Building a Rich Arabic Speech and Language Corpus Based on the Holy Quran

Ali Meftah[1], Yasser Seddiq[1,2(✉)], Yousef Alotaibi[1], and Sid-Ahmed Selouani[3]

[1] College of Computer and Information Sciences, King Saud University, Riyadh, Saudi Arabia
{ameftah, yaalotaibi}@ksu.edu.sa
[2] King Abdulaziz City for Science and Technology (KACST), Riyadh, Saudi Arabia
yseddiq@kacst.edu.sa
[3] LARIHS Lab, Université de Moncton, Campus de Shippagan, Shippagan, Canada
selouani@umcs.ca

Abstract. This paper pursues the goal of creating a reliable speech corpus based on The Holy Quran (THQ) audio recordings. Achieving that goal involves major steps to be done and essential requirements to be considered. With the availability of tremendous amount of recordings nowadays, it is of a fundamental importance to select the ones that feature both high audio quality and perfect reciter performance. Also, since the targeted beneficiaries from the corpus are the digital speech processing research community, it is also very essential to maintain an efficient, a familiar and a convenient way of presenting the audio corpus and other language material, such as the language model. Audio recordings of THQ are selected from four sources having a high standard regarding the reciters' performance. A significant effort is made in phonetical transcription of the audio content such that the written transcript maps perfectly to the uttered phonemes. Furthermore, the corpus dictionary, which is usually required in many fields such as machine learning and datamining, is also created. The first release of the corpus consists of recorded recitations and the necessary metadata of three chapters of THQ of different lengths recited by four reference reciters. Those chapters are selected for this phase based on statistical analysis of the lengths of all chapters and the frequency of occurrence of the Arabic phonemes across all chapters of THQ.

Keywords: The Holy Quran · Speech corpus · Arabic speech processing

1 Introduction

Speech corpora form the solid foundation for any research on data mining and/or speech processing. Speech corpora are language specific and researchers who target a certain language should consider selecting the proper corpus of that language very seriously. They should also give high priority to investigating the available corpora for that language in order to assess accessibility, richness, correctness, and quality of those corpora. Creating new corpora and enhancing the existing ones are both valuable

A. Lachkar et al. (Eds.): ICALP 2017, CCIS 782, pp. 90–101, 2018.
https://doi.org/10.1007/978-3-319-73500-9_7

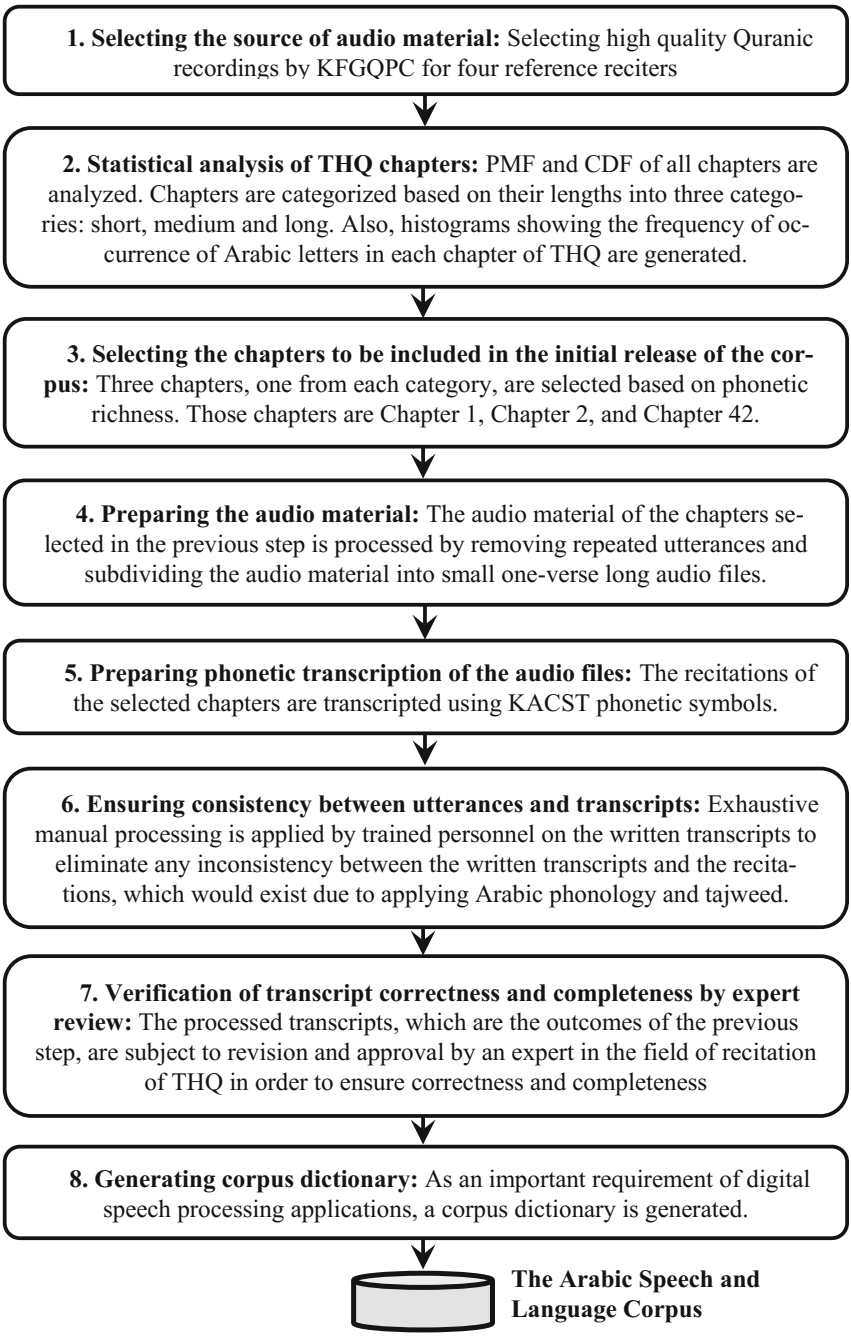


Fig. 1. The process of creating the Arabic Speech and Language Corpus.

contributions that researchers would make to their languages. Building a spoken corpus requires a good audio material, and an efficient approach that allows users to extract a comprehensive language model. The content of a good corpus should also be representative of the phonology and phonetics of a given language. Numerous Arabic speech corpora are available for the research community. For instance, King Abdulaziz City for Science and Technology (KACST) Arabic Phonetic Database [1], The Saudi Accented Arabic Voice Bank [2], The BBN/AUB Corpus of the Levantine dialect [3], and the West Point Corpus of native and non-native speakers [4], just to name a few. However, these corpora were designed for very narrow and specific application. Therefore, the research community in Arabic speech processing field is still looking forward to having access to more comprehensive corpora that would enable researchers to conduct exhaustive studies dedicated to Arabic language.

This work aims at contributing to the enrichment of the Arabic linguistic resources that could be used in various fields of Arabic speech and language processing. This paper presents an Arabic speech corpus based on the recorded recitations of The Holy Quran (THQ) and describes the process and the criteria that we follow to select the most suitable recitations amongst the tremendous amount of recordings that are publicly available nowadays. The process of creating the corpus is illustrated in the chart in Fig. 1. The paper presents detailed description of those steps. The first stage of the corpus creation consists of providing a representative subset of THQ audio and language model and other resources contents. That subset is selected on a statistical basis to ensure audio material adequacy. The outcomes of this first stage of the project are reported.

2 Creating an Arabic Speech Corpus Based on THQ

All audio recordings that we have access to are to be qualified for suitability for the corpus. Two quality criteria are considered: reciter performance and precise written scripts. Therefore, choices are made from the THQ recordings produced by the King Fahd Glorious Quran Printing Complex (KFGQPC) [5], which is an official Saudi government authority responsible for producing authenticated prints and recordings of THQ. Not only those recordings are made under controlled environment to maintain highest acoustic quality, but also the reciters are well selected to ensure correctness of pronunciation performance. Having access to such material is of great value in paving the way for the subsequent activities towards creating the corpus. Four sources from KFGQPC by considering the following reciters: Abdullah Ali Basfar (R01), Mohammed Ayub (R02), Ali Alhuthaifi (R03), and Ibrahim Alakhdhar (R04) are selected in addition to text form reflecting the pronounced text. This audio material is of high quality and deemed very suitable to serve the purpose of creating a corpus.

The four sources are analyzed by collecting statistics about the frequency of occurrence of the Arabic letters. Each audio source should match the histogram illustrated in Fig. 2 that is based on a written script of THQ. Arabic letters and phonemes are transcribed using KACST symbols [6] henceforth as listed in Table 1.

While the corpus ultimatly targets all content of THQ, at this stage, only a part of THQ is covered. Since chapters are not equal in length, we chose to include sample chapters of various lengths for the initial release of the corpus. The distribution of THQ

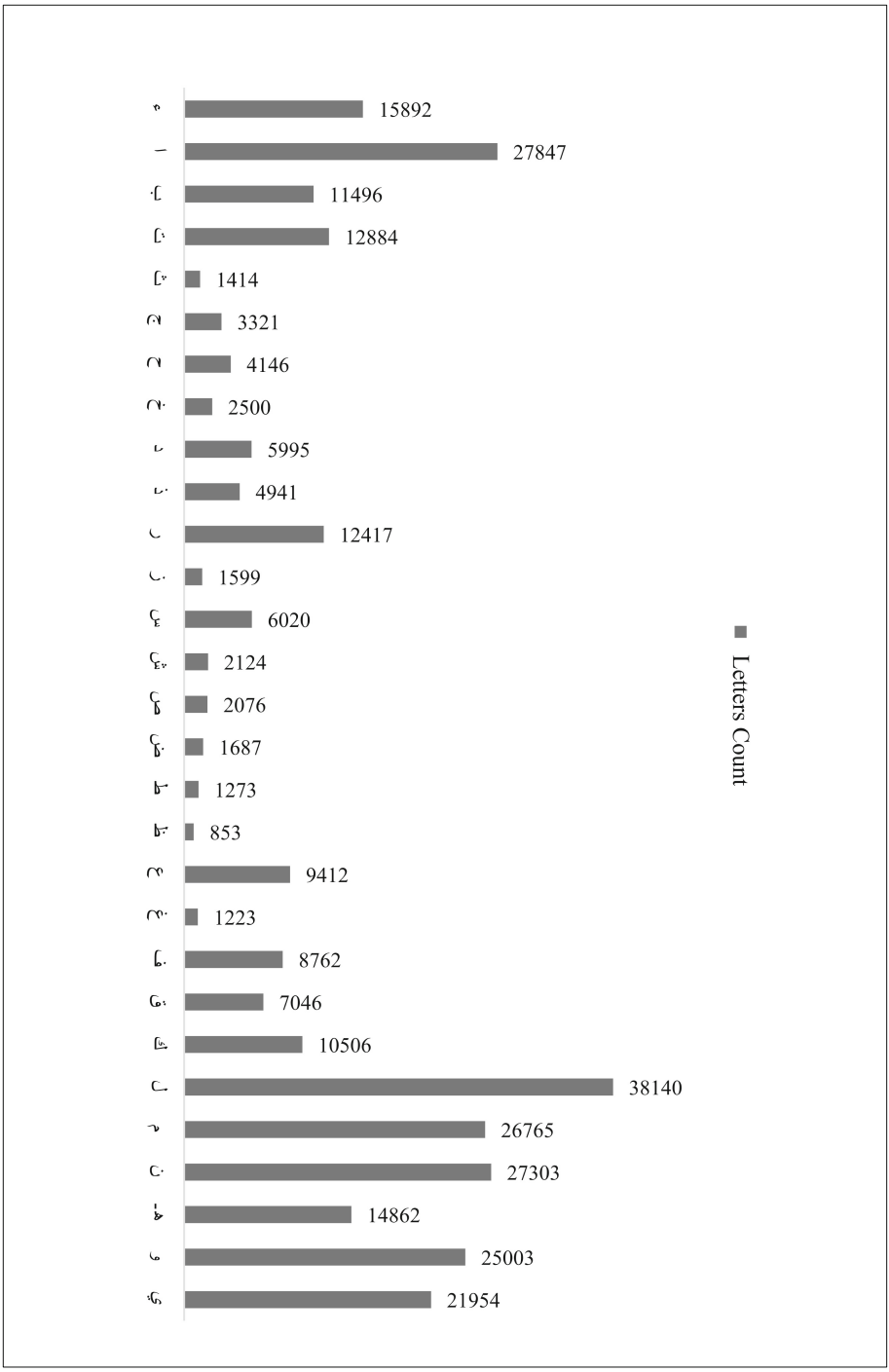
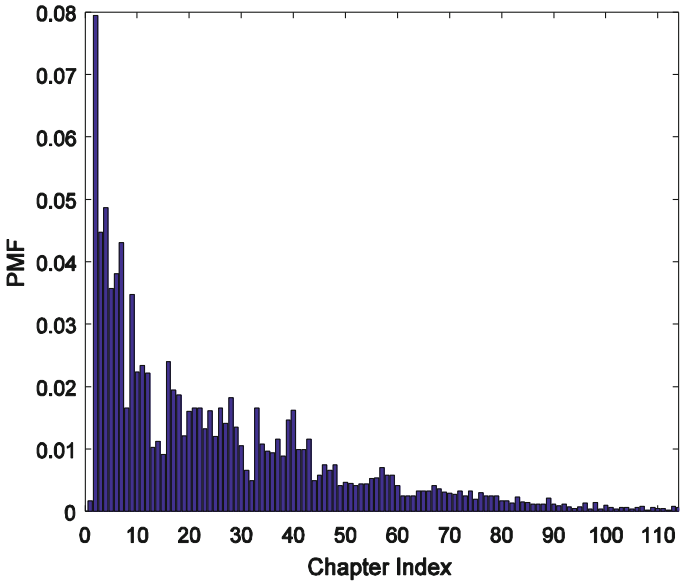


Fig. 2. Statistics of the Arabic letters (written) across all chapters of THQ

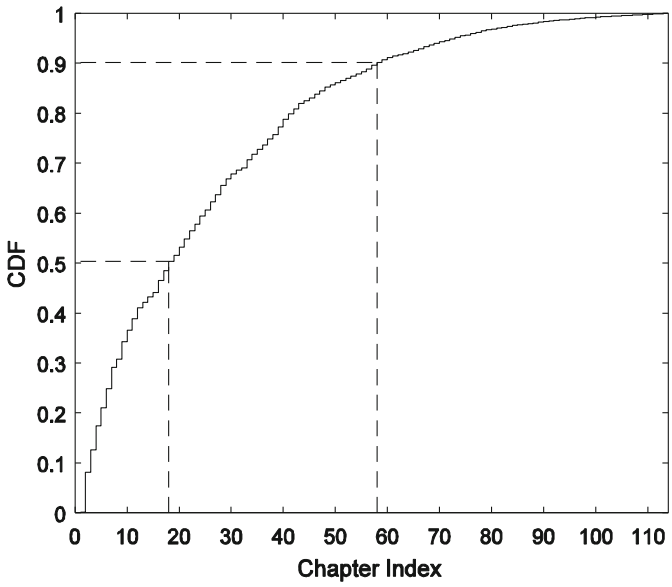
Table 1. KACST phonetic symbols

No.	Arabic Writing	KACST Symbols	IPA Symbols
1	ء	hz10	ʔ
2	ب	bs10	b
3	ت	ts10	t
4	ث	vs10	θ
5	ج	jb10	ʤ
6	ح	hb10	ħ
7	خ	xs10	χ
8	د	ds10	d
9	ذ	vb10	ð
10	ر	rs10	r
11	ز	zs10	z
12	س	ss10	s
13	ش	js10	ʃ
14	ص	sb10	$s^{?}$
15	ض	db10	$d^{?}$
16	ط	tb10	$t^{?}$
17	ظ	zb10	$ð^{?}$
18	ع	cs10	ʕ
19	غ	gs10	ɣ
20	ف	fs10	f
21	ق	qs10	q
22	ك	ks10	k
23	ل	ls10	l
24	لا	lb10	l̪
25	م	ms10	m
26	ن	ns10	n
27	ه	hs10	h
28	و	ws10	w
29	ي	ys10	j
30	ـَ	as10	a
31	ـُ	us10	u
32	ـِ	is10	i
33	ـا	as20	aa
34	ـو	us20	uu
35	ـي	is20	ii

content over all 114 chapters is analyzed by means of the probability mass function (PMF) given in Fig. 3(a) and the cumulative distribution function (CDF) given in Fig. 3(b). According to the CDF, half of THQ content covers only 18 chapters (16% of the chapters). We consider those as long chapters. It is worth mentioning that while this



(a)



(b)

Fig. 3. THQ content with respect to the 114 chapters illustrated using (a) PMF and (b) CDF.

group consists of long chapters, there is an exception of Chapter 1 that is deemed as one of the shortest chapters of THQ. Thus, this chapter will be treated as a member of the group of short chapters to be clarified next. A second group consisting of the last 56 chapters plus Chapter 1 (50% of the chapters) contributes to only 10% of THQ content implying that those are short chapters. In between, there are Chapters 19 to 58 (40 chapters forming 35% of the total chapters) contribute to 40% of THQ, which indicates a medium-length chapters. We decide to select one chapter from each group (long, medium and short) for the initial release of the THQ corpus that we are creating.

From Fig. 2, the least frequent letter in THQ is zb10 (ظ) that occurs 853 times followed by gs10 (غ) and tb10 (ط) that occur 1223 and 1273 times, respectively. Thus, we pay high attention to those least occurring letters when selecting the chapter that we should start with. The occurrence of those letters across the 25 chapters in focus is investigated and summarized in Fig. 4. The selection is made on a chapter that has balanced yet high frequency of occurrence of those three letter, which is Chapter 42 (Alshoura). Beside this chapter, the longest chapter (Chapter 2: Albaqarah) and one of the short chapters (Chapter 1: Alfatehah) are considered.

After selecting the audio material chapters, they are partitioned into reasonably short audio files. Each audio file should contain one complete verse. The maximum length of an audio file is chosen to be the period needed to recite three lines of the written script of THQ based on the KFGQPC print. However, there are verses that are long and some of them could span one page. Such verses are further partitioned such that each part does not exceed the specified maximum length.

The audio files are also processed to ensure consistency of content produced by each reciter. That is, because reciters are allowed to repeat some parts of text whenever appropriate, those repetitions result in inconsistence phoneme histograms across the four sources. Therefore, audio material is traced for repeated text that is eliminated whenever found.

The audio file names are assigned according to this following code: *D06N01SxxxAxxxASxRxxTxx*, where the name is decoded as follows:

- ***D06***: indicates the corpus serial number.
- ***N01***: indicates that the current recitations follow the narration of Assem AlKoofi. This is one of ten different narrations of THQ named after the scholars Assem AlKoofi, Ibn Katheer almakki, Nafea AlMadni, Abu Jaafer AlMadni, Abu Amro AlBassry, Hamzah AlKoofi, Ibn Amer AlShami, AlKessaei AlKoofi, Yaqoob AlBassry, and Khalaf bin Hesham [7].
- ***Sxxx***: is the chapter number.
- ***Axxx:*** is the verse number.
- ***ASx:*** indicates the partition number of a verse in case if it is partitioned due to exceeding maximum length as explained earlier. In case of a one complete verse, this code is set to AS0.
- ***Rxx:*** is the reciter index, where x ranges between 1 and 4.
- ***Txx:*** is the trial number.

For example filename D06N1S042A016AS0R02T01 indicates that the Ayaa (phrase) number 16 from the chapter 42 (Alshoura) read it by the reciter number 2 (Mohammad Ayyub) in trial 1 used the narrator Hafss.

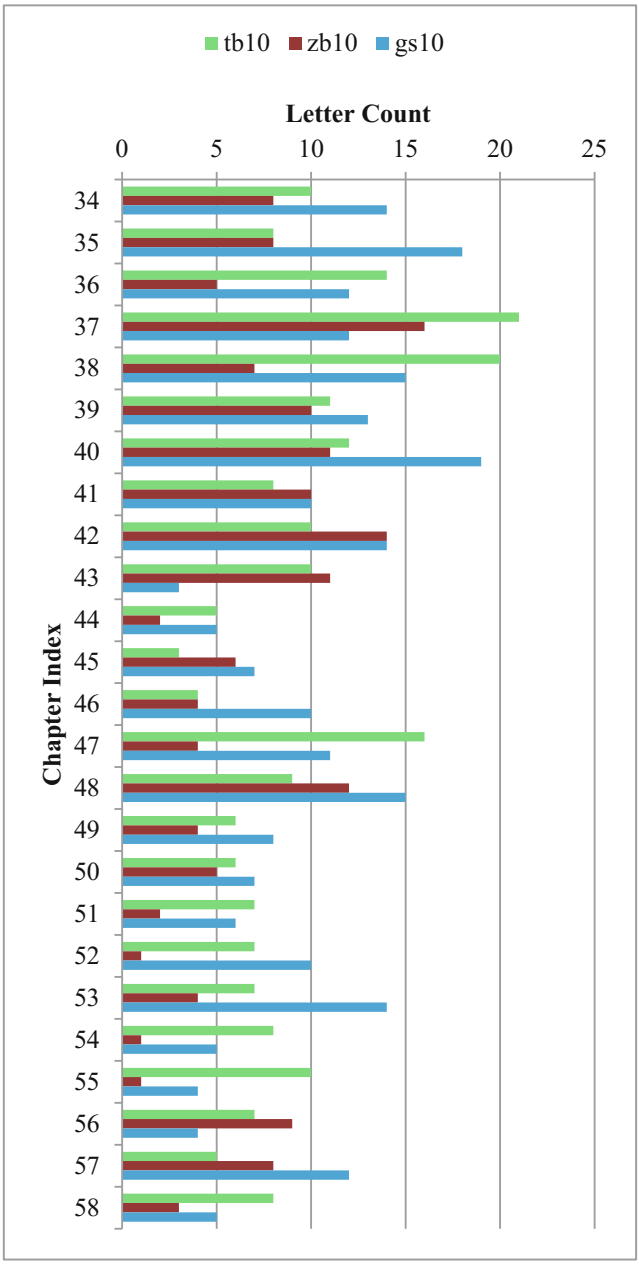


Fig. 4. Histogram of the three least frequent letters in THQ

3 Corpus Metadata

The corpus must contain text-format metadata describing the transcription of the audio material. In this section, we describe the work of preparing the metadata of the corpus. The process starts by using an electronic copy the script of THQ in text format that is published by KFGQPC that is written according to the Othmanic orthography. Indeed, in Arabic, spoken text does not map perfectly to the corresponding written text since there are letters that appear in written text but are not uttered and vice versa. In the case of THQ, there are also the *tajweed* rules that also inforce reciters to alter some written letters. That inconsistency between what is written and what is spoken must be addressed in the corpus because it primary targets speech processing. Therefore, the text metadata of the corpus must adhere perfectly to the uttered speech. It is very important to emphasis on the fact that this way of writing Arabic text is not correct from rules of writing perspective, but in the case of speech corpora this is acceptable because this transcription will be exclusively read by computers not by humans. Moreover, that transcription is written in phonetic alphabets not in ordinary alphabets. In this work, we use the KACST phonetic symbols that are illustrated in Table 1.

A famous inconsistency in Arabic speech is the effect when uttering /hz10 ls10/(ال), which means "the", followed by one of the Solar Letters (also called Sun Letters) [8]. In such case the phoneme /ls10/(ل) is not uttered. For instance, the word (و السماء) meaning "and the sky" is transcripted without the /ls10/(ل) as follows: /ws10 as10 ss20 ms10 as20 hz10 as10/in KACST symbols, which is equivalent to /wassama:ʔ/in IPA. Another case is converting a written (ب) to an uttered /ms10/(م) whenever the former is preceded by (ن). This effect is called *Eqlaab* in Tajweed terminology. Beside the aforementioned two effects, there are many other effects in Arabic and Tajweed such as Tanween, Ghunnah, Edgham and Ekhfaa. All these effects are considered in the process of text transcription of the current corpus.

The transcription process described above is done by a qualified personnel. All the subsequent stages and material are based on the outcomes of this stage. Hence, error-free transcripts must be delivered by end of this fundamental stage. The correctness of the text transcripts is assured by passing an expert review. Finally, the text-format transcripts are read to be added to the corpus material.

4 Corpus Dictionary

An important part of the corpus metadata is the dictionary of the corpus, which is a lookup table listing all unique vocabulary used in the corpus. Each word in the dictionary is transcribed using phonetic symbols and presented in format that can be recognized by ASR systems. In this corpus, the dictionary data is organized in three columns: the original word written in Arabic, the word transcribed using English alphabets, and phonetic transcription of the word using KACST phonetic symbols. A sample of the dictionary is illustrated in Table 2.

Table 2. Sample entries of THQ corpus dictionary

THQ Words		
English	Arabic	Pronunciation
yajtabiii	يَجْتَبِيّ	ys10 as10 jb10 ts10 as10 bs10 is61 sp
yajtanibuwna	يَجْتَنِبُونَ	ys10 as10 jb10 ts10 as10 ns10 is10 bs10 us21 ns10 as10 sp
yajmau	يَجْمَعُ	ys10 as10 jb10 ms10 as10 cs10 us10 sp
yaxtim	يَخْتِمْ	ys10 as10 xs10 ts10 is10 ms10 sp
yaxluqu	يَخْلُقُ	ys10 as10 xs10 ls10 us10 qs10 us10 sp
yashaA	يَشَآ	ys10 as10 js10 as10 hz10 sp
yashaAi	يَشَآءِ	ys10 as10 js10 as10 hz10 is10 sp
yashaaaAu	يَشَآءُ	ys10 as10 js10 as61 hz10 us10 sp
yashaaaAuwna	يَشَآءُونَ	ys10 as10 js10 as61 hz10 us21 ns10 as10 sp
yaGfiruwna	يَغْفِرُونَ	ys10 as10 gs10 fs10 is10 rs10 us21 ns10 as10 sp

The outcome of this work is the first release of the THQ corpus. It contains 346 phrases, with a total of 7,033 spoken words and 44,359 phonemes. Chapter 1 includes seven phrases that contain 29 spoken words, Chapter 42 contains 53 phrases with 860 spoken words, and Chapter 2, which is the longest chapter of THQ, contains 286 phrases containing 6,144 spoken words.

5 Conclusions and Perspective

A phonetically rich Arabic speech corpus was created. Its content is based on the audio material of THQ. Recitations of four reputed reciters were carefully selected. The first release of the corpus consists of a representative subset of THQ audio content and other language resources. Namely, Chapters 1, 2 and 42 were selected for this phase on the basis of a statistical analysis of the CDF of the lengths of all chapters and the occurrence frequency of the Arabic phonemes across all chapters of THQ.

A significant effort was made to achieve perfect phonetical transcription eliminating all inconsistencies between the written script and the uttered phonemes. The existence of such inconsistency is normal due to reciters adherence to Arabic phonology and the tajweed rules. A very important step towards achieving our goal was ensuring that the corpus transcription material passed expert revision before confirming its correctness and validity for digital speech processing applications. Moreover, in order to make the corpus useful for machine learning, data mining applications and ASR, a corpus dictionary was also created. The dictionary entries consist of all unique words in the covered material.

The outcome of this work is a speech corpus consisting of 7,033 spoken words equivalent to 44,359 phonemes. Chapter 1 contains 29 spoken words, Chapter 42 contains 860 spoken words, and Chapter 2, which is the longest chapter of THQ, contains 6,144 spoken words.

This work emphasizes on the process that was applied in creating the current corpus. The steps illustrated in Fig. 1 are of great importance to ensure the soundness and completeness of speech corpora that are based on THQ.

The ultimate goal of this project is to include all chapters of THQ in the corpus. The future work will involve the rest of THQ in a similar procedure that is applied in this phase before introducing the complete corpus.

Acknowledgment. This project was funded by the National Plan for Science, Technology and Innovation (MAARIFAH), King Abdulaziz City for Science and Technology, Kingdom of Saudi Arabia, Award Number (11-INF1968-02).

References

1. Alghamdi, M.: KACST Arabic phonetic database. In: The 15th International Congress of Phonetics Science; 3–9 August 2003; Barcelona, Spain, pp. 3109–3112. Universitat Autònoma de Barcelona, Barcelona (2003)
2. Alghamdi, M., Alhargan, F., Alkanhal, M., Alkhairy, A., Eldesouki, M., Alenazi, A.: Saudi accented Arabic voice bank. J. King Saud Univ. **20**, 45–64 (2008)

3. BBN Technologies (with American University of Beirut a subcontractor), et al.: BBN/AUB DARPA Babylon Levantine Arabic Speech and Transcripts LDC2005S08, 1st edn. Linguistic Data Consortium, Philadelphia (2005)
4. LaRocca, S., Chouairi, R.: West Point Arabic Speech LDC2002S02, 1st edn. Linguistic Data Consortium, Philadelphia (2002)
5. King Fahd Glorious Quran Printing Complex. http://www.qurancomplex.org/
6. Alghamdi, M., Mohamed El Hadj, Y.O., Alkanhal, M.: A manual system to segment and transcribe Arabic speech. In: IEEE International Conference on Signal Processing and Communications (ICSPC), Dubai, United Arab Emirates, pp. 233–236 (2007)
7. AlQahtany, M.O., Alotaibi, Y.A., Selouani, S.-A.: Analyzing the seventh vowel of classical Arabic. In: 2009 International Conference on Natural Language Processing and Knowledge Engineering, NLP-KE 2009, pp. 1–7 (2009)
8. Definition of sun letter, Merriam-Webster Dictionary. https://www.merriam-webster.com/dictionary/sun%20letter. Accessed 24 May 2017