A Pilot Study on Arabic Multi-Genre Corpus Diacritization Annotation

Houda Bouamor,[1] **Wajdi Zaghouani,**[1] **Mona Diab,**[2] **Ossama Obeid,**[1]
Kemal Oflazer,[1] **Mahmoud Ghoneim,**[2] **and Abdelati Hawwari** [2]
[1]Carnegie Mellon University in Qatar; [2]George Washington University
{hbouamor,wajdiz,owo}@qatar.cmu.edu; ko@cs.cmu.edu
{mtdiab,mghoneim,abhawwari}@gwu.edu

Abstract

Arabic script writing is typically underspecified for short vowels and other mark up, referred to as diacritics. Apart from the lexical ambiguity found in words, similar to that exhibited in other languages, the lack of diacritics in written Arabic script adds another layer of ambiguity which is an artifact of the orthography. Diacritization of written text has a significant impact on Arabic NLP applications. In this paper, we present a pilot study on building a diacritized multi-genre corpus in Arabic. We annotate a sample of non-diacritized words extracted from five text genres. We explore different annotation strategies: *Basic* where we present only the bare undiacritized forms to the annotators, *Intermediate* (Basic forms+their POS tags), and *Advanced* (automatically diacritized words). We present the impact of the annotation strategy on annotation quality. Moreover, we study different diacritization schemes in the process.

1 Introduction

One of the characteristics of writing in Modern Standard Arabic (MSA) is that the commonly used orthography is mostly consonantal and does not provide full vocalization of the text. It sometimes includes optional diacritical marks (henceforth, diacritics or vowels). Diacritics are extremely useful for text readability and understanding. Their absence in Arabic text adds another layer of lexical and morphological ambiguity. Naturally occurring Arabic text has some percentage of these diacritics present depending on genre and domain. For instance, religious text such as the Quran is fully diacritized to minimize chances of reciting it incorrectly. So are children's educational texts. Classical poetry tends to be diacritized as well. However, news text and other genre are sparsely diacritized (e.g., around 1.5% of tokens in the United Nations Arabic corpus bear at least one diacritic (Diab et al., 2007)).

From an NLP perspective, the two universal problems for processing language that affect the performance of (usually statistically motivated) NLP tools and tasks are: (1) sparseness in the data where not enough instances of a word type are observed in a corpus, and (2) ambiguity where a word has multiple readings or interpretations. Undiacritized surface forms of an Arabic word might have as many as 200 readings depending on the complexity of its morphology. The lack of diacritics usually leads to considerable lexical ambiguity, as shown in the example in Table 1, a reason for which diacritization, aka vowel/diacritic restoration, has been shown to improve state-of-the-art Arabic automatic systems such as speech recognition (ASR) (Kirchhoff and Vergyri, 2005) and statistical machine translation (SMT) (Diab et al., 2007). Hence, diacritization has been receiving increased attention in several Arabic NLP applications.

In general, building models to assign diacritics to each letter in a word requires a large amount of annotated training corpora covering different topics and domains to overcome the sparseness problem. The currently available diacritized MSA corpora are generally limited to the newswire genres (as distributed by the LDC) or religion related texts such as the Quran or the Tashkeela corpus.[2] In this paper we present a pilot study where we annotate a sample of non-diacritized text extracted from five different text genres. We explore different annotation strategies where we present the data to the annotator in three modes: **Basic** (only forms with no diacritics), **Intermediate** (Basic forms+POS tags), and **Advanced** (a list of forms that is automatically diacritized). We show the impact of the annotation strategy on the annota-

[2]Tashkeela is publicly available at: http://sourceforge.net/projects/tashkeela/

Proceedings of the Second Workshop on Arabic Natural Language Processing, pages 80–88,
Beijing, China, July 26-31, 2015.

Undiacritized	Diacritized	Buckwalter[1]	English
ذكر	ذَكَرَ	/*akara/	he mentioned
ذكر	ذُكِرَ	/*ukira/	it/he was mentioned
ذكر	ذَكَّرَ	/*ak ˜ara/	he reminded
ذكر	ذُكِّرَ	/*uk ˜ira/	it was reminded
ذكر	ذَكَرٌ	/*akaruN/	male
ذكر	ذِكْرٌ	/*ikaruN/	prayer

Table 1: Possible pronunciations and meanings of the undiacritized Arabic word *kr ذكر

tion quality. It has been noted in the literature that complete diacritization is not necessary for readability Hermena et al. (2015) as well as for NLP applications, in fact, (Diab et al., 2007) show that full diacritization has a detrimental effect on SMT. Hence, we are interested in eventually discovering an effective optimal level of diacritization. Accordingly, we explore different levels of diacritization. In this work, we limit our study to two diacritization schemes: FULL and MIN. For FULL, all diacritics are explicitly specified for every word. For MIN, we explore what a minimum and optimal number of diacritics that needs to be added in order to disambiguate a given word in context would be with the objective of making a sentence easily readable and unambiguous for any NLP application.

The remainder of this paper is organized as follows: In Section 2 we describe Arabic diacritics and their usage; In Section 3, we give an overview of the automatic diacritization approaches conducted mainly on news data and for a targeted application; We present the dataset used in our experiments in Section 4, followed by a description of the annotation procedure 5; Our analysis of the fully diacritized data, FULL, is provided in Section 6; In Section 7, we present a preliminary exploration of a MIN diacritization scheme; We finally draw some conclusions in Section 8.

2 Arabic Diacritics

Arabic script consists of two classes of symbols: letters and diacritics. Letters comprise long vowels such as A, y, w as well as consonants. Diacritics, on the other hand, comprise short vowels, gemination markers, nunation markers, as well as other markers (such as hamza, the glottal stop, which appears in conjunction with a small number of letters, e.g., أ, إ, آ, etc., dots on letters, elongation and emphatic markers)[3] which in all, if present, render a more or less precise reading of a word. In this study, we are mostly addressing three types of diacritical marks: short vowels, nunation, and shadda (gemination). Short vowel diacritics refer to the three short vowels in Modern Standard Arabic (MSA)[4] and a diacritic indicating the explicit absence of any vowel. The following are the three vowel diacritics exemplified in conjunction with the letter م/m: مَ/ma (fatha), مُ/mu (damma), مِ/mi (kasra), and مْ/mo (no vowel aka sukuun). Nunation diacritics can only occur word finally in nominals (nouns, adjectives) and adverbs. They indicate a short vowel followed by an unwritten n sound: مًا/mAF,[5] مٌ/mN and مٍ/mK. Nunation is an indicator of nominal indefiniteness. The shadda is a consonant doubling diacritic: مّ/m˜(/mm/). The shadda can combine with vowel or nunation diacritics: مُّ/m˜u or مٌّ/m˜uN.

Functionally, diacritics can be split into two different kinds: **lexical diacritics** and **inflectional diacritics** (Diab et al., 2007) .

Lexical diacritics: distinguish between two lexemes.[6] We refer to a lexeme with its citation

[3]Most encodings do not count hamza as a diacritic and the dots on letters are obligatory, other markers are truly optional hence the exclusion of all these classes from our study.

[4]All reference to Arabic in this paper is specifically to the MSA variant.

[5]Buckwalter's transliteration symbols for nunation, F, N and K, are pronounced /an/, /un/ and /in/, respectively.

[6]A lexeme is an abstraction over inflected word forms which groups together all those word forms that differ only in terms of one of the inflectional morphological categories

form as the lemma. Arabic lemma forms are third masculine singular perfective for verbs and masculine singular (or feminine singular if no masculine is possible) for nouns and adjectives. For example, the diacritization difference between the lemmas كَاتِب/kAtib/'writer' and كَاتَب/kAtab/'to correspond' distinguishes between the meanings of the word (lexical disambiguation) rather than their inflections. Any of diacritics may be used to mark lexical variation. A common example with the shadda (gemination) diacritic is the distinction between Form I and Form II of Arabic verb derivations. Form II, indicates, in most cases, added causativity to the Form I meaning. Form II is marked by doubling the second radical of the root used in Form I: أَكَل/Akal/'ate' *vs.* أَكَّل/Ak˜al/'fed'. Generally speaking, however, deriving word meaning through lexical diacritic placement is largely unpredictable and they are not specifically associated with any particular part of speech.

Inflectional diacritics: distinguish different inflected forms of the same lexeme. For instance, the final diacritics in كِتَابُ/kitAbu/'book [nominative]' and كِتَابَ/kitAba/'book [accusative]' distinguish the syntactic case of 'book' (e.g., whether the word is subject or object of a verb). Additional inflectional features marked through diacritic change, in addition to syntactic case, include voice, mood, and definiteness. Inflectional diacritics are predictable in their positional placement in a word. Moreover, they are associated with certain parts of speech.

3 Related Work

The task of diacritization is about adding diacritics to the canonical underspecified written form. This task has been discussed in several research works in various NLP areas addressing various applications.

Automatic Arabic Diacritization Much work has been done on recovery of diacritics over the past two decades by developing automatic methods yielding acceptable accuracies. Zitouni et al. (2006) built a diacritization framework based on maximum entropy classification to restore missing diacritics on each letter in a given word. Vergyri and Kirchhoff (2004) worked on automatic diacritization with the goal of improving automatic speech recognition (ASR). Different algorithms for diacritization based mainly on morphological analysis and lexeme-based language models were developed (Habash and Rambow, 2007; Habash and Rambow, 2005; Roth et al., 2008). Various approaches combining morphological analysis and/or Hidden Markov Models for automatic diacritization are found in the literature (Bebah et al., 2014; Alghamdi and Muzaffar, 2007; Rashwan et al., 2009). Rashwan et al. (2009) designed a stochastic Arabic diacritizer based on a hybrid of factorized and un-factorized textual features to automatically diacritize raw Arabic text. Emam and Fischer (2011) introduced a hierarchical approach for diacritization based on a search method in a set of dictionaries of sentences, phrases and words, using a top down strategy. More recently, Abandah et al. (2015) trained a recurrent neural network to transcribe undiacritized Arabic text into fully diacritized sentences. It is worth noting that all these approaches target full diacritization.

Impact of Diacritization in NLP Applications Regardless of the level of diacritization, to date, there have not been many systematic investigations of the impact of different types of Arabic diacritization on NLP applications. For ASR, Kirchhoff and Vergyri (2005) presented a method for full diacritization, FULL, with the goal of improving state of the art Arabic ASR. Ananthakrishnan et al. (2005) used word-based and character-based language models for recovering diacritics for improving ASR. Alotaibi et al. (2013) proposed using diacritization to improve the BBN/AUB DARPA Babylon Levantine Arabic speech corpus and increase its reliability and efficiency. For SMT, there is work on the impact of different levels of partial and full diacritization as a preprocessing step for Arabic to English SMT (Diab et al., 2007). Recently, Hermena et al. (2015) examined sentence processing in the absence of diacritics and contrasted it with the situation where diacritics were explicitly present in an eye-tracking experiment for readability. Their results show that readers benefited from the disambiguating diacritics. This study was a MIN scheme exploration focused on heterophonic-homographic target verbs that have different pronunciations in active and

such as number, gender, aspect, voice, etc. Whereas a lemma is a conventionalized citation form.

	Size in words	GOLD annotation
ATB News	2,478	Yes
ATB BN	3,093	Yes
ATB WebLog	3,177	Yes
Tashkeela	5,172	Yes
Wikipedia	2,850	No
Total	16,770	-

Table 2: The size of the data for annotation per corpus genre

passive.

In this work we are interested in two components: annotating large amounts of varied genres type corpora with diacritics as well as investigating various strategies of annotating corpora with diacritics. We also investigate two levels of diacritization, a full diacritization, FULL, and an initial attempt at a general minimal diacritization scheme, MIN.

4 Corpus Description

We conducted several experiments on a set of sentences that we extracted from five corpora covering different genres. We selected three corpora from the currently available Arabic Treebanks from the Linguistic Data Consortium (LDC). These corpora were chosen because they are fully diacritized and had undergone significant quality control, which will allow us to evaluate the annotation accuracy as well as our annotators understanding of the task.

ATB newswire: Formal newswire stories in MSA.[7]

ATB Broadcast news: Scripted, formal MSA as well as extemporaneous dialogue.[8]

We extend our corpus and include texts covering various topics beyond the commonly-used news topics:

ATB Weblog: Discussion forum posts written primarily in MSA and contained in the 70K words Gale Arabic-English Parallel Aligned Treebank.[9]

Tashkeela: a classical Arabic vocalized text corpus, collected using automatic Web crawling methods from Islamic religious heritage (mainly classical Arabic books). This corpus contains over 6 million words fully diacritized. For our study we include a subset of 5k words from this corpus.

Wikipedia: a corpus of selected abstracts extracted from a number of Arabic Wikipedia articles[10].

We select a total of 16,770 words from these corpora for annotation. The distribution of our dataset per corpus genre is provided in Table 2. Since the majority of our corpus is already fully diacritized, we strip all the diacritics prior to annotation.

5 Annotation Procedure and Guidelines

Three native Arabic annotators with good linguistic background annotated the corpora samples described in Section 4 and illustrated in Table 2, by adding the diacritics in a way that helps a reader disambiguate the text or simply articulate it correctly. Diab et al. (2007), define six different diacritization schemes that are inspired by the observation of the relevant naturally occurring diacritics in different texts. We adopt the FULL diacritization scheme, in which all the diacritics should be specified in a word (e.g., سَتُرَمَّمُ الجُدرَانُ/saturammu AlojidorAnu/"The walls will be restored").

5.1 Annotation Procedure

We design the following three strategies: (i) **Basic**, (ii) **Intermediate**, and, (iii) **Advanced**. These strategies are defined in order to find the best annotation setup that optimizes the annotation efforts and workload, as well as assessing the annotator skills in building reliable annotated corpora.

Annotators were asked to fully diacritize each word. They were assigned different tasks in which

[7] ATB Part 1 Version 4.1 Catalog No: LDC2010T1

[8] Arabic Treebank Broadcast News v1.0 Catalog No: LDC2012T07

[9] Catalog No : LDC2014T08.

[10] http://ar.wikipedia.org/

English	The ITU is the second oldest international organization that still exists.				
Buckwalter	*AlAtHAd Aldwly llAtSAlAt hw vAny >qdm tnZym EAlmy mA zAl mwjwdA.*				
Basic	الاتحاد الدولي للاتصالات هو ثاني أقدم تنظيم عالمي ما زال موجودا.				
Intermediate	الاتحاد/NN الدولي/Adj ل/Prep الاتصالات/NN هو/Pron ثاني/Adj أقدم/VV تنظيم/NN عالمي/Adj مَا/Pron زَال/VV موجودَا/Adj ./Punc				
Advanced	**Word**	**MADAMIRA candidates**		**Word**	**MADAMIRA candidates**
	الاتحاد	→ [الَاِتِّحَادُ، الَاِتِّحَادِ، الَاِتِّحَادَ]		تنظيم	→ [تَنظِيمٍ ، تَنظِيمٍ]
	الدولي	→ [الدُّوَلِيُّ، الدَّوَلِيُّ، الدُّوَلِيِّ]		عالمي	→ [عَالَمِيٍّ، عَالَمِيٌّ، عَالَمِيّ]
	للاتصالات	→ [لِلَاِتِّصَالَاتِ، لِلَاِتِّصَالَاتِ، لِلَاِتِّصَالَات]		ما	→ [مَا، مَا]
	هو	→ [هُوَ، هُو]		زال	→ [زَالَ]
	ثاني	→ [ثَانِي، ثَانِي، ثَانِي]		موجودا	→ [مَوجُوداً، مَوجُودَا]
	أقدم	→ [أَقدَمَ، أُقَدِّمُ، أُقَدَّمُ]		.	→ [.]

Table 3: Examples of a sentence (along with its English translation and Buckwalter transliteration) as presented to the annotator, in the Basic, Intermediate and Advanced annotation modes.

we vary the level and/or the text genre as follows:

	Annot$_1$	**Annot$_2$**	**Annot$_3$**
Text1	Basic	Advanced	Intermediate
Text2	Advanced	Basic	Intermediate
Text3	Basic	Advanced	Intermediate
Text4	Basic	Intermediate	Advanced
Text5	Intermediate	Advanced	Basic

Table 4: Data distribution per annotator and per annotation strategy.

Basic: In this mode, we ask for annotation of words where all diacritics are absent, including the naturally occurring ones. The words are presented in a raw tokenized format to the annotators in context. An example is provided in Table 3.

Intermediate: In this mode, we provide the annotator with words along with their POS information. The intuition behind adding POS is to help the annotator disambiguate a word by narrowing down on the diacritization possibilities. For example, the surface undiacritized spelling consonantal form for the Arabic word بين/byn could have the following possible readings: بَيَّنَ/bay˜ina/'made clear|different', when it is a verb or بَينَ/bayona/'between' when it corresponds to the adverb. We use MADAMIRA (Pasha et al., 2014), a morphological tagging and disambiguation system for Arabic, for determining the POS tags.

Advanced: In this mode, the annotation task is formulated as a selection task instead of an editing task. Annotators are provided with a list of automatically diacritized candidates and are asked to choose the correct one, if it appears in the list. Otherwise, if they are not satisfied with the given candidates, they can manually edit the word and add the correct diacritics. This technique is designed in order to reduce annotation time and especially reduce annotator workload. For each word, we generate a list of vowelized candidates using MADAMIRA (Pasha et al., 2014). MADAMIRA is able to achieve a lemmatization accuracy 99.2% and a diacritization accuracy of 86.3%.

We present the annotator with the top three candidates suggested by MADAMIRA, when possible. Otherwise, only the available candidates are provided, as illustrated in Table 3. Each text genre (Text1→5) is assigned to our annotators (Annot$_1$, Annot$_2$ and Annot$_3$) in the three different modes. Table 4 shows the distribution of data per annotator and per mode. For instance, Text1 is given to Annot$_1$ in Basic mode, to Annot$_2$ in Advanced mode and to Annot$_3$ in Advanced mode. Hence, each text genre is annotated 3 times in 3 modes by the 3 annotators.[11]

[11] Different tasks were assigned based on the availability of the annotators since some annotators can afford more hours per week than others.

	News	BN	WebLog	Tashkeela	Wiki
Basic	32.23	33.59	37.13	42.86	46.16
Intermediate	31.86	33.07	35.02	39.79	39.00
Advanced	5.58	4.36	3.16	4.92	1.56

Table 5: IAA in terms of WER

	News	BN	Weblog	Tashkeela	Wiki
Basic	68.36	69.01	62.50	68.03	66.14
Intermediate	78.05	76.31	73.77	69.25	71.48
Advanced	**98.00**	**94.59**	**88.88**	**73.10**	**95.23**

Table 6: Annotations accuracy for the different corpora per mode

5.2 Guidelines

We provided annotators with detailed guidelines, describing our diacritization scheme and specifying how to add diacritics for each annotation strategy. We described the annotation procedure and specified how to deal with borderline cases. We also provided in the guidelines many annotated examples to illustrate the various rules and exceptions.

We extended the LDC guidelines (Maamouri et al., 2008) by adding some diacritization rules: The shadda mark should not be added to the definite article (e.g., اللَيمون'lemon' and not اللّيمون); The sukuun sign should not be indicated at the end of silent words (e.g., مِن'from'); The letters followed by a long Alif, should not be diacritized as it is a deterministic diacritization (القَوَاعِد'the rules'); Abbreviations are not diacritized (كم'km', كغم'kg'). We also added an appendix that summarized all Arabic diacritization rules.[12]

6 Annotation Analysis and Results

In order to determine the most optimized annotation setup for the annotators, in terms of speed and efficiency, we test the results obtained following the three annotation strategies. These annotations are all conducted for the FULL scheme. We first calculated the number of words annotated per hour, for each annotator and in each mode. As expected, following the Advanced mode, our three annotators could annotate an average of 618.93 words per hour which is double those annotated in the Basic mode (only 302.14 words). Adding POS tags to the Basic forms, as in the Intermediate mode, does not accelerate the process much. Only +90 more words are diacritized per hour compared to the basic mode.

Then, we evaluated the Inter-Annotator Agreement (IAA) to quantify the extent to which independent annotators agree on the diacritics chosen for each word. For every text genre, two annotators were asked to annotate independently a sample of 100 words. We measured the IAA between two annotators by averaging WER (Word Error Rate) over all pairs of words. The higher the WER between two annotations, the lower their agreement. The results given in Table 5, show clearly that the Advanced mode is the best strategy to adopt for this diacritization task. It is the less confusing method on all text genres (with WER between 1.56 and 5.58). We note that Wiki annotations in Advanced mode garner the highest IAA with a very low WER.

We measure the reliability of the annotations by comparing them against gold standard annotations. In order to build the gold Wiki annotations, we hired two professional linguists, provided them with guidelines and asked them to fully diacritize the sentences. We compute the accuracy of the annotations obtained in each annotation mode and report results in Table 6 by measuring the pairwise similarity between annotators and the gold annotations.

The best result is obtained on the ATB-news dataset using the Advanced mode (annotation based on MADAMIRA's output). This is not surprising as MADAMIRA is partly trained on this corpus for diacritization. The accuracy of 98.0 obtained on this corpus validates our intuition be-

[12]The guidelines are available upon request.

hind using this annotation strategy. It is not surprising that Basic is the most difficult mode for our annotators. These are not trained lexicographers, though they possess an excellent command of MSA they are at a level where they need the Advanced mode. Furthermore, adding the POS information in the Intermediate mode helps significantly over the Basic mode, but it is still less accurate than annotations obtained in the Advanced mode.

The accuracy of the annotations for Tashkeela corpus in all the modes is very low compared to the other corpora, especially in the Advanced mode. Tashkeela was parsed with MADAMIRA and the annotations were presented to the annotators. So the results of MADAMIRA tagging are lower, hence the choice was among bad diacritized candidates. By observing the the number of edits done in the Advanced mode, we realize that annotators tend to not to edit (only 194 edits in total) in order to render a correct form of diacritization, this fits perfectly with the notion of tainting in annotation. It is always a trade off between quality and efficiency.

It is worth noting that the Basic mode shows that the Weblog corpus was the hardest one for the annotators in terms of raw accuracy. Further analysis is needed to understand why this is the case.

7 MIN annotation scheme: Preliminary study

This is a diacritization scheme that encodes the most relevant differentiating diacritics to reduce confusability among words that look the same (homographs) when undiacritized but have different readings. Our hypothesis in MIN is that there is an optimal level of diacritization to render a text unambiguous for processing and enhance its readability.

Annotating a word with the minimum diacritics needed to render it readable and unambiguous in context is subjective and depends on the annotator's understanding of the task. It also depends on the definition of the MIN scheme in the guidelines. We describe here a preliminary study aiming at exploring this diacritization scheme and measuring Inter-annotator agreement between annotators for such a task using the Basic mode.

We select a sample of 100 sentences (comprising 3,527 words) from the ATB News corpus and processed them with MADAMIRA. We, then assign it to four annotators including a lead annotator for providing a gold standard.[13] This task is done using the advanced mode.

We measure the IAA for this task using WER. We obtain an average WER of 27%, which reflects a high disagreement between annotators in defining the minimum number of diacritics to be added. The WER are shown in Table 9.

Annot$_1$	27.44
Annot$_2$	24.74
Annot$_3$	27.92
Average	27.15

Table 9: IAA WER scores against gold (Annot$_4$) for the MIN annotation scheme

An observation of some cases of disagreement of the examples in Table 7 and Table 8 shows a variable interpretation of what should be the MIN diacritization scheme. For Example, there is clear confusion about the letters to diacritize in the case of conjunctions and prepositions (such as: كَمَا/'as well' and عَلَى/'on'). In some other cases there is a disagreement of which diacritics to mention such as the word بحمامات/'with baths' in Table 7 written in four different ways by the four annotators (بحمَامَات, بِحَمَّامَاتٍ, بحمَامَات, بِحَمَّامَاتِ, بِحَمَامَاتِ).

The outlier annotator (Annot$_1$) has been detected based on a large number of cases in which he disagree with the rest. For example, the words ضفاف/'banks' and خصوصا/'especially' in the sentence given in Table 7, were erroneously fully diacritized, while adding a fatha on the second letter is enough to disambiguate these words.

By design we meant for the guidelines to be very loose in attempt to discover the various factors impacting what a possible MIN could mean to different annotators. The main lessons learned from this experiment is: first, this is a difficult task since every annotator can have a different interpretation of what is a minimum diacritization. Second, we also noticed that the same annotator could be inconsistent in his interpretation. Third, we believe that the educational and cultural background of the annotator plays an important role in the various MIN scheme interpretations. However,

[13] Annot$_4$ is the lead annotator

English	And the spread of the phenomenon of building chalets equipped with steam baths especially on lake banks.
Annot$_1$	كَمَا اِنتَشَرَت ظاهرة بنَاء شاليهات مُجَهَّزَةٍ بِحَماماتِ بُخارٍ علَى ضِفافِ البُحَيراتِ خُصوصَاً .
Annot$_2$	كَمَا اِنتَشَرَت ظاهرة بنَاء شاليهَات مجهزة بِحَمّاماتِ بخار علَى ضفاف البحيرات خصوصَا .
Annot$_3$	كمَا انتشرت ظاهرة بنَاء شاليهات مجهزة بحمامات بخار علَى ضفاف البحيرات خصوصَا .
Annot$_4$	كَمَا اِنتَشَرَت ظاهرة بِنَاءِ شاليهات مُجَهَّزَةٍ بِحَمّاماتٍ بخار عَلَى ضفاف البُحَيراتِ خصوصَا .

Table 7: An example showing a sentence with low average IAA (WER: 44.87).

English	And Dick Brass promised the readers by saying: we will put in your hands story books. And you will find in it the sound, the image and the text.
Annot$_1$	وَوَعَدَ ديك برَاس القرَاء بقوله : سنضع بَينَ ايديكم كتبَا تَحكي . وستجدون فيهَا الصوت وَالصورة وَالنص .
Annot$_2$	وّعد ديك برَاس القرَاء بقوله : سنضع بين ايديكم كتبَا تحكي . وستجدون فيهَا الصوت وَالصورة وَالنص .
Annot$_3$	وَوَعَدَ ديك برَاس القرَاء بقوله : سنضع بَينَ أَيدِيكُم كتبَا تحكي . وَسَتَجِدُونَ فيهَا الصوت وَالصورة وَالنص .
Annot$_4$	وَوَعَدَ ديك برَاس القرَاء بقوله : سنضع بَينَ ايديكم كتبَا تَحكِي . وَسَتَجِدُونَ فيهَا الصوت وَالصورة وَالنص .

Table 8: An example showing a sentence with higher average IAA (WER: 16.66).

this provides an interesting pilot study into creating guidelines for this task.

8 Conclusion

We described a pilot study to build a diacritized multi-genre corpus. In our experiments, we annotated a sample of non-diacritized words that we extracted from five text genres. We also explored different annotation strategies, and we showed that generating automatically the diacritized candidates and formulating the task as a selection task, accelerates the annotation and yields more accurate annotations. We also conducted a preliminary study for a minimum diacritization scheme and showed the difficulty in defining such a scheme and how subjective this task can be. In the future, we plan to explore the minimum scheme more deeply.

Acknowledgements

We thank Nizar Habash and anonymous reviewers for their valuable comments and suggestions. We also thank all our dedicated annotators: Noor Alzeer, Anissa Jrad, Samah Lakhal, Jihene Wafi, and Hoda Ibrahim. This publication is made possible by grant NPRP-6-1020-1-199 from the Qatar National Research Fund (a member of the Qatar Foundation).

References

Gheith A Abandah, Alex Graves, Balkees Al-Shagoor, Alaa Arabiyat, Fuad Jamour, and Majid Al-Taee. 2015. Automatic Diacritization of Arabic Text using Recurrent Neural Networks. *International Journal on Document Analysis and Recognition (IJDAR)*, 18(2):1–15.

Mansour Alghamdi and Zeeshan Muzaffar. 2007. KACST Arabic Diacritizer. In *The First International Symposium on Computers and Arabic Language*, pages 25–28.

Y.A. Alotaibi, A.H. Meftah, and S.A. Selouani. 2013. Diacritization, Automatic Segmentation and Labeling for Levantine Arabic Speech. In *Digital Signal Processing and Signal Processing Education Meeting (DSP/SPE), 2013 IEEE*, pages 7–11, Napa, CA.

Sankaranarayanan Ananthakrishnan, Shrikanth Narayanan, and Srinivas Bangalore. 2005. Automatic Diacritization of Arabic Transcripts for Automatic Speech Recognition. In *Proceedings of the 4th International Conference on Natural Language Processing*, pages 47–54.

Mohamed Bebah, Amine Chennoufi, Azzeddine Mazroui, and Abdelhak Lakhouaja. 2014. Hybrid Approaches for Automatic Vowelization of Arabic Texts. *International Journal on Natural Language Computing (IJNLC)*, 3(4).

Tim Buckwalter. 2002. Buckwalter Arabic Morphological Analyzer Version 1.0. Technical Report LDC2002L49, Linguistic Data Consortium.

Mona Diab, Mahmoud Ghoneim, and Nizar Habash. 2007. Arabic Diacritization in the Context of Statistical Machine Translation. In *Proceedings of MT-Summit*, Copenhagen, Denmark.

Ossama Emam and Volker Fischer. 2011. Hierarchical Approach for the Statistical Vowelization of Arabic Text. US Patent 8,069,045.

Nizar Habash and Owen Rambow. 2005. Arabic Tokenization, Part-of-Speech Tagging and Morphological Disambiguation in One Fell Swoop. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics*, pages 573–580, Ann Arbor, Michigan.

Nizar Habash and Owen Rambow. 2007. Arabic Diacritization through Full Morphological Tagging. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Companion Volume, Short Papers*, pages 53–56, Rochester, New York.

Ehab Hermena, Denis Drieghe, Sam Hellmuth, and Simon P Liversedge. 2015. Processing of Arabic Diacritical Marks: Phonological–Syntactic Disambiguation of Homographic Verbs and Visual Crowding Effects. *Journal of Experimental Psychology. Human Perception and Performance*, 41(2):494–507.

Katrin Kirchhoff and Dimitra Vergyri. 2005. Cross-Dialectal Data Sharing for Acoustic Modeling in Arabic Speech Recognition. *Speech Communication*, 46(1):37–51.

Mohamed Maamouri, Ann Bies, and Seth Kulick. 2008. Enhancing the Arabic Treebank: a Collaborative Effort toward New Annotation Guidelines. In *LREC*. Citeseer.

Arfath Pasha, Mohamed Al-Badrashiny, Ahmed El Kholy, Ramy Eskander, Mona Diab, Nizar Habash, Manoj Pooleery, Owen Rambow, and Ryan Roth. 2014. MADAMIRA: A Fast, Comprehensive Tool for Morphological Analysis and Disambiguation of Arabic. In *In Proceedings of the 9th International Conference on Language Resources and Evaluation*, Reykjavik, Iceland.

Mohsen Rashwan, Mohammad Al-Badrashiny, Mohamed Attia, and Sherif Abdou. 2009. A Hybrid System for Automatic Arabic Diacritization. In *The 2nd International Conference on Arabic Language Resources and Tools*, Cairo, Egypt.

Ryan Roth, Owen Rambow, Nizar Habash, Mona Diab, and Cynthia Rudin. 2008. Arabic Morphological Tagging, Diacritization, and Lemmatization Using Lexeme Models and Feature Ranking. In *Proceedings of ACL-08: HLT, Short Papers*, pages 117–120, Columbus, Ohio.

Dimitra Vergyri and Katrin Kirchhoff. 2004. Automatic Diacritization of Arabic for Acoustic Modeling in Speech Recognition. In *Proceedings of the Workshop on Computational Approaches to Arabic Script-based Languages*, pages 66–73. Association for Computational Linguistics.

Imed Zitouni, Jeffrey S. Sorensen, and Ruhi Sarikaya. 2006. Maximum Entropy Based Restoration of Arabic Diacritics. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 577–584, Sydney, Australia.