Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 142 (2018) 72–82

Procedia Computer Science

www.elsevier.com/locate/procedia

The 4th International Conference on Arabic Computational Linguistics (ACLing 2018), November 17-19 2018, Dubai, United Arab Emirates

SUAR: Towards Building a Corpus for the Saudi Dialect

Nora Al-Twairesh1, Rawan Al-Matham, Nora Madi, Nada Almugren, Al-Hanouf Al-Aljmi, Shahad Alshalan, Raghad Alshalan, Nafla Alrumayyan, Shams Al-Manea, Sumayah Bawazeer, Nourah Al-Mutlaq, Nada Almanea, Waad Bin Huwaymil, Dalal Alqusair, Reem Alotaibi, Suha Al-Senaydi and Abeer Alfutamani

Information Technology Department, King Saud University, Riyadh, Saudi Arabia

Abstract

This paper presents the preliminary results of the construction of a morphologically annotated corpus for the Saudi dialect. We call the corpus SUAR (SaUdi corpus for NLP Applications and Resources). The corpus consists of around 104,079 words collected from different online sources. The linguistic features of the Saudi dialect are elaborated and compared with Modern Standard Arabic and other Arabic dialects. This paper conducts a pilot study to explore possible directions to facilitate the morphological annotation of the Saudi corpus. The corpus was automatically annotated using the MADAMIRA tool, after which it was manually inspected to validate the resulting analysis.

Keywords: Arabic NLP, Saudi Arabic, Saudi corpus, morphological annotation, Arabic dialects;

1. Introduction

Modern Standard Arabic (MSA) was the dominant written language in the Arab World. However, with the emergence of social media platforms, people started using Dialectal Arabic (DA) more frequently than MSA. While

* Corresponding author.*E-mail address: twairesh@ksu.edu.sa*

10.1016/j.procs.2018.10.462

MSA and DA have some similarities, they differ phonologically, morphologically, and syntactically [1]. Furthermore, morphological analysis of DA using natural language processing (NLP) tools designed for MSA presents inaccurate results because it has been reported that MSA morphological analyzers cover only 60% of Levantine Arabic verb forms [2], as other dialects have more complex morphological variations than MSA [3]. Moreover, dialects do not have standard orthographies. This makes the task of building morphological analyzers and Part of Speech (POS) taggers for dialects immensely challenging.

It is essential to build NLP tools that can accurately process the vast amount of dialectal Arabic text on the web. Most NLP applications such as machine translation, sentiment analysis, information extraction, and dialogue systems need enabling technologies such as morphological analyzers, POS taggers, and tokenizers to function correctly. As stated before, because DA differs dramatically in all linguistic features from MSA, tools designed for MSA perform poorly when applied to DA. Moreover, Arabic dialects differ to the extent that they can be considered different languages in their own right [4].

Accordingly, extensive efforts have been made to build tools tailored to specific dialects. The Egyptian dialect (EGY) and Levantine dialect (LEV) have received much attention [1,3], and recent work has focused on the Palestinian Dialect (PAL) [5] and the Gulf Dialect (GLF) [6]. However, the Saudi dialect (SD) has received less attention; no previous study has highlighted the linguistic features of SD when compared to MSA and other dialects. Moreover, there exists no corpora for SD that annotates its morphology.

This paper is the first step towards building NLP tools for the Saudi dialect. We discuss the process of collecting and building a corpus of text written in the Saudi dialect. Then, following the work of Jarar et. al [7], we perform a pilot study to investigate the relevance of the MADAMIRA tool [3] for morphological analysis of SD. We run the tool on the corpus then carry out a manual inspection to validate the analysis. We call the corpus the SaUdi corpus for NLP Applications and Resources (SUAR).

The paper is structured as follows. Section 2 contains a discussion of related work on Arabic dialects. Section 3 presents the linguistic variations in the Saudi dialect. Section 4 contains a description of the corpus collection. Section 5 presents the corpus annotation details. Finally, Section 6 presents the conclusion and future recommendations.

2. Related Work

2.1. Arabic Dialectal morphological analyzers

Most of the existing Arabic morphological analyzers were dedicated to serve MSA [8,9]. However, there is a considerable gap on the side of dialectal morphological analyzers. One of the Arabic MSA and dialectal morphological analyzers is CALIMA [1], which was specifically created for the Egyptian dialect. A subsequent study by Habash et al. [10] introduced MADA-ARZ, which is a version of MADA [11], developed specifically for the Egyptian dialect. In the following year, MADAMIRA, which is used to analyze Arabic MSA and EGY dialects, was introduced by [3]. MADAMIRA merges MADA [10] and AMIRA [12]. Like most morphological analyzers, MADAMIRA can be used for tokenization, lemmatization, and identification of morphological features such as parts-of-speech, stems, base phrase chunks, named entities, and diacritic forms. Another morphological analyzer and tagger was introduced by [13] for Egyptian and Levantine dialects. Recent work by Khalifa et al. [14] created the CALIMAGLF morphological analyzer for Gulf dialects as an extension of CALIMAEGY [1].

2.2. Corpus collection and annotation

A considerable amount of research has been conducted to build corpora for the Arabic language. Such work mainly aims to facilitate developing Arabic NLP applications. Contributions in this regard mostly serve the processing of MSA Arabic such as [15,16].

In [17], COLABA which is an Arabic corpus that was built for NLP resources covering four Arabic dialects—Levantine, Egyptian, Moroccan, and Iraqi—was introduced. The authors utilized MAGEAD [2] and the Buckwalter morphological analyzer and generator (BAMA) [8].

Another contribution in the area of dialects was the Gumar corpus, which was compiled by Khalifa et al. [6] for Gulf Arabic dialects. It contains 110 million words that were collected from forum novels. They annotated the corpus at the

document level with Gulf dialect labels (i.e., there is no morphological annotation). In addition, a recent study by Khalifa et al. [18] was an extension of the Gumar corpus. Around 200,000 Emirati Arabic dialect words were selected, after which the corpus was annotated manually to identify tokenization, POS, lemmas, and English glosses. During the manual annotation phase, spelling conventionalization and dialect identification were also taken into consideration.

Curras was built by [5] for the Palestinian Dialect. They collected 43,000 words of the Palestinian Arabic dialect from social media. The annotation process for the corpus was conducted using the MADAMIRA tool [3]. In addition, the authors identified a standard form to orthographically annotate Levantine dialects. This form is used as an extension of CODA (Conventional Orthography for Dialectal Arabic), which was proposed by [19]. CODA was intended to be used as a unified framework for the conventional orthography of Arabic dialects. Habash et al. [19] described the CODA guidelines for the EGY dialect in detail. In addition, in a recent effort, [20] extended the guidelines to cover the dialects of 25 Arabic cities.

Another recent project in the area of Arabic dialects is the MADAR project, built by [21]. Their aim was to develop one framework with unified annotation guidelines for applications of Dialect Identification (DID) and Machine Translation (MT). They created a parallel corpus for the dialects of 25 Arabic cities by translating a set of selected sentences (2000) from the Basic Traveling Expression Corpus (BTEC) [22] in French, English, and MSA. In addition, they created a lexicon containing 1,045 concepts from 25 cities.

As for the dedicated corpora that serve NLP applications for the Saudi dialect, there have been efforts to build corpora from Twitter data for sentiment analysis [23,24]. However, to the best of our knowledge, no morphologically annotated corpus that is dedicated to the Saudi dialect exists. Therefore, we planned to build a corpus and conduct a preliminary study to investigate the linguistic features of the Saudi dialect. This corpus will facilitate the future construction of an efficient and effective morphological analyzer.

3. Saudi Dialect Linguistic Variations

In this section, we discuss some of the most prominent linguistic variations of the Saudi dialect that distinguish it from MSA and other Arabic dialects. The main four variants within Saudi Arabia are: Najdi (the middle part of Saudi Arabia), Hijazi (the western part of Saudi Arabia), Gulf Arabic (the eastern part of Saudi Arabia) and southern dialects (the southern part of Saudi Arabia). In this paper, we concentrate on two subdialects of the Saudi dialect: Hijazi and Najdi. This focus was due to the fact that most of the social media content that was collected was written in these two subdialects (we reached this result after manually inspecting the data). In this section, we will review some of the distinguishing features of the Saudi dialect in comparison to MSA and other dialects.

3.1. Morphological variations

Important differences exist between MSA and SD in terms of morphology. First, like many other dialects, SD lost the feminine and masculine plurals and duals in verbs and most nouns. Some specific inflections that are clear in MSA are ambiguous in SD. For example, حسيت Hset ‘I felt’ in SD is written as أحسستُ aHsast in MSA.

Second, SD uses almost all the attached clitics in MSA (e.g., the definite article +ال /Al+/). SD also has many clitics that do not exist in MSA; for example, the future marker in MSA is +س /sa+/ but in SD it is +ح /Ha+/ (as in حاخذهم ‘I will take them’). Other articles in SD include the progressive particle +ب /b+/ (as in بتجلس ‘she sits’), the demonstrative particle +ها/ha+/ (as in هالمسجد ‘this mosque’), and the interrogative proclitic +ش/š+/ (as in شسالفه ‘what happened?’ and شرايكم ‘what do you think?’).

Third, like several other dialects, SD includes the proclitics +ع /a+/ and +ف /f+/, a shortened form of the prepositions على and في (as in عاليسار ‘on the left’ and فالشنطة ‘in the bag’).

3.2. Orthographic variations

All Arabic dialects, including SD, suffer from orthographic variations due to the lack of standardized orthographic guidelines. Words are normally spelled as they are pronounced, and phonological variations have influenced Saudi dialect orthography.

The Hijazi and Najdi subdialects of Saudi agree on the following orthography: Hamza's writing (ؤ,ء,ئ,أ) is turned into the respective letter corresponding to the pronounced sound. For example, ئ is turned into ي as seen in the word مئة in MSA that is written as مية, the word كفؤ is written as كفو, and the word جاءت in MSA is written as جات or جت.

Hijazi subdialect orthography differs in:

- The letter ث writing which is pronounced as ت. An example of this is in the word اثنين that is written as اتنين.
- The letter ذ writing which is pronounced as د. An example of this is in the word اخذ that is written as خد.

3.3. Phonological variations

Like other Arabic dialects, e.g. PAL [7], SD consists of several distinct subdialects that are phonologically different from MSA and from each other. The most distinguishing pronunciation feature is the phoneme /q/ (corresponding to MSA ق), which is pronounced as /g/ in almost all Saudi subdialects. This feature causes the word /قلب/ to be pronounced as /galb/ instead of /qalb/. Another major difference is the pronunciation of the MSA phoneme /D/ (corresponding to ض), which is normalized to /Dˇ/ in all subdialects. Similar to most other Arabic dialects, such as Egyptian and Palestinian, the MSA glottal stop phoneme has disappeared from Saudi dialects in most cases (i.e. if not preceded or followed by a vowel). For instance, the word ذئب /ði'b/ in MSA is pronounced as /ðīb/, but the glottal stop phoneme in the word بيئةBiŷah is preserved.

The most prominent distinctive phonological feature among Saudi subdialects is the phoneme /k/. In most subdialects, this phoneme is transformed to different pronunciations to distinguish between masculine and feminine singular object and possessive pronouns. For instance, in Najd, the word لك lak in MSA, which means 'to you,' is pronounced as /lis/ or /litš/ in the feminine case and is preserved as /lak/ in the masculine case. It is also pronounced as /ج/ in Gulf Arabic and /ش/ in Southern dialect for feminine pronouns, and /كبدة/ is pronounced as جبدة ,تسبدة, or شبدة. Another example is the difference in the phonemes /ð/ and /θ/, which in Hijazi become /d/ and /t/ (or /s/), respectively. Further, short vowels that appear in MSA are omitted in most Saudi subdialects, which makes بيوت /biyot/ pronounced as /byot/. Examples of phonological variations are presented in Table 1.

Table 1. Phonological variations of Saudi dialect

Phoneme feature	In Saudi dialect	Example
/q/ phoneme (MSA ق)	/g/ in almost all Saudi subdialects	قلب /qalb/ becomes /galb/
/D/ phoneme (MSA ض)	/Dˇ/	ضرس /Dirs/ becomes /Dˇirs/
MSA glottal stop phoneme	Disappeared in most cases (i.e. if not preceded or followed by a vowel)	ذئب /ði'b/ becomes /ðīb/
/k/ phoneme (MSA ك)	In some cases, is transformed to: • /s/ or /tš/ in Najd • /ج/ in Gulf Arabic • /ش/ in Southern dialect	لك/lak/ becomes: • /lis/ or /litš • /lij/ • /lish/
/θ/ phoneme (MSA ث)	/s/ or /t/ in Hijazi	ثاني/θani/ becomes /sani/ or /tani/
/ð/ phoneme (MSA ذ)	/d/ in Hijazi	كذب/kaðib/ becomes /Kdb/
Short vowels	SD omits many short vowels that appear in the MSA	بيوت /biyot/ becomes /byot/

3.4. Lexical variations

Lexically, most SD words are cognate and semantically identical with MSA words. However, there are considerable variations between Saudi and MSA lexicons. Some words used in Saudi dialect are compound of two or more MSA words, and their combinations introduce new forms and sometimes new meanings. For example, the SD word عشان 'because' is created from the combination of two MSA words: شأن 'matter' and على 'preposition.' Another word is كلش 'everything,' which corresponds to MSA words كل 'all' and شيء 'thing.' Similarly, the word ايش 'what' is

formed from combining the MSA word أي 'which' and the word شيء 'thing.' The Saudi lexicon also introduces new forms of words that are not used in MSA such as برضو, 'also,' which corresponds to أيضا in MSA, and حليل 'nice,' which corresponds to لطيف in MSA. Another class of words is that of homonyms—words used in both MSA and SD but with different meanings. For example, راح, which means 'went to' in MSA, is used in Saudi dialect to mean سوف 'will.' Another example is the word 'بكره,' which means 'tomorrow' in SD, while in MSA it means 'early morning.' In addition, many commonly used words in SD are borrowed from different languages, such as بس, which is a Persian word that means 'only' or 'enough'; and دريل from English word 'drill.' Examples of lexical variations of SD are presented in Table 2.

Table 2. Lexical variations of Saudi dialect

Word in SD	Translation	Note
عشان	Because	The word is formed from combining two MSA words: شأن 'matter' and على 'preposition.'
كلش	Everything	The word is formed from combining two MSA words: كل 'all' and شيء 'thing.'
ايش	What	The word is formed from combining two MSA words: أي 'which' and شيء 'thing.'
برضو	Also	The word corresponds to أيضا in MSA.
حليل	Nice	The word corresponds to لطيف in MSA.
راح	Will	The word means 'went to' in MSA.
بكره	Tomorrow	The word means 'early morning' in MSA.
بس	Only/Enough	Borrowed word from Persian.
دريل	Drill	Borrowed word from English.

4. Corpus Collection

The SUAR corpus is a Saudi dialect corpus that contains 104,079 words from different social media sources and includes different Saudi dialects such as Najdi, Hijaz, and Gulf. The multiple stages in the process of its creation are presented in Fig. 1.

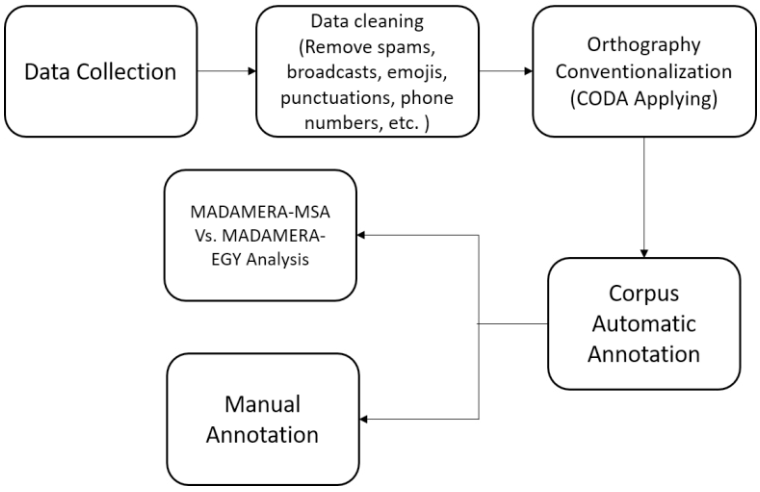

Fig. 1. SUAR Corpus Building Stages.

The process of collecting the corpus data and preparing it for the annotation process is reported in the following sections.

4.1. Corpus collection

Text written in the Saudi Dialect was collected from different social media resources including YouTube, Twitter, forums, blogs, WhatsApp, and Instagram. The following points describe the collection process:

- **Twitter:** Tweets were extracted from the most trending Saudi hashtags at the time of data collection; namely: '#متي_وصلتك_اخر_هديه' '#When_did_you_last_receive_a_gift,' '#حياتك16' '#YourLife16,' and '#اصدقاء_الدراسه_وينهم' '#Where_are_your_school_friends.'
- **YouTube:** videos of Saudi people speaking specific dialects were selected and transcribed.
- **WhatsApp:** Text was collected from 27 WhatsApp groups. The group chats were extracted as a text file, excluding the groups' media.
- **Blogs:** Saudi blogs were selected following several approaches, such as Google searches and through Twitter profiles. We also referred to blogs that had been referenced by other blogs. After accumulating a large list of blogs, we extracted data using two main methods: 1) a Python module for fetching URLs (urllib.request) and 2) a scraper tool[2].
- **Instagram:** Instagram posts were collected from five Saudi accounts that were selected based on popularity, number of posts, and length of posts (accounts with longer posts were preferred). We used the Scrapy framework[3] with Python to crawl Instagram pages.
- **Forums:** Content was extracted from different pages of the Saudi forum vb.eqla3.com. URLs from which data was extracted were chosen manually to ensure text appropriateness.

Table 3 presents the different types of Saudi dialects as well as the number of words and word-types in the collected data. Collected text includes words, punctuation marks, and digits. In addition, Table 4Table presents sentence samples from the collected data.

Table 3. Statistics about the Collected Data.

Data Source	Word Types	Words	Dialect
Twitter	4019	11807	Najdi
YouTube	4503	11795	Gulf, Najdi
WhatsApp	13371	46800	Najdi
Blogs	4822	10907	Najdi, Hijazi
Instagram	5107	12409	Najdi, Hijazi
Forums	4994	10361	Najdi, Hijazi
Total	25998	104079	-

Table 4. SUAR Sentence Samples.

Source	Sample	Translation
Twitter	#متى_وصلتك_اخر_هديه قبل فتره بسيطه يمكن اسبوعين كذا الله لا يحرمنا من الناس اللي تحبنا	#When_did_you_last_receive_a_gift A short while ago, probably about two weeks ago, I ask God not to deprive us from those who love us
Instagram	السلام عليكم شباب	Hi guys
		The most frequent question I ask my classmates is: If you could wake up knowing three languages that you

[2] https://chrome.google.com/webstore/detail/scraper/mbigbapnjcgaffohmbkdlecaccepngjd. [Accessed: 25-Feb-2018]
[3] https://scrapy.org/. [Accessed: 24- Aug- 2018]

	هذا أكثر سؤال اسأله الطلاب الي معي في الصف: لو قالوا لك تخيل تصحى من النّوم و تقدر تتكلم ثلاث لغات بدون ما تتعلمهم ولا شيء فجأة انت تقدر تتكلم لغات جديدة، ايش الثلاث اللغات الي راح تختارهم؟! بشوف ايش اكثر اللغات الي ودكم تتقنونها	have never learned before, what would those languages be?! I want to know the languages that you would like to master the most.
YouTube	السلام عليكم أنا قبل فترة في أحد الكلاسات سمعت المدرس يقول إن الورود تتأثر بالكلام السلبي و بالكلام الإيجابي	Salam, a while ago in class I heard the teacher say that roses are affected by negative speech and positive speech
Forums	حتى اللي تتعود عليه تشتري منه تخاف منه بعد الحين صار الغش صعب كشفه	(Discussing counterfeit products) Even if you were used to buying and using it, you fear consuming it now that fraud has become difficult to detect
Blogs	ما في اهداف معينه كتبتها الى الان بس ناويه على اشياء معينه لكن الى الان ما كتبتها بالشكل اللي حابته	I haven't written any specific goals yet, I intend to do certain things but still haven't written them down the way I would like to
WhatsApp	دكتورة اماني دايم تقول الدورات اهم من المحاضرة اذا شفتي دورة رهيبة غيبي عادي	Dr. Amani always says that courses are more important than lectures, so if you found a really good one, it would be fine to skip class

4.2. Corpus cleaning and preprocessing

Data cleaning and preprocessing is one of the most important processes that need to be performed on social media text since it includes noisy and unnecessary information. We removed URLs, emails, phone numbers, hashtags, emojis, punctuations, and duplicate posts such as retweets in Twitter and broadcasts in WhatsApp. In addition, we performed normalization for the Arabic letters (و, ي, ة, أ) by converting the different forms of 'alif' (أ، آ،إ) to (ا), the letter 'ta'a' (ة) to (ه), the different forms of 'ya'a' (ى،ي) to (ي) and the letters (ؤ،ئ) to (ء). Moreover, we deleted repeated letters from some words, such as changing (مبرووووك), which means 'congratulations,' to (مبروك).

4.3. Orthography conventionalization

Saudi dialects are similar to other Arabic dialects in that they lack standardized orthography guidelines, whereas MSA has an extant orthographic standard. There is a great variation in orthography between Arabic speaker's writing in different Saudi dialects and even within the dialects themselves. Therefore, there are inconsistencies between written texts, even when they are written by the same author. This spelling inconsistency may reflect the phonology of the words or the way of writing the words that are derived from MSA orthography. For example, in MSA, the number three is written as **ثلاثه** thlathah; most speakers of Saudi dialects spelled it the same, but Hijazis pronounce and write it as **تلاتا** talata. Another example is the demonstrative pronoun **هذا** hatha, which means 'this is.' It is spelled and pronounced by all Saudis as '**هذا**' except Hijazis, who write and pronounce it as '**هادا**,' 'hada.' In addition, the feminine subject pronoun, which is known as the letter '**ك**' in Arabic MSA, has different forms in different Saudi dialects. It is written as **تس** ts, or **س** s in the Najdi dialect, **ج** g, or **تش** tsh in Gulf Arabic dialect, and **كي** kee in southern dialects. Consequently, these orthography variations present numerous challenges for computational models in effectively identifying and analyzing dialect words.

In order to overcome these variations and to prepare the data for the automatic annotation that is discussed in the next section, we applied manual Conventional Orthography (CODA) to the corpus text, which was performed by annotators who followed a recent version of CODA that was introduced by [20]. The authors in [20] identified guidelines for 28 Arabic cities including Riyadh and Jeddah. In addition, in this phase, misspelled words were corrected.

5. Corpus Annotation

This section presents our annotation methodology for SUAR. First, we performed automatic annotation using MADAMIRA. Manual annotation was then conducted on a sample of 8,000 words from the SUAR corpus. The approaches we followed in the automatic annotation are described in the next subsections.

5.1. Automatic annotation

For the automatic annotation, we used the existing morphological tagging tool MADAMIRA [3] to expedite the annotation process. This tool was used in similar previous works that were dedicated to build Arabic dialect corpora, such as the Curras corpus built by Jarrar et al. [7] for the PAL dialect. Another work is the Gummar corpus built by Khalifa et al. [6] for the Gulf dialect. They evaluated the use of MADAMIRA in EGY mood to annotate their corpora and concluded that the use of MADAMIRA-EGY to annotate PAL dialect and Gulf dialect is efficient and can be used as an initial annotation process. In addition, we assumed that the Saudi dialects and EGY/MSA share many characteristics and morphological and orthographic features. Therefore, we used MADAMIRA with two dialect models: MADAMIRA-MSA and MADAMIRA-EGY. MADAMIRA has a list of analyses that specify the morphological interpretation per word in-context. Most of the features are selected for the analysis of each word such as part-of-speech (POS), diacritization, lemma, stem, the word proclitic, the word enclitic, and the Buckwalter tag. In addition, we selected the word type to indicate whether each word exists in the MADAMIRA dictionary. The existing words were classified into type 'ARABIC' and the new words into type 'NO_ANALYSIS.' Table 5 presents a sample of MADAMIRA_MSA output and Table 6 presents a sample of MADAMIRA_EGY output.

Table 5. Sample Output from MADAMIRA-MSA.

Row	lemma	Buckwalter tag (BW)	POS	Stem	Type
قاعدين	قاعِد qAEid	qAEid/NOUN+iyona/NSUFF_MASC_PL_ACC	noun	قاعِد	ARABIC
تسافرون	سافَر sAfar	tu/IV2MP+sAfir/IV+uwna/IVSUFF_SUBJ:MP_MOOD:I	verb	سافِر	ARABIC
اوريكم	وَرَّى war~aY	>u/IV1S+war~iy/IV+kum/IVSUFF_DO:2MP	verb	وَرّي	ARABIC
ماشين	ماشِي mA$iy	mA$/ADJ+iyn/NSUFF_MASC_PL	noun	ماش	ARABIC
بزران	بزران bzrAn	-	noun	-	NO_ Analysis
يبونك	يبونك ybwnk	-	verb	-	NO_ Analysis
تتصدم	تتصدم tnSdm	-	noun	-	NO_ Analysis

Table 6. Sample Output from MADAMIRA-EGY.

Row	Lemma	Buckwalter tag (BW)	POS	Stem	Type
قاعدين	قاعِد qAEid	qAEod/ADJ+iyn/NSUFF_MASC_PL	adj	قاعْد	Arabic
تسافرون	سافَر sAfar	ti/IV2P+sAfir/IV+uwna/IVSUFF_SUBJ:MP	verb	سافِر	
اوريكم	وَرَّى war~aY	>u/IV1S+war~iy/IV+kum/IVSUFF_DO:2MP	verb	وَرّي	
ماشين	ماشِي mA$iy	mA$/ADJ+iyn/NSUFF_MASC_PL	adj	ماش	
بزران	بزران bzrAn	-	Adj _comp	-	NO_ Analysis
يبونك	يبونك ybwnk	-	Adj _comp	-	NO_ Analysis
تتصدم	تتصدمtnSdm	-	Adj _comp	-	NO_ Analysis

5.2. MSA vs. EGY analysis

A comparison between the analysis output of MADAMIRA-MSA and MADAMIRA-EYG was conducted to evaluate the use of MADAMIRA with SD. We counted all the 'NO_ANALYSIS' words for the EGY and MSA output files and the results are reported in Table 7.

Table 7. MSA Vs. EGY 'No_ANALYSIS' words.

	MSA	EGY
Words	104,079	104,079
Types	25,998	25,998
No Analysis Types	3,122 (30%)	1,694 (16%)

Overall, the MADAMERA-EGY model performs better with SD with which it gained 16% 'NO_ANALYSIS' types, which is less than the MADAMERA-MSA model, which gained 30%.

The MADAMIRA-MSA was unable to analyze certain words such as 'حبيت,' which has been analyzed as a noun, but in MADAMIRA-EGY 'حبيت' is considered as a verb that means 'I liked something.' Another example, the word, عائشة, was analyzed incorrectly as a verb in MADAMERA-MSA. In contrast, MADAMIRA-EGY analyzed the word correctly within the context as a proper noun. More examples are presented in Table 8 and Table 9.

Table 8. Words Annotated by MADAMIRA-MSA as 'NO_ANALYSIS.'

Raw	POS	Lemma	Status
حبيت	noun	Hbyt	NA
يجيني	noun	Yjyny	NA
عايشة	noun	EAya$	NA
مسوية	noun	Mswyp	NA
حاولو	noun	HAwlw	NA

Table 9. Words Annotated by MADAMIRA-EGY correctly.

Raw	POS	Lemma	Status
حبيت	verb	~Hab	Analyzed
يجيني	verb	jA	Analyzed
عايشة	verb	EA}i$	Analyzed
مسوية	verb	saw~aY	Analyzed
حاولو	verb	HAwil	Analyzed

This confirmed our assumption that the use of MADAMIRA-EGY model is more suitable for analyzing SD. In particular, one of our corpus dialects, the Hijazi dialect, is similar in nature to the EGY dialect. In addition, the high error ratio of the MADAMIRA-MSA model may be a result of the non-existence of these vocabulary in the MADAMIRA-MSA lexicon, i.e. Out of Vocabulary (OOV).

5.3. Manual annotation

We conducted a manual annotation as a pilot study for 8,000 words selected from 8 different MADAMERA-EGY output files since they had the least 'no analysis' types ratio. For each word, the annotators validated the POS, then counted the number of incorrect POS tags. The results of the manual annotation are reported in Table 10.

Table 10. Manual Annotation Results.

Statistics	MADAMIRA-EGY
Correctly analyzed	82.5%
Incorrectly analyzed	18.9%

MADAMIRA was able to assign correct POS tags for some of the 'No-ANALYSIS' words, so we included these tags in our experiment results. Correct POS tags in context were given to the words correctly recognized by the

MADAMIRA-EGY model and the annotators evaluated them as true. By contrast, incorrect POS tags in context were given to the wrongly analyzed words and the annotators assigned them as false. An example of wrongly assigned POS tags, the word مرة mart, in Table 9 which means 'his wife' is known and MADAMERA-EGY tagged it as a verb. This error may be due to the different orthography forms between this word in EGY dialect since it is spelled as مراة marat. Other examples are illustrated in Table 11.

Table 11. Sample of Manual Annotation.

Sentence	وعازمة مرة السفير				
Tokens	الكويتي	السفير	مرة	عازمة	و
Pos	adj	noun	verb	verb	conj
Evaluation	True	True	FALSE	FALSE	True

Overall, the ratio of the correctly analyzed POS tags obtained 82.5% indicates that the use of MADAMIRA-EGY is almost accurate with the Saudi dialect. Therefore, it can be used to accelerate the annotation process.

6. Conclusion and Future Work

In this paper, we conducted a pilot study and presented our preliminary results in building an annotated corpus of the Saudi dialect. We gathered a Saudi corpus that consists of 104,079 words from social media websites such as Twitter, YouTube, WhatsApp, blogs, Instagram, and forums. We compared the linguistic variations and challenges of the Saudi dialect with Modern Standard Arabic regarding lexicon, morphology, phonology, and orthography.

We discussed and compared our annotations to the automatic annotations of MADAMIRA-MSA and MADAMIRA-EGY. The results suggest that using MADAMIRA-EGY automatic annotations as a starting point for manual annotation of the Saudi dialect speeds up the process.

In the future, we intend to expand our corpus to include more texts, and we plan to build a morphological analyzer for the Saudi dialect. We plan to make the corpus publicly available.

Acknowledgment

The authors thank the Deanship of Scientific Research and RSSU at King Saud University for their technical support.

References

1. Habash N, Eskander R, Hawwari A. A morphological analyzer for Egyptian Arabic. In: *Proceedings of the Twelfth Meeting of the Special Interest Group on Computational Morphology and Phonology*. Association for Computational Linguistics; 2012, p. 1–9.
2. Habash N, Rambow O. MAGEAD: A Morphological Analyzer and Generator for the Arabic Dialects. In: *Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics.* Association for Computational Linguistics 2006; p. 681–688
3. Pasha A, Al-Badrashiny M, Diab M, Kholy AE, Eskander R, Habash N, et al. MADAMIRA: A Fast, Comprehensive Tool for Morphological Analysis and Disambiguation of Arabic. *Proc Ninth Int Conf Lang Resour Eval LREC-2014*; 2014.
4. Zaidan OF, Callison-Burch C. Arabic dialect identification. *Comput Linguist* 2014;40(1):171–201.
5. Jarrar M, Habash N, Alrimawi F, Akra D, Zalmout N. Curras: an annotated corpus for the Palestinian Arabic dialect. *Lang Resour Eval.* 2017 Sep 1;51(3):745–75.
6. Khalifa S, Habash N, Abdulrahim D, Hassan S. A Large Scale Corpus of Gulf Arabic. In: *Proceedings of the 10th International Conference on Language Resources and Evaluation, LREC 2016.* European Language Resources Association (ELRA) 2016; p. 8.
7. Jarrar M, Habash N, Akra D, Zalmout N, Bank W. Building a Corpus for Palestinian Arabic : a Preliminary Study. *Assoc Comput Linguist* 2014;18–27.
8. Buckwalter T. Buckwalter Arabic Morphological Analyzer Version 2.0. *Linguistic Data Consortium*, University of Pennsylvania, 2002. LDC Cat alog No.: LDC2004L02. ISBN 1-58563-324-0; 2004.
9. Sawalha M, Atwell E, Abushariah MA. SALMA: standard Arabic language morphological analysis. In: *Communications, Signal Processing, and their Applications (ICCSPA)*, 2013 1st International Conference on. IEEE 2013; p. 1–6.
10. Habash N, Roth R, Rambow O, Eskander R, Tomeh N. Morphological Analysis and Disambiguation for Dialectal Arabic. In: *HLT-*

NAACL. Citeseer; 2013; p. 426–432.
11. Habash N, Rambow O, Roth R. 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 2009; p. 102–109.
12. Diab M, Hacioglu K, Jurafsky D. Automated methods for processing arabic text: from tokenization to base phrase chunking. *Arab Comput Morphol Knowl-Based Empir Methods KluwerSpringer*; 2007.
13. Eskander R, Habash N, Rambow O, Pasha A. Creating resources for dialectal arabic from a single annotation: A case study on egyptian and levantine. In: *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers.* 2016; p. 3455–3465.
14. Khalifa S, Hassan S, Habash N. A morphological analyzer for Gulf Arabic verbs. In: *Proceedings of the Third Arabic Natural Language Processing Workshop*; 2017, p. 35–45.
15. Habash N, Roth RM. CATiB: The Columbia Arabic Treebank. In: *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers* [Internet]. Stroudsburg, PA, USA: Association for Computational Linguistics; 2009 [cited 2018 May 12]. p. 221–224. (ACLShort '09).
16. Arts T, Belinkov Y, Habash N, Kilgarriff A, Suchomel V. arTenTen: Arabic Corpus and Word Sketches. *J King Saud Univ - Comput Inf Sci* 2014; Dec 1;26(4):357–71.
17. Diab M, Habash N, Rambow O, Altantawy M, Benajiba Y. COLABA: Arabic dialect annotation and processing. In: *LREC Workshop on Semitic Language Processing* 2010; p. 66–74.
18. Khalifa S, Habash N, Eryani, F, Obied O, Abdulrahim D, AlKaabi M. A Morphologically Annotated Corpus of Emirati Arabic. In: *Proceedings of the 11th International Conference on Language Resources and Evaluation, LREC 2018*; 2018.
19. Habash N, Diab MT, Rambow O. Conventional Orthography for Dialectal Arabic. In: *LREC 2012*; 2012, p. 711–718.
20. Habash N, Eryani F, Khalifa S, Rambow O, Abdulrahim D, Erdmann A, et al. Unified Guidelines and Resources for Arabic Dialect Orthography. In: *Proceedings of the 11th International Conference on Language Resources and Evaluation, LREC 2018;* 2018, p. 10.
21. Bouamor H, Habash N, Salameh M, Zaghouani W, Rambow O, Abdulrahim D, et al. The MADAR Arabic Dialect Corpus and Lexicon. In: *Proceedings of the 11th International Conference on Language Resources and Evaluation, LREC 2018*; 2018, p. 10.
22. Takezawa T, Kikui G, Mizushima M, Sumita E. Multilingual spoken language corpus development for communication research. *Int J Comput Linguist Chin Lang Process* Vol 12 Number 3 Sept 2007 Spec Issue Invit Pap ISCSLP 2006 2007;12(3):303–324.
23. Al-Twairesh N, Al-Khalifa H, Al-Salman A, Al-Ohali Y. AraSenTi-Tweet: A Corpus for Arabic Sentiment Analysis of Saudi Tweets. In *ACLing 2017*, Dubai, United Arab Emirates: Elsvier; 2017.
24. Assiri A, Emam A, Al-Dossari H. Saudi twitter corpus for sentiment analysis. *World Acad Sci Eng Technol Int J Comput Electr Autom Control Inf Eng.* 2016;10(2):272–275.