Arabic (MSA), Egyptian (EGY), Levantine (LEV), North African/Maghrebi (NOR), and Gulf (GLF). (For more details about this process, see Wray and Ali (2015).) For the current study, we used audio segments which had been classified as EGY with at least 75% agreement between annotators.

Egyptian data was chosen as a test case to perform experiments and determine best practices for transcription for several reasons. First, EGY as a category consists of a smaller and potentially less diverse set of dialects than a more geographically spread category. For example, the category NOR contains speech from Morocco, Libya, Tunisia, Algeria, and Mauritania. Because CF allows restriction of tasks to users in specific countries, by selecting Egypt using the platform and presenting annotators with audio from the EGY category, there was a greater chance of the transcriber speaking the same dialect as the speaker in the audio clip when compared to other dialect categories. Secondly, the demographics of the classification task (Wray and Ali, 2015) showed that approximately 40% of users of CF in the Arab world who participated were located in Egypt, meaning that focusing on EGY audio and Egyptian annotators allowed us to complete multiple iterations of transcription tasks with a quick turnover. We found that the amount of users in Egypt contributed to an average of 287 transcripts per hour as opposed to an average of 3 transcripts per hour for users in the Levant, for example. Finally, there were significantly more audio segments classified with high levels of inter-annotator agreement as EGY when compared to other dialect categories.

4.2 ASR Transcripts

All of the speech data was automatically transcribed using the QCRI Advanced Transcription System (QATS) (Ali et al., 2014c). This Arabic ASR system is a grapheme-based system with sequential Deep Neural Network (DNN) for Acoustic Modeling (AM) with feature space Maximum Likelihood Linear Regression (fMLLR) adaption for the first pass. A tri-gram Language Model (LM) is used for recognition and a four-gram LM is used for LM rescoring. The AM was trained using about 400 hours Broadcast News (BCN) data, containing a mix of MSA and DA (Walker et al., 2013). The LM was trained using a six year worth archive of Al Jazeera print news as well as some web crawling. The lexicon for the ASR is 1.2M words, with an Out Of Vocabulary (OOV) rate of 2.5% on the test set.

We evaluated the ASR system using a test set taken from Ali et al. (2014b), with the resulting Word Error Rate (WER) shown in Table 1. The WER for report data is 12.35% which is largely MSA data, and 29.8% for conversational speech containing a mix of DA and MSA. The combined WER for the mix of both report and conversational data is 25.4%. More details about the grapheme dialectal Arabic system can be found here (Ali et al., 2014a).

Rep.	Conv.	Comb.
12.35%	29.8%	25.4%

Table 1: Grapheme Arabic ASR System WER

Once automatic transcriptions were obtained, we also generated phoneme-level transcriptions using a phoneme-based Arabic ASR (Ali et al., 2014b) in order to split the audio into short segments. We have found for human transcription, it is better to keep speech segments short (3-6 sec) for a transcriber not to get confused or discouraged with a long segment. To split the audio, we used the phoneme-level output and cut at periods of silence of at least 300 milliseconds.

5 Transcription Experiments

To guide our ideas for the development of possible protocols for quality control to test, we first submitted approximately two hours of EGY audio to CF for transcription by users in Egypt over the course of a month and observed what kinds of errors existed in poor quality transcripts in the results. Examples of poor quality transcripts are shown in Table 2.

After development of potential quality control methods (covered in detail in subsection 5.1), we ran experiments to test their efficiency. To de-

Transcription	Case
ليهتحليهتنحليتنحليتنل	word len >MAX_LEN
تتتتتتتت	repeated letters
قغف غعفغ فغق	same keyboard row
ا ثغ اق اق	word len <MIN_LEN
بلاتؤبغ خسعء	invalid char n-gram
ةكنثال	letter ة must appear at the end of word
نعم	# words <MIN_WORDS
...... ——	non-Arabic characters
قم بكتابة رقم الملف	copy from job instructions
والله منور يا باشا	irrelevant text not related to audio segment

Table 2: Types of poor transcription

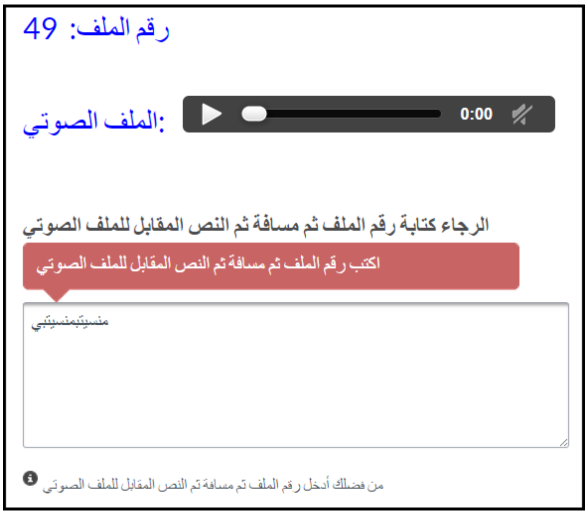
termine the highest performing protocol for quality control, we sampled 100 new audio segments of the EGY data described in Section 4 and submitted them to CF for transcription by users in Egypt. The 100 segments were submitted eight times: once for both High Quality and High Speed users for each of the four conditions described in the following section.

For each segment, five separate transcripts were collected from five different users. Users were presented with an audio button which they could press to listen to the audio an unlimited number of times, and a text box for entering the transcript. Users were directed to write as precisely as possible, to heed the item ID number, and to avoid using non-Arabic characters. Five items were presented per page, and completion of a page resulted in USD .05 compensation. An example of a single item as viewed by a user is shown in Figure 2.

5.1 Quality Control Parameters

Baseline Under the baseline condition, no quality control was performed. Users received direction on completion of the transcription tasks, but the input box did not issue a warning regardless of what the user typed into the box. Any text input was accepted by the system and users did not have a minimum set time required to be spent on the page, so they could submit after only a few seconds on the page.

Surface checks For this condition, we enabled a validation system that served two purposes: 1) a red notification flag with a warning to carefully


رقم الملف: 49

الملف الصوتي:

الرجاء كتابة رقم الملف ثم مسافة ثم النص المقابل للملف الصوتي

اكتب رقم الملف ثم مسافة ثم النص المقابل للملف الصوتي

منسيتبمنسيتبي

من فضلك أدخل رقم الملف ثم مسافة ثم النص المقابل للملف الصوتي

Figure 2: User view of single audio file and text box. The red flag contains a warning against writing gibberish, which has been entered in the input box.

follow directions appeared above the input box 2) the user was prevented from submitting the information entered on the page until whatever had triggered the warning was rectified. We accomplished this by using Javascript in CF's customization window to repurpose an existing CF validator (of which there are many, for example: must be a phone number) in order to satisfy our own conditions and display our own warning message. An example warning flag is shown in Figure 2.

The checks which triggered a flag under this condition were:

- 4 or more identical characters in a row
- fewer than 15 total characters
- url (to prevent copying and pasting of url into input box)
- lack of space character
- text from question display text (to prevent copying and pasting of task text into input box)

In addition to these checks, the user was required to spend a minimum of 40 seconds on the page before being allowed to submit. This time minimum was determined by observing the speed of completion of good quality transcripts from our original two hours of audio submitted during pilot work. We observed that users who submitted a page any quicker than this tended to submit spammy transcripts. Note that the Surface Checks condition did not rely on the existence of any gold standard or expected transcripts.

Expert-annotated checks Adopting traditionally-used methods of gold standard questions in crowdsource tasks, we obtained transcripts of 20% of the audio from a native speaker of Egyptian. These transcripts were incorporated into a validator which would issue a warning flag as described in the previous condition. We used Equation 1 in order to determine when to raise a warning flag and alert the user to be more careful:

$$diff(transcription) = \frac{dist}{refLen} \cdot 100 \qquad (1)$$

where $dist$ refers to Levenshtein edit distance[1] between the transcript and the reference (spaces are treated as characters), and $refLen$ refers to the length (in characters) of the expert-provided reference.

If $diff(transcription) \leq Threshold$, the transcript will be accepted.

$$Threshold = \begin{cases} 70\% & \text{for human transcript} \\ 80\% & \text{for ASR transcript} \end{cases}$$

When $Threshold$ = 70%, this means there should be at least 30% overlap with the reference. These thresholds were selected empirically based on observations of the number of different variations of writing words in DA.

Users were not aware which items would be compared to an existing transcript. If the item did not have an existing transcript (the remaining 80% of the data), the **surface checks** previously outlined were utilized.

ASR checks Recall that word-level transcripts were produced automatically by ASR (see Section 4). The ASR check condition also involved issuing a warning flag, but in contrast with the previous condition, every audio segment was compared to an expected input, and this time the expected transcript was produced by ASR. String overlap was also calculated using Equation 1, but to account for the higher WER for the ASR output than a human transcript, we lowered the threshold of overlap to 20% in comparison with the 30% overlap for expert-produced transcripts.

Because every item was compared to an automatic transcript, no other checks were utilized in this condition.

5.2 Results

A total of 149 users participated in the transcription tasks of EGY audio. The average WER for each user was calculated based on comparing each transcript to the four other user-provided transcripts for each item. As shown in Figure 3, there were different distributions of above-average and below-average users across conditions. In Figure 3, users were binned based on their personal average compared to the the averages of the whole sets.

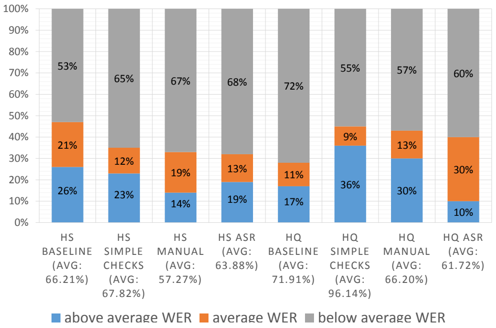


Figure 3: Average user WER for EGY audio across conditions

Given that the average WER for all conditions was very high, it was necessary to get a more complete picture about the true quality of the transcripts. (For comparison, typical WER for crowdsourced transcripts written in languages with standard orthography are around 5-25% (Parent and Eskenazi, 2010)). Therefore, the performance of the four quality control methods was evaluated by manually counting the number of poor quality transcripts accepted after five transcriptions from five different users had been collected for each of the 100 segments. Poor quality for our purposes was defined in the following two ways:

- Error: a transcript which was irrelevant, or gibberish. An irrelevant transcript contained valid Arabic text with no relation to the audio segment, and a gibberish transcript is one which contained strings of characters not considered to be a legal Arabic sequence.

[1] JavaScript implementation taken from: `https://gist.github.com/andrei-m/982927`

- Partial: a transcript which was truncated with respect to the appropriate output, for example a user who only wrote the first 3 words of a 6 word utterance.

Results comparing the four possible methods are shown in Table 3.

These results show that using ASR output as a comparison for every item outperformed other quality control methods for both High Speed and High Quality transcribers. In comparison to the baseline no quality control condition, the ASR check with only 20% string overlap between the transcription and the ASR output resulted in a total gain of 12.2% in error-free transcripts in the High Speed condition. The ASR comparison method is also a far more effective quality control method than using a human-annotated gold standard. Not only does ASR require less effort on the part of the researcher because it is automatically produced and does not require consulting a native speaker, it also outperforms the traditional use of interspersing gold standard questions which have been annotated by an expert (1% of transcripts with errors vs. 7.4% of transcripts with errors for High Speed users.) Overall, the highest performing option was using High Quality transcribers and the ASR output check, which resulted in 0.4% total errors, a reduction of 7.2% when compared to the baseline and 14.6% when compared to the worst performing condition.

It is interesting to note also that Surface Checks did not always result in cleaner data. Although for High Speed transcriptions, the total errors were reduced from 13.2% in the baseline to 10% by using Surface Checks, this trend did not continue for the High Quality condition. In fact, the total percentage of poor quality transcripts increased from 7.6% to 15%. Recall that 23% of users had above-average WER in the Simple Checks condition. However, further analysis showed that these users contributed 27% of the data. If an error-prone user happens to be prolific, and checks are not sufficient to stop the user's submissions, their errors may be over-represented.

5.3 NOR Speech Replication

To test the possibility of generalizing this method and utilizing it outside of Egypt, we replicated the experiments on another dialect group with a larger geographic spread. We selected 100 segments of NOR and submitted it under the four conditions on the High Speed option. The expert-annotated transcripts were written by a native speaker of Algerian. The CF task was then restricted to users in Morocco, Algeria, and Tunisia. Results of this replication are shown in Table 4.

As shown in Table 4, using the ASR output and comparing to every user input as a method of quality control shows that ASR still outperforms other methods of quality control for NOR audio just as for EGY audio. Compared to a baseline rate of 7.8% of poor quality transcripts, quality control using the ASR transcripts resulted in reduction to 4% total errors. Note again that Surface Checks resulted in a higher total percentage of poor quality transcripts from the 7.8% baseline to 9.4%. Higher still is the traditionally employed method of inserting random human-annotated transcripts for comparison as a gold standard, which has a total of 13.6% total of poor transcripts. As expected, these iterations happened to exhibit prolific above-average WER users (contributing 17% of the data for the Simple Checks condition and 19% of the data for the manual test questions condition, compared to z 15% of the data for the baseline condition and 15% of the data for the ASR condition.) However, even taking user variation into account, the ASR condition outperformed the baseline by 3.8%.

6 Best Practices

Based on the results presented in subsections 5.2 and 5.3, we have determined a working list of best practices for using a crowdsourcing platform such as CF for transcription of DA data:

- Segment audio files into smaller segments (from 3 seconds to 6 seconds each) such that transcription of each audio segment has a few words (more than 2 words, but less than 1 line of text).

- Restrict tasks to users in specific countries to match the required language skills needed for dialectal transcription.

- Perform dialect classification tasks or start with data for which the dialect is already known. When coupled with targeting users in a particular region, this will decrease the likelihood that a user is transcribing a dialect they are unfamiliar with.

EGY speech - 100 segments				
	Baseline	**Surface Checks**	**Manual edit distance**	**ASR edit distance**
High Speed				
Errors	1.6%	2.8%	1.2%	**0.6%**
Partial	11.6%	7.2%	6.2%	**0.4%**
Total	13.2%	10.0%	7.4%	**1.0%**
High Quality				
Errors	4.0%	12.2%	1.2%	**0.0%**
Partial	3.6%	2.8%	3.4%	**0.4%**
Total	7.6%	15.0%	4.6%	**0.4%**

Table 3: Percent of low quality transcripts across automatic quality control conditions

NOR speech - 100 segments				
	Baseline	**Surface Checks**	**Manual edit distance**	**ASR edit distance**
High Speed				
Errors	5.2%	3.6%	7.6%	**2.6%**
Partial	2.6%	5.8%	6.0%	**1.4%**
Total	7.8%	9.4%	13.6%	**4.0%**

Table 4: Percent of low quality transcripts across automatic quality control conditions

- Assign each audio segment an ID and ask annotators to write the ID and the transcript.
- Use JavaScript-defined or similar code for validation to check user input. First, check that the input ID is a valid ID. Then ASR output matched with the ID can be used to detect invalid transcription. In using this option, the acceptance threshold when using string matching should be lower than of human-written gold transcriptions to accommodate any limitations in ASR.
- Utilize automatic feedback to warn users to be more careful when they do not submit text that conforms to desired norms. In addition to simply warning, utilize automatic methods of preventing submission of poor data.
- Do not completely rely on quality control messages which do not refer to the content of the audio. Usage of quality control checks which aim to restrict input to a possible string of Arabic without consideration for the audio segment itself may result in the propagation of errors from irrelevant text
- After each job, generate statistics about the quality of all users (for example, how much agreement with other transcribers by calculating WER across transcriptions) and use the results to block low quality users from participating in future transcription tasks.

7 Summary

In this paper, we have shown that using the output of a publicly available ASR system trained on MSA and DA with an edit distance algorithm with a low threshold is an effective form of quality control in crowdsourcing transcriptions of a non-standard variety, namely Egyptian DA. We have also demonstrated the ability of using the same methodology on another dialect group, specifically North African DA. Currently, we are working to replicate our methods across all DA dialect groups to create a multi-dialectal DA speech corpus with both automatic and manual transcriptions.

References

Ahmed Ali, Hamdy Mubarak, and Stephan Vogel. 2014a. Advances in dialectal arabic speech recognition: A study using twitter to improve egyptian asr. In *International Workshop on Spoken Language Translation (IWSLT 2014)*, pages http–workshop2014.

Ahmed Ali, Yifan Zhang, Patrick Cardinal, Najim Dahak, Stephan Vogel, and Jim Glass. 2014b. A Complete Kaldi Recipe For Building Arabic Speech Recognition Systems. In *Spoken Language Technology Workshop (SLT), 2014 IEEE.*

Ahmed Ali, Yifan Zhang, and Stephan Vogel. 2014c. Qcri advanced transcription system (qats). In *Spoken Language Technology Workshop (SLT), 2014 IEEE.*

Kartik Audhkhasi, Panayiotis Georgiou, and Shrikanth S Narayanan. 2011. Accurate transcription of broadcast news speech using multiple noisy transcribers and unsupervised reliability metrics. In *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*, pages 4980–4983. IEEE.

Kareem Darwish, Hassan Sajjad, and Hamdy Mubarak. 2014. Verifiably Effective Arabic Dialect Identification. *EMNLP-2014.*

Ramy Eskander, Nizar Habash, Owen Rambow, and Nadi Tomeh. 2013. Processing spontaneous orthography. In *HLT-NAACL*, pages 585–595.

Keelan Evanini, Derrick Higgins, and Klaus Zechner. 2010. Using Amazon Mechanical Turk for transcription of non-native speech. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, pages 53–56. Association for Computational Linguistics.

Hadrien Gelas, Solomon Teferra Abate, Laurent Besacier, and François Pellegrino. 2011. Quality assessment of crowdsourcing transcriptions for African languages. In *Interspeech.*

Nizar Habash, Mona T Diab, and Owen Rambow. 2012. Conventional Orthography for Dialectal Arabic. In *LREC*, pages 711–718.

Annika Hämäläinen, Fernando Pinto Moreira, Jairo Avelar, Daniela Braga, and Miguel Sales Dias. 2013. Transcribing and annotating speech corpora for speech recognition: A three-step crowdsourcing approach with quality control. In *First AAAI Conference on Human Computation and Crowdsourcing.*

Chiaying Lee and James Glass. 2011. A transcription task for crowdsourcing with automatic quality control. In *Interspeech*, pages 3041–3044.

Matthew Marge, Satanjeev Banerjee, and Alexander I Rudnicky. 2010a. Using the Amazon Mechanical Turk for transcription of spoken language. In *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on*, pages 5270–5273. IEEE.

Matthew Marge, Satanjeev Banerjee, and Alexander I Rudnicky. 2010b. Using the Amazon Mechanical Turk to transcribe and annotate meeting speech for extractive summarization. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, pages 99–107. Association for Computational Linguistics.

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

Scott Novotney and Chris Callison-Burch. 2010. Cheap, fast and good enough: Automatic speech recognition with non-expert transcription. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, HLT '10, pages 207–215, Stroudsburg, PA, USA. Association for Computational Linguistics.

Gabriel Parent and Maxine Eskenazi. 2010. Toward better crowdsourced transcription: Transcription of a year of the let's go bus information system data. In *Spoken Language Technology Workshop (SLT), 2010 IEEE*, pages 312–317. IEEE.

Rachele Sprugnoli, Giovanni Moretti, Matteo Fuoli, Diego Giuliani, Luisa Bentivogli, Emanuele Pianta, Roberto Gretter, and Fabio Brugnara. 2013. Comparing two methods for crowdsourcing speech transcription. In *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, pages 8116–8120. IEEE.

Kevin Walker, Christopher Caruso, Kazuaki Maeda, Denise DiPersio, and Stephanie Strassel. 2013. *GALE Phase 2 Arabic Broadcast Conversation Speech.* Linguistics Data Consortium.

Jason D Williams, I Dan Melamed, Tirso Alonso, Barbara Hollister, and Jay Wilpon. 2011. Crowdsourcing for difficult transcription of speech. In *Automatic Speech Recognition and Understanding (ASRU), 2011 IEEE Workshop on*, pages 535–540. IEEE.

Samantha Wray and Ahmed Ali. 2015. Crowdsource a little to label a lot: Labeling a Speech Corpus of Dialectal Arabic. In *Interspeech.* (in press).

Omar F Zaidan and Chris Callison-Burch. 2011. Crowdsourcing translation: Professional quality from non-professionals. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pages 1220–1229. Association for Computational Linguistics.

Ines Zribi, Rahma Boujelbane, Abir Masmoudi, Mariem Ellouze, Lamia Belguith, and Nizar Habash. 2014. A Conventional Orthography for Tunisian Arabic. In *Proceedings of the Language Resources and Evaluation Conference (LREC), Reykjavik, Iceland.*