The 2nd Workshop on Arabic Corpora and Processing Tools 2016 Theme: Social Media

Workshop Programme

Date 24 May 2016

09:00 – 09:20 – Welcome and Introduction by Workshop Chairs

09:20 – 10:30 – Session 1 (Keynote speech)

Nizar Habash, *Computational Processing of Arabic Dialects: Challenges, Advances and Future Directions*

10:30 – 11:00 Coffee break

10:30 – 13:00 – Session 2

Soumia Bougrine, Hadda Cherroun, Djelloul Ziadi, Abdallah Lakhdari and Aicha Chorana, *Toward a rich Arabic Speech Parallel Corpus for Algerian sub-Dialects*

Maha Alamri and William John Teahan, *Towards a New Arabic Corpus of Dyslexic Texts*

Ossama Obeid, Houda Bouamor, Wajdi Zaghouani, Mahmoud Ghoneim, Abdelati Hawwari, Mona Diab and Kemal Oflazer, *MANDIAC: A Web-based Annotation System For Manual Arabic Diacritization*

Wajdi Zaghouani and Dana Awad, *Toward an Arabic Punctuated Corpus: Annotation Guidelines and Evaluation*

Muhammad Abdul-Mageed, Hassan Alhuzali, Dua'a Abu-Elhij'a and Mona Diab, *DINA: A Multi-Dialect Dataset for Arabic Emotion Analysis*

Nora Al-Twairesh, Mawaheb Al-Tuwaijri, Afnan Al-Moammar and Sarah Al-Humoud, *Arabic Spam Detection in Twitter*

Editors

Hend Al-Khalifa	King Saud University, KSA
Abdulmohsen Al-Thubaity	King Abdul Aziz City for Science and Technology, KSA
Walid Magdy	Qatar Computing Research Institute, Qatar
Kareem Darwish	Qatar Computing Research Institute, Qatar

Organizing Committee

Hend Al-Khalifa	King Saud University, KSA
Abdulmohsen Al-Thubaity	King Abdul Aziz City for Science and Technology, KSA
Walid Magdy	Qatar Computing Research Institute, Qatar
Kareem Darwish	Qatar Computing Research Institute, Qatar

Workshop Programme Committee

Abdullah Alfaifi	Imam University, KSA
Abdulrhman Almuhareb	King Abdul Aziz City for Science and Technology, KSA
Abeer ALDayel	King Saud University, KSA
Ahmed Abdelali	Qatar Computing Research Institute, Qatar
Areeb AlOwisheq	Imam University, KSA
Auhood Alfaries	King Saud University, KSA
Hamdy Mubarak	Qatar Computing Research Institute, Qatar
Hazem Hajj	American University of Beirut, Lebanon
Hind Al-Otaibi	King Saud University, KSA
Houda Bouamor	Carnegie Mellon University, Qatar
Khurshid Ahmad	Trinity College Dublin, Ireland
Maha Alrabiah	Imam University, KSA
Mohammad Alkanhal	King Abdul Aziz City for Science and Technology, KSA
Mohsen Rashwan	Cairo University, Egypt
Mona Diab	George Washington University, USA
Muhammad M. Abdul-Mageed	Indiana University, USA
Nizar Habash	New York University Abu Dhabi, UAE
Nora Al-Twairesh	King Saud University, KSA
Nouf Al-Shenaifi	King Saud University, KSA
Tamer Elsayed	Qatar University, Qatar
Wajdi Zaghouani	Carnegie Mellon University in Qatar, Qatar

Table of contents

Author Index

Preface

Given the success of our first Workshop on Free/Open-Source Arabic Corpora and Corpora Processing Tools in LREC 2014 where three of the presented papers received 15 citations up to now. The second workshop on Free/Open-Source Arabic Corpora and Corpora Processing Tools (OSACT2) with special emphasis on Arabic social media text processing and applications aims to encourage researchers and developers to foster the utilization of freely available Arabic corpora and open source Arabic corpora processing tools and help in highlighting the drawbacks of these resources and discuss techniques and approaches on how to improve them.

OSACT2 had an acceptance rate of 55%, where we received 11 papers from which 6 papers were accepted. We believe the accepted papers are high quality and present mixture of interesting topics. Three papers are about corpus development and annotation guidelines for different domains such as speech and dyslexic texts, two papers about spam detection and emotion analysis, and finally, a paper presenting a web-based system for manual annotation of Arabic diacritization.

We would like to thank all people who in one way or another helped in making this workshop a success. Our special thanks go to Dr. Nizar Habash for accepting to give the workshop keynote talk, to the members of the program committee who did an excellent job in reviewing the submitted papers, and to the LREC organizers. Last but not least we would like to thank our authors and the participants of the workshop.

Hend Al-Khalifa, Abdulmohsen Al-Thubaity, Walid Wagdy and Kareem Darwish
Portorož (Slovenia), 2016

(Keynote Speech)
Computational Processing of Arabic Dialects: Challenges, Advances and Future Directions

Nizar Habash
New York University, Abu Dhabi, UAE
nizar.habash@nyu.edu

The Arabic language consists of a number of variants among which Modern Standard Arabic (MSA) has a special status as the formal, mostly written, standard of the media, culture and education across the Arab World. The other variants are informal, mostly spoken, dialects that are the languages of communication of daily life. Most of the natural language processing resources and research in Arabic have focused on MSA. However, recently, more and more research is targeting Arabic dialects. In this talk, we present the main challenges of processing Arabic dialects, and discuss common solution paradigms, current advances, and future directions.

Toward a Rich Arabic Speech Parallel Corpus for Algerian sub-Dialects

Soumia BOUGRINE*, **Hadda CHERROUN***, **Djelloul ZIADI****
Abdallah LAKHDARI*, **Aicha CHORANA***
* Laboratoire d'informatique et Mathématiques LIM - Université Amar Telidji Laghouat, Algérie
** Laboratoire LITIS - Université Normandie Rouen, France
{sm.bougrine, hadda_cherroun, a.lakhdari, a.chorana}@mail.lagh-univ.dz, djelloul.ziadi@univ-rouen.fr

Abstract

Speech datasets and corpora are crucial for both developing and evaluating accurate Natural Language Processing systems. While Modern Standard Arabic has received more attention, dialects are drastically underestimated, even they are the most used in our daily life and the social media, recently. In this paper, we present the methodology of building an Arabic Speech Corpus for Algerian dialects, and the preliminary version of that dataset of dialectal arabic speeches uttered by Algerian native speakers selected from different Algeria's departments. In fact, by means of a direct recording way, we have taken into acount numerous aspects that foster the richness of the corpus and that provide a representation of phonetic, prosodic and orthographic varieties of Algerian dialects. Among these considerations, we have designed a rich speech topics and content. The annotations provided are some useful information related to the speakers, time-aligned orthographic word transcription. Many potential uses can be considered such as speaker/dialect identification and computational linguistic for Algerian sub-dialects. In its preliminary version, our corpus encompasses 17 sub-dialects with 109 speakers and more than 6 K utterances.

Keywords: Algerian Arabic sub-dialects, Speech Corpus, Algerian speakers, Parallel Corpus, Syllable segmentation, orthographic transcription.

1. Introduction

Language corpus is necessary for most of the front-end Natural Language Processing (NLP) researches. In fact, it is important for the evaluation of NLP approaches. These corpora can be written or spoken only, or both (Lindquist, 2009). Spoken Language corpus (SL) is any collection of speech recording increased by means of some annotation files. In addition, each corpus have to include documentation that permits to re-use these data. Applications that use SL can be grouped into four major categories: speech recognition, speech synthesis, speaker recognition/verification and spoken language systems.

Spoken language data varies in four dimensions, read or spontaneous, formal or casual, monologue or dialogue, and standard or dialect. The latter dimension is one of the most used data type in speech research. The common methods to collect data are by writing questionnaire, investigator observations or speech corpora. Speech Dialect corpora have to include variant speaker in term of age, gender and social status (Gibbon et al., 1998).

For many languages, the state of the art of designing and developing speech banks and corpora has achieved a mature situation. For instance, the collection of English corpora have started since the $1960s$. On the other extreme, there is few corpora for Arabic, which is considered as under resourced language (Mansour, 2013).

Arabic is a semitic language, which is ranked in the first five spoken languages in the world (Lewis et al., 2015). Geographically, it is one of the most widespread languages of the world (Behnstedt and Woidich, 2013). Actually, it has two major variants: Modern Standard Arabic (MSA), and Dialectal Arabic (DA) (Embarki, 2008). In fact, while MSA vehicles formal communications, DA is often referred to colloquial Arabic, or vernaculars; which is more largely used than MSA. Indeed, it represents the most common way of communication in every day life and recently in social media.

Arabic language and its dialectal variations have very distinctive characteristics concerning its phonetic system which makes the task of automatic speech processing very challenging. Thus, the necessity of accurate Arabic speech corpora, especially for Arabic dialects, as their number is very important. An Arabic Spoken Corpus (ASC) can include MSA or DA, or both variants.

Many reasons lead us to consider that building Arabic dialect corpus is an increase necessity. This task is more challenging than building MSA corpus, as there is large number of dialects in Arabic countries, which are different according to many features: morphology, rhythm and lexical feature (Habash, 2010). Arabic dialects are grouped into four huge categories: Arabian Peninsula, Levantin, Mesopotamian, Egyptian and Maghrebi (Versteegh, 1997). However, we can distinguish many dozen of sub-dialects within the same country or region.

For Arabic dialects, yet there is minor speech datasets production. In addition, while they exist, they are incomplete or designed for a specific research purpose. Moreover, most of them are not publicly available.

In this paper, we focus on building a spoken corpus for Algerian Arabic sub-dialects which are part of Maghrebi dialect group. Our investigation, is driven by many reasons. First, there is no available Arabic dialect speech database that represents consistently Algerian dialect profile. Second, the Algerian dialect is less studied despite it represents a specific Arabic dialect as it contains numerous linguistic variations due to both arabization phases and deep colonization history.

The rest of this paper is organized as follows. In the next section, we review some related work that have built corpora for both MSA and DA. In Section 3. we give a brief

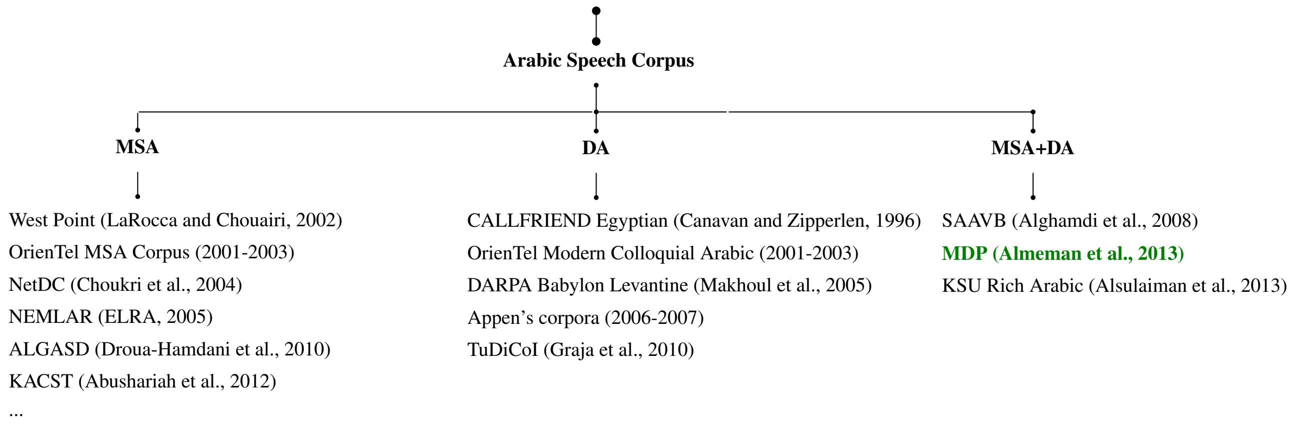

Arabic Speech Corpus
MSA
DA
MSA+DA
West Point (LaRocca and Chouairi, 2002)
OrienTel MSA Corpus (2001-2003)
NetDC (Choukri et al., 2004)
NEMLAR (ELRA, 2005)
ALGASD (Droua-Hamdani et al., 2010)
KACST (Abushariah et al., 2012)
...
CALLFRIEND Egyptian (Canavan and Zipperlen, 1996)
OrienTel Modern Colloquial Arabic (2001-2003)
DARPA Babylon Levantine (Makhoul et al., 2005)
Appen's corpora (2006-2007)
TuDiCoI (Graja et al., 2010)
SAAVB (Alghamdi et al., 2008)
MDP (Almeman et al., 2013)
KSU Rich Arabic (Alsulaiman et al., 2013)

Figure 1: Main Spoken Arabic Corpora.

description of Algerian dialects features. Section 4. is dedicated to our contribution. We start by explaining the followed methodology for designing the corpus text, selecting target speakers and choosing material and environment of recording. The preliminary version of the outcome corpus is described in Section 5. We enumerate some potential uses of this corpus in Section 6.

2. Related Work

We have classified most important Arabic Spoken Corpora according to the fact that they are built for MSA or DA, or both of them. Figure 1. illustrates the studied existing corpora according to this taxonomy where the green color is used to indicate free corpus. Following another classification way, the Arabic speech corpora can be grouped into four categories according to the collecting method. Indeed, it can be done by recording broadcast news, spontaneous telephone conversations, telephone response of questionnaire or by direct recording.

First, let us briefly review some MSA corpora which are not directly concerned by our study. It is important to mention that most existing corpora are collected early in 2000s. For instance, LaRocca and Chouairi built the West Point Arabic Speech corpus, which is collected by direct recording method (LaRocca and Chouairi, 2002). It is available via Linguistic Data Consortium LDC catalogue [1]. The speech data are collected from 110 speakers, which are native and non-native. In addition, speech corpora for MSA are collected by using telephone response of questionnaire through the *OrienTel* [2] project funded by European Commission. It focuses on the development of language resources for speech-based telephony applications across the area between Morocco and the Gulf States. These corpora are available for MSA uttered by speakers from Egypt, Jordan, Morocco, Tunisia, and United Arab Emirates country, which are available via the European languages resources Association ELRA catalogue [3].

Concerning Arabic corpus collected by recording Broadcast News from radio, we have *NetDC Arabic* (Choukri et al., 2004) and *NEMLAR* (ELRA, 2005) corpus.

About MSA corpus uttered by Algerian speakers, *ALGASD* corpus is collected by direct recording (Droua-Hamdani et al., 2010). The speakers are selected from 11 regions from Algeria. This corpus contains 300 speakers and the total number of utterances is 1080.

More recently, *KACST* Arabic phonetics database (Abushariah et al., 2012) has collected MSA corpus by using direct recording. It encompasses from 11 Arab countries of three different Arab regions: Levant, Gulf and Africa. The speakers read 415 sentences, which are phonetically rich and balanced.

In contrast to this relative abundance of speech corpora for MSA, very few attempts have tried to collect Arabic Speech corpora for dialects. Table 1. reports some features of the studied *DA* and *MSA+DA* corpora. The first set of corpora has exploited the limited solution of telephony conversation recording. In fact, as far as we know, the pioneer DA corpus has begun in the middle of the nineties and it is *CALLFRIEND Egyptian* (Canavan and Zipperlen, 1996). Another part of *OrienTel* project, cited below, has been dedicated to collect speech corpora for Arabic dialects of Egypt, Jordan, Morocco, Tunisia, and United Arab Emirates countries. In these corpora, the same telephone response of questionnaire method is used. These corpora are available via ELRA catalogue [4].

The *DARPA Babylon Levantine* Arabic speech corpus gathers some Levantine dialects spoken by speakers from four Arab countries: Jordan, Syria, Lebanon, and Palestine (Makhoul et al., 2005). This corpus is available via LDC catalogue [5].

Appen company has collected three Arabic dialects corpora by means of spontaneous telephone conversations method. These corpora uttered by speakers from Gulf, Iraqi and Levantine are available via LDC catalogue [6]. With a more guided telephone conversation recording protocol, *Fisher Levantine Arabic* corpus is available via LDC catalogue [7].

[1] Code product is LDC2002S02.

[2] `http://www.speechdat.org/ORIENTEL/`

[3] Respectives code products are ELRA-S0222, ELRA-S0290, ELRA-S0184, ELRA-S0187 and ELRA-S0259.

[4] Respectives code products are ELRA-S0221, ELRA-S0289, ELRA-S0183, ELRA-S0186 and ELRA-S0258.

[5] Code product is LDC2005S08.

[6] Respectives code products are LDC2006S43, LDC2006S45 and LDC2007S01.

[7] Code product is LDC2007S02.

Table 1: Speech Corpora for Arabic dialects.

Corpus	**# Dialects**	**Recording Method**	**Corpus Details**
CALLFRIEND	1 dialect	Spontaneous telephone conversations	60 conversations, lasting between 5-30 minutes.
OrienTel MCA	5 dialects	Telephone response of questionnaire	Number of speakers: 750 Egyptian, 757 Jordanian, 772 Moroccan, 792 Tunisian and 880 Emirates.
DARPA Babylon Levantine	4 dialects	Direct recording of spontaneous speech	164 speakers, 75900 Utterances, Size: 6.5 GB, 45 hours.
Appen's corpora	3 dialects	Spontaneous telephone conversations	Gulf: 975 conver, ∽ 93 hours ; Iraqi: 474 conver, ∽ 24 hours; Levantine: 982 conver, ∽ 90 hours.
Fisher Levantine	5 dialects	Guided telephone conversations	279 conversations, 45 hours.
TuDiCoI	1 dialect	Spontaneous dialogue	127 Dialogues, 893 utterances.
SAAVB	1 dialect + MSA	Selected speaker before telephone response of questionnaire	1033 speakers; 83% MSA utterances, 17% DA utterances, Size: 2.59 GB.
MDP	3 dialects + MSA	Direct Recording	52 speakers; 23% MSA utterances, 77% DA utterances, 32 hours.
KSU Rich Arabic	9 dialects + MSA	Guided telephone conversations and Direct recording.	201 speakers from nine Arab countries.

The speakers are selected from Jordan, Lebanon, Palestine, Lebanon, Syria and other Levantine countries.

TuDiCoI (Graja et al., 2010) is a spontaneous dialogue speech corpus dedicated to Tunisian dialect, which contains recorded dialogues between staff and clients in the railway of Sfax town, Tunisia.

Concerning corpora that gather MSA and Arabic dialect, we have studied three corpora. *SAAVB* corpus is dedicated to speakers from all the cities of Saudi Arabia country using telephone response of questionnaire method (Alghamdi et al., 2008). The main characteristic of this corpus is that before recording, a speaker and environment selection are performed. The selection aims to control speaker age and gender and telephone type. Almeman and al. (2013) have compiled a *Multi-Dialect Parallel (MDP)* corpus which gathers MSA and three Arabic dialects. Namely, the dialect are from Gulf, Egypt and Levantine. The speech data is collected by direct recording method.

KSU Rich Arabic corpus encompasses speakers by different ethnic groups, Arabs and Non-Arabs (Africa and Asia). Concerning Arab speakers in this corpus, they are selected from nine Arab countries: Saudi, Yemen, Egypt, Syria, Tunisia, Algeria, Sudan, Lebanon and Palestine. This corpus is rich in many aspects. Among them, richness of the recording text. In addition, different recording sessions, environments and systems are taken into account (Alsulaiman et al., 2013).

According to our study of these major corpora specially those dedicated to dialects, we make some remarks. First, we should mention the fact that these corpora are mainly fee-based and the free ones are extremely rare. We can enumerate only one free corpus, namely "*MDP corpus*", as confirmed by Zaghouani in his recent critical survey on freely available corpora (Zaghouani, 2014). Second, let us also observe that the direct recording method combined with a chosen text, is less used in spite of that it exhibits better the language/dialect features. Third, concerning our purpose, we can observe that there are no corpus dedicated to Algerian dialect variety. In fact, just KSU corpus represents Algerian dialect by means of few number of speakers (four). Concerning Algerian MSA, only ALGASD and KACST corpora have considered Arabic Algerian speakers, however these corpora are not free.

In what follows, we give some information about Algerian dialects before describing the followed methodology and the collected corpus.

3. A Glance at Algerian Sub-Dialects

Algeria is a large country, with a total area of about 2.4 million km^2. The country is bordered by mainly three Arabic countries, in the North-East by Tunisia, in the East by Libya, in the West by Morocco. It is administratively divided into 48 departments. Algeria's official language is MSA like in all Arab countries. In every day life, Algerian dialect is the most used compared with MSA.

Algerian Arabic dialect is one of Maghrebi dialect group. It has many variations which are mainly developed as a result of both phases of arabization and French deep colonizations. According to the arabization phases, we can historically classify Algerian dialects into three major groups: pre-Hilālī, Hilālī and the mixed dialects. The pre-Hilālī are called rural, sedentary dialects; which are spoken in areas that are affected by the expansion of Islam in the 7^{th} century. The Hilālī dialects take its name from the Banu Hilāl tribu, it is named Bedouin dialects; spoken in areas which are influenced by the Arab immigration in the 11th century. Called urban pre-Hilālī, the mixed dialects are spoken in regions that are affected by both arabization phases (Palva, 2006), (Pereira, 2011). Furthermore, Algerian dialects are also influenced by the long period of a deep colonization.

Compared with the commonly spoken Arabic dialects in other parts of the arabic world, Algerian one is somewhat different. It is considerably varietal from a region/department to another. In fact, the language has been greatly affected by Berber and other languages, such as Turkish, French, Italian, and Spanish (Leclerc, 2012), (Chami, 2009). In fact, Algerian sub-dialects have many borrowed words (Guella, 2011). For instance, it is significantly affected by the French language. Furthermore, it is important to note the omnipresence of the code-switching phenomenon between the two languages.

For more details on Algerian dialects characteristics refer to (Maïri, 1987), (Taleb-Ibrahimi, 1995), (Caubet, 2000 2001), (Pereira, 2011).

4. Methodology

The crucial points to be taken into consideration when designing and developing relevant speech corpus are numerous. We can mention some of them: scope and the size of the corpus, richness of speech topics and content, number of speakers, gender, regional dialects, recording environment and materials. We have attempted to cover a maximum of these considerations. We will underline each considered point in what follows.

In order to build our dialect corpus, we have to resolve and fix some choices, namely which text (words, sentences and paragraphs) we have to put in the corpus that foster its richness?, which method to use in order to collect speech data?, and finally which speaker profile we have to adopt?

First, let us recall that collecting dialect speech data can be done by means of four ways: straightforward recording some reports and telecasts from regional radios and TV, by telephone response of questionnaire, by spontaneous telephone conversations, or by direct recording. Obviously, it can be performed also by combining at least two ways.

In our case, we have selected *Direct Recording* method to collect speech data in order to provide a rich corpus that can be used to display differences and similarities between dialects. Furthermore, a direct recording compared with broadcast news-based method, allows us to control speaker's profile.

4.1. Corpus Text Content

Concerning the speech material, we have adopted a handmade text. This latter with selected and guided content allows the corpus to be parallel. So, it allows highlighting fairly dialect features and catching speaker's finger, . . .

We prompt the speaker to utter utterance through a question sheet. Our question sheet is composed of many words, sentences and paragraphs. We have categorized them into four sections according to their purpose.

4.1.1. Spontaneous Speech –SS part–

Where the speaker answers 05 questions, which are "Where are you from?", "What is the time?", "Tell us about what do you like in your city?", "Tell us about the weather today?", and "Describe to us the last meal you have eaten?". The aim of this data is to record conversations.

4.1.2. Read Sentences in MSA –RS part–

In this category, we have selected ten sentences, which are phonetically rich and balanced. We have taken the tenth list from the research of Boudraa et al. (Boudraa et al., 2000). This part has twofold purpose. First, MSA speech are represented in the corpus. Second, it is shown by Ammar et al. (Ammar et al., 2014) that the accent in MSA utterance is widely influenced by the dialect of the speaker.

4.1.3. Translate Text –TT part–

The speaker have to translate 18 common words, 16 sentences from MSA to his dialect:

- *Digits:* from 0 through 10 and *days of the week.*
- *Accent sentences/paragraphs:* this part includes four sentences, which contain the pronunciation of the main five discriminative sounds in Arabic dialects. This part gathers between the three forms of sentences, which are normal, negative and interrogative.

  The aim of adding this part is justified by the studies of Taine-Cheikh (Taine-Cheikh, 2000) and Holes (Holes, 2004). Taine-Cheikh proves that the pronunciation of [q] and [ð], which are in Arabic the letter "ق" *qaf* and "ذ" *d̠al* respectively, are discriminative sounds in Arabic dialects, especially they are quite discriminative when we deal with Algerian dialects. In addition, Holes adds to this list three other discriminative sounds. Two interdentals sounds [θ], and [ðˤ], which are in Arabic the letter "ث" *t̠a* and "ظ" *Ẓa* letters respectively, and an alveolar affricate sound [ʤ] which is "ج" *ǧim* in Arabic letter.

  The texts of this part are selected from two different corpora. From SAAVB corpus (Alghamdi et al., 2008), we have selected two accent variation sentences, this is due to two reasons: i) The first sentence includes the five discriminative sounds in MSA words that rarely change their nucleus in the dialect of a speaker, it is:

  "جَاءَ الضُيُوف الثَلاثَة بالذَهَب قَبلَ الظُهر"

  /ʒaːʔ aldˤyuːf alθalaːθah bialðahab qabl alðˤuhr/

  "The three guests came with the gold before noon"

  ii) The second sentence is an interrogative sentence. It is an indirect question introduced by "why" which expresses the interrogative style.

  "لِمَاذَا سَافَرتَ إِلَى الخَارِجِ فِي العِيد؟"

  /limaːðaː saːfart ʔilaː alχaːriʒ fiː alʕiːd/

  "Why did you travel abroad during the festival?"

From another linguistic corpus (Ammar et al., 2014), we have selected two sentences. They include interdental fricative sounds, with a short sentence in negative style.

- *A short text story:* this part includes 12 sentences, which are selected from the well known story "The North Wind and the Sun". We have intentionally added this part as this story represents a reference text used by the Association International Phonetics (AIP) for phonetic description of world languages[8]. AIP provides the record of this story in many languages including MSA. It represents de facto standard in most corpora that are used in language processing for phonetic research. Furthermore, it is widely used for Arabic dialect processing. It makes performing comparative studies more easier.

  Specifically for Arabic dialects, this story is presented in the corpus of Hamdi et al. (Hamdi et al., 2004) and also in Araber Corpus for Arabic dialects collected by Barkat-Defradas available through *Corpus Pluriels* of praxiling laboratory [9].

4.1.4. Sub-Spontaneous Sentences –SSS part–

In this category, the speaker has to narrate, in his dialect, an image story by interpreting a set of 24 ordered pictures selected from kid story "Frog, where are you?, which is called in linguistic literature "The Frog Story" (Mayer, 1969). This part leads to 24 utterances. It is added to the corpus text content for many purposes. The first, is that, this story is near to daily speech in its grammatical and lexical varieties. In addition, it contains different levels and types of linguistic expression. Like "The North Wind and the Sun" story, it is frequently used in main corpora. Furthermore, it can be used to perform prosodic measurements due to the call of the narration style as proven by (Himmelmann and Ladd, 2008). In fact, for highlighting prosodic features, a semi-guided narration style allows two antagonist criteria. First, like in a spontaneous speech, it gives freedom to the speaker to narrate in his style. Second, it guides the narrator to follow story events and to use a specific vocabulary, essentially the motion verbs.

In Arabic, this story is used by (Barkat et al., 1999) to sketch up the fact that the prosody features can be discriminative for Arabic dialects. In addition, it is presented in Araber corpus, cited bellow, which has been used in prosodic research of Rouas (Rouas, 2007).

Table 2. gives a summary of the corpus's speech material described below. *Ratio* column expresses the ratio of each part in the whole text content in term of sentences.

4.2. Speaker Profile

The speakers are chosen from adult population with an age in the range 18-50 years. We have made sure that speakers and their parents are native from the department of the corresponding dialect.

Table 2: Corpus Speech Material.

Part	#Utterance	Ratio (%)
SS	05	8.8 %
RS	10	17.5 %
TT	18	31.6 %
SSS	24	42.1 %
Total	57	100

Let us mention that when the department is a great metropolis, we have made sure that the speakers are from the same area. In fact, we have noticed that for Algiers, Oran and Constantine departments, we can have more than one subdialect.

The speaker's personal information are provided by the speaker him-self. These information are namely: first and last name (optional), date and place of birth, education level, origin of his/her parents and their education level. The applicant has to inform about the different places/dates where/when he lived since he/she has 2 years old.

4.3. Material and Environment Recording

Concerning the recording environment, some conditions are respected. In fact, all recordings are performed in a quiet environment and when it is possible, we have performed recording in a sound proof room. In order to avoid noise in the recorded speeches, all the recording are done in nice days without wind, thunder and rains.

Concerning the hardware configuration of the recording material, we have use a dictaphone *TASCAM DR-05* sets. The sample rate of recording is 48K sample/sec with 16 bits resolution.

In order to explain how we have got the speech data, let us sketch the procedure in these two steps:

Before Recording Step

First, we explain to each applicant the purpose and objectives of the targeted speech corpus in order to reassure and give him confidence. In addition, the applicant fills in an information form that will provides useful information related to the speaker profile.

Second, each applicant has to take notes in advance about the whole question sheet because of the translation of some sentences in his dialect.

During Recording Step

Many verbal trials are allowed until we and the speaker are satisfied about the quality of dialect utterance, specially RS, TT and SSS parts. Among the satisfaction criteria that we have looked after that the speaker avoids using French terminology as possible as he can. However, he can use lexical borrowing, where the speaker use a foreign word, which is adapted to his syntactic/morphology.

Generally, each recording is performed twice. However, when the speaker feels uncomfortable or shy, we repeat the recording until we get an acceptable quality one.

[8] https://www.internationalphoneticassociation.org

[9] http://www.praxiling-corpuspluriels.fr

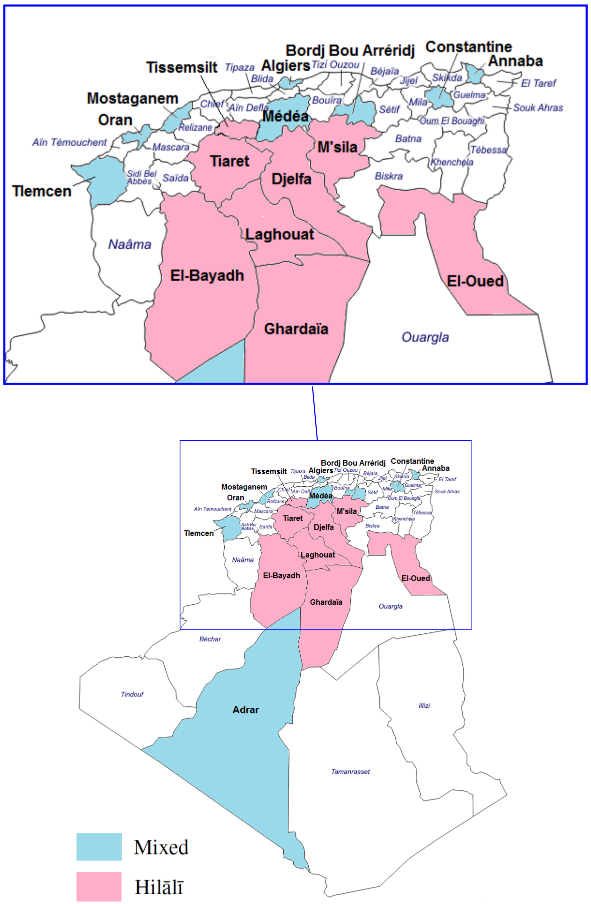


Figure 2: Distribution of covered departments in ALG-DARIDJAH corpus.

5. Our Spoken Corpus

In this section, we describe the current version of the collected corpus. Let us mention that the task is very challenging and it is time consuming. While the recording operation is still ongoing, what we describe here is the preliminary version. We have baptised our corpus **ALG-DARIDJAH** for **ALG**erian **Darijah**. In the Maghreb, the term Daridjah الدارِجَة means dialectal arabic, in the Est, they use the vocable Al-Amia العامية.

5.1. Corpus Description

Due to the lack of efficient linguistic atlas for Algeria, we have chosen to collect spoken data from chef-lieu of departments where there are prominent population density. This fairy fine-grained division favors efficient dialect feature captures.

Within this preliminary version, ALG-DARIDJAH covers more than one third of the whole sub-dialects of Algeria. The covered sub-dialects are from departments of Adrar, Algiers, Annaba, Bordj Bou Arréridj, M'sila, Constantine, Djelfa, El-Bayadh, El-Oued, Ghardaïa, Laghouat, Média, Mostaganem, M'sila, Tiaret, Tissemsilt, Tlemcen and Oran. Figure 2 illustrates their geographical distribution. Let us note that dialects spoken in these departements are quite close to each other. They differ mainly in pronunciation and some local words. For this reason, we talk about sub-dialects.

Table 3: ALG-DARIDJAH Speakers Distribution.

Departement	Gender		Total
	#Male	#Female	
Adrar	04	02	06
Algiers	04	04	08
Annaba	05	05	10
Bordj Bou Arréridj	-	01	01
Constantine	01	01	02
Djelfa	05	05	10
El-Bayadh	01	03	04
El-Oued	05	05	10
Ghardaïa	05	05	10
Laghouat	05	05	10
Médéa	04	05	09
Mostaganem	-	01	01
M'sila	01	09	10
Tiaret	02	04	06
Tissemsilt	02	02	04
Tlemcen	02	02	04
Oran	02	02	04
Total	48	61	109

Table 4. gives a summary of some statistics of the built corpus. A sample of ALG-DARIDJAH corpus is available online [10]. ALG-DARIDJAH corpus encompassed 17 sub-dialects, an average of more than 6 speakers by sub-dialect and 109 in total. The male speaker ratio is about 44%. Table 3. gives more details on the distribution of speakers for each sub-dialect.

Concerning provided annotations, each utterance in ALG-DARIDJAH has time-aligned orthographic word transcription. In addition to this annotation, we have performed automatic syllable segmentation of all utterances by using Praat tool enhanced by the script Prosogram V 2.9 (Mertens, 2004). The segmentation is performed in the context of a system design that identifies dialect based on prosody. In Figure 3, we report a sample of an orthographic transcription of one utterance among Translate Text part (TT) sentences.

As a first corpus analysis, there is a lot of lexical differences between sub-dialects. We have captured some of them which are illustrated in Table 4.

5.2. Corpus Package Organization

In order to facilitate the deployment of ALG-DARIDJAH corpus, we have adopted the same packaging method as TIMIT Acoustic-Phonetic Continuous Speech Cor-

[10] http://perso.lagh-univ.dz/~hcherroun/Alg-Daridja.html

MSA	جَاءَ الضُيُوف الثَلاثَة بالـذَهَـب قَبـلَ الظُـهر /ʒaːʔ aldˤyuːf alθalaːθah bialðahab qabl alðˤuhr/
Laghouat dialect	جَاونَا ثْلاَثْ ضْيَافْ وْ جَابُو مْعَاهُمْ ذْهَبْ قُدَامْ الظُهُرْ / ʒawnaː tlaːθ dˤiaːf w ʒaːbuː mʔˤhum ðhab gudam alðˤhur/
Oran dialect	جَاو تِلْتْ دْيَافْ بِدْهَبْ قْبَلْ الدُهُرْ / ʒaw tilt diaːf bidhab qbal alduhur/
El-Bayadh dialect	ضْيَافْ ثْلاَتَ جَابُو ذْهَبْ قُدَامْ الظُهُرْ /dˤiaːf tlaːta ʒaːbuː ðhab gudaːm alðˤhur/

Figure 3: A sample orthographic transcription.

Table 4: Corpus Statistics and Details.

Number of targeted departments	17
Number of speakers	109
Number of utterances	6213
Average recording by speaker	2.50 mn
Total recording	∽ 4 h 30 mn
Size of corpus	1.8 Gb

pus (Garofolo et al., 1993). In fact, TIMIT is considered as de facto standard of speech corpora. In the root directory of ALG-DARIDJAH package, we provide some textual files with .txt extension:

```
- Prompt.txt: Table of sentence prompts.
- Speakerinfo.txt: Table of speaker
  description.
- Orthocode.txt: Table of symbols
  used in orthographic transcriptions.
```

For each speaker, we have four folders, each one contains a part of text corpus. In each folder, there are a wave file and two transcription files for each utterance.

6. Potential Uses

The built corpus, is the first of its kind in term of that it is the first rich recorded corpus for Algerian Arabic dialects. It can be useful for many purposes both for NLP and computational linguistic communities. In fact, it can be used for building efficient models for both speaker and dialect identification systems for Algerian dialects. For linguistic and sociolinguistics communities, it can serve as base for capturing dialects characteristic. It can also be used for test and evaluation of such systems by splitting the corpus into both training and test parts.

In addition, linguists can use our fine-grained spoken corpus to create contemporary dialect atlases by gathering similar sub-dialects.

7. Conclusion

In this paper, we have presented ALG-DARIDJAH, a speech corpus dedicated to a part of Algerian Arabic sub-dialects. We have focused on describing the methodology of its design. In fact, we have justified the made choices on the text content, selected speakers, collection method and the recording material and environment. The followed methodology is performed respect to the wide literature in the domain of building speech corpora. The rich text content of the speech is designed such that it highlights the features of the dialect. Indeed, a well chosen words to express all discriminative phones. Text narrations in dialect variations to reflect features of dialects.

Englais	MSA	Sub-dialects	
Look for	يبحثا	Adrar: يْحَوطُو	/jħawtˤuː/
		Ghardaïa: يْحَوسُو	/jħawsuː/
		Laghouat: يْدَورُو	/jdawruː/
		Tiaret: يْرَڨْبُو	/jragbuː/
Child	الطفل	El-Bayadh: بَزْ	/baz/
		Tlemcen: لِولِدْ	/wild/
		Oran: غُريَانْ	/ɣurjaːn/
Dispute	تنازعتا	Algiers: دَارْبُو	/daːrbuː/
		Annaba: تْعَاركُو	/tʔˤaːrkuː/
		Constantine: تْفَاتْنُو	/tfaːtnuː/
		Laghouat: دَوسُو	/dawsuː/
		Médéa: ضَارْبُو	/dˤaːrbuː/
		Tiaret: دَاڨُو	/daːguː/
		Tlemcen: دَابْزُو	/daːbzuː/

Figure 4: Illustrate of some lexical differences in ALG-DARIDJAH.

The first version of ALG-DARIDJAH corpus contains 6213 utterance spoken by 109 native speakers. It covers 17 Algerian dialects. The annotations give some information about speakers and transcription.

Mainly developed to be used in dialect identification, ALG-DARIDJAH can serve as a testbed supporting evaluation of wide spectrum of natural language processing approaches and applications.

Actually, we are working on spreading the corpus to all Algerian sub-dialects. Indeed, the recording is ongoing to cover the remain dialects of other departments. In addition, we work on other corpora engineering aspects such valida-

tion, more annotations, and manual transcription.
In future work, we will extend the corpus by collecting Algerian sub-dialects uttered by berber native speakers.

8. References

Abushariah, M. A. M., Ainon, R. N., Zainuddin, R., Elshafei, M., and Khalifa, O. O. (2012). Phonetically rich and balanced text and speech corpora for Arabic language. *Language Resources and Evaluation*, 46(4):601–634.

Alghamdi, M., Alhargan, F., Alkanhal, M., Alkhairy, A., Eldesouki, M., and Alenazi, A. (2008). Saudi Accented Arabic Voice Bank. *Journal of King Saud University-Computer and Information Sciences*, 20:45–64.

Almeman, K., Lee, M., and Almiman, A. A. (2013). Multi Dialect Arabic Speech Parallel Corpora. In *Communications, Signal Processing, and their Applications (ICCSPA)*, pages 1–6, Feb.

Alsulaiman, M., Muhammad, G., Bencherif, M. A., Mahmood, A., and Ali, Z. (2013). KSU Rich Arabic Speech Database. *Journal of Information*, 16(6).

Ammar, Z., Fougeron, C., and Ridouane, R. (2014). A la recherche des traces dialectales dans l'arabe standard: production des voyelles et des fricatives inter-dentales par des locuteurs tunisiens et marocains. In *JEP, Le Mans, France*, pages 684–693.

Barkat, M., Ohala, J., and Pellegrino, F. (1999). Prosody as a Distinctive Feature for the Discrimination of Arabic Dialects. In *Eurospeech, Budapest, Hungary*, pages 395–398.

Behnstedt, P. and Woidich, M. (2013). Dialectology. In Owens, J., editor, *The Oxford Handbook of Arabic Linguistics*, pages 300–325.

Boudraa, M., Boudraa, B., and Guerin, B. (2000). Twenty Lists of Ten Arabic Sentences for Assessment. *Acta Acustica united with Acustica*, 86(5):870–882.

Canavan, A. and Zipperlen, G. (1996). CALLFRIEND Egyptian Arabic LDC96S49. Philadelphia: Linguistic Data Consortium.

Caubet, D. (2000-2001). Questionnaire de Diactologie du Maghreb (D'après les Travaux de W. Marçais, M. Cohen, G. S. Colin, J. Cantineau, D. Cohen, P. Marçais, S. Levy, etc.) . *Estudios de Dialectologia Norteafricana y Andalusi*, 5:73–92.

Chami, A. (2009). A Historical Background of the Linguistic Situation in Algeria. مجلة المواقف للبحوث و الدراسات في المجتمع و التاريخ, 4:387–395.

Choukri, K., Nikkhou, M., and Paulsson, N. (2004). Network of Data Centres (NetDC): BNSC - An Arabic Broadcast News Speech Corpus. In *LREC*. European Language Resources Association.

Djellab, M., Amrouche, A., Bouridane, A., and Mehallegue, N. (2016). Algerian Modern Colloquial Arabic Speech Corpus (AMCASC): regional accents recognition within complex socio-linguistic environments. *Language Resources and Evaluation*, pages 1–29.

Droua-Hamdani, G., Selouani, S., and Boudraa, M. (2010). Algerian Arabic Speech Database (ALGASD): Corpus Design and Automatic Speech Recognition Application. *Arabian Journal for Science and Engineering*, 35:158.

ELRA. (2005). NEMLAR Broadcast News Speech Corpus. ELRA catalog ELRA-S0219.

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

Garofolo, J. S., Lamel, L. F., Fisher, W. M., Fiscus, J. G., Pallett, D. S., and Dahlgren, N. L. (1993). DARPA TIMIT Acoustic Phonetic Continuous Speech Corpus CDROM.

Gibbon, D., Moore, R., and Winski, R. (1998). *Spoken Language System and Corpus Design*. Handbook of Standards and Resources for Spoken Language Systems. Bod Third Party Titles.

Graja, M., Jaoua, M., and Hadrich-Belguith, L. (2010). Lexical Study of A Spoken Dialogue Corpus in Tunisian Dialect. In *The International Arab Conference on Information Technology (ACIT), Benghazi, Libya*.

Guella, N. (2011). Emprunts lexicaux dans des Dialectes Arabes Algériens. *Synergies Monde arabe*, 8:81–88.

Habash, N. (2010). *Introduction to Arabic Natural Language Processing*. Synthesis Lectures on Human Language Technologies. Morgan and Claypool Publishers.

Hamdi, R., Barkat-Defradas, M., Ferragne, E., and Pellegrino, F. (2004). Speech Timing and Rhythmic structure in Arabic dialects: a comparison of two approaches. In *InterSpeech*.

Himmelmann, N. P. and Ladd, D. R. (2008). Prosodic Description: An Introduction for Fieldworkers.

Holes, C. (2004). *Modern Arabic: Structures, Functions, and Varieties*. Georgetown University Press.

LaRocca, S. and Chouairi, R. (2002). West Point Arabic Speech LDC2002S02. Philadelphia: Linguistic Data Consortium.

Leclerc, J. (2012). Algérie dans l'aménagement linguistique dans le monde. Web, 30 avril.

Lewis, M. P., Gary, F. S., and Charles, D. F. (2015). Ethnologue: Languages of the World, Eighteenth edition. Web.

Lindquist, H. (2009). *Corpus Linguistics and the Description of English*. Edinburgh University Press Series. Edinburgh University Press.

Maïri, L. (1987). A Bibliography of Algerian Arabic Linguistics. *Zeitschrift für Arabische Linguistik*, (17):96–107.

Makhoul, J., Zawaydeh, B., Choi, F., and Stallard, D. (2005). BBN/AUB DARPA Babylon Levantine Arabic Speech and Transcripts. Linguistic Data Consortium (LDC). LDC Catalog Number LDC2005S08.

Mansour, M. A. (2013). The Absence of Arabic Corpus Linguistics: A Call for Creating an Arabic National Corpus. *International Journal of Humanities and Social Science*, 3(12):81–90.

Mayer, M. (1969). *Frog, where are you?* New York: Dial Press.

Mertens, P. (2004). The Prosogram: Semi-Automatic Transcription of Prosody based on a Tonal Perception Model. In *Speech Prosody, Nara, Japan*, pages 23–26.

Palva, H. (2006). Dialects: Classification. *Encyclopedia of Arabic Language and Linguistics*, 1:604–613.

Pereira, C. (2011). Arabic in the North African Region. In Weniger, S., Khan, G., Streck, M. P., and Watson, J. C. E., editors, *Semitic Languages. An International Handbook*, pages 944–959. Berlin.

Rouas, J. L. (2007). Automatic Prosodic Variations Modeling for Language and Dialect Discrimination. *Audio, Speech, and Language Processing, IEEE Transactions on*, 15(6):1904–1911, Aug.

Taine-Cheikh, C. (2000). Deux macro-discriminants de la dialectologie arabe (la réalisation du qâf et des interdentales). *Matériaux arabes et sudarabiques (GELLAS)*, 9:11–51.

Taleb-Ibrahimi, K. (1995). *Algériens et leurs langues*. Collection Connaissance de l'Algérie contemporaine. Editions el Hikma.

Versteegh, K. (1997). *The Arabic Language*. Edinburgh University Press, Cambridge.

Zaghouani, W. (2014). Critical Survey of the Freely Available Arabic Corpora. In *LREC'14 Workshop on Free/Open-Source Arabic Corpora and Corpora Processing Tools (OSACT). Reykjavik, Iceland*, pages 1–8.

Towards a New Arabic Corpus of Dyslexic Texts

Maha M Alamri, William J Teahan
School of Computer Science, Bangor University
Bangor, United Kingdom
elp003@bangor.ac.uk, w.j.teahan@bangor.ac.uk

Abstract

This paper presents a detailed account of the preliminary work for the creation of a new Arabic corpus of dyslexic text. The analysis of errors found in the corpus revealed that there are four types of spelling errors made as a result of dyslexia in addition to four common spelling errors. The subsequent aim was to develop a spellchecker capable of automatically correcting the spelling mistakes of dyslexic writers in Arabic texts using statistical techniques. The purpose was to provide a tool to assist Arabic dyslexic writers. Some initial success was achieved in the automatic correction of dyslexic errors in Arabic text.

Keywords: Arabic, Corpus, Dyslexia, Errors, Spelling

1. Introduction

Given that almost one fifth of the population suffers from some form of a language-related disability and based on the fact that an overwhelming majority of them (70–80%) are probably dyslexics, studying dyslexia and an array of related matters is of a particular importance (International Dyslexia Association, 2011).

However, dyslexia is mostly related to the inability of a person to master the utilization of written language, including issues with comprehension. It is vital to emphasise that dyslexia is not an illness insofar as it cannot be cured, nor will it just disappear as an affected child gets older. There are ways of helping these people to ease their difficulties and improve their quality of life in the pertinent areas of using language. One of the tools used in this regard are computers (Pedler, 2007).

Notwithstanding the application of corpora consisting of the most common errors made by dyslexics in diagnosing dyslexia (Schulte-Körne et al., 1996), or in creating new and better spellcheckers, the availability of these corpora is limited and their number is very low (Pedler, 2007).

2. Spelling Errors

2.1. Common Spelling errors

Damerau argued based on his research that 80% to 95% of all mistakes made in spelling result in a mis-spelt word that comprises usually a similar amount of letters to the correct version of the word (Damerau, 1964). In his study, Damerau also pointed out that there are four possible categories according to which spelling errors can occur: additional letters e.g. *unniverse*; omitted letters e.g. *univrse*; substituted letters e.g *umiverse*; and swapped letters e.g. *uinverse*. Moreover, real-word errors and non-word errors are two kinds of spelling errors. A non-word error does not have any meaning (Mishra and Kaur, 2013) and is not found in a dictionary (Samanta and Chaudhuri, 2013). Real-word errors happen when someone mistakenly types a correctly spelt word but another was intended, which could be meaningful but is not appropriate for the sentence structure (Islam and Inkpen, 2009).

2.2. Spelling errors caused by dyslexia

The mistakes in spelling made by dyslexic writers are of a significantly higher level of severity than those of non-dyslexic writers (Coleman et al., 2008). There are several studies, for instance that of Fisher et al. (Fischer et al., 1985), that used the categorization of words according to the relation between their pronunciation and their spelling. Their findings indicate that people suffering from dyslexia are challenged more by words that necessitate knowledge of a complex word structure than with words that are learned through repetition or words whose spelling is possible to predict (Moats, 1993).

Very often, words contain certain letters that are silent; therefore, the correct form of the word needs to be first learnt in order to produce the correct spelling of the word on subsequent occasions. Examples of such words include knife, knight, hour, etc. Another example might be the word ‘musician’, which is derived from the word ‘music’, yet the pronunciation of the ‘c’ in both words differs substantially. There are also certain words that modify the spellings of their morphemes in the case of when affixes are added: explain–explanation, miracle–miraculous, etc. (Bourassa and Treiman, 2008).

The struggle of dyslexic writers with the relationship between the sound of a word and its spelling represents a considerable hindrance in the way they acquire the ability to write in a systematic manner. In other words, their preoccupation with the morphology and phonetics of language often prevents them from dealing with the peculiarities of the language’s orthography (Korhonen, 2008). Regardless, the degree to which dyslexic writers make errors is largely determined by the nature of the writing system of the particular language in which they are writing (Lindgrén and Laine, 2011).

It has been suggested that people suffering from dyslexia face a substantial challenge in terms of orthographic spellings. A particular challenge for dyslexic students at universities in this regard relates to words that are exceptional or rare and thus do not belong to commonly used vocabulary (Meyler and Breznitz, 2003). Even highly skilled dyslexic students, who otherwise manage to do well utilizing various simple phonological strategies, struggle with words that necessitate memorization (Kemp et al., 2009).

It is assumed that this might be caused by an absence of reading experience with dyslexic students, or by their inability to preserve in memory orthographic symbols. Such inability has been suggested to be connected to an inadequate visual memory insofar as dyslexic students find it difficult to remember the right order of letters in a word. Research into this area has sometimes concluded that dyslexia across languages and linguistic systems shares the same difficulty in phonological decoding (Aaron, 1989). However, researching dyslexia across other orthographies may help to understand how works across other languages, as well as dyslexia on the whole (Abu-Rabia, 2001; Abu-Rabia et al., 2003).

2.3. Spelling errors by Arabic writers with dyslexia

Within research into dyslexia, there are a variety of linguistic challenges that are particular to Arabic. Firstly, there have been limited studies into dyslexia in Arabic, due to the fact that dyslexia is not recognised in many Arabic cultures to be a particular type of reading issue, and academic research and interest in this area has been minimal. This is despite a substantial effort on the side of educational figures and organisations to bring more attention to the existence of learning difficulties and special needs (Elbeheri et al., 2006).

There have also been limited studies on spellings of both regular readers and dyslexic readers in Arabic (Abu-Rabia and Sammour, 2013).

One study by Abu-Rabia and Taha (2004) examined the spelling mistakes observed in speakers and writers in Arabic, both dyslexic and aged-matched readers. This showed seven types of errors that may be observed in the three different types of participants (dyslexics, reading level matched readers, and aged-matched readers) phonetic errors. Alongside this, students may spell an Arabic word according to how they hear it in the local spoken dialect of Arabic which they use in their day-to-day life, rather than using the correct Arabic spelling of it. Furthermore, semi-phonetic errors, dysphonetic errors, visual letter confusion, irregular spelling rules, word omission and functional word omission.

They found that the misspellings and errors made by the dyslexic group were comparable to those of the readers who were matched by reading level, both in terms of frequency and of type. The most frequent spelling errors were phonetic. They also found that the types of spelling errors made were related to the orthography being Arabic.

Abu-Rabia and Sammour (2013) examined the errors of Arabic dyslexic students in comparison with age-matched and spelling-level-matched regular students in two languages, Arabic and English. The spelling errors analysis was based on four criteria: phonetic, semiphonetic, dysphonetic and word omission errors. In Arabic, the lack of knowledge of spelling rules led to most of the phonetic errors. The spelling mistakes consisted of the inability to specify the correct form of the Hamzah, difficulty in writing the letters in the correct shape, and Hamzat-lwasl was the most common spelling error. In addition, long vowel errors and exchanging consonants were common. As a result, the phonetic error type is the most notable error in Arabic.

3. Related Corpora

To date, the produced corpora of high quality are commonly found in most Latin-based languages, but similar corpora based on Arabic remain a rarity (AbdelRaouf et al., 2010). Furthermore, there is a noticeable absence of corpora designed specifically for the needs of dyslexics. To the best of our knowledge, there are two corpora used primarily for dyslexic texts:

- The corpus employed by Pedler (2007) constructed a spelling correction program that focuses on errors in words made by those with dyslexia. This version comprises 3,314 English words accompanied by 363 corresponding errors (Pedler, 2007). Structurally, this corpus consists of homework, completed via Microsoft Word, produced by a student in the third year of secondary school and saved prior to being spellchecked. Furthermore, it includes two samples with errors that were employed in a test comparing spellcheckers (Mitton, 1996). Finally, this corpus comprises pieces of creative writing composed during the 1960s by secondary school students with low levels of academic ability (Holbrook, 1964). Developing a program capable of correcting errors in the writing of dyslexics necessitated increasing the size of the abovementioned corpus to 21,524 words, with 2,654 errors and 800 real-word errors.

- The Spanish corpus (Dyscorpus), which was created by Rello (2014), was collected from dyslexic children aged between 6 and 15 years. The total of the texts is 83: 54 from school essays and homework exercises and 29 from parents of dyslexic children, with a total of 1,171 errors. Moreover, the corpus was annotated and creates a list of the unique errors from Dyscorpus.

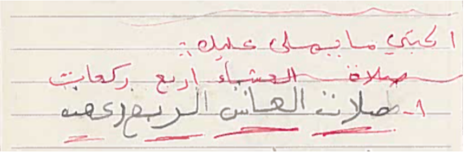
It is worth highlighting at this point, a corpus that takes into account the needs of Arabic dyslexics is missing.

4. Arabic Corpus of Dyslexic Texts

According to Sterling et al. (1998), the rate of misspellings in the text is noticeably higher in the case of children. Therefore, the texts were collected from female primary school students with dyslexia who have been taught in resource rooms [1], been professionally diagnosed with dyslexia and given an appropriate learning environment. The collection resulted in Arabic texts composed by female dyslexic pupils aged 8-10 years. These texts were initially produced for native Arab-speaking students as writing exercises. These texts served as a platform for forming a corpus comprising 1,067 words, with 694 errors. Thus, this

[1] "A room in an ordinary school which students with special needs attend for a period of not more than a half of the school day for the purpose of receiving special education services from a special education teacher." (Ministry of Education of Saudi Arabia, 2002)

attempt has the character of a preliminary version of the Bangor Dyslexic Arabic Corpus (BDAC), which aims to investigate the possibility of a corpus being used as an aid for Arabic dyslexic writers. Figure 1 [2] is an example of a portion of one of the texts used in this research.



ا كتبي مايملى عليك:

صلاة العشاء أربع ركعات

١- صلات العاش الربع ركعت

Figure 1: Screenshot of a scanned image of one of the texts written by a dyslexic female child (nine years old).

The abovementioned example included basic errors. Table 1 shows the wrong and correct forms of words as taken from Figure 1:

Type of error	Error	Correct	Explanation
Substitution	صلات	صلاة	(ت) Instead (ة)
Insertion	الربع	اربع	Added (ل)
Omission	ركعت	ركعات	Deleted (ا)
Transposition	العاش	العشاء	(ـاشـ) Instead (ـشا)

Table 1: Errors and correct words from Figure1.

Table 2 presents numerical data regarding the corpus and its composition. It lists the number of documents, characters, words, sentences, paragraphs and errors for the dyslexic texts.

Category	Docs	Chars	Words	Sent.	Para.	Errors
Words only	20	1646	357	–	–	357
Sentences only	45	1577	380	64	–	192
Paragraphs	7	1292	330	30	7	145
Total	72	4515	1067	94	7	694

Table 2: Total number of documents, characters, words, sentences, paragraphs and errors in the dyslexic texts.

Having analysed the BDAC corpus, the following types of mistakes as detailed below were evident. The first four are basic spelling errors and the remaining errors were also, found in the corpus.

1. Omission had the highest number of occurrences (191 times). The common omission word is hamzah . Students often forget to write hamzah (ء) either on the top of the letter(أ) or at the end of a word, such as in the example of (بناء). In total, there were 168 mistakes of this type.

2. Insertion (64 times).

3. Substitution (47 times). The replacement of (ه) to (ة), which is due to the similarity in their sounds, for example the incorrect form (كثيره) and its correct form (كثيرة). In total, this type of error was registered 79 times in the corpus. This error is again based on mistakenly changing (ة) or (ه) with the letter (ت), presumably caused by the similarity in their sounds. An example of this error can be seen in (غاليت), whereas the correct form is (غالية). There were 28 errors of this type in the corpus. This was followed by the replacing of (ت) with the letters (ة) or (ه), which is caused by the similarity in their sounds. For illustration of this type of error, one can look at (أصبحة), whose correct form is (أصبحت). This kind of error occurred nine times in the corpus represented by exchanging the letter (ض) with (ظ) or vice versa. An example of this type of error is (مضهر) and its correct form is (مظهر). This type of error occurred eight times in the corpus..

4. Transposition (19 times).

   The Arabic system of writing necessitates the use of a group of symbols in a form of diacritics that are found either on top of or below letters to express a particular kind of gemination, case, or silence (Béland and Mimouni, 2001)

5. Long vowel ‘ـا ، ـى ’ (both pronounced as /a:/), ‘ـي’ (pronounced as /i:/) and‘ـو ’ (pronounced as /u:/)(Zitouni et al., 2006). In the case of this corpus, an error of this type was observed 18 times. It should be noted that this type of error is critical as long vowels are formed through a combination of ‘ـا ، ـي ’and ‘ـو’ .

   There are two types of short vowels:

6. **Damma** stands for the /u/ sound and is marked by a symbol (اُ)(Zitouni et al., 2006). The analysis of this corpus revealed that students often mistakenly used the letter (و) to indicate damma, like for example (لوغتي) instead of (لُغتي) . The total number of words containing this type of error was 11.

7. **Kasra** indicates the /i/ sound and and is marked by a symbol (اِ) (Zitouni et al., 2006).Through analyzing the corpus, it was possible to observe that some students made an error in using the letter (ي) in place of kasra, for example (ثيمار) instead of (ثِمار). In total, there were 18 words that included this type of mistake.

8. The term **Tanween** denotes a situation where a short vowel is put on the last letter of a word to indicate the **N** sound (Zitouni et al., 2006). Arabic distinguishes between three types of tanween diacritics: tanween

[2] Approximate translation : Isha prayer consists of four rakats

al-fatha, tanween al-damma and tanween al-kasra. It should be highlighted that it is quite common for a dyslexic student to replace (tanween) (ةٌ) with the letter N, as in (ثقتن) where the correct form is (ثقة). There were 34 instances of this type of error.

5. Towards Automatic Correction of Dyslexic Errors

This study also attempted to investigate the possibility of utilizing automatic natural language processing techniques as a form of assistance for Arabic dyslexic writers.

TMT is a software package designed specifically to conduct tasks revolving around compression, text categorisation and correction, and segmentation of the text (Teahan, 2016). The toolkit was used to correct a small number of the dyslexic errors using a method that was similar to the method described by Alhawiti (Alhawiti, 2014) found effective for the correction of errors in Arabic OCR text. First, it was crucial to choose a large training corpus of Arabic text to train the compression-based language model created by the toolkit.

After researching suitable corpora, the Bangor Arabic Compression Corpus (BACC) created by Alhawiti (Alhawiti, 2014) was chosen. Due to the current limitations of the TMT software, the correction of the dyslexic texts was applied just for one-to-one character errors using the toolkit's markup correction capabilities that was able to find the most probable corrected sequence given the compression-based language model. Full details of the correction technique that was used is provided in Alhawiti's thesis (Alhawiti, 2014).

5.1. Experimental Results

As stated, based on the limitations regarding the use of the statistical TMT software, the results for this method are based just on one-to-one character errors. Prior to this experiment, all errors containing more than one character were removed with the result that the total number of errors in the BDAC corpus that were able to be corrected dropped from 694 to 280.

The documents in the BDAC are divided into word, sentence and paragraph types. For word type documents, the total numbers of errors are equal to 153. The results show that the TMT software was able to correct 99 of the one-to-one character errors in these documents. In the case of sentence type documents, there are 80 errors and the TMT software was able to correct 49 of them. For paragraph type documents, there are 47 errors and the software was able to correct 39 of them. The overall results for corrections was 67% (correcting 187 out of 280 errors) meaning that the TMT software was able to correct more than half of the one-to-one character errors.

6. Future work

Currently, work is continuing on extending the corpus. In order to minimize difficulties faced when collecting the preliminary version of Bangor Arabic Dyslexia Corpus, fieldwork was undertaken in KSA and currently around 9000 words of text written by dyslexia students has been collected. Some of the texts come from tests whilst other texts were taken from children's homework.

All of these texts were collected from the Resource room. With the analysis of the further texts, there are errors similar to those mentioned in section 4. whilst there are also further odd errors such as:

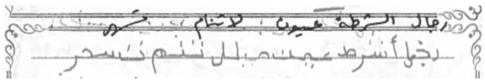


Figure 2: Further errors caused by dyslexia in Arabic texts.

As shown in Figure 2, the student wrote the shape of the letter (ي) as if it was at the end of the word. The correct shape should be (يـ) because the letter is in the middle of the word.

7. Conclusion

This paper illustrated the process of building a new corpus, which is now available for public use via the author's blog `mahaalamri.wordpress.com`. The research in this area is compounded by a general lack of any Arabic corpus specifically comprising dyslexic errors. The overall size of the BDAC corpus has currently reached 9000 words. Regarding the relative small size of this corpus, this can be explained by the absence of easily available texts composed by dyslexics. Regardless of its limited size, the corpus used in this study offers a useful platform for analysing dyslexic errors made in the Arabic language whilst providing a better understanding of the occurrence of these errors and the factors determining such occurrences and therefore it is suitable for assisting dyslexic writers. Furthermore, this corpus can serve as a platform for other researchers to build upon. It can be used as a first step in developing a much larger corpus.

This study also attempted to investigate the possibility of utilizing natural language processing techniques as a form of assistance for Arabic dyslexic writers and some initial success was achieved in the automatic correction of dyslexic errors in Arabic text. In future work, it requires considerably more resources and effort to extend the corpus to include more text for analysis (such as to include Arabic texts written by people of different genders, ages and education).

8. Bibliographical References

Aaron, P. (1989). Orthographic systems and developmental dyslexia: A reformulation of the syndrome. In *Reading and writing disorders in different orthographic systems*, pages 379–400. Springer.

AbdelRaouf, A., Higgins, C. A., Pridmore, T., and Khalil, M. (2010). Building a multi-modal Arabic corpus (MMAC). *International Journal on Document Analysis and Recognition (IJDAR)*, 13(4):285–302.

Abu-Rabia, S. and Sammour, R. (2013). Spelling errors' analysis of regular and dyslexic bilingual Arabic-English students. *Open Journal of Modern Linguistics*, 3(01):58.

Abu-Rabia, S. and Taha, H. (2004). Reading and spelling error analysis of native Arabic dyslexic readers. *Reading and Writing*, 17(7-8):651–690.

Abu-Rabia, S., Share, D., and Mansour, M. S. (2003). Word recognition and basic cognitive processes among reading-disabled and normal readers in Arabic. *Reading and writing*, 16(5):423–442.

Abu-Rabia, S. (2001). The role of vowels in reading semitic scripts: Data from Arabic and Hebrew. *Reading and Writing*, 14(1-2):39–59.

Alhawiti, K. M. (2014). *Adaptive Models of Arabic Text*. Ph.D. thesis, The School of Computer Science, Bangor University.

Béland, R. and Mimouni, Z. (2001). Deep dyslexia in the two languages of an Arabic/French bilingual patient. *Cognition*, 82(2):77–126.

Bourassa, D. C. and Treiman, R. (2008). Morphological constancy in spelling: A comparison of children with dyslexia and typically developing children. *Dyslexia*, 14(3):155–169.

Coleman, C., Gregg, N., McLain, L., and Bellair, L. W. (2008). A comparison of spelling performance across young adults with and without dyslexia. *Assessment for effective intervention*.

Damerau, F. J. (1964). A technique for computer detection and correction of spelling errors. *Communications of the ACM*, 7(3):171–176.

Elbeheri, G., Everatt, J., Reid, G., and Mannai, H. a. (2006). Dyslexia assessment in Arabic. *Journal of Research in Special Educational Needs*, 6(3):143–152.

Fischer, F. W., Shankweiler, D., and Liberman, I. Y. (1985). Spelling proficiency and sensitivity to word structure. *Journal of memory and language*, 24(4):423–441.

Holbrook, D. (1964). English for the rejected: Training literacy in the lower streams of the secondary school.

International Dyslexia Association. (2011). Frequently asked questions about dyslexia.

Islam, A. and Inkpen, D. (2009). Real-word spelling correction using Google Web IT 3-grams. *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing,EMNLP 2009*, pages 1241–1249.

Kemp, N., Parrila, R. K., and Kirby, J. R. (2009). Phonological and orthographic spelling in high-functioning adult dyslexics. *Dyslexia*, 15(2):105–128.

Korhonen, T. (2008). *Adaptive spell checker for dyslexic writers*. Springer.

Lindgrén, S.-A. and Laine, M. (2011). Multilingual dyslexia in university students: Reading and writing patterns in three languages. *Clinical linguistics & phonetics*, 25(9):753–766.

Meyler, A. and Breznitz, Z. (2003). Processing of phonological, orthographic and cross-modal word representations among adult dyslexic and normal readers. *Reading and Writing*, 16(8):785–803.

Ministry of Education of Saudi Arabia. (2002). Regulations of special education institutions and programmes . Kingdom of Saudi Arabia, Riyadh,administration of special education.

Mishra, R. and Kaur, N. (2013). A survey of spelling error detection and correction techniques. *International Journal of Computer Trends and Technology*, 4(3).

Mitton, R. (1996). *English spelling and the computer*. Longman Group.

Moats, L. C. (1993). Spelling error interpretation: Beyond the phonetic/dysphonetic dichotomy. *Annals of Dyslexia*, 43(1):174–185.

Pedler, J. (2007). *Computer correction of real-word spelling errors in dyslexic text*. Ph.D. thesis, Birkbeck College, University of London.

Rello, L. (2014). *A Text Accessibility Model for People with Dyslexia*. Ph.D. thesis, Department of Information and Communication Technologies, University Pompeu Fabra.

Samanta, P. and Chaudhuri, B. B. (2013). A simple real-word error detection and correction using local word bigram and trigram. In *ROCLING*.

Schulte-Körne, G., Deimel, W., Müller, K., Gutenbrunner, C., and Remschmidt, H. (1996). Familial aggregation of spelling disability. *Journal of Child Psychology and Psychiatry*, 37(7):817–822.

Sterling, C., Farmer, M., Riddick, B., Morgan, S., and Matthews, C. (1998). Adult dyslexic writing. *Dyslexia*, 4(1):1–15.

Teahan, W. J., (2016). *An improved interface for probabilistic models of text.* Bangor University. Technical Report.

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 the 44th annual meeting of the Association for Computational Linguistics*, pages 577–584. Association for Computational Linguistics.

MANDIAC: A Web-based Annotation System For Manual Arabic Diacritization

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

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

Abstract

In this paper, We introduce MANDIAC, a web-based annotation system designed for rapid manual diacritization of Standard Arabic text. To expedite the annotation process, the system provides annotators with a choice of automatically generated diacritization possibilities for each word. Our framework provides intuitive interfaces for annotating text and managing the diacritization annotation process. In this paper we describe both the annotation and the administration interfaces as well as the back-end engine. Finally, we demonstrate that our system doubles the annotation speed compared to using a regular text editor.

Keywords: Arabic Diacritization, Annotation, Tool

1. Introduction

Modern Standard Arabic (MSA) script employs a writing system that is typically under-specified for short vowels and diacritical marks (for ease of exposition, both are referred to here as diacritics), for the most part, words are written as a sequence of consonants and long vowels.[1] The absence of diacritics in text adds another layer of lexical and morphological ambiguity to the natural ambiguity present in language. While such ambiguity rarely impedes proficient speakers, it can certainly be a source of confusion for beginning readers (Abu-Rabia, 1999). Hence, having explicit diacritics on text is useful for several Natural language Processing (NLP) applications as well as text readability and understanding.

From statistical NLP perspective, the two universal problems that affect the performance of several tools and tasks are: (1) sparseness in the data, where not enough instances of a word type is observed in a corpus, and (2) ambiguity, where a word has multiple readings or interpretations. The examples given in Table 1 show the six possible pronunciations and meanings for the undiacritized form of the Arabic word ذكر.

Much work has been done on automatic Arabic diacritization (Vergyri and Kirchhoff, 2004; Kirchhoff and Vergyri, 2005; Zitouni et al., 2006; Diab et al., 2007; Rashwan et al., 2011; Shahrour et al., 2015; Abandah et al., 2015). Designing and implementing methods to automatically assign diacritics to each letter in a word requires a large amount of manually annotated training data. Most of the reported systems are trained using Arabic Treebanks (Maamouri et al., 2010). However, to adapt for different genre and dialects there is a need for creating new datasets which is time-consuming and arduous as it places heavy demands on human annotators for maintaining annotation quality and consistency (Stenetorp et al., 2012).

To alleviate some of the burden, we describe the design and implementation of our web-based annotation system, MANDIAC, developed to help expedite the manual diacritization task and reduce manual annotation errors. The MANDIAC framework provides intuitive interfaces for both managing the annotation process and performing the annotation tasks.[3]

To the best of our knowledge, there exist no tools or web interfaces dedicated for large-scale manual diacritization annotation. As opposed to raw text editors, our system provides facilities for managing thousands of documents, distributing tasks to tens of annotators and evaluating inter-annotator agreement (IAA), thereby allowing for seamless quality control of the annotation process. Additionally, we demonstrate that our system doubles the annotation speed compared to using a basic text editor. In these respects, MANDIAC architecture is based on QAWI (Obeid et al., 2013) a token-based editor, developed to manually correct spelling errors in Arabic text for the Qatar Arabic Language Bank (QALB) project, a large-scale manually annotated Arabic text correction project (Zaghouani et al., 2014; Zaghouani et al., 2015; Zaghouani et al., 2016b; Mohit et al., 2014; Rozovskaya et al., 2015).

The remainder of this paper is organized as follows. In Section 2., we present our annotation web interface; Then,we describe the design and architecture of our annotation system in Section 3.; We demonstrate in Section 4. that our system provides a significant increase in annotator productivity.

[1] Only Classical Arabic texts such as religious texts (e.g., Quranic texts) are written with full explicit diacritic specification. Diacritics are used to minimize chances of reciting them incorrectly.

[3] This annotation interface will be made available

Undiacritized	Diacritized	Buckwalter[2]	English
ذكر	ذَكَرَ	/*akara/	He mentioned
ذكر	ذُكِرَ	/*ukira/	It was mentioned
ذكر	ذَكَّرَ	/*ak ˜ara/	He reminded
ذكر	ذُكِّرَ	/*uk ˜ira/	It was reminded
ذكر	ذَكَرٌ	/*akaruN/	Male
ذكر	ذِكْرٌ	/*ikoruN/	Prayer

Table 1: Possible pronunciation and meanings of the undiacritized Arabic word ذكر

2. Annotation Web Interface

In this section we present the annotation web interface which will be used to carry out the annotation process.

2.1. Annotation Interface

The MANDIAC interface was designed to accommodate large scale manual and semi-automatic diacritization task. The system was created to be a token-based editing tool (see Figure 1). However, instead of performing corrections on text directly, we use MADAMIRA (Pasha et al., 2014), a system for morphological analysis and disambiguation of Arabic texts, to compute a ranked list of diacritizations for each token. MADAMIRA uses a morphological analyzer to produce, for each input word, a list of all possible analyses. Then it applies a set of models to produce a prediction, per word in context, for different morphological features, such as POS, lemma, gender, number or person. A ranking component scores the analyses produced by the morphological analyzer using a tuned weighted sum of matches with the predicted features. For MADAMIRA top-scoring analysis, the word error rate for the diacritization task is 5.5% if we ignore the diacritization error on the last word letters. Last word letters diacritics are related to case and mood. For most of the cases, it can be one of three possible diacritics; 'a','u' or 'i' for definite noun case; 'F' ,'N' or 'K' for indefinite noun case; or 'a', 'u' or 'o' for verb mood. Accordingly, we restrict the diacritization choices presented to the annotators to the top three unique MADAMIRA analyses as this guarantees the correct diacritization will be one of the candidates for 94.5% of the cases. An option to manually edit a token is available in the case where none of the precomputed candidates is correct (see Figure 2).

Furthermore, the token editor (Figure 3), using regular expressions, insures only Arabic diacritics are added, removed, or modified. This added a significant consistency checks in terms of annotators usage.

Finally, the annotation interface, illustrated in Figure 1, provides annotators with a few extra utilities such as:

- Undo and redo buttons;
- A timer to help the annotator measure her/his speed;
- A link to the annotation guidelines;
- A counter that tracks the number of words that are yet to be annotated;
- Highlighting to differentiate between tokens that have been or are yet to be annotated and those that should not be annotated (such as digits, punctuation);
- The option to flag documents with issues such as bad format, poor writing quality or dialectal usage. This will alert the annotation manager about the issue.

2.2. Annotation Management Interface

The annotation management interface is used by the annotation manager to organize the entire annotation workflow such as adding user accounts, uploading and assigning files, tracking the annotation and annotation evaluation using automatic Inter-annotator agreement measures.

User Accounts Management: The user accounts management menu allows the annotation workflow manager to add and remove user accounts, add users to specific annotation groups, compute the user activity log with detailed statistics as shown in Figure 4.

Annotation workflow Management: The annotation workflow management interface is designed to assist the annotation manager in the following annotation workflow operations: (a) uploading files, (b) organizing files by project, (c) file assignment to a designated group of annotators, (d) file management operations such as adding files, removing files and viewing files.

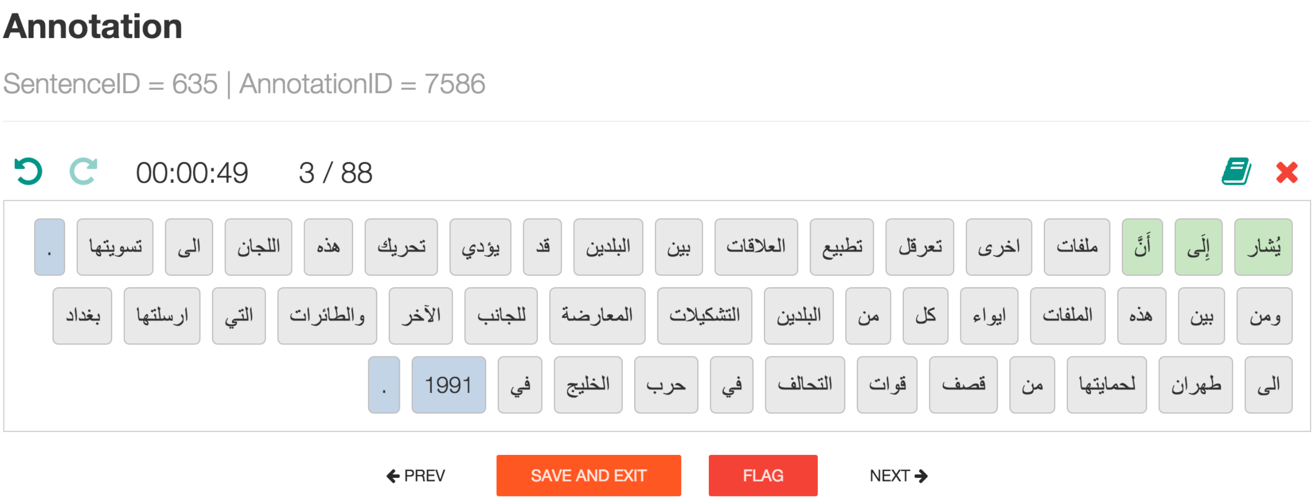
Finally, in order to help the annotation manager to track the annotation progress, completed tasks are highlighted in green while pending tasks are highlighted in yellow as shown in Figure 5 and Figure 6.

Evaluation and Monitoring: In addition to the annotator and file management, MANDIAC's management interface allows to evaluate inter-annotator agreement and compare the annotations produced by each annotator to a gold reference, in order to monitor the quality of annotations during the life cycle of each project. We use various evaluation metrics such as the Word Error Rate (WER) and the Diacritics Error Rate (DER) on a randomly assigned blind files on 10% of the data.

3. System Design and Architecture

3.1. General Architecture

The MANDIAC system is a Web application composed of four major components: the back-end server, the MADAMIRA toolkit, the annotation interface, and the management interface. The back-end server provides an



Annotation

SentenceID = 635 | AnnotationID = 7586

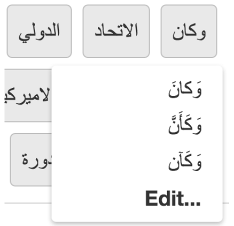
00:00:49 3 / 88

يُشار إِلَى أَنَّ ملفات اخرى تعرقل تطبيع العلاقات بين البلدين قد يؤدي تحريك هذه اللجان الى تسويتها .

ومن بين هذه الملفات ايواء كل من البلدين التشكيلات المعارضة للجانب الآخر والطائرات التي ارسلتها بغداد

الى طهران لحمايتها من قصف قوات التحالف في حرب الخليج في 1991 .

Figure 1: An example of a sentence, as displayed in our annotation interface



وكان الاتحاد الدولي

لاميركي

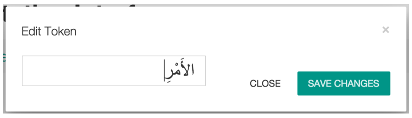
ورة

وَكانَ

وَكَأَنَّ

وَكَآن

Figure 2: An example of a token with possible diacritizations, in our interface



Edit Token

الأَمْرِ

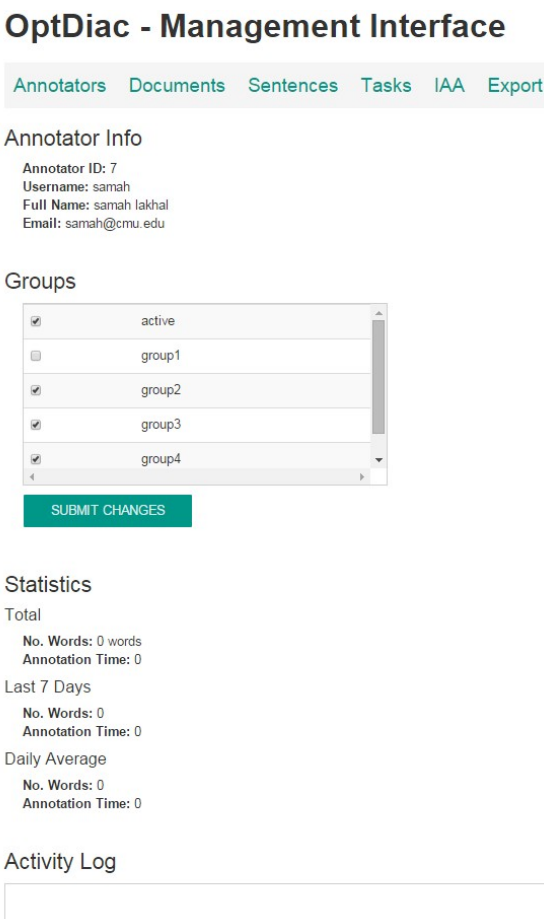
Figure 3: A screenshot of the token editor in MANDIAC

HTTP REST API through which the management and annotation interfaces interact with the server. These interactions all occur using JSON objects allowing for great flexibility (discussed in Section 3.2.).

The back-end server also uses HTTP requests to obtain precomputed diacritizations from MADAMIRA. MADAMIRA has the advantage of running as a server allowing us to process documents faster than using the command line interface as MADAMIRA's large models only need to be loaded once on start-up.

3.2. Data Storage

All data in MANDIAC is stored on a single relational database. Our database design uses a JSON blob to store content, while utilizing various fields to store meta-data. This allows for quick data search and retrieval and ensures more flexibility in the content that can be stored. Consequently, the front-end can be extended to support different annotation modes without having to modify the back-end



OptDiac - Management Interface

Annotator Info

Annotator ID: 7
Username: samah
Full Name: samah lakhal
Email: samah@cmu.edu

Groups

☑ active
□ group1
☑ group2
☑ group3
☑ group4

Statistics

Total

No. Words: 0 words
Annotation Time: 0

Last 7 Days

No. Words: 0
Annotation Time: 0

Daily Average

No. Words: 0
Annotation Time: 0

Activity Log

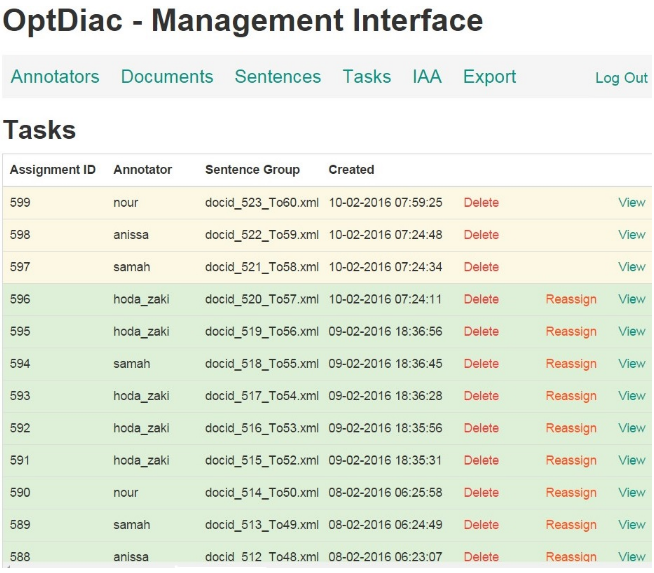
Figure 4: User management interface in MANDIAC

significantly, if at all.

4. Experimental setup and evaluation

4.1. Annotation task description

Annotators are asked to fully diacritize each word. Our annotation task is formulated primarily as a selection task



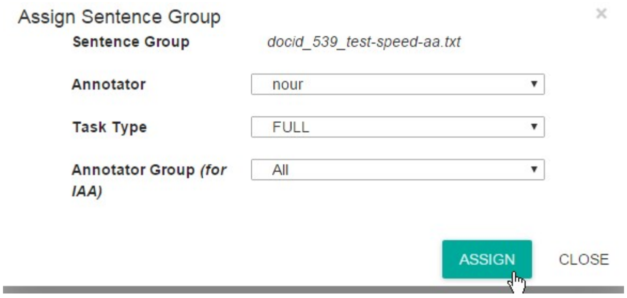
OptDiac - Management Interface

Annotators Documents Sentences Tasks IAA Export Log Out

Tasks

Assignment ID	Annotator	Sentence Group	Created			
599	nour	docid_523_To60.xml	10-02-2016 07:59:25	Delete		View
598	anissa	docid_522_To59.xml	10-02-2016 07:24:48	Delete		View
597	samah	docid_521_To58.xml	10-02-2016 07:24:34	Delete		View
596	hoda_zaki	docid_520_To57.xml	10-02-2016 07:24:11	Delete	Reassign	View
595	hoda_zaki	docid_519_To56.xml	09-02-2016 18:36:56	Delete	Reassign	View
594	samah	docid_518_To55.xml	09-02-2016 18:36:45	Delete	Reassign	View
593	hoda_zaki	docid_517_To54.xml	09-02-2016 18:36:28	Delete	Reassign	View
592	hoda_zaki	docid_516_To53.xml	09-02-2016 18:35:56	Delete	Reassign	View
591	hoda_zaki	docid_515_To52.xml	09-02-2016 18:35:31	Delete	Reassign	View
590	nour	docid_514_To50.xml	08-02-2016 06:25:58	Delete	Reassign	View
589	samah	docid_513_To49.xml	08-02-2016 06:24:49	Delete	Reassign	View
588	anissa	docid_512_To48.xml	08-02-2016 06:23:07	Delete	Reassign	View

Figure 5: Annotation task assignment interface



Assign Sentence Group ×

Sentence Group: *docid_539_test-speed-aa.txt*

Annotator: nour

Task Type: FULL

Annotator Group *(for IAA)*: All

Figure 6: User-Task assignment interface

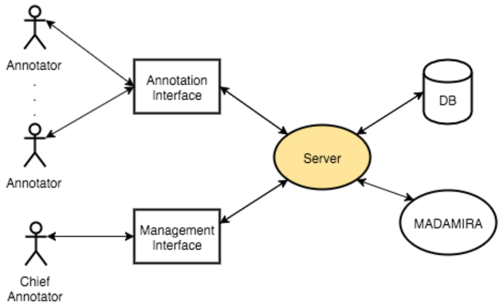


Figure 7: Component interaction diagram.

with minimal editing. 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 provided candidates, they are allowed to manually edit the word and add/replace the correct diacritics. This technique is designed in order to reduce annotation time, reduce annotator workload and especially reduce diacritization errors (Bouamor et al., 2015; Zaghouani et al., 2016a).

For each word, we generate a list of vocalized candidates using MADAMIRA. MADAMIRA achieves a lemmatization accuracy of 99.2% and a full diacritization accuracy of 86.3%. We present the annotator with the top three candidates suggested by MADAMIRA, where available, as illustrated in Figure 2.

4.2. Evaluation

To evaluate the efficiency of the MANDIAC system, we assigned five annotators around 1,500 words each extracted from the Penn Arabic Treebank (Maamouri et al., 2010). Half of the words were annotated using a text editor while the other half were annotated using our system. Table 2 shows the average speeds of both approaches in words per hour. The results obtained clearly indicate that our system allows annotators to double their annotation speed. This is mainly due to the availability of the correct diacritization as one of the offered options in most of the cases, in addition to the improved visual presentation compared to a classical editor. Moreover, the annotators who used a text editor introduced some errors and typos while editing the text, on the other side, those who used our tool were prevented of freely editing the text and produced a consistent annotation since they are using a dedicated tool with a designated pull down menu showing a list of options to select from or manually adding the diacritics while ensuring that only diacritics are being added or removed.

Approach	Speed (words/hour)
Classic text editor	302
MANDIAC system	618

Table 2: Comparison of average annotation speed.

Furthermore, we conducted several experiments (Zaghouani et al., 2016a) to compute inter-annotator agreement (IAA) to evaluate the extent to which our trained annotators agree on the diacritics added for each word using the MANDIAC tool on a portion of the corpus of contemporary Arabic (Al-Sulaiti and Atwell, 2006).

We measured IAA scores between two annotators by averaging WER (Word Error Rate) over all pairs of words as defined in (Snover et al., 2006). In this experiment, should a single letter in a given word has a diacritization error, then the whole word is considered as erroneous. Overtime, we conducted three IAA iterations to check for possible annotation consistency improvement using our tool. The results given in Table 3 show a regular IAA improvement after each iteration with a WER reduced to 9.31%. Note that the higher the WER between two annotations, the lower their agreement.

5. Conclusion

In this paper, we demonstrated a new token-based annotation system for manual Arabic diacritization, MANDIAC. We have shown that the system allows annotators to be more productive, doubling their annotation speed, by providing them with precomputed diacritizations. In future, we plan to release the tool and make it freely available to the research community so it can be used in other related

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

Table 3: Average WER obtained after each annotation iteration on the CCA corpus (Zaghouani et al., 2016a).

annotation tasks that could benefit from precomputed annotations.

6. Acknowledgements

We thank the three anonymous reviewers for their valuable comments and suggestions. This publication was made possible by grant NPRP 6-1020-1-199 from the Qatar National Research Fund (a member of Qatar Foundation). The statements made herein are solely the responsibility of the author[s].

7. 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.

Abu-Rabia, S. (1999). The effect of Arabic vowels on the reading comprehension of second-and sixth-grade native Arab children. *Journal of psycholinguistic research*, 28(1):93–101.

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.

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 annotation. *ANLP Workshop 2015*, page 80.

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

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., Seth Kulick, S. K., Gaddeche, F., and Zaghouani, W. (2010). Arabic Treebank: Part 3 v 3.2 LDC2010T08.

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., Zaghouani, W., Mohit, B., Habash, N., Oflazer, K., and Tomeh, N. (2013). A web-based annotation framework for large-scale text correction. In *The Companion Volume of the Proceedings of IJCNLP 2013: System Demonstrations*, pages 1–4, Nagoya, Japan, October. Asian Federation of Natural Language Processing.

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.

Rashwan, M. A., Al-Badrashiny, M. A., Attia, M., Abdou, S. M., and Rafea, A. (2011). A stochastic arabic diacritizer based on a hybrid of factorized and unfactorized textual features. *Audio, Speech, and Language Processing, IEEE Transactions on*, 19(1):166–175.

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.

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, September. Association for Computational Linguistics.

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

Stenetorp, P., Pyysalo, S., Topić, G., Ohta, T., Ananiadou, S., and Tsujii, J. (2012). BRAT: A Web-based Tool for NLP-assisted Text Annotation. In *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 102–107, Avignon, France.

Vergyri, D. and Kirchhoff, K. (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.

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., Bouamor, H., Hawwari, A., Diab, M., Obeid, O., Ghoneim, M., Alqahtani, S., and Oflazer, K. (2016a). Guidelines and framework for a large scale arabic diacritized corpus. In *International Conference on Language Resources and Evaluation (LREC 2016).*

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

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.

Toward an Arabic Punctuated Corpus: Annotation Guidelines and Evaluation

Wajdi Zaghouani[1] **and Dana Awad**[2]

[1]**Carnegie Mellon University in Qatar**
wajdiz@qatar.cmu.edu
[2]**The Lebanese University**
danaawad@hotmail.fr

Abstract

We present our effort to build a large scale punctuated corpus for Arabic. We illustrate in details our punctuation annotation guidelines designed to improve the annotation work flow and the inter-annotator agreement. We summarize the guidelines created, discuss the annotation framework and show the Arabic punctuation peculiarities. Our guidelines were used by trained annotators and regular inter-annotator agreement measures were performed to ensure the annotation quality. We highlight the main difficulties related to the Arabic punctuation annotation that arose during this project.

Keywords: Punctuation Annotation, Arabic Language, Guidelines, Evaluation

1. Introduction

Punctuation can be defined as the use of spacing and conventional signs to help the understanding of handwritten and printed texts. Punctuation marks are used to create sense, clarity and stress in sentences and also to structure and organize text.

Punctuation rules vary with language and register. Some punctuation aspects are of stylistic choices. For a language such as Arabic, punctuation marks are relatively a modern innovation since Arabic did not use punctuation and it was mainly introduced via translation and through borrowing from European languages as stated in (AlQinai, 2015) and therefore punctuation rules in Arabic are not consistently used although its becoming more popular in recent years. Asfour (2009) discussed this issue in his book on Arabic punctuation from which we cite the following quote:

> *"There is no doubt that Arabic has borrowed certain writing styles from foreign languages through translation, but the establishment of rules for Arabic punctuation does not depend on the various translations done by such and such translators, but mainly on what is commonly accepted by the Arabic writers."* Asfour (2009)

Attia et al. (2014) investigated the punctuation use in the parallel English/Arabic Gigaword corpus and found that around 10% of tokens in the English Gigaword corresponded to punctuation marks, while it is only 3% tokens in the in the Arabic counterpart.

From a Natural Language Processing (NLP) perspective, punctuation marks can be useful in the automatic sentence segmentation tasks, since sentence boundaries and phrase boundaries can be estimated based on punctuation marks. Moreover, as shown by a number of studies e.g., (Furui et al., 2004; Jones et al., 2003; Matusov et al., 2007), the absence of punctuation can be confusing both for humans and computers. Jones et al. (2003), for example, showed that sentence breaks are critical for text legibility. Furthermore, many NLP systems trained on well-formatted text often have problems when dealing with unstructured texts. Furui et al. (2004) showed that speech summarization improves when the sentence boundaries were available. Also, Matusov et al. (2007) showed that use of punctuation is beneficial for machine translation.

In order to build robust automatic punctuation systems, large scale manually punctuated corpora are usually needed. In this paper, we present our effort to build a large scale punctuated corpus for Arabic. We present our punctuation annotation guidelines designed to improve the inter-annotator agreement. Our guidelines were used by trained annotators and regular inter-annotator agreement measures were done to ensure the annotation quality during the lifespan of the project.

In the next sections, we briefly review the related work (Section 2), describe our corpus and the punctuation guidelines (Sections 3 and 4), and review our annotation procedure (Section 5), then we present the evaluation in Section 6.

2. Related Work

Large scale manually punctuated corpora are not yet widely available due to the high cost related to building such resources. We were able to locate some related works on punctuation annotation systems and resources such as the lightweight punctuation annotation system for the Arabic language built by Beeferman et al. (1998) to automatically insert intra-sentence punctuation into Arabic text. More recently, Attia et al. (2014) developed a punctuation errors correction tool based on a CRF (Conditional Random Fields) classifier, and Mubarak and Darwish (2014) used two approaches to restore punctuation marks in Arabic text, in the first approach they used a simple statistical word-based system, and in the second approach a conditional random fields (CRF) sequence labeler that tries to recover punctuation based on Part-of-Speech (POS) of the current word and also of the two previous and following words. They used also word sequences such as the current word, the previous or the next word in the sentence.

For English, Spitkovsky et al. (2011) showed how punctuation can be used to improve unsupervised dependency parsing. They presented a linguistic analysis to confirm the connection between English punctuation and phrase boundaries in the Penn Treebank. The Arabic punctuated corpus we present in this paper is a part of a large scale manually error annotated project described in the next section.

3. Corpus Description

In this work, we describe a 2 million words corpus developed for the Qatar Arabic Language Bank (QALB) project, a large-scale annotation effort created to deliver a manually corrected corpus of errors including punctuation errors for a variety of Arabic texts (Zaghouani et al., 2014b; Zaghouani et al., 2015; Zaghouani et al., 2016). The overreaching goal of the punctuation annotation sub-task in this project is twofold: correct the existing punctuation found in text, then add the missing necessary punctuation when needed. The corpus presented contains a variety of text genres and sources: (a) user comments on news websites (L1 corpus), (b) native speaker essays (L1 corpus) (c) non-native speaker essays (L2 corpus), (d) machine translation output (MT corpus) as shown in Table 1 . The *user comments* portion is 1.5 million words corpus collected from AlJazeera.Net news.[1] The comments are selected from the related news stories. The *native student essays corpus* includes 151,000 words extracted from the Arabic Learners Corpus (ALC) Alfaifi and Atwell (2012) and 66,000 words from the University Students Essays Corpus (Alkanhal et al., 2012).

The *non-native student essays corpus* has 131,000 words selected from the Arabic Learners Corpus (ALC) by Alfaifi and Atwell (2012) and 51,000 words from the Arabic Learners Written Corpus (ALWC) compiled by Farwaneh and Tamimi (2012) . The data is organized by the students level (beginner, intermediate, advanced), learner type (L2 vs. heritage),[2] and essay type (description, narration, instruction).

Finally, the *machine translation output corpus* was collected from English news articles covering 100,000 words extracted from the collaborative journalism Wikinews website[3]. The corpus includes 520 articles with an average of 192 words per article. The original English files were in HTML format and later on exported to a UTF-8 plain Text standard format so it can be uploaded in the annotation tool. Afterwards, the collected corpus was automatically translated from English to Arabic using the Google Translate API paid service [4].

4. Punctuation Annotation Guidelines

Punctuation errors are estimated to constitute 40% of the errors in the QALB corpus, that is 10 times higher than the 4% of punctuation errors found in the English data used in CoNLL 2013 Shared Task on English Grammatical Error Correction (Ng et al., 2013). As mentioned earlier in this article, punctuation use in English or French language is guided by a series of grammar-related rules, while in other languages such as Arabic, punctuation is a recent innovation as premodern Arabic did not use punctuation (Zaki, 1995). According to Awad (2013), there is an inconsistency in the punctuation rules and use in Arabic, and omitting the punctuation marks is very frequent. To create our guidelines, we use the Arabic standard punctuation rules commonly used today and described in Awad (2013).

The annotation guidelines typically document the core of the annotation policy. Our punctuation guidelines focus on the types of punctuation errors that are targeted and they describe how the correction should be done and also when new punctuation marks should be added. Many annotated examples are provided in the

[1] We thank AlJazeera.Net for granting the rights to use the user comments from their website in our corpus.

[2] Heritage learner is generally used to describe a person learning a language and who has some proficiency in or he is culturally connected to that language through country of origin or family.

[3] https://en.wikinews.org

[4] https://cloud.google.com/translate

Corpus Type	Source	Size
News Comments (L1)	Aljazeera.Net	1500
Native Students Essays (L1)	Arabic Learners Corpus (Alfaifi and Atwell, 2012)	151
Native Students Essays (L1)	University Students Essays (Alkanhal et al., 2012)	66
Non-Native Students Essays (L2)	Arabic Learners Corpus (Alfaifi and Atwell, 2012)	131
Non-Native Students Essays (L2)	Arabic Learners Written Corpus (Farwaneh and Tamimi, 2012)	51
Machine Translation Output	Wikinews (English-Arabic MT)	100

Table 1: Corpus types and sources used in the QALB project. The size is displayed in K-words.

guidelines to illustrate the various annotation rules and exceptions.

Since the Arabic punctuation rules are not always clearly defined, we adopted an iterative approach for developing the guidelines, which includes multiple revisions and updates needed for the different rounds of of annotation in order to reach a consistent set of instructions. In order to help the annotators deal with some complex punctuation rules, we wrote a summary of the most common punctuation marks rules in Arabic as an appendix to the guidelines.

Furthermore, we instructed the annotators to convert any Latin punctuation sign to the equivalent Arabic sign when applicable. In particular, we instructed them to use always the Arabic comma ،, the Arabic Semicolon ؛ and the Arabic question marks ؟ instead of the English equivalent punctuation marks. The punctuation annotation guidelines are published as a technical report (Zaghouani et al., 2014a) and they are currently available for download from the Qatar Arabic Language Bank project web page.[5]

The rules of punctuation vary with the language and the register. Moreover, some aspects of punctuation use vary from author to author, and can be considered a stylistic choice. Punctuation errors are especially present in student essays and online news comments. This is mainly due to the fact that some punctuation mark rules are still not clearly defined in Arabic writing references as explained previously. Table 2 shows an example of two punctuation errors and their corrections.

We created a set of simple rules for correcting punctuation and adding the missing ones. Below, we list a portion of the punctuation guidelines to illustrate some of the most important punctuation marks used in Arabic.

The Comma: The Arabic comma correction rules state the following four uses as valid: (1) to separate coordinated and main-clause sentences, usually between short sentences, in order to specify that there is a continuation in the topic; (2) during enumeration to avoid repetition; (3) to provide an explanation or a definition of the previous word; and (4) to separate between parts of the conditional sentences. Furthermore, while English places commas between adjectives of equal weight, Arabic uses zero punctuation or the optional coordination conjunction wa و ‘and’. In other cases, a lexical substitute is inserted instead of the comma. For example the Arabic expression: إلا أن ’except that’ can be used to replace the comma.

Error	أحب السفر كل، صيف ولكن لن أسافر هذا العام
Edit	أحب السفر كل صيف ولكن لن أسافر هذا العام.
English	I like to travel every summer but I won’t this year.

Table 2: Example of two punctuation errors. A comma is used in the wrong place and the sentence does not end with a period.

The Period: The Arabic period or full stop is used at the end of declarative sentences and it is similar to the common use in other languages such in English.

Semi-Colon: In Arabic the semicolon is called Fasila Manquta, literally meaning a dotted comma and is written inverted. In general, Arabic semi-colons are used to indicate that what follows the semicolon is explaining, elaborating, or justifying what precedes it. It can also be used in the following two cases: It can be used between two phrases, in which the first phrase causes the second. And it can be used in two phrases, where the second is a reason for the first. In our guidelines and according to Newmark (1984), Arabic semicolon is also used between two parallel sentences bearing similarity or contrast. In a similar way to the Arabic comma, the Arabic semicolon can be substituted to a coordination conjunction wa و ’and’.

The Colon: The most common use of the colon is to inform the reader that what follows the colon proves, explains, or lists elements of what preceded it. Be-

[5] http://nlp.qatar.cmu.edu/qalb/

low is a restricted list of when the colon may be used during the annotation.

1. Following a dialogue or a conversation as in سألته: من أين لك هذا ؟ 'I asked him: from where you got this?'.

2. To enumerate the classes or the types of related objects as in the following example أيام الدهر ثلاثة : يوم مضى.... 'the days of life are of three types : a day that passed...'.

3. Following some specific Arabic expressions such as التالية 'the next' and ما يلى 'what comes next'.

4. In the case of citations introduced in the text.

The Question mark: This punctuation mark replaces the period at the end of an interrogative sentence. Arabic question mark may be used when there is an Arabic interrogative particle in the sentence such as من 'who', متى 'when', كيف 'how' and أين 'where' as in the following examples: من قال كذا ؟ 'Who said such and such?' متى تصل ؟ 'When will you arrive?' كيف تعمل ؟ 'How it works?' أين أخوك ؟ 'Where is your brother?'.

The Exclamation Mark: The exclamation mark is usually used after an interjection or exclamation to indicate strong feelings or high volume (shouting), and often marks the end of a sentence. In Arabic, the exclamation mark may be used in a similar way to the English as in the following example: كم هذا رائع ! 'How wonderful!'.

5. Annotation Procedure

The lead annotator acts as the annotation work-flow manager in this project. He frequently evaluates the quality of the annotation, monitor and reports on the annotation progress. A clearly defined protocol is set, including a routine for the annotation job assignment. The lead annotator is also responsible for the corpus selection and normalization process, besides the annotation of the gold standard files to be used to compute the Inter-Annotator Agreement (IAA) portion of the corpus needed to control the quality of the annotation. The annotators in this project are five Arabic native speakers university graduates with good Arabic language background. To ensure the annotation consistency, an extensive training phase for each annotator was conducted. During the training phase, various meetings were organized to have the annotators read the guidelines in groups and practice various annotation tasks before discussing issues raised during the training.

Afterwards, the annotator's performance is closely monitored during the initial period, before allowing the annotator to join the official annotation production phase. Furthermore, a dedicated on-line discussion group is frequently used by the annotation team to keep track of the punctuation questions and issues raised during the annotation process, this mechanism, proved to help the annotators and the lead annotator to have a better communication.

The annotation itself is done using a web annotation framework built originally for the manual correction annotation of errors in the QALB project (Obeid et al., 2013). This project framework includes two major components:

1. The annotation management interface used to assist the lead annotator in the general work-flow process. The interface allows the user to upload, assign, monitor, evaluate and export annotation tasks.

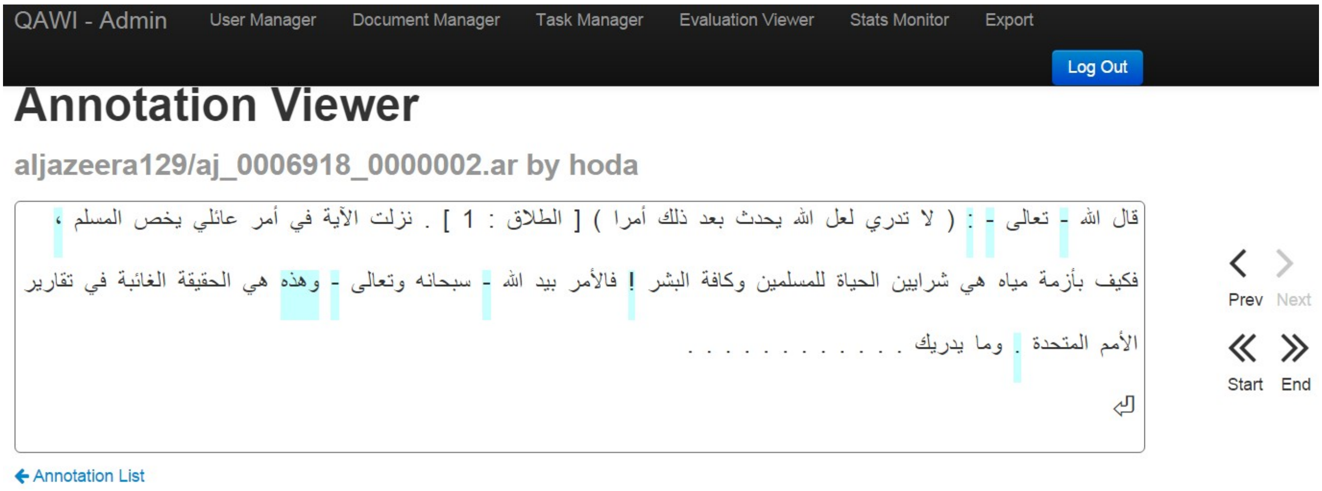
2. The annotation interface is the actual annotation tool (Figure 1), used by the annotators to do the manual correction of the Arabic text and to add the missing punctuation or correct the existing ones. As shown in Figure 1, edited words and punctuation marks are highlighted in blue.

All the annotation history is recorded in a database and can be exported to an XML export file to keep a trace of the entire correction actions for a given file as shown in Figure 2.

6. Evaluation

To evaluate the punctuation annotation quality, we measure the inter-annotator agreement (IAA) on randomly selected files to ensure that the annotators are consistently following the annotation guidelines. A high annotation agreement score is a good indicator of the data quality.

The IAA is measured by averaging WER over all pairs of annotations to compute the AWER (Average Word Error Rate). In this evaluation, the WER measures the punctuation errors against all existing punctuation marks in the text. The IAA results shown in Table 3 are computed over 10 files from each corpus (30 files and 4,116 words total) annotated by at least three different annotators. As seen in the table 3, we noticed an improvement over time of the agreement, since the project started with the L1 corpus, then the L2 corpus and finally the MT corpus. This can be explained in a way by the experience gained overtime by the annotators during the project and also by the frequent guidelines updates done to simplify the punctuation guide-

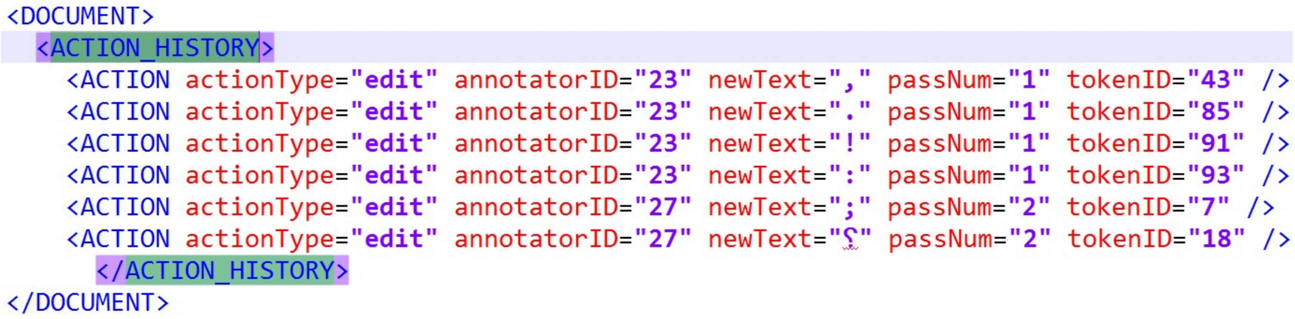


Annotation Viewer

aljazeera129/aj_0006918_0000002.ar by hoda

قال الله - تعالى - : (لا تدري لعل الله يحدث بعد ذلك أمرا) [الطلاق : 1] . نزلت الآية في أمر عائلي يخص المسلم ،
فكيف بأزمة مياه هي شرايين الحياة للمسلمين وكافة البشر ! فالأمر بيد الله - سبحانه وتعالى - وهذه هي الحقيقة الغائبة في تقارير
الأمم المتحدة . وما يدريك . . . .

Figure 1: The annotation interface. Edited words are highlighted in blue.



```
<DOCUMENT>
  <ACTION_HISTORY>
    <ACTION actionType="edit" annotatorID="23" newText="," passNum="1" tokenID="43" />
    <ACTION actionType="edit" annotatorID="23" newText="." passNum="1" tokenID="85" />
    <ACTION actionType="edit" annotatorID="23" newText="!" passNum="1" tokenID="91" />
    <ACTION actionType="edit" annotatorID="23" newText=":" passNum="1" tokenID="93" />
    <ACTION actionType="edit" annotatorID="27" newText=";" passNum="2" tokenID="7" />
    <ACTION actionType="edit" annotatorID="27" newText="؟" passNum="2" tokenID="18" />
      </ACTION_HISTORY>
</DOCUMENT>
```

Figure 2: Extract of output file showing some punctuation correction action history

lines. Moreover, we noticed a much higher agreement with the MT corpus. A closer analysis revealed that this was caused by a much less occurrence of the comma in the MT corpus as compared to L1 and L2 corpus as seen in Table 4. Indeed the comma punctuation mark is not always clearly marked in many cases and it can be substituted by the Arabic conjunction wa و 'and'. Furthermore, the MT corpus has a much higher occurrence of the period mark which is considered easier to edit. Overall, the results obtained showed that the annotators are consistently following the punctuation guidelines provided to them.

Corpus	Average IAA	Average WER
L1 Corpus	89.84	10.16
L2 Corpus	91.9	8.1
MT Corpus	93.78	6.22

Table 3: Average percent Inter-Annotator Agreement (IAA) and the Average WER (AWER) obtained for the three corpora.

Punctuation Mark	L1	L2	MT
Comma	44.77%	45%	38%
Period	37%	40%	50%
Question Mark	8.79%	7.38%	4.50%
Exclamation Mark	4.89%	4%	3%
Colon	2.56%	2.12%	2.50%
Semi Colon	2%	1.50%	2%
Total	100%	100%	100%

Table 4: Distribution of each punctuation mark in the three corpora.

7. Conclusions

We presented our method to create an Arabic manually punctuated corpus, including the writing of the guidelines as well as the annotation procedure and the quality control method used to verify the annotation quality. We showed that there is a high variety in the use of punctuation in Arabic texts and despite the existence of punctuation rules, the use of punctuation in Arabic is highly individual and it depends on the style of the author who may define his/her own use of punctuation. The specific punctuation annotation guidelines presented in this paper helped to reduce the variations in the punctuation use during the annotation experiment.

The created resource has been already freely released for the research community during the related shared tasks we organized in recent years (Mohit et al., 2014; Rozovskaya et al., 2015). We hope to receive feedback from the community on the usefulness and potential related applications of the resource.

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 Fathy, Hoda Ibrahim, Anissa Jrad and Jihene Wafi. We thank Nizar Habash, Behrang Mohit, Kemal Oflazer and Ossama Obeid for their valuable contribution in this project. This publication was made possible by grants NPRP-4-1058-1-168 from the Qatar National Research Fund (a member of the Qatar Foundation). The statements made herein are solely the responsibility of the authors.

8. References

Alfaifi, A. and Atwell, E. (2012). Arabic Learner Corpora (ALC): A Taxonomy of Coding Errors. In *The 8th International Computing Conference in Arabic (ICCA 2012)*, Cairo, Egypt.

Alkanhal, M. I., Al-Badrashiny, M. A., Alghamdi, M. M., and Al-Qabbany, A. O. (2012). Automatic Stochastic Arabic Spelling Correction With Emphasis on Space Insertions and Deletions. *IEEE Transactions on Audio, Speech, and Language Processing*, 20(7):2111–2122.

AlQinai, J. (2015). Mediating punctuation in english arabic translation. *Journal of Applied Linguistics and Professional Practice*, 5(1).

Asfour, M. (2009). *at-Tarqim fi l-lugati al-arabiyya*. Dar al-Baraka, Amman.

Attia, M., Al-Badrashiny, M., and Diab, M. (2014). Gwu-hasp: Hybrid arabic spelling and punctuation corrector. In *Proceedings of the EMNLP 2014 Workshop on Arabic Natural Langauge Processing (ANLP)*, pages 148–154.

Awad, D. (2013). La ponctuation arabe : histoire et règles.

Beeferman, D., Berger, A., and Lafferty, J. (1998). Cyberpunc: a lightweight punctuation annotation system for speech. In *Acoustics, Speech and Signal Processing, 1998. Proceedings of the 1998 IEEE International Conference on*, volume 2, pages 689–692 vol.2, May.

Farwaneh, S. and Tamimi, M. (2012). Arabic Learners Written Corpus: A Resource for Research and Learning. *The Center for Educational Resources in Culture, Language and Literacy.*

Furui, S., Kikuchi, T., Shinnaka, Y., and Hori, C. (2004). Speech-to-text and speech-to-speech summarization of spontaneous speech. *IEEE Transactions on Speech and Audio Processing*, 12(4):401–408.

Jones, D. A., Wolf, F., Gibson, E., Williams, E., Fedorenko, E., Reynolds, D. A., and Zissman, M. A. (2003). Measuring the readability of automatic speech-to-text transcripts. In *INTERSPEECH*. ISCA.

Matusov, E., Hillard, D., Magimai-doss, M., Hakkanitur, D., Ostendorf, M., and Ney, H. (2007). Improving speech translation with automatic boundary prediction. In *Proceedings of Interspeech*, pages 2449–2452.

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.

Mubarak, H. and Darwish, K. (2014). Automatic correction of arabic text: a cascaded approach. In *Proceedings of the EMNLP 2014 Workshop on Arabic Natural Langauge Processing (ANLP)*, pages 148–154.

Newmark, P. (1984). *Approaches to Translation*. Oxford.

Ng, H. T., Wu, S. M., Wu, Y., Hadiwinoto, C., and Tetreault, J. (2013). The conll-2013 shared task on grammatical error correction. In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning: Shared Task*, pages 1–12, Sofia, Bulgaria, August. Association for Computational Linguistics.

Obeid, O., Zaghouani, W., Mohit, B., Habash, N., Oflazer, K., and Tomeh, N. (2013). A Web-based Annotation Framework For Large-Scale Text Correction. In *The Companion Volume of the Proceedings of IJCNLP 2013: System Demonstrations*, Nagoya, Japan, October.

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.

Spitkovsky, V. I., Alshawi, H., and Jurafsky, D. (2011). Punctuation: Making a point in unsupervised dependency parsing. In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning*, CoNLL '11, pages 19–28, Stroudsburg, PA, USA. Association for Computational Linguistics.

Zaghouani, W., Habash, N., and Mohit, B. (2014a). The qatar arabic language bank guidelines. Technical Report CMU-CS-QTR-124, School of Computer Science, Carnegie Mellon University, Pittsburgh, PA, September.

Zaghouani, W., Mohit, B., Habash, N., Obeid, O., Tomeh, N., Rozovskaya, A., Farra, N., Alkuhlani, S., and Oflazer, K. (2014b). Large scale arabic error annotation: Guidelines and framework. In *International Conference on Language Resources and Evaluation (LREC 2014)*.

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)*.

Zaki, A. (1995). at-tarqim wa alamatu-hu fi al-lugati al-arabiyya, maktabat al-matbuat al islamiyya bi halab.

DINA: A Multi-Dialect Dataset for Arabic Emotion Analysis

Muhammad Abdul-Mageed[†‡], **Hassan AlHuzli**[†], **Duaa' Abu Elhija, Mona Diab**[‡]
School of Informatics, Computing[†]; Department of Linguistics
Indiana University, Bloomington, IN
The Department of Computer Science, The George Washington University, Washington, DC[‡]
{mabdulma,halhuzal,dabuelhi}@indiana.edu; mtdiab@gwu.edu[‡]

Abstract

Although there has been a surge of research on sentiment analysis, less work has been done on the related task of emotion detection. Especially for the Arabic language, there is no literature that we know of for the computational treatment of emotion. This situation is due partially to lack of labeled data, a bottleneck that we seek to ease. In this work, we report efforts to acquire and annotate a multi-dialect dataset for Arabic emotion analysis.

Keywords: Arabic, emotion analysis, Arabic sentiment

1. Introduction

Compared to the related task of sentiment analysis, emotion detection has not witnessed as much surge of research. Detection of emotion is a useful task, as it can have many practical applications. The following are only some examples:

- **Opinion Mining:** Emotion detection can enhance opinion mining and sentiment analysis, providing more nuanced insights into what social media users feel about a given product, person, or organization. As such, emotion detection provides an enriching component beyond the mere binary valence (i.e. positive and negative) of most sentiment analysis systems. Used for mining user opinions, emotion analysis can be valuable for industries doing market research, politicians running campaigns, organizations' expansion of online user base, etc.

- **Health and Wellness/Forensics:** Emotion analysis can be useful from a health and wellness perspective in various ways. For example, it can be used for early detection of certain emotional disorders such as depression. In addition, since emotions have been shown to be contagious (Kramer et al., 2014), emotion generation can be used to improve the overall well being of people by exposing them to desired emotions such as happiness (e.g., over social networks) for example.

- **Education:** Synthetizing learners to the emotional aspects of automatically generated language units (e.g., words and phrases) can be instructive. Integrating emotionally-aware agents in intelligent computer-assisted language learning, for example, should prove useful and enhance the naturalness of the pedagogical experience.

- **Marketing:** Since emotions play a significant role in decision making (Bechara et al., 2000; Bechara, 2004; Schwarz, 2000; Sanfey et al., 2003) and at least some role in message propagation (Heath et al., 2001; Tan et al., 2014), emotion-sensitive language generation should be useful in advertisement in that it can create messages with higher likelihood of appealing to audiences.

- **Security:** Knowledge about the emotional stability of online users can be used to deflect potential hazards and anticipate potentially dangerous behaviors.

- **Author Profiling:** Emotion words can also be used for author profiling. For example, (Meina et al., 2013; Flekova and Gurevych, 2013; Farias et al., 2013; Bamman et al., 2014; Forner et al., 2013) have used emotion words for predicting age and gender. In addition, (Mohammad and Kiritchenko, 2013) have used emotion-based features to predict personality type.

Although there has been some work on detecting emotion in English (e.g., (Strapparava and Mihalcea, 2007; Aman and Szpakowicz, 2007)), there is currently no work that we know of on treating emotion in Arabic. Detecting emotion in Arabic should prove attractive, due to the strategic importance of the language and its various varieties (Diab and Habash, 2007; Diab et al., 2010; Habash, 2010) but also its morphological richness (Abdul-Mageed et al., 2013; Diab et al., 2004; Habash et al., 2009). One of the problems that hinder progress toward building emotion detection systems for Arabic is unavailability of labeled data. In this paper, we seek to partially solve this issue: We present DINA, multi-dialect, dataset for Arabic emotion analysis. In this context, we report efforts to acquire and annotate a number of data subsets for the six basic Ekman emotions (Ekman, 1992) of *anger,disgust, fear, happiness, sadness*, and *surprise*.
The rest of the paper is organized as follows: Section 2. is a review of related litrature, Section 3. is an overview of DINA and its development, Section 4. is a detailed description of DINA and the various emotion categories, with illustrating examples, Section 5. discusses contexts with no emotion as well as those with mixed emotions, and Section 6. is a conclusion.

2. Related Work

In addition to Ekman's (Ekman, 1992) 6 basic emotions of *anger,disgust, fear, happiness, sadness*, and *surprise*, there

Class	ENG	Arabic
Anger	angry, resentful, etc.	غاضب، ناقم، ساخط، حانق
Disgust	disgusted, nauseated, etc.	متقزز، مشمئز، قرفان، ارفان
Fear	fearful, scared, etc.	خائف، مرعوب، مرتاع، مذعور
Happiness	happy, joyful, etc.	سعيد، فرحان، مسرور، مبتهج
Sadness	sad, sorrowful, etc.	حزين، تعيس، مبتئس، مغموم
Surprise	surprised, taken, etc.	مندهش، متفاجيء، مذهول، مشدوه

Table 1: Emotion seeds

are other classifications of emotions in the psychological literature. For example, (Plutchik, 1980; Plutchik, 1985; Plutchik, 1994) adds *trust* and *anticipation* to Ekman's 6 basic types. (Francisco and Gervás, 2006) report work on marking the attributes of *pleasantness, activation,* and *dominance* in the genre of fairy tales. There is also a line of work on 'mood.' For example, (Bollen et al., 2011) reports research on attributes like *tension, depression, anger, vigor, fatigue,* and *confusion*, relating these to predicting the stock market. (Pearl and Steyvers, 2010) describes work on identifying attitudes and intentions, and discuss attributes like *politeness, rudeness, embarrassment, formality, persuasion, deception, confidence,* and *disbelief.* (Mohammad and Kiritchenko, 2015) report crawling a Twitter corpus with 585 hashtags related to emotional states (e.g., *excitement, guilt, yearning,* and *admiration*) and describe experiments using the data in the context of improving personality detection.

There is some literature describing emotion data collection focused at English. For example, the SemEval-2007 Affective Text task (Strapparava and Mihalcea, 2007) [SEM07] focused on classification of emotion and valence (i.e., positive and negative texts) in news headlines, with the assumption that since news headlines are short and are usually written creatively to capture reader attention, they will have emotions expressed in them. The headlines comprising the data set are drawn from newspapers such as New York Times, CNN, and BBC News, as well as from the Google News search engine. The total size of the data set is at 1,250 labeled headlines and are split into a development data set consisting of 250 data points, and a test data set of 1,000 data points. The data were labeled via a Web-based interface displaying one headline at a time, along with with six slide bars for emotions. Given an emotion, a slide bar has intervals set to [0, 100], where 0 means "no-emotion" and 100 represents "maximum emotional" load. The interval for the valence annotations was set to [-100, 100], where 0 represents a neutral headline, -100 represents a highly negative headline, and 100 corresponds to a highly positive headline.

In addition, (Aman and Szpakowicz, 2007) describe an emotion annotation task of identifying emotion category, emotion intensity and the words/phrases that indicate emotion in blog post data of 4090 sentences and a system exploiting the data. To collect data where emotion is more likely to be observed, they use a seed list of words 'commonly' employed in the context of each of Ekman's six basic emotions. For example, they use the words "happy", "enjoy", and "pleased" as seeds for the "happiness" emotion, and "afraid", "scared", and "panic" for the fear category. They then crawl blog posts containing one or more of each of the words in their seed list and present the data for double annotations by four judges[1] (one judge hence labeled all the data and each of the three others labeled a subset of the data). (Aman and Szpakowicz, 2007) point out that the annotators received no training, but were given samples of annotated sentences to illustrate the different types of emotions and annotations required. In addition to the six categories of emotions, annotators were asked to assign two more tags (i.e., *mixed-emotion* and *no-emotion* in a non-overlapping fashion (i.e., each sentence is assigned only one of the eight tags). In addition, annotators were required to assign emotion intensity tags from the set low, medium, high to all emotion-carrying sentences (thus excluding sentences tagged with *no-emotion*). (Aman and Szpakowicz, 2007) also ask annotators to identify *emotion indicators*(i.e., spans of text of any length). The data set ends up with 1,466 emotion-bearing sentences and 2,800 assigned "no-emotion."

Moreover, some researchers, e.g., (Qadir and Riloff, 2014; Mohammad, 2012; Wang et al., 2012), use hashtags as an approximation of emotion categories to collect emotion data. Our approach is closest to this literature. However, we use phrases instead of hashtags.

Finally, there has been a fair amount of work on the related task of Arabic sentiment analysis and other social media analysis tasks. For example, (Abdul-Mageed et al., 2011b; Abdul-Mageed et al., 2011a) report social media analyses of Arabic Twitter and YouTube data. (Abdul-Mageed and Diab, 2011; Abdul-Mageed and Diab, 2012a; Refaee and Rieser, 2014) describe efforts to collect and/or label social media data for Arabic sentiment detection. (Abdul-Mageed and Diab, 2012b; Abdul-Mageed and Diab, 2014; Badaro et al., 2014) describe Arabic sentiment lexica, and (Abdul-Mageed et al., 2011c; Abdul-Mageed et al., 2012; Abdul-Mageed et al., 2014) describe Arabic sentiment detection systems.

3. Development of DINA

The DINA corpus comprises Modern Standard Arabic (MSA) and dialectal Arabic (DA) data crawled from Twitter [2] between July and October, 2015. We use the Python

[1] We employ the two terms "annotator" and "judge" interchangeably.

[2] https://twitter.com

Library Tweepy [3] to collect a corpus of tweets with phrases from a seed set of emotion words. More specifically, we create a seed set of size < 10 phrases for each of the six Ekman emotion types. In this approach, we collect only tweets where a seed phrase occurs in the tweet body text. Note this approach does not depend on hashtags and is only conditioned on a given phrase occurring in the tweet text as captured by a regular expression. Each phrase is composed of the first person pronoun انا (Eng. ”I”) + a seed word expressing emotion (e.g., سعيد [Eng. ”happy”]). Examples of the seeds words for each emotion type are provided in Table 1. The seed phrases are chosen such that they capture data representing various Arabic dialects, which is what we observe in the collected data set [4]. For the current work, we select 500 tweets with seeds from each of the 6 Ekman categories of emotion, for a total of 3,000 tweets that are chosen from a wider pool of in-house twitter corpus. It is noteworthy that we remove duplicates from the wider pool of tweets using a simple Python script that 1) compares the tweet ID, and 2) removes all no alphabetical character, white space, ”rt” characters, and usernames, and finally compares the identity of the tweet body. A manual evaluation shows this method to successfully solve the problem of tweet duplicates to a considerable extent. To ensure acquiring a duplicate-free dataset, however, all remaining duplicates are removed manually before we choose the 500 tweets belonging to each class. To test the utility of this phrase-based approach for collecting emotions, we ask each of two judges to label the data set of 3,000 tweets. We describe the annotation task next.

Since we wanted to capture emotion intensity in the data, we decided to include three intensity tags similar to (Aman and Szpakowicz, 2007)’s ”Low,” ”Medium,” and ”High” tags; we call them ”weak,” ”fair,” and ”strong.” Realizing that there could be cases that carry no emotion, even with the existence of our phrase seeds, we needed to include a ”no-emotion” tag. In addition, we suspected we may need a ”mixed-emotion” category, since there could be mixed emotions in a single tweet. In order to test the need for both the ”no-emotion” and ”mixed-emotions” tags, we labeled 100 random data points and found that these included 7% that would be tagged with either of these tags and all of the 7% indeed carry no emotion at all. For this reason,and in order to reduce cognitive overload during the annotation process, we combined the ”no-emotion” and ”mixed-emotion” tags as a ”zero” tag. [5] The tagset employed is thus the set {*zero, weak, fair, strong*}. Given a tweet with an emotion belonging to one of the 6 Ekman categories, annotators were asked to decide whether the tweet belongs to the respective emotion (and hence assign an intensity tag) or not (and hence assign the ”zero” tag). As such, for cases where the respective emotion the seed phrase is meant to capture is not correct, the judge should assign the ”zero” tag. This includes examples where a tweet possibly expresses an emotion other than what the seed phrase is intended to capture or sarcastic tweets. We provided annotators with guidelines covering these cases as well.

The two annotators are native speakers of Arabic with post-graduate education, both with strong proficiency in Modern Standard Arabic (MSA) and Egyptian Arabic. In addition one of the annotators have native fluency in Levantine Arabic and the other has native proficiency in Gulf Arabic. Annotators were advised to consult each other, ask online, and in their circles of friends on cases where a given dialect was not intelligible. They were also provided several examples illustrating each tag. We measure overall agreement between the two annotators in terms of Cohen’s (Cohen, 1960) Kappa and also calculate the percentage of per-class agreement. Cohen’s Kappa is calculated via defining a k by k confusion matrix, in which an element f_{ij} defines the number of cases Annotator A assigned a particular case to category i and Annotator B assigned to j. As such, f_{jj} is the number of agreements for category j. Then (from (Altman, 1991)) suppose:

$$P_o = \frac{1}{N}\sum_{j=1}^{k} f_{jj}, \tag{1}$$

$$r_i = \sum_{j=1}^{k} f_{ij}, \forall i, \text{ and } c_j = \sum_{i=1}^{k} f_{ij}, \forall j, \tag{2}$$

$$P_e = \frac{1}{N^2}\sum_{i=1}^{k} r_i c_i, \tag{3}$$

where P_o is the observed proportional agreement, r_i and c_j are the row and column totals for category i and j, and P_e is the expected proportion of agreement. Cohen’s Kappa *K* is then calculated as: (4).

$$\kappa = \frac{P_o - P_e}{1 - P_e}. \tag{4}$$

4. DINA Properties

	Kappa *(K)*	**% Emotion**
Anger	0.68	83.50%
Disgust	0.33	92.30%
Fear	0.23	98.90%
Happiness	0.71	82.50%
Sadness	0.54	98.90%
Surprise	0.55	98.80%
Average	0.51	92.48%

Table 2: Agreement in fine-grained annotation and average percentage of emotion

As Table 2 shows, annotators agreed to assign one or another of the emotion intensity tags (i.e., ”weak,” ”fair,” and ”strong”) in 92.48% of the cases. This 92.48% agreement reflects the effectiveness of the phrase-based seed approach

[3] http://www.tweepy.org

[4] Another possibility would be combine this approach with the tweet geo-location anduse a dialect identification tool to recognize the dialect of a given tweet, and we plan to attempt this in the future.

[5] This is different from (Strapparava and Mihalcea, 2007) who drop the ”mixed” category altogether, although they instruct annotators with examples of mixed emotions.

for automatically acquiring emotion data. Table 2 also shows that the judges disagreed to varying degrees when choosing the specific positive emotion tag to assign, with agreement as high as 0.71% in the case of *happiness* and as low as 0.23% in the case of *fear*, and an average of 0.51%. We now discuss further details related to each of the 6 types of emotions.

4.1. Anger

Our annotators agree with a Cohen Kappa (K)= 0.68% on the *anger* data. Related agreement on the different anger categories is shown in Table 3. [6] Table 3 shows that the most confused label is ”strong,” and that it is either confused with ”fair” or ”weak.”
Examples 1-3 below illustrate the *anger* class:

	ZERO	WEAK	FAIR	STR	Total
ZERO	78.75	20.00	1.25	0.00	16.00
WEAK	6.23	85.46	7.12	1.19	67.40
FAIR	1.35	9.46	81.08	8.11	14.80
STR	0.00	11.11	22.22	66.67	1.80
Total	17.00	62.40	17.40	3.20	100

Table 3: Agreement in *anger* annotation

- **(١)** انا اسف ماكان قصدي اعصب عليج وخليج تنامين
وانتي متضايقة بس شسوي اذا انا عصبي ومشكلجي ،
اسف للمرة الثانيةاصلن عيب تزعلين مني

(1) Eng. ”I’m sorry. I didn’t mean to get nervous with you and have you go to bed upset. It’s that I’m a bit nervous and trouble-maker. I’m sorry for the second time; I’d actually hope you won’t be upset because of me.” (**Tags:** *weak,weak*)

- **(٢)** أنا غاضب جدا من رأسي إلى أصابع قدمي ،
بسبب كل الأيام التي أضعتها مخدوع منك.

(2) Eng. ”I’m very angry, from head to toes, because of all the days I have wasted [for you]. I was deceived by you.” (**Tags:** *fair,weak*)

- **(٣)** انا ساخط علي الهلال لاني اشوفه يعيش اخر مواسمه
ان لم يتم انتشاله كما انتشل فيصل بن تركي النصر

(3) Eng. ”I’m displeased with Al-Hilal and I’m seeing it living its last season, unless it is saved from this destiny the same way Faisal Ibn Turky has saved Al-Nasr.” (**Tags:** *fair,weak*)

4.2. Disgust

For data collected with the disgust seeds, the judges agree with a Cohen Kappa (K)= 0.33%. Table 4 below shows the related per-tag agreement percentages.

	ZERO	WEAK	FAIR	STR	Total
ZERO	33.33	63.33	0.00	3.33	6.00
WEAK	8.55	83.85	5.23	2.38	84.20
FAIR	0.00	29.73	59.46	10.81	7.40
STR	8.33	50.00	16.67	25.00	2.40
Total	9.40	77.80	9.20	3.60	100

Table 4: Agreement in *disgust* annotation

Examples 4-8 illustrate the *disgust* class. It can be seen from Table 4 that the most confused tag is ”strong” and that it is confused with ”weak” and ”fair,” but also with ”zero.”

- **(٤)** انتي اللي داخله بالمنشن حقي تحتكين بجد انا
قرفانه من سواليفكم مافيها ملح سامجات بالعربي

(4) Eng. ”It was you who are calling for trouple by mentioning me, really. I’m disgusted with all your discourse, as they are clearly all boring anecdotes.” (**Tags:** *weak,weak*)

- **(٥)** من كتر ما تخصصي جايبلي القرف في عمري ،
اتردّ وانا ادعي لنفسي بالتوفيق كدا من كتر ما انا
قرفانة مابا انجح فيلو

(5) Eng. ”Since my major is bothering me non-stop, I hesitate when I pray to succeed. I’m too disgusted to the extent that I don’t want to succeed.” (**Tags:** *fair,fair*)

- **(٦)** انا مش مرعوبه خالص كان لازم يحصل كل ده علي
الاقل نشوف حقيقة ناس كتير انا قرفانه اوّوّوّي بس
لامرعوبه ولامصدومه

(6) Eng. ”I’m not scared at all. All this should have happened so that we at least see the true faces of many people. I’m very disgusted, but neither scared nor shocked.” (**Tags:** *fair,fair*)

- **(٧)** أخيرا .. ياراجل دا أنا كنت قرفان منه قرف
... ياساتر أعوذ بالله من هذه الحثالة إمتى مصر
تنظف من القذارة دى بقا!

(7) Eng. ”Finally.. Man, I was truly disgusted of him. Oh my God, may we not come to involve with this mean people. When will Egypt be purged off these people!” (**Tags:** *strong,strong*)

[6] The columns and rows labeled with ”Total” in the confusion matrices show percentages of tweets assigned a given tag by the respective judge.

- (٨) انا مش قرفان منكم بس...
انا قرفان من العالم كله...
منظر الكره الارضيه مش عاجبنى..

  (8) Eng. ”I’m not disgusted with you. I’m disgusted with the whole world. I don’t like the look and feel of the whole earth planet.” (**Tags:** *strong,strong*)

4.3. Fear

The judges agree with a Cohen Kappa (*K*)= 0.23% on the *fear* data. The related agreement on the different classes is shown in Table 5. As Table 5 shows, that ”zero” is the most confused label and that it is confused with ”weak” in most of the cases, but also with ”strong.” Table 5 shows the per-tag agreement on the *fear* data. It can be seen from the Table that the most confused category is ”zero” and that it is only confused with ”weak.” It can also be seen from Table 5 that there is perfect agreement on assigning the ”strong” label.

	ZERO	**WEAK**	**FAIR**	**STR**	**Total**
ZERO	11.11	88.89	0.00	0.00	1.80
WEAK	0.21	96.47	3.32	0.00	96.40
FAIR	0.00	62.50	37.50	0.00	1.60
STR	0.00	0.00	0.00	100	0.20
Total	0.40	95.60	3.80	0.20	100

Table 5: Agreement in *fear* annotation

Examples 9-11 illustrate the *fear* category.

- (٩) و الله العظيم انا خايف فشخ الريال يندم ع
الفرص ال بتضيع دى .. ده انا مّكن اتشل فيها و ربنا !!

  (9) Eng. ”I swear by the Almighty God, I am really scared Reyal [Madrid] will come to lamenting all these wasted opportunities.. I could get a coma as a result of this, I swear!!” (**Tags:** *weak,fair*)

- (١٠) ده انا خايف خايف و حاسس بالخطر الامتحانات

  (10) Eng. ”I’m scared, scared and under the threat of the upcoming examinations.” (**Tags:** *weak,fair*)

- (١١) انا مرعوبه والله اول مرة اخاف كدة. نسمع
اغانى عشان ننسى الحلقه

  (11) Eng. ”.I swear, I’m really scared. We should listen to some songs so that we forget the soap opera.” (**Tags:** *strong,strong*)

4.4. Happiness

For the happiness fine-grained tags, annotators agree with a Cohen Kappa (*K*)= 0.71%. Table 6 below shows agreement between the annotators on labeling the automatically acquired happiness data. Table 6 also shows that the ”zero” is the most confused label and that it is confused with all other labels by one judge and by all other labels except ”weak” by the other judge.

	ZERO	**WEAK**	**FAIR**	**STR**	**Total**
ZERO	84.34	0.00	1.20	14.46	16.60
WEAK	6.12	46.94	0.00	46.94	9.80
FAIR	8.33	0.00	75.00	16.67	2.40
STR	5.06	0.28	0.56	94.10	71.20
Total	18.40	4.80	2.40	74.40	100

Table 6: Agreement in *happiness* annotation

Examples 12-16 below illustrate the different tags assigned to the *happiness* data.

- (١٢) تسألني عن حالتي فالبعد وشلوني؟
انا سعيد بغيابك وانت شخبارك ؟

  (12) Eng. ”Asking about me and how I am while you’re away [?] I’m happy you’re away; and how are *you*[?].”(**Tags:** *weak,weak*)

- (١٣) انا مبسوطه اني مصاحباكي ي حيااتي

  (13) Eng. ”I am happy I have you as a friend, my dear.” (**Tags:** *weak,weak*)

- (١٤) انا فرحانه و مبسوطه و طايره كده
و مستعده اعمل أي حاجه دلوقتي
انشالّه أقوم انضف البيت لماما

  (14) Eng. ”I’m happy and glad and feel like flying, and I’m like ready to do anything even if it is to tidy the house for mom.” (**Tags:** *fair,fair*)

- (١٥) يا حبيب ضحكاتي يا حبيب قلبي يا بابا يا غالي؛
شكرا لك على تفضيل تغريدتي انا فرحانة
قوييييييي قويييييييي

  (15) Eng. ”Oh, the love of my laughter and sweetheart, dad; thank you for favoriting my tweet. I’m veryyyy veryyyy happy.” (**Tags:** *strong,strong*)

- (١٦) فرحانه باليوم اللي جمّعنا،فرحانه بالساعات
اللي تلمّنا ،فرحانه بالحظ الّي جابّك لقلبي
انا فرحانه فيك..

  (16) Eng. ”[I’m] happy with the day that got us to meet; [I’m] happy with the hours we spend together. [I’m] happy with the luck that brought you to my heart. I’m happy with you..” (**Tags:** *strong,strong*)

4.5. Sadness

For the sadness class labeling, annotators agree with a Cohen Kappa (K)= 0.54%. Table 7 below shows agreement on sadness annotation. As can be seen from Table 7, the most confused label is ”zero” and it is only confused with the category ”weak.” The Table also shows perfect agreement on the label ”strong.”

	ZERO	WEAK	FAIR	STR	Total
ZERO	11.11	88.89	0.00	0.00	1.80
WEAK	0.21	96.47	3.32	0.00	96.40
FAIR	0.00	62.50	37.50	0.00	1.60
STR	0.00	0.00	0.00	100	0.20
Total	0.40	95.60	3.80	0.20	100

Table 7: Agreement in *sadness* annotation

Examples (17) and (18) illustrate the *sadness* class.

- (١٧) لوكا: أنا حزين جدا لأني أصبت مرة أخرى وبهذا الوقت الأهم بالموسم.

  (17) Eng. ”Luke: I’m very sad I got an injury once again, especially that this is the most important time of the year.” (**Tags:** *fair,fair*)

- (١٨) ناس حكيره مرا انا حزينه قدا قدا

  (18) Eng. ”Very mean people! I’m very very sad.” (**Tags:** *fair,strong*)

4.6. Surprise

The judges agree with a Cohen Kappa (K)= 0.55% on the *surprise* data. The related agreement on the different classes is shown in Table 8. Table 8 shows that the label ”strong” is only assigned 0.20% of the cases and is never agreed upon. This sets *surprise* aside from the other emotions, where is is some agreement on assigning ”strong.” This may imply that perception of what constitutes a ”strong” emotion varies across individuals in the case of *surprise* more than it does for other emotions. [7]

	ZERO	WEAK	FAIR	STR	Total
ZERO	10.00	90.00	0.00	0.00	2.00
WEAK	0.00	98.54	1.46	0.00	95.60
FAIR	0.00	0.00	90.91	9.09	2.20
STR	0.00	100	0.00	0.00	0.20
Total	0.20	96.20	3.40	0.20	100

Table 8: Agreement in *surprise* annotation

Examples 19-23 illustrate the *surprise* data.

- (١٩) انا مستغربة ازاي الناس اللي عيالها بتموت زي الفراخ في سينا ماحدش منهم اتكلم على ان ابنه اتاخد اجباري واتحط في مكان خطر من غير تدريب مناسب

  (19) Eng. ”I’m surprised none of the people whose children are killed indifferently in Sinai has questioned how it is that these kids were taken compulsively, without training, and dispatched in such a dangerous place without due training.” (**Tags:** *weak,weak*)

- (٢٠) راحت العندية والدماغ الناشفة .. !!
  راحت الهيبة و الشخصية .. !!
  انا مستغربة نفسى بجد !!

  (20) Eng. ”Stubbornness and bullheadedness are gone!! Air of awe and hotshotness are gone!! I’m surprised of myself.” (**Tags:** *weak,weak*)

- (٢١) انا مستغرب فشخ الناس اللي بتصحى دلوقت من النوم انا كنت من شهر زيهم بس مستغربهم فشخ يعني.

  (21) Eng. ”I find it strange there are people who get up at this time; I was like them a month ago, but it’s really strange to me.” (**Tags:** *weak,fair*)

- (٢٢) انا مستغرب كثير من الجمهور يسال ماسبّ شحن ومحاربه الشبابين بينهم البين ! لهدرجه خايفين تقولون طارق النوفل ! ماسك عليكم شي !

  (22) Eng. ”I’m really surprised a lot of the fans ask what is the reason of emotional charging and fighting. AlShababeen have are at odds. Is it to that extent you are worried to say it’s Tariq AlNawfal?! Does he hold something against you?” (**Tags:** *weak,fair*)

- (٢٣) قط بنت الجيران اكل معى سمك ومن وقتها رفض
  تمام يروح لها تانى انا مستغرب جداااااااااااااااااااااااااا

  (23) Eng. ”The cat of our neighbor’s daughter ate fish with me, and from that point onward refused totally to go back to her. Really surprised!!” (**Tags:** *weak,strong*)

5. Contexts of No- and Mixed- Emotion

Emotion is similar to other pragmatic-level phenomena where no one-to-one mapping between what is *said* and what is being truly communicated or *meant*. For this reason, even with a crafted list of seeds intended to capture emotion expression, inter-annotator agreement shows that in 7.50% of the cases both annotators assign a ”zero” tag. A consideration of the data shows that there are some contexts where this is specially the case. We describe some of these contexts in the subsections below.

[7] Interestingly, (Ekman, 1992) discusses ways in which *surprise* differs from other emotions. These, however, are beyond our current focus.

5.1. Reported Speech

In some cases, Twitter users employ reported speech to describe a situation where someone is expressing one emotion or another. In many of these cases, the main goal of the tweet is not to express emotion *per se*, but rather to discuss something else. For instance, example (24) below is educational in purpose, describing what a parent should do if their child expresses anger, and examples (25) and (26) are more of pieces of advice on perspective on difficulties in life and interpersonal relationships, respectively.

- (٢٤) لا تتجاهل مشاعر طفلك أذ قال لك مثلا انا غضبانْ
  يجب ان ترد انا اقدر مشاعرك لماذا انت غضبان ؟
  من كتاب كيف تربي شخصية طفلك

  (24) Eng. ”Don’t ignore your child’s emotion. If he tells you ’I’m angry,’ ask him why he is angry. Quote from the book ”How to Raise your Child”.”

- (٢٥) لاتقل أنا متعب فالكل متعب، ولاتقل أنا حزين
  فالكل حزين، بل قل الحمدلله صبحا ومساء
  ليخف أنينك وتنتهي أوجاعك.

  (25) Eng. ”Don’t say I’m tired, because everyone is tired. And don’t say I’m sad, as everyone is sad. Rather, say Thank God every morning and evening so that He relieves your trouble and removes your pains.”

- (٢٦) لازم تقدروا الناس الكتومة تقدروهم بجد ..
  الناس دي مبتحبش تعد تقول انا زعلان مضايق .
  مخنوق . قرفان .. و كلام كلام

  (26) Eng. ”You must really value reticent people.. This type of people don’t keep saying I’m annoyed, I’m upset, frustrated, disgusted.. and all that.”

5.2. Sarcasm

Sarcastic language is known to have the effect of reversing the polarity of texts (Bamman and Smith, 2015; Davidov et al., 2010; González-Ibánez et al., 2011; Tsur et al., 2010) and we believe it has the same effect as to emotion expression. We describe how sarcasm interacts with our data collection method and its distribution in the data set, illustrating with examples. For this reason, we sensitized the judges to the concept of sarcasm and initially asked them to label sarcastic tweets in the initial stage of annotation. After labeling a total of 600 tweets (100 from each emotion type), however, we found that almost none of these data points carried sarcasm. For this reason, and since our current interest is not to capture sarcasm in data, we decided not to carry on with sarcasm annotation. The decision was made to also reduce cognitive overload and help expedite the annotation process. As such, a decision was simply made to assign sarcastic tweets that carry an emotion opposite to the seed it contains a ”no-emotion” tag. An analysis of the data points that ended up being assigned a ”no-emotion” label show that the data set has only one sarcastic tweet, one we provide in example (27) below:

- (٢٧) أنا كويس.. أنا مبسوط.. أنا زي الفل..
  أنا مش مضايق.. الدنيا تمام جدا..
  أنا مرتاح.. أنا مش متوتر..
  أنا فاصل ومش بافكر.. أنا تمام.. أنا تمام

  (27) Eng. ”I’m fine.. I’m happy.. I’m really good.. I’m not upset.. Life is pretty good.. I’m relaxed.. I’m not nervous.. I’m unwired and not thinking.. I’m OK.. I’m OK.”

We note that an ideal emotion classification system should incorporate sarcasm detection, and that an automatic approach to emotion data acquisition using the seeds we have employed could be negatively impacted by assigning the wrong tags to a very small, and perhaps negligible, fraction of the data.

5.3. Objective Evaluations

Sometimes, users provide evaluations of their emotional states in an air of objectivity. In cases like this, the two annotators assigned a ”no-emotion” tag. Judges also assigned ”no-emotion” labels to tweets where users refute what others think these users’ emotional states are. We illustrate these cases with examples:

- (٢٨) بس انا من يوم ما قررت ابقي سعيده ، وانا
  ببقي سعيده حتي لو بعض الاوقات بكتئب
  بس ف المجمل انا سعيده

  (28) Eng. ”But since I decided to become happy, I’m happy. Even though sometimes I’m depressed, but in general I’m happy.”

- (٢٩) طب لو انا حزين جلال يبقى ايه ؟
  والله انا مش كئيب ولا
  حاجة.. بس هفرحك حاضر
  :D @mohamedayman20

  (29) Eng. ”So, what if I’m sad, Jalal? I swear I’m not depressed or anything.. but, fine, I’ll make you happy :D @username. [8]”

6. Conclusion

In this paper, we presented DINA, a multi-dialect data set for Arabic emotion analysis. We described an automatic approach of acquiring emotion data in MSA and dialectal Arabic from the Twitter genre, and evaluated that approach via human annotation. Our analysis shows the feasibility of the automatic acquisition of emotion data using the phrase-seed approach we employ. In the future, we plan to extend DINA to other genres, crawl more data using the same seed set (and possibly other seed sets that we may come to find useful), and develop machine learning classifiers exploiting the data.

[8] The real username is replaced by the token @username here.

7. References

Abdul-Mageed, M. and Diab, M. (2011). Subjectivity and sentiment annotation of modern standard arabic newswire. In *Proceedings of the 5th Linguistic Annotation Workshop*, pages 110–118, Portland, Oregon, USA, June. Association for Computational Linguistics.

Abdul-Mageed, M. and Diab, M. (2012a). *AWATIF*: A multi-genre corpus for modern standard arabic subjectivity and sentiment analysis. In *Proceedings of LREC*, volume 12.

Abdul-Mageed, M. and Diab, M. (2012b). Toward building a large-scale arabic sentiment lexicon. In *Proceedings of the 6th International Global WordNet Conference*, Matsue, Japan, January.

Abdul-Mageed, M. and Diab, M. T. (2014). Sana: A large scale multi-genre, multi-dialect lexicon for arabic subjectivity and sentiment analysis. In *LREC*, pages 1162–1169.

Abdul-Mageed, M., AlAhmed, A., and Korayem, M. (2011a). Linguistic features, language variety, and sentiment in Arabic comments on Aljazeera and Alarabiya YouTube Videos. In *Georgetown University Round Table on Languages and Linguistics (GURT2011). Language and New Media: Discourse 2.0.* Georgetown University.

Abdul-Mageed, M., Albogmi, H., Gerrio, A., Hamed, E., and Aldibasi, O. (2011b). Tweeting in Arabic: What, How and Whither. In *The 12th annual conference of the Association of Internet Researchers (Internet Research 12.0 ? Performance and Participation). Seattle, USA*. Association of Internet Researchers.

Abdul-Mageed, M., Diab, M., and Korayem, M. (2011c). Subjectivity and sentiment analysis of modern standard arabic. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 587–591, Portland, Oregon, USA, June. Association for Computational Linguistics.

Abdul-Mageed, M., Kübler, S., and Diab, M. (2012). Samar: a system for subjectivity and sentiment analysis of arabic social media. In *Proceedings of the 3rd Workshop in Computational Approaches to Subjectivity and Sentiment Analysis*, WASSA '12, pages 19–28, Stroudsburg, PA, USA. Association for Computational Linguistics.

Abdul-Mageed, M., Diab, M. T., and Kübler, S. (2013). Asma: A system for automatic segmentation and morpho-syntactic disambiguation of modern standard arabic. In *RANLP*, pages 1–8.

Abdul-Mageed, M., Diab, M., and K S. (2014). Samar: Subjectivity and sentiment analysis for arabic social media. *Computer Speech Language*, 28(1):20 – 37.

Altman, D. G. (1991). *Practical Statistics for Medical Research*. Chapman & Hall / CRC, London.

Aman, S. and Szpakowicz, S. (2007). Identifying expressions of emotion in text. In *Text, Speech and Dialogue*, pages 196–205. Springer.

Badaro, G., Baly, R., Hajj, H., Habash, N., and El-Hajj, W. (2014). A large scale arabic sentiment lexicon for arabic opinion mining. *ANLP 2014*, 165.

Bamman, D. and Smith, N. A. (2015). Contextualized sarcasm detection on twitter. In *Ninth International AAAI Conference on Web and Social Media*.

Bamman, D., Eisenstein, J., and Schnoebelen, T. (2014). Gender identity and lexical variation in social media. *Journal of Sociolinguistics*, 18(2):135–160.

Bechara, A., Damasio, H., and Damasio, A. R. (2000). Emotion, decision making and the orbitofrontal cortex. *Cerebral cortex*, 10(3):295–307.

Bechara, A. (2004). The role of emotion in decision-making: evidence from neurological patients with orbitofrontal damage. *Brain and cognition*, 55(1):30–40.

Bollen, J., Mao, H., and Zeng, X. (2011). Twitter mood predicts the stock market. *Journal of Computational Science*, 2(1):1–8.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20:37–46.

Davidov, D., Tsur, O., and Rappoport, A. (2010). Semi-supervised recognition of sarcastic sentences in twitter and amazon. In *Proceedings of the Fourteenth Conference on Computational Natural Language Learning*, pages 107–116. Association for Computational Linguistics.

Diab, M. and Habash, N. (2007). Arabic dialect processing tutorial. In *Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Tutorial Abstracts*, pages 5–6. Association for Computational Linguistics.

Diab, M., Hacioglu, K., and Jurafsky, D. (2004). Automatic tagging of arabic text: From raw text to base phrase chunks. In *Proceedings of HLT-NAACL 2004: Short Papers on XX*, pages 149–152. Association for Computational Linguistics.

Diab, M., Habash, N., Rambow, O., Altantawy, M., and Benajiba, Y. (2010). Colaba: Arabic dialect annotation and processing. In *LREC Workshop on Semitic Language Processing*, pages 66–74.

Ekman, P. (1992). An argument for basic emotions. *Cognition & emotion*, 6(3-4):169–200.

Farias, D. I. H., Guzman-Cabrera, R., Reyes, A., and Rocha, M. A. (2013). Semantic-based features for author profiling identification: First insightsnotebook for pan at clef 2013. *Forner et al.*

Flekova, L. and Gurevych, I. (2013). Can we hide in the web? large scale simultaneous age and gender author profiling in social media. In *CLEF 2012 Labs and Workshop, Notebook Papers*.

Forner, P., Navigli, R., and Tufis, D. (2013). Clef 2013 evaluation labs and workshop–working notes papers, 23-26 september. *Valencia, Spain*.

Francisco, V. and Gervás, P. (2006). Automated mark up of affective information in english texts. In *Text, speech and dialogue*, pages 375–382. Springer.

González-Ibánez, R., Muresan, S., and Wacholder, N. (2011). Identifying sarcasm in twitter: a closer look. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language*

Technologies: short papers-Volume 2, pages 581–586. Association for Computational Linguistics.

Habash, N., Rambow, O., and Roth, R. (2009). Mada+tokan: A toolkit for arabic tokenization, diacritization, morphological disambiguation, pos tagging, stemming and lemmatization. In *Proceedings of the 2nd International Conference on Arabic Language Resources and Tools (MEDAR)*, Cairo, Egypt.

Habash, N. (2010). Introduction to arabic natural language processing. *Synthesis Lectures on Human Language Technologies*, 3(1):1–187.

Heath, C., Bell, C., and Sternberg, E. (2001). Emotional selection in memes: the case of urban legends. *Journal of personality and social psychology*, 81(6):1028.

Kramer, A. D., Guillory, J. E., and Hancock, J. T. (2014). Experimental evidence of massive-scale emotional contagion through social networks. *Proceedings of the National Academy of Sciences*, 111(24):8788–8790.

Meina, M., Brodzinska, K., Celmer, B., Czoków, M., Patera, M., Pezacki, J., and Wilk, M. (2013). Ensemble-based classification for author profiling using various features. *Notebook Papers of CLEF*.

Mohammad, S. M. and Kiritchenko, S. (2013). Using nuances of emotion to identify personality. *arXiv preprint arXiv:1309.6352*.

Mohammad, S. M. and Kiritchenko, S. (2015). Using hashtags to capture fine emotion categories from tweets. *Computational Intelligence*, 31(2):301–326.

Mohammad, S. M. (2012). #emotional tweets. In *Proceedings of the First Joint Conference on Lexical and Computational Semantics-Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation*, pages 246–255. Association for Computational Linguistics.

Pearl, L. and Steyvers, M. (2010). Identifying emotions, intentions, and attitudes in text using a game with a purpose. In *Proceedings of the naacl hlt 2010 workshop on computational approaches to analysis and generation of emotion in text*, pages 71–79. Association for Computational Linguistics.

Plutchik, R. (1980). *Emotion: A psychoevolutionary synthesis*. Harpercollins College Division.

Plutchik, R. (1985). On emotion: The chicken-and-egg problem revisited. *Motivation and Emotion*, 9(2):197–200.

Plutchik, R. (1994). *The psychology and biology of emotion*. HarperCollins College Publishers.

Qadir, A. and Riloff, E. (2014). Learning emotion indicators from tweets: Hashtags, hashtag patterns, and phrases. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics*, pages 1203–1209.

Refaee, E. and Rieser, V. (2014). An arabic twitter corpus for subjectivity and sentiment analysis. In *LREC*, pages 2268–2273.

Sanfey, A. G., Rilling, J. K., Aronson, J. A., Nystrom, L. E., and Cohen, J. D. (2003). The neural basis of economic decision-making in the ultimatum game. *Science*, 300(5626):1755–1758.

Schwarz, N. (2000). Emotion, cognition, and decision making. *Cognition & Emotion*, 14(4):433–440.

Strapparava, C. and Mihalcea, R. (2007). Semeval-2007 task 14: Affective text. In *Proceedings of the 4th International Workshop on Semantic Evaluations*, pages 70–74. Association for Computational Linguistics.

Tan, C., Lee, L., and Pang, B. (2014). The effect of wording on message propagation: Topic-and author-controlled natural experiments on twitter. *arXiv preprint arXiv:1405.1438*.

Tsur, O., Davidov, D., and Rappoport, A. (2010). Icwsm-a great catchy name: Semi-supervised recognition of sarcastic sentences in online product reviews. In *ICWSM*.

Wang, W., Chen, L., Thirunarayan, K., and Sheth, A. P. (2012). Harnessing twitter" big data" for automatic emotion identification. In *Privacy, Security, Risk and Trust (PASSAT), 2012 International Conference on and 2012 International Confernece on Social Computing (SocialCom)*, pages 587–592. IEEE.

Arabic Spam Detection in Twitter

Nora Al Twairesh*, Mawaheb Al Tuwaijri, Afnan Al Moammar, Sarah Al Humoud
College of Computer and Information Science
Imam Muhammad Bin Saud Islamic University
*King Saud University
Riyadh, Saudi Arabia
E-mail: twairesh@ksu.edu.sa, mituwaijri@sm.imamu.edu.sa, aaalmoamar@imamu.edu.sa,
s.alhumoud@ccis.imamu.edu.sa

Abstract

Spam in Twitter has emerged due to the proliferation of this social network among users worldwide coupled with the ease of creating content. Having different characteristics than Web or mail spam, Twitter spam detection approaches have become a new research problem. This study aims to analyse the content of Saudi tweets to detect spam by developing both a rule-based approach that exploits a spam lexicon extracted from the tweets and a supervised learning approach that utilizes statistical methods based on the bag of words model and several features. The focus is on spam in trending hashtags in the Saudi Twittersphere since most of the spam in Saudi tweets is found in hashtags. The features used were identified through empirical analysis then applied in the classification approaches developed. Both approaches showed comparable results in terms of performance measures reported reaching an average F-measure of 85% for the rule based approach and 91.6% for the supervised learning approach.

Keywords: Twitter Spam, Arabic, Content-Based.

1. Introduction

Although a phenomenon on the Web and email for decades, spam in social networks or Twitter differs from web or mail spam in different ways hence the techniques for detecting it also differ. There is no unified definition of Twitter spam and spammers in the literature. Spam could be any message containing a malicious URL or advertising content that is not related to the hashtag. Most studies consider any automation of tweets as spam (Verma and Sofat 2014). Therefore Twitter spam has been manipulated in different ways accordingly. Although Twitter does apply several spam detection techniques to detect and suspend spam accounts. As stated in ("The Twitter Rules" 2016), more than 15 factors are listed as indications of spamming and can result in the suspension of such accounts. In this work, we consider spam that is manifested in trending hashtags with the aims of advertising unrelated content or disseminating malicious or unrelated URLs.

Spam detection techniques for Twitter can be classified into content-based, which analyse the characteristics of the content of the tweet, user-based which analyse the spam accounts through their social networks and behaviour or a combination of both (Verma and Sofat 2014). The focus of this paper is on content based spam.

As spamming techniques in Twitter get more sophisticated, a need for revising the methods and features used in detecting spam emerges. In the case of spam in trending hashtags, it has been proven more expedient to exploit natural language processing methods e.g. language models to analyse the content of the tweets in relation to the topic of the hashtag (Martinez-Romo and Araujo 2013). Also number of hashtags in a tweet (Benevenuto et al. 2010) and URL presence (Grier et al. 2010) are good indicators of spam.

Empirical analysis of a small set of Saudi tweets extracted from trending hashtags was performed first to understand the nature of spam in Saudi tweets. Specific features were identified and used in both a rule-based approach and supervised learning approach. The features are very simple compared to what is found in the literature; hence the aim here is to add this spam detecting component to a larger framework that will be developed to perform sentiment analysis of trending Saudi hashtags. Using simple methods and features has the advantage of less processing time, a very important factor when performing real-time detection.

This paper is organized as follows: in section 2, a review of the literature on spam in Twitter is cited. Section 3, presents the details of data collection while Section 4 identifies the features to be used in classification through empirical analysis. In Section 5, the preprocessing steps of the dataset are listed. The rule-based approach and supervised learning approach are both presented in Section 6. In Section 7, results the classifiers have reached are reported and discussed. Finally, Section 8, concludes the study findings.

2. Related Work

The majority of studies on Twitter spam started by detecting spam accounts using different features such as followers and followees numbers and the ratio between them, tweets frequency, account age, the device used for tweeting, behavioral aspects, and interaction rate (Chu et al. 2010; Benevenuto et al. 2010; Stringhini, Kruegel, and Vigna 2010; Wang 2010; Wang 2012). In addition to these features, some studies exploit social honeypots to attract spammers, and then thoroughly analyze these accounts such as (Lee, Eoff, and Caverlee 2011).

Although earlier studies on twitter spam focus on detecting spam accounts, (Martinez-Romo and Araujo

2013) however proposed a method to detect spam in the content of tweets regardless of the source account. They used statistical analysis of language to detect spam in trending hashtags. Santos et al. (2014) also present a study based on content of tweets by using machine learning and compression techniques similar to what is used in email spam detection. In means of real-time spam detection (W. Chen et al. 2015; C. Chen et al. 2015) focus on light features for detecting Twitter spam in favor of online detection. A survey of these studies can be found in (Verma and Sofat 2014). While a good review of the features used in most Twitter spam studies can be found in (Al-Khalifa 2015).

In the expanse of Arabic Tweets, not much work has been done to detect spam. Al-Khalifa (2015), conducts an empirical analysis of Twitter spam accounts in the Saudi Arabian Twittersphere. The total number of 2187 spam accounts were collected and analyzed for their behavior, content and network properties. The findings of Al-Khalifa (2015) distinguish three types of spam accounts: re-tweeting spammers, news promoters and trending topics polluters which the author relates to the political nature of the country. As for network analysis the main observation is that the spam accounts are socially disconnected but the occurrence of spam farms is noticed. A machine learning based system was developed in (Mawass and Alaboodi 2015) to detect spam and spammers in trending Saudi hashtags. Features for both user-based and content-based are exploited in this study, although they are very similar to what was applied in previous studies. The classifier was also augmented with a hunter unit to find more spammers. An attempt to detect twitter accounts that disseminate Arabic abusive content is presented in (Abozinadah, Mbaziira, and Jones Jr 2015).

3. Data Collection

A dataset of around 40K tweets was collected using the Twitter Search API during December 2015 and January 2016. The tweets were collected from trending hashtags in Saudi Arabia in different domains (social, political, sports, technology). As illustrated in (Golbeck, Grimes, and Rogers 2010) manual content analysis has been proven useful to understand different characteristics of twitter accounts. In this line, we randomly chose 995 tweets to manually analyze their content in order to identify the features that distinguish spam content in Saudi tweets. Description of this development dataset is illustrated in Table 1.

After that we randomly chose 15,000 tweets from the larger dataset of 40,000 tweets (excluding the above development dataset). One of the main characteristics of spam tweets is that spammers tend to retweet the same tweet to the extent that a hashtag gets flooded with the same tweet, this complicates collecting unique spam tweets. As such duplicated tweets and retweets were removed. After preprocessing the dataset, we were left with a dataset of 3064 tweets, which we will call the unbalanced dataset since the number of spam and non-spam tweets was not balanced. We augmented this dataset with more tweets with the aim of reaching a balanced dataset, the result was 5000 tweets that were manually labeled as spam or non-spam, in total there were 2500 spam tweets and 2500 non-spam tweets this constitutes the balanced training dataset. Another dataset for testing was constructed, it contains 740 tweets: 370 spam and 370 non-spam. The datasets are illustrated in Table 2. We plan to release these labelled datasets for the research community.

Hashtag	Number of Tweets	Spam	Non Spam
#اوامر_ملكيه	199	55	144
#من_عيوب_تويتر	200	69	131
#الخليج_النصر	199	42	157
# الهلال_الاهلي	99	35	64
# سي_ان_ان_للسعوديه	199	40	159
#إعفاء_عزام_الدخيل	99	4	95
Total	995	245	750

Table 1: Hashtags used to construct the development dataset and number of spam and non-spam tweets in each.

Dataset	Spam	Non-spam	Total
UnBalanced dataset	1054	1992	3046
Balanced training dataset	2500	2500	5000
Test dataset	370	370	740

Table 2: Datasets used in the experiments.

4. Features

First, we analyzed the presence of URLs in a tweet as an indicator of spam following (Grier et al. 2010), we found that 232 of the non-spam tweets in this dataset contained URLs and 211 of the spam tweets contained URLs. As a conclusion URL presence is not a good indicator of spam but to validate this finding we added URL presence as a feature to our classification model. Further analysis of this dataset also revealed that tweets containing ads usually contain phone numbers; this was also added as a feature. The analysis of the vocabulary found in the spam tweets revealed the presence of certain words as indicators of spam. Consequently we extracted these words and constructed an Arabic Spam Detecting Lexicon (ASDL). The lexicon contains 108 words; examples of these words are illustrated in Table 3. We plan to make the lexicon publically available for the research community. The last phenomenon observed is the presence of more than one hashtag in the spam tweets. We hypothesize that tweets containing more than four hashtags are usually spam. To validate this we apply this rule on the dataset and perform error analysis. The result is that this rule does indicate spam except in two cases: first, tweets in the sports domain can contain more than four hashtags. Second, tweets that present positive/ inspirational quotes and proverbs can contain more than four hashtags as positive words, examples of tweets in this category are illustrated in Table 4. Therefore, using a sentiment lexicon of positive words from, (a previous paper by the authors), we first check that the hashtag word is not in this lexicon.

According to what was found in (Al-Khalifa 2015) that the source application of a tweet (i.e. what application was used to post the tweet), which is included in the tweet entity when downloading it from the Twitter API, can be an indicator of spam, since some spammers use automated sources such as (Yoono [1], HootSuite [2], TweetDeck[3], Twitterfeed[4], IFTTT[5]).

Al-Khalifa (2015) illustrates that through examining the source applications of tweets in her study's dataset, spammers usually use third party tools (Yoono and IFTT) while regular users use mobile phones or the web. In an attempt to explore whether this observation can be used in this study, we examined the sources of the tweets in a subset of this study's dataset (1500 tweets, 1405 non-spam, 95 spam) the confusion matrix for the manual classification we performed is in Table 5. Noticeably, 10% of non-spam tweets were classified as spam due to the source being one of the third party tools mentioned before. Consequently, the source application of a tweet cannot be used as an indicator of spam.

After this empirical analysis, four features were identified to be used in the classification model (URL, phone number, number of hashtags, spam lexicon), which we chose to experiment with two approaches: rule-based method and a supervised learning method using machine learning classifiers.

متجر
اقوى العروض
#تبادل_رتويت
#تابعني_اتابعك
#زيادة_متابعين
فلو
ضيفوني
رتوت

Table 3: Examples of words in the Arabic Spam Detection Lexicon.

Tweet	Category
لمن تتوقع الفوز في مباراه# الهلال_التعاون في# كاس_ولي_العهد# الهلال# التعاون	four hashtags but not spam, from the sports domain.
"هناك كلام لا يقول شىء، وهناك صمت يقول كل شىء"نجيب محفوظ #بوح_حرف #بوح #مما_قرأت #أعجبني	four hashtags but not spam, positive/inspirational quotes.

Table 4: Examples of tweets containing four hashtags and are not spam.

	Tweet source from third-party	
	Spam	Non-spam
Spam	72	23
Non-spam	140	1265

Table 5: Confusion matrix of tweets classified as spam or non-spam according to the source of tweet.

5. Preprocessing

The informal nature of tweets' text necessitates preprocessing the tweets before performing the classification. First, tweets were cleaned from unrelated content such as user mentions and non-Arabic characters except for the hashtag sign and URLs. Next, content filtering was applied to remove content that does not affect the text or meaning. Filtering step contains four stages: normalization, removing repeated letters and elongation, removing special characters and punctuation and stemming. Following is a list of the preprocessing steps performed:

1. Normalization: The Arabic letters (أ, ة, ي ,و) are manually normalized to convert multiple shapes of the letter to one shape, the different forms of "alif" (أ,آ,إ) are converted into (ا) , the letter "ta'a" (ة) is converted to (ه), the different forms of "ya'a " (ي,ى) are converted into (ي),and the letters (ؤ, ئ) are converted to (ء).
2. Removing repeated letters and elongation: by removing repeated letters such as "جددددا" to become"جدا", and removing the elongation character " ـ " that is used in Arabic to elongate words such as (أنـــــا) to (أنا).
3. Removing special characters, punctuation and diacritics.
4. Stemming: stemming was performed using AraNLP (Althobaiti, Kruschwitz, and Poesio 2014).

In validating the developed approaches we will report the performance measures of precision (P) and recall (R) for spam and non-spam and the average F-measure (F1) of both.

$$P= TP/(TP+FP)$$

$$R=TP/(TP+FN)$$

$$F1=2PR/(R+P)$$

Where in the case of the spam class: TP is the number of spam tweets classified correctly as spam (true positive), FP is the number of non-spam tweets falsely classified as spam (false positive) and FN is the number of spam tweets falsely classified as non-spam (false negatives). The same holds for the non-spam class and the average F1 is calculated as

$$F1_{avg}=(F1_{spam}+F1_{non\text{-}spam})/2$$

6. Classification

6.1 Rule-Based Approach

The rule-based algorithm exploits the ASDL lexicon that was extracted from the development dataset. If any of the words in ASDL is present in the tweet, then the tweet is considered spam. Then the second feature to be tested is

[1] http://www.yoono.com
[2] www.hootsuite.com
[3] www.tweetdeck.com
[4] http://twitterfeed.com
[5] http://ifttt.com

the presence of a phone number using a regular expression that recognizes phone numbers either local (Saudi) or international. If a phone number is found, then the tweet is classified as spam. Next, the algorithm counts hashtags in the tweet and checks if they are non-sentiment hashtags. If there are more than four non- sentiment hashtags the tweet will be classified as spam. When the algorithm does not find any advertisement or spam words, or phone numbers, and the tweet has a sentiment hashtag or less than four hashtags it will be labeled as non-spam. Algorithm 1 presents the details of the rule-based approach in pseudo code.

Algorithm1: Rule-based Classifier

INPUT: Tweets *T*, Arabic Spam Detection Lexicon *ASDL*, Sentiment Hashtags Lexicon *SHL*.
OUTPUT: P = (Spam, Non-Spam).
INITIALIZATION: Counter=0, *where Counter: Count the number of hashtags in each tweet.*

```
FOR each $T_i$ ϵ T DO
    Counter =CountHashTags in $T_i$
    IF phoneNumberFound in $T_i$
        $P_i$ = Spam
    ELSE
        IF $W_j$ (the word j in the tweet $T_i$) ϵ ASDL THEN
            $P_i$ = Spam
        ELSE
            IF counter >= 4 THEN
                IF $W_j$ ϵ SHL THEN
                    $P_i$ = Non-spam
                ELSE
                    $P_i$ = Spam
                END IF
            ELSE
                $P_i$ = Non-spam
            END IF
        END IF
    END IF
END FOR
Return Pi
```

6.2 Supervised Learning Approach

The supervised learning algorithm was evaluated using two machine learning classifiers (ML) Naïve-Bayes (NB) and Support Vector Machines (SVM) since these two algorithms have proven to give better results in most text classification problems. The Weka tool (Hall et al., 2009) was used for the implementation of the classifiers. Several experiments were done with the balanced dataset and with the unbalanced dataset with and without the features. The features are: Phone number presence, number of hashtags, URL presence in addition to the bag of words. Further experiments were performed on the balanced dataset, to test the usefulness of stemming and to test each feature alone and a combination of features.

All experiments were done on the training dataset (see Table 2), and then the resulting classifier was further applied on the test set. The reported measurements are on the test set.

7. Results and Discussion

Although the features used are simple, they were able to detect spam with promising measures. The rule-based approach reported an average F-measure of 85% on the training and test dataset which is decent given the simple approach used. The advantage of using this approach would be that there is no need to have training data and rules can be added as new features are found. The supervised learning approach was validated first on an unbalanced dataset; the results are reported in Table 6. In Table 7 the ML classifiers were applied without stemming, it is shown that the SVM classifier with all features gave the best results 90.2%. Although the average F1 was close in the balanced and unbalanced datasets but as displayed in Table 6 and Table 7, the precision and recall of spam in the unbalanced dataset was very low compared to the balanced dataset, hence a balanced dataset is necessary to build a classifier that can correctly identify spam.

Method	Spam		Non-Spam		Avg. F1
	P	R	P	R	
SVM (bag of words)	87.5	89.2	89	87.3	88.2
NB (bag of words)	87.3	78.4	80.4	88.6	83.5
SVM (All features)	88.6	92.2	91.8	88.1	90.1
NB (All features)	86.6	83.8	84.3	87	85.4

Table 6: Measurements of the unbalanced dataset for NB and SVM with and without features.

Method	Spam		Non-Spam		Avg. F1
	P	R	P	R	
SVM (bag of words)	83.1	97.3	96.7	80.3	88.7
NB (bag of words)	85.9	89.2	88.8	85.4	87.3
SVM (All features)	84.3	98.9	98.7	81.6	**90.2**
NB (All feature)	83.9	93.2	92.4	82.2	87.7

Table 7: Measurement of the balanced dataset for NB and SVM with and without features, without stemming.

The next sets of experiments were done on the balanced dataset to measure the performance of the classifiers in three cases: first, with and without stemming. Second, with and without features. Third, a combination of features as illustrated in Table 8. The highest average F-measure 91.6% was reached when using NB classifier with stemming and the phone number only as a feature or with the number of hashtags. The SVM classifier gave better results without stemming (90.2%) while the NB classifier gave better results with stemming. The effect of features varied, but the phone number feature seems to be the best feature; to the best of our knowledge no previous study used the phone number as a feature in detecting spam. The number of hashtags also performed very well, while the URL presence as we hypothesized before is not a good indicator of spam since non-spam tweets can contain URLs.

All the performance measures of the important experiments are reported in Table 8. Some of the marked experiments are presented, due to space limit.

Method	Spam		Non-Spam		Avg.
	P	R	P	R	F1
Rule-Based	90.8	78.1	80.8	92.1	**85**
SVM (bag of words)	82.9	95.9	95.2	80.3	88
NB (bag of words)	87	92.4	91.9	86.2	89
SVM (All features)	83.4	98.9	98.7	80.3	**89.5**
NB (All features)	86.2	94.9	94.3	84.9	89.8
SVM (Phone+Hash)	83	98.9	98.7	79.7	89.2
NB (Phone+Hash)	90.2	91.6	91.5	90	90.8
SVM (Phone+URL)	84.2	97.8	97.4	81.6	89.7
NB (Phone+URL)	88.1	95.7	95.3	87	**91.3**
SVM (URL+Hash)	83.3	97	96.4	80.5	88.7
NB (URL+Hash)	84.1	94.3	93.5	82.2	88.2
SVM (Phone)	83.6	97.8	97.4	80.8	89.2
NB (Phone)	90.7	92.7	92.5	90.5	**91.6**
SVM (Hash)	82.7	97	96.4	79.7	88.3
NB (Hash)	86.4	91.4	90.8	85.7	88.5
SVM (URL)	82.4	95.9	95.1	79.5	87.6
NB (URL)	83.3	93	92	81.4	87.1

Table 8: Experiments on the balanced dataset with stemming for the two approaches, and with, without features and combination of features.

8. Conclusion

In this study, we tried to identify the best features that can detect spam in Saudi tweets through empirical and experimental analysis. We adhered to using simple features with the aim of decreasing the processing time in favor of performing spam detection in real-time when analyzing hashtags. Although the ML classifiers gave better results than the rule-based classifier as expected but the simplicity of the rule-based approach makes it a good candidate with lack of enough training data. We identified four features: URL presence, phone number, number of hashtags and tweet source. In addition to building and constructing a spam lexicon in the rule-based approach and using the bag of words model in the supervised learning approach. The strongest features to identify spam were the presence of a phone number and number of hashtags per tweet while the URL presence and source application caused most of the classification errors. Also stemming was tested, it gave better results with the NB classifier while it didn't improve the SVM classifier.

As spam detection improves, likewise spamming techniques will, this requires periodically updating training sets and revisiting new features to adhere to the evolving techniques of Twitter spam. Furthermore, for future work we aim to identify opinion spam, which is spam that is related to the topic of the hashtag but is added with the aim of driving people to a certain opinion. This study focused on content based spam detection techniques, the incorporation of user account-based spam detection can also be a venue for future work.

9. Acknowledgements

This work was done by the Arabic Sentiment Analysis Group supported by the Deanship of Scientific Research at Imam Muhammad ibn Saud Islamic University, Riyadh, Saudi Arabia.

10. Bibliographical References

Abozinadah, E. A., Mbaziira, A. V., & Jones Jr, J. H. (2015). Detection of Abusive Accounts with Arabic Tweets. *International Journal of Knowledge Engineering*, *1*(2).

Al-Khalifa, H. S. (2015). On the Analysis of Twitter Spam Accounts in Saudi Arabia. *International Journal of Technology Diffusion (IJTD)*, *6*(1), 46–60.

Althobaiti, M., Kruschwitz, U., & Poesio, M. (2014). AraNLP: A Java-based Library for the Processing of Arabic Text. Presented at the Proceedings of the 9th Language Resources and Evaluation Conference (LREC)), Rekjavik.

Benevenuto, F., Magno, G., Rodrigues, T., & Almeida, V. (2010). Detecting spammers on twitter. In *Collaboration, electronic messaging, anti-abuse and spam conference (CEAS)* (Vol. 6, p. 12).

Chen, C., Zhang, J., Chen, X., Xiang, Y., & Zhou, W. (2015). 6 million spam tweets: A large ground truth for timely Twitter spam detection. In *Communications (ICC), 2015 IEEE International Conference on* (pp. 7065–7070). IEEE.

Chen, W., Yeo, C. K., Lau, C. T., & Lee, B. S. (2015). Real-Time Twitter Content Polluter Detection Based on Direct Features. In *Information Science and Security (ICISS), 2015 2nd International Conference on* (pp. 1–4). IEEE.

Chu, Z., Gianvecchio, S., Wang, H., & Jajodia, S. (2010). Who is tweeting on Twitter: human, bot, or cyborg? In *Proceedings of the 26th annual computer security applications conference* (pp. 21–30). ACM.

Golbeck, J., Grimes, J. M., & Rogers, A. (2010). Twitter use by the U.S. Congress. *Journal of the American Society for Information Science and Technology*, *61*(8), 1612–1621. http://doi.org/10.1002/asi.21344

Grier, C., Thomas, K., Paxson, V., & Zhang, M. (2010). @ spam: the underground on 140 characters or less. In *Proceedings of the 17th ACM conference on Computer and communications security* (pp. 27–37). ACM.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I. H. (2009). The WEKA data mining software: an update. *ACM SIGKDD Explorations Newsletter*, *11*(1), 10–18.

Lee, K., Eoff, B. D., & Caverlee, J. (2011). Seven Months with the Devils: A Long-Term Study of Content Polluters on Twitter. In *ICWSM*. Citeseer.

Martinez-Romo, J., & Araujo, L. (2013). Detecting malicious tweets in trending topics using a statistical analysis of language. *Expert Systems with Applications*, *40*(8), 2992–3000.

Mawass, N. E., & Alaboodi, S. (2015). Hunting for Spammers: Detecting Evolved Spammers on Twitter. *arXiv Preprint arXiv:1512.02573*.

Santos, I., Miñambres-Marcos, I., Laorden, C., Galán-García, P., Santamaría-Ibirika, A., & Bringas, P. G. (2014). Twitter content-based spam filtering. In *International Joint Conference SOCO'13-CISIS'13-ICEUTE'13* (pp. 449–458). Springer.

Stringhini, G., Kruegel, C., & Vigna, G. (2010). Detecting spammers on social networks. In *Proceedings of the 26th Annual Computer Security Applications Conference* (pp. 1–9). ACM.

The Twitter Rules. (n.d.). Retrieved from https://support.twitter.com/articles/18311

Verma, M., & Sofat, S. (2014). Techniques to Detect Spammers in Twitter-A Survey. *International Journal of Computer Applications*, *85*(10).

Wang, A. H. (2010). Don't follow me: Spam detection in twitter. In *Security and Cryptography (SECRYPT), Proceedings of the 2010 International Conference on* (pp. 1–10). IEEE.

Wang, A. H. (2012). Machine Learning for the Detection of Spam in Twitter Networks. In *e-Business and Telecommunications: 7th International Joint Conference, ICETE, Athens, Greece, July 26-28, 2010, Revised Selected Papers* (Vol. 222, p. 319). Springer.