Guidelines and Framework for a Large Scale Arabic Diacritized Corpus

Wajdi Zaghouani[1], **Houda Bouamor**[1], **Abdelati Hawwari**[2], **Mona Diab** [2], **Ossama Obeid**[1], **Mahmoud Ghoneim**[2], **Sawsan Alqahtani**[2] **and Kemal Oflazer**[1]

[1]**Carnegie Mellon University in Qatar**
{wajdiz,hbouamor,owo}@cmu.edu, ko@cs.cmu.edu
[2]**George Washington University**
{mdiab,abhawwari,ghoneim,sawsanq}@gwu.edu

Abstract

This paper presents the annotation guidelines developed as part of an effort to create a large scale manually diacritized corpus for various Arabic text genres. The target size of the annotated corpus is 2 million words. We summarize the guidelines and describe issues encountered during the training of the annotators. We also discuss the challenges posed by the complexity of the Arabic language and how they are addressed. Finally, we present the diacritization annotation procedure and detail the quality of the resulting annotations.

Keywords: Arabic Diacritization, Guidelines, Annotation

1. Introduction

Written Modern Standard Arabic (MSA) poses many challenges for natural language processing (NLP). Most written Arabic text lacks short vowels and diacritics rendering a mostly consonantal orthography (Schulz, 2004). Arabic diacritization is an orthographic way to describe Arabic word pronunciation, and to avoid word reading ambiguity. Arabic writing system has two classes of symbols: letters and diacritics. diacritics are the marks that reflect the phonological, morphological and grammatical rules. Diacritization may be classified according to the linguistic rules they are representing, into two types: a) word form diacritization, which shows how a word is pronounced, except the last letter diacritization, and b) case and mood diacritization, which exists above or below the last letter in each word, indicating its grammatical function in the sentence. There are three types of diacritics: vowel, nunation, and shadda (gemination).

The lack of diacritics leads usually to considerable lexical and morphological ambiguity as shown in the example in Table 1.[1] Full diacritization has been shown to improve state-of-the-art Arabic automatic systems such as speech recognition (ASR) systems (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 (Zitouni et al., 2006; Shahrour et al., 2015; Abandah et al., 2015; Belinkov and Glass, 2015).

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 MSA diacritized corpora are generally limited to newswire stories (those distributed by the LDC), religious texts such as the Holy Quran, or educational texts.

Undiacritized	Diacritized	Buckwalter	English
وعد	وَعَدَ	/waEad/	he promised
وعد	وَعْدٌ	/waEodN/	it/a promise
وعد	وُعِدَ	/wuEida/	he was promised
وعد	وَعَدَّ	/waEad ˜a/	and he counted
وعد	وَعُدَّ	/waEud ˜a/	and he was counted

Table 1: Possible pronunciations and meanings of the undiacritized Arabic word وعد /wEd/

This paper presents the work carried out within the optimal diacritization scheme for Arabic orthographic representation (OptDiac) project. The focus of this work is to manually create a large-scale annotated corpus with the diacritics for a variety of Arabic texts covering more than 10 genres. The target size of the annotated corpus is 2 million words. The present work was mainly motivated by the lack of equivalent multi-genres large scale annotated corpus. The creation of manually annotated corpus usually presents many challenges and issues. In order to address those challenges, we created comprehensive and simplified annotation guidelines that were used by a team consisting of five annotators and an annotation manager. The guidelines were defined after an initial pilot annotation experiment described in Bouamor et al. (2015). In order to ensure a high annotation agreement between the annotators, multiple training sessions were held and a regular inter-annotator agreement (IAA) measures were performed to check annotation quality. To the best of our knowledge, this is the first Arabic diacritized multi-genres corpus.

The remainder of this paper is organized as follows. We review related work in section 2. Afterwards, we discuss the challenges posed by the complexity of the Arabic diacritization process in section 3. Then, we describe our corpus and the development of the guidelines in sections 4 and 5.

[1] We use the Buckwalter transliteration encoding system to represent Arabic in Romanized script (Buckwalter, 2002)

We present the diacritization annotation procedure in section 6 and analyze the quality of the resulting annotations in 7. Finally we conclude and present our future work in section 8.

2. Related Work

Since, our paper is mainly about the creation and evaluation of a large annotated corpus, we will focus mostly on this aspect in the previous works. There have been numerous approaches to build an automatic diacritization system for Arabic using rule-based, statistical and hybrid methods. We refer to the recent literature review in Abandah et al. (2015) for a general overview of these methods and tools.

The most relevant resource to our work is the Penn Arabic Treebank (PATB), a large corpus annotated by the Linguistic Data Consortium (Maamouri et al., 2010). Most of the LDC Treebank corpora are also manually diacritized, but they cover mainly news and weblog text genres. The PATB served later to build the first Arabic Proposition Bank (APB) using the fully specified diacritized lemmas (Diab et al., 2008; Zaghouani et al., 2010).

The Tashkeela classical Arabic vocalized corpus (Zerrouki, 2011) is another notable dataset covering six million words. Tashkeela was compiled from various web sources covering Islamic religious heritage (mainly classical Arabic books). Moreover, Dukes and Habash (2010), created the Quranic Arabic Corpus, a fully diacritized annotated linguistic resource which we used later on to build the first Quranic Arabic Proposition Bank Zaghouani et al. (2012).

The Qatar Arabic Language Bank (Zaghouani et al., 2014; Zaghouani et al., 2015; Zaghouani et al., 2016) is another relevant work that aims to build a large corpus of manually corrected Arabic text for building automatic correction tools for three Arabic text genres: native, non-native and machine translation post-edited text.

Recently, in Bouamor et al. (2015), we conducted various annotation experiments to find the most suitable and efficient annotation procedure in creating a large scale diacritized corpus.

3. 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)[2] which in all, if present, render a more or less exact precise reading of a word. In this study, we are mostly addressing three types of diacritical marks: short vowels, nunation, and shadda (gemination). In this study, short vowel diacritics refer to the three short vowels in Modern Standard Arabic (MSA)[3] and a diacritic indicating the 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,[4] 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.[5] We refer to a lexeme with its citation 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: أَكَلَ /Akal/'ate' *vs.* أكَّل /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.

4. Corpus Description

We use the corpus of contemporary Arabic (CCA) compiled by Al-Sulaiti and Atwell (2006). It is a balanced corpus divided into the following genres: autobiography, short stories, children's stories, economics, education, health and medicine, interviews, politics, recipes, religion, sociology,

[2] 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.

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

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

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

science, sports, tourism and travel. The CCA corpus text genres were carefully selected by its compilers since the target users of the corpus were mostly language teachers and teachers of Arabic as a foreign language. Various metadata information are included in the corpus such as the information about the text, the author and the source. In order to use the CCA corpus, a normalization effort was done to produce a consistent XML mark-up format to be used by our annotation tool.

5. Development of the Guidelines

We provided the annotators with detailed guidelines, describing our diacritization scheme and specifying when and where to add the diacritics. We describe the annotation procedure and explained how to deal with borderline cases. We also include several annotated examples to illustrate the specified rules.

Our guidelines are mostly inspired from the LDC POS annotation guidelines (Maamouri et al., 2008). Since, the LDC guidelines are mainly designed for the POS annotation and not specifically for the diacritization per se, we created a simplified version and added some specific diacritization rules to make the annotation process consistent. Below we provide some examples of diacritization exceptions and specific rules.

The Shadda: The shadda mark should be added in all cases specified in the guidelines except the following in the definite artilce, where it should not be added to the letter ل /l/ of the definite article (e.g. الليمون /Allymwn/ 'lemon' and not اللّيمون /Aĩlymwn/). Moreover, the shadda should be added to the first letter that follows the definite article with a solar letter construction such as in النّاس /AlñAs/ 'The people' and not الناس /AlnAs/.

The Soukoun: The sukuun sign should not be indicated at the end of silent words (e.g., من /mn/ 'from').

The Proper Nouns: The proper noun case endings are not to be added as they are defined by their nature with the exception of an accusative proper noun of Arabic origin as in قابَلتُ عَلِيّاً /qAbaltu EaliÃF/ 'I met Ali'.

Abbreviations: Abbreviations are not to be diacritized (كم 'km' /km/, كغ /kg/ 'kg').

Nunations: In the case of nunation at the end of a word, if the word ends with an accusative nunation as in عُدْواناً /EudowAnAF/ 'Hostile', the nunation -an signs are placed on the letter Alif and not on the nunated letter as in عُدْواناً /EudowAnFA/ 'Hostility'.[6]

Deterministic Diacritization: In some cases, the diacritization is deterministically found in the case of letters followed by a long letter Alif should not be diacritized as in مِيثاقُ /miyvAqu/ 'Treaty' and not مِيثَاقُ /miyvaAqu/.

A summary of the common Arabic diacritization rules is also added as a reference in the guidelines.[7]

6. Annotation Framework

As a large scale corpus annotation project, this project involves a team of annotators, lead annotation managers and consists of five annotators and a programmer.

6.1. Annotation Management

The lead annotation manager is responsible for the whole annotation workflow. This includes corpus selection and normalization and the annotation of the gold standard used to compute the Inter-Annotator Agreement (IAA) level. Moreover, the lead annotator is responsible for writing and updating the annotation guidelines when deemed necessary, evaluate the quality of the annotation, monitor and report on the annotation progress.

To control the quality of the annotation of each newly hired annotator, we proceed as follows. After an initial training phase, the annotator's work is closely monitored during the initial weeks, afterwards, the annotator can join the official production phase. Recently, a dedicated on-line discussion group was created to keep track of the issues raised during the annotation so that the annotators and the lead annotator can have a better communication.

6.2. Annotator Training

The annotators in this project are mostly university graduates who are native Arabic speakers.[8] During the hiring phase, the annotators were tested on an Arabic language screening test (syntax and Arabic diacritization related questions). Once selected, the annotators were trained as a group for the task. The training consisted of various annotation tasks to be done by all the participants, guidelines reading and meetings with the annotation manager and the other annotators.

6.3. The Annotation Tool

We designed and implemented a web-based annotation tool and a work-flow management interface (Obeid et al., 2016). Our online interface allows annotators to select from an automatically generated diacritized words and/or edit words manually as shown in Figure 1. The annotation interface allows users to undo/redo actions, and the history is kept over multiple sessions. The interface includes a timer to keep track of how long each sentence annotation has taken. The annotation work-flow management interface is used by the lead annotator to organize the annotation pipeline including: (i) the organization of the corpus to be annotated, (ii) tasks assignment, (iii) tracking the annotation progress and (iv) measuring automatically and regularly the agreement between annotators. Once an annotator submits his task, the annotation manager is alerted through the interface by a green highlight of the task, as shown in Figure 2. Then, the annotation manager can view and check on the quality of the annotation (Figure 3).

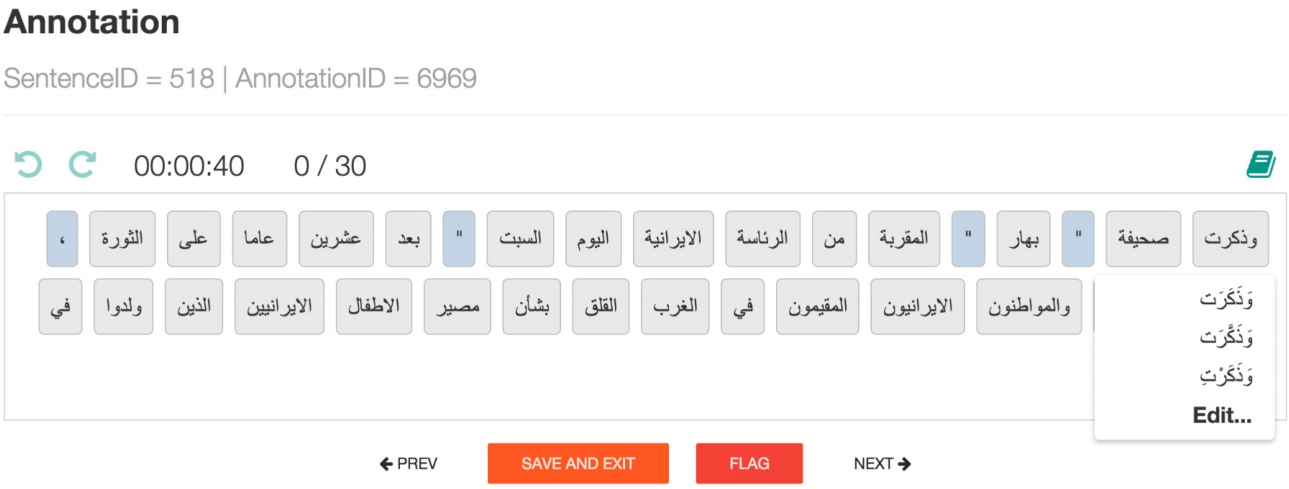
6.4. The Annotation Procedure

Following the recommendations obtained from the pilot study conducted in Bouamor et al. (2015), we formulated the diacritization annotation as a selection task. Annotators

[6] The addition of a final Nun sound to a noun or adjective indicates that it is a declinable and unmarked for definiteness.

[7] The guidelines will be soon made publicly available.

[8] Some annotators have Arabic teaching experience.



Annotation

SentenceID = 518 | AnnotationID = 6969

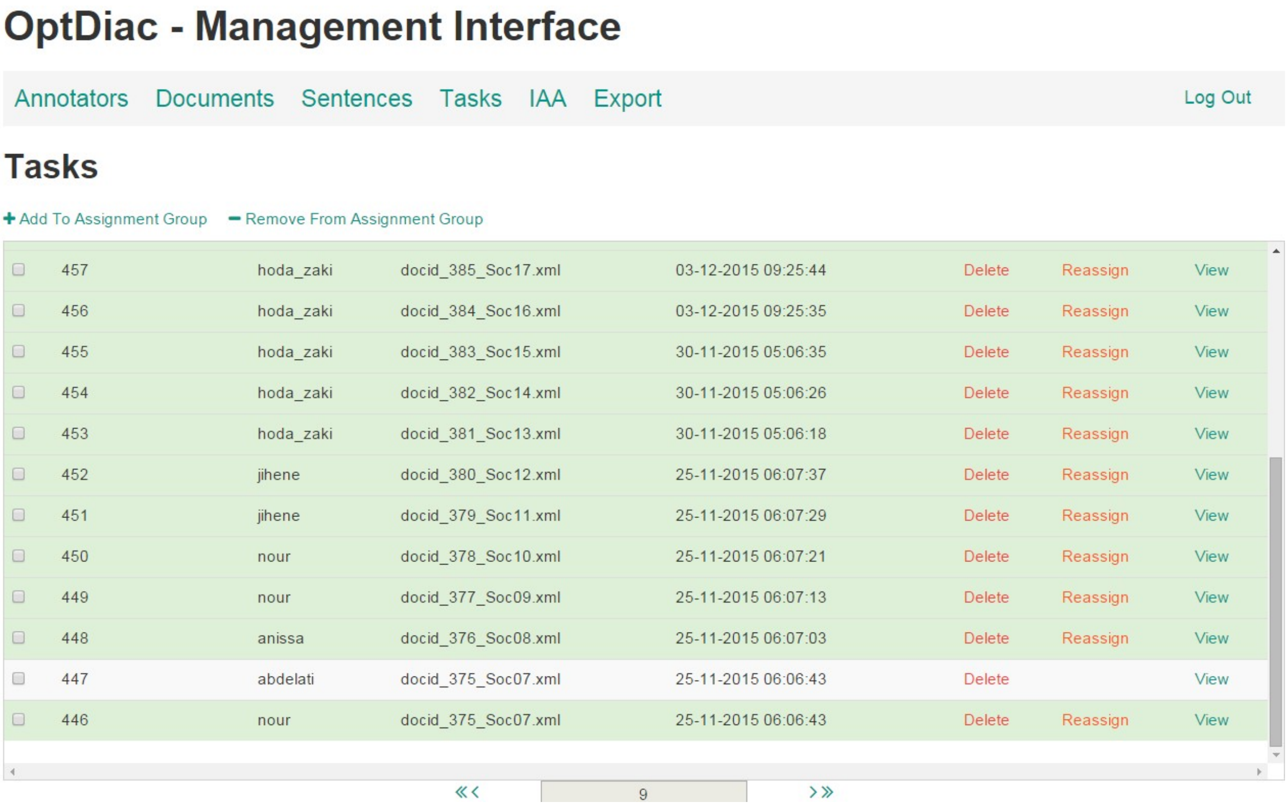
00:00:40 0 / 30

وذكرت صحيفة " بهار " المقربة من الرئاسة الايرانية اليوم السبت " بعد عشرين عاما على الثورة ،

والمواطنون الايرانيون المقيمون في الغرب القلق بشأن مصير الاطفال الايرانيين الذين ولدوا في

وَذَكَرَت
وَذَكَّرَت
وَذَكَرْتِ
Edit...

Figure 1: The annotation interface showing a selection action



OptDiac - Management Interface

Tasks

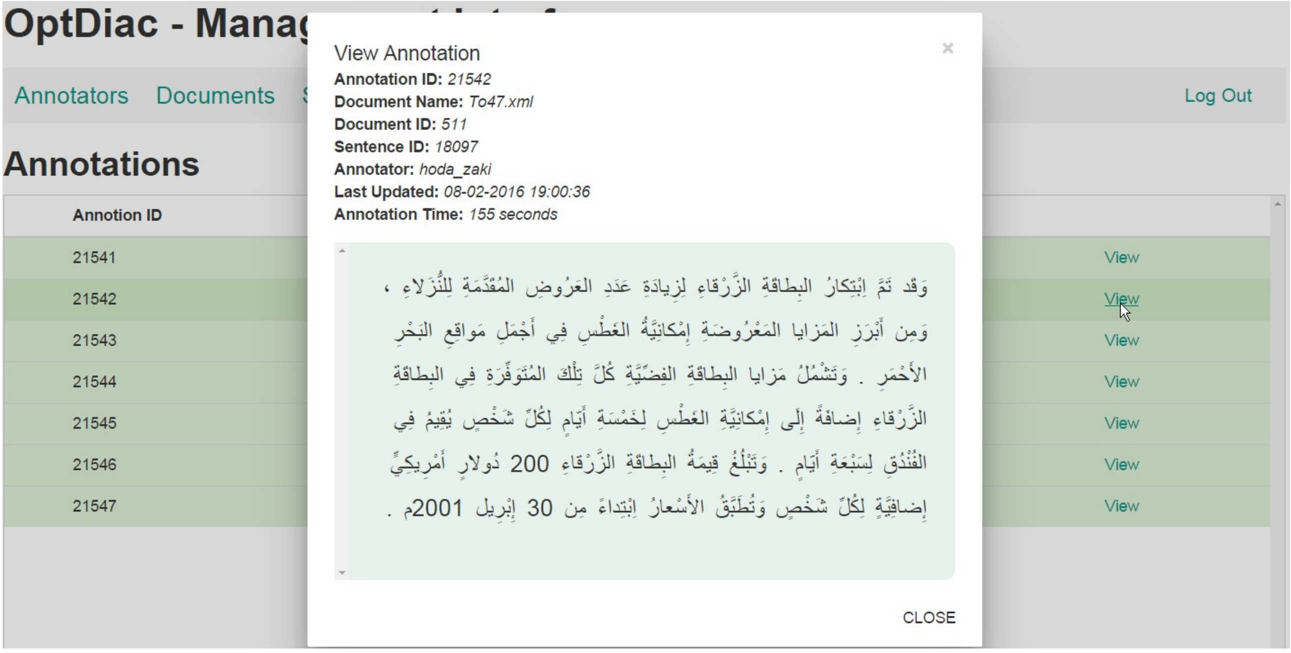
□	457	hoda_zaki	docid_385_Soc17.xml	03-12-2015 09:25:44	Delete	Reassign	View
□	456	hoda_zaki	docid_384_Soc16.xml	03-12-2015 09:25:35	Delete	Reassign	View
□	455	hoda_zaki	docid_383_Soc15.xml	30-11-2015 05:06:35	Delete	Reassign	View
□	454	hoda_zaki	docid_382_Soc14.xml	30-11-2015 05:06:26	Delete	Reassign	View
□	453	hoda_zaki	docid_381_Soc13.xml	30-11-2015 05:06:18	Delete	Reassign	View
□	452	jihene	docid_380_Soc12.xml	25-11-2015 06:07:37	Delete	Reassign	View
□	451	jihene	docid_379_Soc11.xml	25-11-2015 06:07:29	Delete	Reassign	View
□	450	nour	docid_378_Soc10.xml	25-11-2015 06:07:21	Delete	Reassign	View
□	449	nour	docid_377_Soc09.xml	25-11-2015 06:07:13	Delete	Reassign	View
□	448	anissa	docid_376_Soc08.xml	25-11-2015 06:07:03	Delete	Reassign	View
□	447	abdelati	docid_375_Soc07.xml	25-11-2015 06:06:43	Delete		View
□	446	nour	docid_375_Soc07.xml	25-11-2015 06:06:43	Delete	Reassign	View

Figure 2: The annotation management interface showing completed tasks highlighted in green

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 reduces annotation time, increases annotation quality and especially reduce annotator workload. For each word, we generate a list of diacritized candidates using MADAMIRA (Pasha et al., 2014). MADAMIRA is able to achieve a lemmatization accuracy of 99.2% and a diacritization accuracy of 86.3%. An example of diacritization candidates is given in Figure 1.

7. Annotation Analysis and Results

To quantify the extent to which independent annotators agree on the diacritics chosen for each word. We compute the inter-annotator agreement (IAA) to evaluate the extent to which our independant trained annotators agree on the diacritics added for each word. We measured the IAA between two annotators by averaging WER (Word Error Rate) over all pairs of words. We define the WER as the percentage of the incorrectly diacritized words (Snover et al., 2006), if a single letter in a given word has a diacritization error, then the whole word is considered as incorrect. Note that the higher the WER between two annotations, the



View Annotation

Annotation ID: *21542*
Document Name: *To47.xml*
Document ID: *511*
Sentence ID: *18097*
Annotator: *hoda_zaki*
Last Updated: *08-02-2016 19:00:36*
Annotation Time: *155 seconds*

وَقد تَمَّ اِبْتِكارُ البِطاقَةِ الزَّرْقاءِ لِزِيادَةِ عَدَدِ العَرُوضِ المُقَدَّمَةِ لِلنُّزَلاءِ ،
وَمِن أَبْرَزِ المَزايا المَعْرُوضَةِ إِمْكانِيَّةُ الغَطْسِ فِي أَجْمَلِ مَواقِعِ البَحْرِ
الأَحْمَرِ . وَتَشْمُلُ مَزايا البِطاقَةِ الفِضِّيَّةِ كُلَّ تِلْكَ المُتَوَفِّرَةِ فِي البِطاقَةِ
الزَّرْقاءِ إِضافَةً إِلَى إِمْكانِيَّةِ الغَطْسِ لِخَمْسَةِ أَيّامٍ لِكُلِّ شَخْصٍ يُقِيمُ فِي
الفُنْدُقِ لِسَبْعَةِ أَيّامٍ . وَتَبْلُغُ قِيمَةُ البِطاقَةِ الزَّرْقاءِ 200 دُولارٍ أَمْرِيكِيٍّ
إِضافِيَّةٍ لِكُلِّ شَخْصٍ وَتُطَبَّقُ الأَسْعارُ اِبْتِداءً مِن 30 إِبْرِيل 2001م .

Figure 3: The view annotation menu in the annotation management interface.

lower their agreement.
During the annotation of the CCA corpus, we conducted three iterations to improve and simplify and update our guidelines and also to address borderline annotation issues. After each iteration, we measure the IAA to check for possible consistency improvement. The results given in Table 2 show a steady IAA improvement after each iteration with a WER reduced to 9.31%.

	CCA Corpus
WER$_{iteration1}$	16.59
WER$_{iteration2}$	12.09
WER$_{iteration3}$	**09.31**

Table 2: Average WER obtained after each annotation iteration on the CCA corpus.

7.1. Error Analysis

During the multiple IAA evaluations, we observed various sources of inconsistent annotation between the annotators. In some cases, there was no agreement on whether to add the diacritics or not, while in in other cases, the annotators disagreed on the syntactic interpretation of the word. We compiled below the list of the most important cases of disagreement sorted from the most frequent to the less frequent.

1. Disagreement due to two possible sentence interpretations.
2. foreign words and proper noun diacritics disagreement.
3. Diacritization disagreement in misspelled words.
4. Case endings disagreement.
5. Shadda diacritization disagreement.
6. Soukoun diacritization disagreement.
7. Dialectal Arabic expressions diacritization.

We will continue our annotator training and update our guidelines in order to reach better IAA scores.

8. Conclusion and future work

In this paper, we presented large-scale diacritizaton annotation effort for multi-genre Arabic texts, including guideline development and the annotation framework. We discussed the challenges inherent in corpus diacritization including the most frequent cases of annotation disagreement. The results obtained during the evaluation suggest that the annotation consistency improved overtime following the guidelines updates.
We will continue working to improve the inter-annotator agreement and we plan to make the annotated data available soon for the research community to develop related natural language processing applications. Finally, we hope that the annotated data could be used as part of a shared task to build automatic diacritization tools for Arabic in a similar way to the shared tasks we organized in recent years on automatic text correction of Arabic (Mohit et al., 2014; Rozovskaya et al., 2015).

Acknowledgements

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

9. References

Abandah, G. A., Graves, A., Al-Shagoor, B., Arabiyat, A., Jamour, F., and Al-Taee, M. (2015). Automatic Diacritization of Arabic Text using Recurrent Neural Networks. *International Journal on Document Analysis and Recognition (IJDAR)*, 18(2):183–197.

Al-Sulaiti, L. and Atwell, E. S. (2006). The design of a corpus of Contemporary Arabic. *International Journal of Corpus Linguistics*, 11(2):135–171.

Belinkov, Y. and Glass, J. (2015). Arabic Diacritization with Recurrent Neural Networks. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2281–2285, Lisbon, Portugal.

Bouamor, H., Zaghouani, W., Diab, M., Obeid, O., Oflazer, K., Ghoneim, M., and Hawwari, A. (2015). A Pilot Study on Arabic Multi-Genre Corpus Diacritization. In *Proceedings of the Association for Computational Linguistics Second Workshop on Arabic Natural Language Processing*, pages 80–88, Beijing, China.

Buckwalter, T. (2002). Buckwalter Arabic Morphological Analyzer Version 1.0. Technical Report LDC2002L49, Linguistic Data Consortium.

Diab, M., Ghoneim, M., and Habash, N. (2007). Arabic Diacritization in the Context of Statistical Machine Translation. In *Proceedings of MT-Summit*, Copenhagen, Denmark.

Diab, M., Mansouri, A., Palmer, M., Babko-Malaya, O., Zaghouani, W., Bies, A., and Maamouri, M. (2008). A pilot arabic propbank. In *Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC)*.

Dukes, K. and Habash, N. (2010). Morphological annotation of quranic arabic. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*.

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

Maamouri, M., Bies, A., and Kulick, S. (2008). Enhancing the arabic treebank: a collaborative effort toward new annotation guidelines. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC 2008)*, Marrakech, Morocco.

Maamouri, M., Bies, A., Kulick, S., Zaghouani, W., Graff, D., and Ciul, M. (2010). From speech to trees: Applying treebank annotation to arabic broadcast news. In *Proceedings of International Conference on Language Resources and Evaluation (LREC 2010)*.

Mohit, B., Rozovskaya, A., Habash, N., Zaghouani, W., and Obeid, O. (2014). The first qalb shared task on automatic text correction for arabic. In *Proceedings of the EMNLP Workshop on Arabic Natural Language Processing*, page 39.

Obeid, O., , Bouamor, H., Zaghouani, W., Ghoneim, M., Hawwari, A., Diab, M., and Oflazer, K. (2016). MANDIAC: A Web-based Annotation System For Manual Arabic Diacritization. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC-2016) Workshop on Free/Open-Source Arabic Corpora and Corpora Processing Tools (OSACT2)*.

Pasha, A., Al-Badrashiny, M., Kholy, A. E., Eskander, R., Diab, M., Habash, N., Pooleery, M., Rambow, O., and Roth, R. (2014). MADAMIRA: A Fast, Comprehensive Tool for Morphological Analysis and Disambiguation of Arabic. In *Proceedings of the 9th International Conference on Language Resources and Evaluation*, Reykjavik, Iceland.

Rozovskaya, A., Bouamor, H., Habash, N., Zaghouani, W., Obeid, O., and Mohit, B. (2015). The second qalb shared task on automatic text correction for arabic. In *Proceedings of the ACL-IJCNLP Workshop on Arabic Natural Language Processing*, page 26.

Schulz, E. (2004). *A Student Grammar of Modern Standard Arabic*. Cambridge University Press.

Shahrour, A., Khalifa, S., and Habash, N. (2015). Improving Arabic Diacritization through Syntactic Analysis. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1309–1315, Lisbon, Portugal.

Snover, M., Dorr, B., Schwartz, R., Micciulla, L., and Makhoul, J. (2006). A study of translation edit rate with targeted human annotation. In *Proceedings of AMTA*, pages 223–231.

Zaghouani, W., Diab, M., Mansouri, A., Pradhan, S., and Palmer, M. (2010). The revised arabic propbank. In *Proceedings of the Association for Computational Linguistics Fourth Linguistic Annotation Workshop*, pages 222–226. Association for Computational Linguistics.

Zaghouani, W., Hawwari, A., and Diab, M. (2012). A pilot propbank annotation for quranic arabic. In *Proceedings of the Fourth Workshop on Computational Linguistics for Literature co-located with the North American Association Computational Linguistics conference (NAACL-HLT 2012*, page 78.

Zaghouani, W., Mohit, B., Habash, N., Obeid, O., Tomeh, N., Rozovskaya, A., Farra, N., Alkuhlani, S., and Oflazer, K. (2014). Large scale arabic error annotation: Guidelines and framework. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014), Reykjavik, Iceland, May 26-31, 2014.*, pages 2362–2369.

Zaghouani, W., Habash, N., Bouamor, H., Rozovskaya, A., Mohit, B., Heider, A., and Oflazer, K. (2015). Correction annotation for non-native arabic texts: Guidelines and corpus. In *Proceedings of the Association for Computational Linguistics Fourth Linguistic Annotation Workshop*, pages 129–139.

Zaghouani, W., Habash, N., Obeid, O., Mohit, B., Bouamor, H., and Oflazer, K. (2016). Building an arabic machine translation post-edited corpus: Guidelines and annotation. In *International Conference on Language Resources and Evaluation (LREC 2016)*.

Zerrouki, T. (2011). Tashkeela: Arabic vocalized text corpus.

Zitouni, I., Sorensen, J. S., and Sarikaya, R. (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.