خصائص اللهجة الكويتية المكتوبة واستخدامها في إنشاء موارد للتحليل الصرفي الآلي

بشاير عبد الله العتيبي *
إيمان توفيق الشرهان **

الملخص

إن اللهجة الكويتية كبقية لهجات العربية لهجة متداولة شفهيًا، ولا تمتلك معايير مكتوبة موحدة على خلاف اللغة العربية الفصحى. وبعد ظهور منصات التواصل الاجتماعي وانتشارها وجدت اللهجات طريقها إلى الوسائط المكتوبة، وبرزت الحاجة لمعالجتها آليًا جنبًا إلى جنب مع اللغة العربية الفصحى. ولعل أبرز مشكلة واجهت المعالجات الآلية أن اللهجات لا تتمتع بمعايير كتابية ثابتة كالفصحى، وعادة ما يَتّبع الكتّاب باللهجة نظام الكتابة الصوتية؛ أي كتابة الكلمات كما تنطق، مما فتح المجال لوجود تباين في كتابة اللهجة الواحدة وبين اللهجات والفصحى. ولعل أهم المتطلبات التي تحتاجها المعالجات الآلية لمعالجة اللغة الطبيعية هي وجود معايير كتابية واضحة للغة أو اللهجة المراد معالجتها وتحليلها، وقد توالت الجهود لضبط معايير كتابة اللهجات العربية، إلا أن اللهجة الكويتية لم تلقَ الاهتمام المطلوب. ويقدم البحث الحالي حلاً عمليًا لمعالجة اللهجة الكويتية المكتوبة آليًا، فقد تضمنت الدراسة تحديد واستخراج أهم معايير اللهجة الكويتية المكتوبة من بيانات طبيعية جُمعت من تغريدات مغردين كويتيين في تويتر بوصفها نموذجًا من الاستخدام الحقيقي والطبيعي للهجة المكتوبة، تجاوزت مئة ألف تغريدة، ثم تعزيز المحلل الصرفي (MADAMIRA) – وهو محلل صرفي مخصص للغة العربية الفصحى – بهذه المعايير المستخلصة للهجة الكويتية. كما تضمن العمل إثراء المحلل الصرفي بقاموس من المصطلحات والمفردات الكويتية التي جمعت من موسوعة اللهجة الكويتية، ومن أكثر الكلمات الكويتية شيوعًا في تويتر؛ حتى يتعرف المحلل الآلي على هذه المفردات ويحللها تحليلاً سليمًا. وتعد النسخة الموسعة من المحلل الصرفي (MADAMIRA-KA) الأولى من نوعها المخصصة كليًا لمعالجة اللهجة الكويتية، وقد حققت أداء متميزًا في تحليل أكثر من مئة ألف تغريدة كويتية بنجاح. وتكمن أهمية هذه الدراسة في توفير هذا المعالج الصرفي الذي يمكن استخدامه في برامج الترجمة الآلية، والتعرف الآلي على اللهجات، والاستقراء الآلي للرأي والانطباعات.

الكلمات مفتاحية: العربية المكتوبة، المحلل الصرفي، المعالجة الآلية للغة الطبيعية، التواصل الاجتماعي، الكتابة الصوتية، معايير الكتابة.

* أستاذ مساعد، قسم اللغة العربية، كلية الآداب، جامعة الكويت. bashayer.alotaibi@ku.edu.kw
** أستاذ مشارك، قسم اللغة العربية، كلية الآداب، جامعة الكويت. eiman.alsharhan@ku.edu.kw
الاستلام: 2023/4/4، التعديل النهائي: 2023/7/30، إجازة النشر: 2023/9/18.
https://doi.org/10.34120/0117-042-166-009

To cite this article: Alotaibi, Bashayer & Alsharhan, Eiman: "Characteristics of Written Kuwaiti Arabic and their use in Creating Resources for Morphological Analysis". *Arab Journal for the Humanities*, 42, 166, 2024, 275 -306 .

Characteristics of Written Kuwaiti Arabic and their use in Creating Resources for Morphological Analysis

Bashayer Abdullah Alotaibi *
Eiman Alsharhan **

Abstract

Kuwaiti Arabic (KA), like other Arabic dialects, is a spoken variety of Arabic that does not have a standardized written convention contrary to Modern Standard Arabic (MSA). With the emergence and spread of social media platforms, Arabic dialects have found their way into the written medium, and hence a need arose to process them alongside MSA. The biggest challenge facing NLP tools is that dialects do not have consistent written conventions contrary to MSA, and writers expressing their dialects usually follow a phonetic writing system, or they write words as they pronounce them. This has opened the door for variations within the same dialect and between dialects and MSA. Furthermore, a prerequisite for analysing any language or dialect is the presence of clear written conventions. Therefore, efforts have been made to establish written conventions for Arabic dialects, but the Kuwaiti dialect has not received the required attention. The current study offers a practical solution for processing written KA. It identified and extracted the written conventions of KA from natural data collected from over 100K Kuwaiti tweets since they represent a good model of natural language. The morphological analyzer (MADAMIRA) - which is designed to process MSA - was enhanced with the extracted criteria. Furthermore, the study involved enriching the analyzer with a dictionary of Kuwaiti terms and vocabulary 'lemmas' collected from the Encyclopaedia of Kuwaiti Arabic and from the most used Kuwaiti words on Twitter (currently X). Providing the analyzer with this dictionary of KA words helps it process KA more accurately. The expanded version of the analyzer (MADAMIRA-KA) is the first of its kind designed entirely to process the Kuwaiti dialect and has achieved excellent performance in analyzing over 100K Kuwaiti tweets successfully. The importance of this study lies in developing such a morphological analyzer, which can be used for automated translation, dialect recognition and sentiment analysis.

Keywords: written Arabic, morphological analyzer, NLP, social media, phonemic writing, written convention.

* Assistant professor, Department of Arabic Language, College of Arts, Kuwait University, Kuwait.
bashayer.alotaibi@ku.edu.kw
** Associate Professor, Department of Arabic Language, College of Arts, Kuwait University, Kuwait.
eiman.alsharhan@ku.edu.kw

Submitted: 4/4/2023**, Revised:** 30/7/2023**, Accepted:** 18/9/2023**.**

https://doi.org/10.34120/0117-042-166-009

To cite this article: Alotaibi, Bashayer & Alsharhan, Eiman: "Characteristics of Written Kuwaiti Arabic and their use in Creating Resources for Morphological Analysis". *Arab Journal for the Humanities*, 42, 166, 2024, 275 -306 .

1. Introduction

Modern Standard Arabic (MSA) is a formal variety of Arabic used in written contexts. It is usually taught in schools and used in government transactions, and media.(1) However, owing to the spread of social media and the introduction of Arabic text on social media platforms, many Arabic users tend to convey their dialects in their writings. Written Arabic is governed by the rules and conventions of MSA, but there are no widely accepted standardized conventions for Arabic dialects (AD henceforth).

Despite many resources, such as dictionaries and grammars of several Arabic dialects, written Arabic dialects do not have standardized conventions, in contrast to written MSA.(2) Nevertheless, users on social media platforms appear to express their dialects using MSA's orthography and write in a way that represents the phonemes of their dialect, as much as the orthography of MSA would allow. In addition, users express their AD via distinctive vocabulary items of high-frequency nouns or verbs that are different from MSA and from one another. For example, functional words such as 'I want' have different forms depending on the Arabic variety used, such as *ʔarːd* (3) < أريد >, *ʔabi*: < أبي >, *ʔabɣi*: < أبغي >, *ʕaːjiz* < عايز >, *baddi*: < بدّي >, *bɣi*:t < بغيت >, and *baːya*: < بايا >, which are typical of specific dialects (discussed further in the end of section 2.3). Furthermore, other prominent features are reflected in the written orthography, which help distinguish MSA from AD and different dialects from one another; however, they have not been the subject of documentation and examination in their 'written' form, especially not for the case of Kuwaiti Arabic (henceforth KA).

KA is a cover term for several closely related urban dialects spoken in Kuwait. It is usually described as one of many Gulf Arabic Dialects. Its main phonological characteristics and lexical inventory have been documented and described in many references.(4) However, it is still considered an under-resourced dialect, especially in the domain of natural language processing (NLP)(5). One reason pertains to its written system, and another to its morphology. In written Arabic, short vowels may be written as diacritics above or under the main consonant grapheme, but in practice, they are not used; hence, only the consonants are represented in addition to long vowels. Speakers of Arabic can read Arabic texts efficiently depending on their competence and prior knowledge of how those written words are pronounced. However, in some words, a change of vowel can result in two different words. For example, a word such as < كسر > can mean either /kasara/ 'to break' or /kusira/ 'broken' since they are written the same way (without diacritics), although that they have different vowels and meanings. As for Arabic derivational morphology, it follows a root/pattern system, where roots are usually triliteral consonants, and patterns are simple vowels, or a combination of consonants and vowels affixed between the consonants of the root. Returning to the previous example, < كسر > /*kasara*/ 'to break' is a verb that has the following three consonants: *k*,

s, and r. The vowels in between reflect the pattern that can be represented as follows: $C_1aC_2aC_3a$. The verb < كسّر >/*kassara*/ 'to break repeatedly or forcefully, to smash' has the same root consonants but a different pattern: $C_1aC_2C_2aC_3a$. A morphological analyzer of Arabic should be able to differentiate roots from pattern consonants and vowels to easily process words, which is a task that has been advanced in NLP. However, a morphological analyzer for Arabic dialects would have an additional layer of complexity since it must be able to analyze the different phonological, morphological, and syntactic features distinctive of AD, in addition to those of MSA.

1.1. Related Work in NLP

In the domain of natural language Processing (NLP), there is a need to develop language processors and morphological analyzers that can analyze natural written Arabic of different varieties. Work on developing morphological analyzers for MSA has been ongoing for the past thirty years, resulting in numerous resources with varying degrees of coverage and accuracy.(6) However, it has been reported that using morphological analyzers designed for MSA for analyzing dialectal Arabic shows imprecise results due to the significant variations between these dialects and MSA.(7) As a result, there was a substantial change in research in regard to building different morphological analysis tools adapted to specific dialects. Egyptian and Levantine Arabic have received the greatest attention recently compared to other dialects.(8)

There are some efforts to provide sufficient corpora for AD such as project MADAR which aims at dialectal identification.(9) However, to our knowledge, KA does not enjoy rich corpora resources, dictionaries, and written conventions, and consequently, no dedicated morphological analyzer to process its data proficiently.

Improving the performances of morphological analyzers involves enriching them with written conventions and a sufficient dictionary. As mentioned above, MSA has a well-documented and described written convention, in contrast to DA. Furthermore, written KA has not been well-described, and accordingly, it does not have a written convention. Thus, this study aims to develop a morphological analyzer capable of handling KA data by describing the prominent features of KA conveyed in the 'written' domain, based on actual data collected from KA users of Twitter. The description was then used as input for developing a morphological analyzer capable of handling written KA. Thus, the research aimed to answer the following questions: 1) what are the features of written KA on Twitter platform? and 2) Can these features be used to substantially improve the performance of a morphological analyzer, initially designed to handle MSA, for analyzing written KA data?

We anticipate that the expanded analyzer will prove to be a useful tool in developing

most NLP applications for KA such as machine translation, part of speech tagging, sentiment analysis, information retrieval, and speech recognition systems. All these applications depend on a well-designed morphological analyzer, that is enriched with the necessary written convention and rules of written KA.

The following section provides a general description of Kuwaiti Arabic, focusing mainly on elements that distinguish KA from other ADs and from MSA, and how these characteristics may see their way into written KA. Section (3) presents the methodology used in this research, where we relied on natural written data collected from Twitter, and the steps taken to develop the morphological analyzer. Section (4) discusses the results. Finally, we conclude the paper with recommendations for further research.

2. Characteristics of Kuwaiti Arabic

In this section, we present a description of the main features of KA and focus mainly on those we found apparent in the written form, based on data collected from 100,000 tweets by KA users (see Section 4). These features can be categorized as phonological, morphological, syntactic, and vocabulary items, which we present separately.

2.1. Distinctive Phonological Features of KA

Some of the obvious phonological differences between MSA and KA include consonant substitutions, epenthetic vowel insertion, and hamza alleviation. These differences may affect the way KA users reflect written KA on Twitter. These features are discussed subsequently.

Consonant substitutions are one of the main phonological features of KA. The following consonants usually undergo substitution in KA when compared to MSA: < ك، ق، ج، ض >. They not only undergo substitution when spoken, but also when written as the results shown in Section (5).

First, it is common in KA for words containing /k/ < ك > to be affricated and pronounced as /ʧ/. However, this change is not random nor consistent in all words containing /k/. Some words still maintain the original /k/ such as kuwajt < كويت > 'Kuwait,' kala:m < كلام > 'speech,' and kawkab < كوكب > 'planet.' Furthermore, some words are pronounced with /ʧ/ instead, such as ʧalb < چلب > 'dog,' ʧaðða:b < چذّاب > 'liar,' and ʧabri:t < چبريت > 'match sticks.' Affrication of /k/ is not consistent with all words including /k/, nor is it consistent among all speakers of KA. Nevertheless, with functional morphemes or functional verbs, the situation is more stable and predictable. For example, the pronominal suffix /k/ is affricated for the feminine but not for the masculine, as in ʃiftik < شفتك > 'I saw you' ('you' being a masculine object)' vs. ʃift-iʧ < شفتـچ > 'I saw you'

(feminine object). Also, the functional verb ka:n < كان > 'was' is pronounced as ʧa:n in certain contexts indicating that the two versions have distinguishable functions.(10) Interestingly, in written Arabic, the fricative /ʧ/ does not have a corresponding letter in the alphabet, and writers do not always write it with the alphabet letter /k/ as the results show in Section (5).

Also, many words that originally contain < ج > in MSA are pronounced with /dʒ/ in KA. For example, dʒiri:ʃ < جريش > 'crushed wheat' is pronounced as /jiri:ʃ/, whereas dʒimʕa < جمعة > 'Friday' is pronounced as /jimʕa/. Again, this substitution is not consistent throughout all words containing the consonant < ج >. The conditions of this substitution are not entirely clear and may be a result of old occurrences in the Kuwaiti dialect that have been inherited transgenerationally. However, in educated speech, most of these cases are returned to their original form /dʒ/.(11) Therefore, it is expected that writing such words would deviate from the standard.

As for /q/ it undergoes affrication and fronting resulting in a /g/ /dʒ/ split. This is evident in words such as qalb < قلب > 'heart,' sa:q < ساق > 'leg,' and qamar < قمر > 'moon,' which are pronounced in KA as /galb/, /sa:g/, and /gumar/ respectively. Other words, such as qidr < قدر > 'pot,' ri:q < ريق > 'saliva,' and qali:b < قليب > 'well' are pronounced as /dʒidir/, /ri:dʒ/, and /dʒili:b/. This split is expected to appear in the written context as well.

Finally, almost all spoken Ads do not distinguish between the consonants < ض > /dˁ/ and < ظ > /ðˁ/ in pronunciation. In KA, as in many other Gulf dialects, words that contain < ض > in MSA are pronounced with /ðˁ/ such as /ðˁa:bi tˁ/ < ضابط > 'officer' or /ðˁifdaʕ/ < ضفدع > 'frog'. In written Arabic, it is expected that users would get confused between the letters representing /dˁ/ ض and /ðˁ/ ظ.

Vowel epenthesis is another phonological process that may be reflected in written Ads. Vowel epenthesis is the addition of a vowel in some contexts caused by an underlying phonological process such as syllabification. When the vowel is added to the beginning of the word, it is expected to appear in the written form, not as a diacritic but as a connective hamza. One phonological process common to KA and Najdi-type dialects that can trigger such an effect is the gahawa-syndrome phenomenon. It relates to $C_1aC_2aC_3V$ sequences in MSA that are re-syllabified as $C_1C_2VC_3V$ in Najdi-type dialects.(12) It consists of the deletion of /a/ in C_1aC_2 first and non-final syllables when C_2 is a guttural consonant, and the epenthesis of an /a/ is after C_2.(13) For example, the word ʃadʒara < شجرة > 'tree,' which has a $C_1aC_2aC_3V$ sequence in MSA, is pronounced as /iʃjara/ with a $iC_1C_2aC_3V$, where the /a/ in the first syllable is dropped and substituted with an epenthetic vowel at the beginning of the word, which helps break the consonant cluster(14).

Another instance of vowel epenthesis appears with imperfective verbs when the root of the verb starts with a guttural consonant < ع، غ، هـ، ح، خ >. For example, the imperfective

verb *jaʕrif* < يعرف > 'knows' is pronounced with three syllable /ijʕarif/, whereas a regular imperfective verb with a non-guttural consonant as its first root would have two syllables such as jadris < يدرس > 'study' pronounced as /jadris/.

Furthermore, with perfective verbs in KA, an epenthetic vowel is introduced when the verb shows 3rd person plural agreement and 3rd person singular feminine agreement only. For example, the verb *kataba* < كَتَب > 'he wrote' starts with the short syllable /ki/ as it is pronounced /kitab/ in all its inflectional paradigm in KA, except for the two cases *katabat* < كتبت > 'she wrote' /iktibat/ and *katabu*: < كتبوا > 'they wrote' /iktibaw/.

The addition of a bound object pronoun that starts with a vowel may also trigger an epenthetic vowel at the beginning of the perfective verb. For example, an epenthetic vowel is used with the following object pronouns: *samiʕa-**hu*** < سمعه > 'he heard him' / ismaʕ-**a**/, *samiʕa-**ka*** < سمعك > 'he heard you' /ismaʕ-**ik**/, and /ismaʕ-**iʧ**/ 'he heard you', but not with object pronouns that start with a consonant such as *samiʕa-**ha***: < سمعها > 'he heard her' /simaʕ-ha/ or *samiʕa-**kum*** < سمعكم > 'he heard you all' /simaʕ-**kum**/, for example.

Finally, hamza, which is a glottal stop, is generally alleviated or changed into a vowel KA. For example, the word *raʔs* < رأس > 'head' contains a vowel-less hamza preceded by the vowel /*a*/ in MSA is changed into a long vowel /ra:s/ in KA. Hamza, at the end of the word such as *sama:ʔ* < سماء > 'sky' is usually deleted as in /sima/. A connective hamza may also be deleted in the written data, as will be shown in examples in Section (5).

2.2. Distinctive Morphological Features of KA

This section presents a description of KA's prefixes, suffixes, and clitics that are distinctive of KA and not present in MSA. They include case, mood, and agreement suffixes, definitive and future tense prefixes, and finally functional clitics.

Case markers in Arabic are either simply short vowels or a complex of a vowel and a consonant suffixed to the noun. Short vowels are -*u* ó-for the nominative, -*a* ó-for the accusative, and -i -ó-for the genitive suffixed to the nominal. Complex case markers, for example, are *a:n* < ان > for nominative dual nouns or *u:n* < ون > for nominative masculine plurals. When the dual noun is accusative or dative, the suffix used is *ajn* < يْن >, whereas with the accusative or dative masculine plural, the suffix *i:n* < ين > is used. In the Arabic dialects, these simple-case markers are always dropped. As for the complex case morphemes, KA does not make the distinction between nominative and accusative/dative cases. It uses one form consistently, as shown in example (2) compared to MSA in (1):

(1) MSA

a. daxala	l-muslim.u:n	/ l-muslim.a:n	[Nominative]
entered	the-muslim.P.NOM	/ the-muslim.D.NOM	
'The Muslims entered'		/ 'The two Muslims entered'	
b. samiʔ.tu	l-mudarris.i:	/ l-mudarris.ajn	[Accusative]
heard.1S	the-teacher.P.ACC	/ the-teacher.D.ACC	
'I heard the teachers'		/ 'I heard the two teachers'	

(2) KA

a.idxal.aw	l-muslim.i:n	/ l-muslim.e:n	[Nominative]
entered	the-muslim.P	/ the-muslim.D	
'The Muslims entered'		/ 'The two Muslims entered'	
b. simaʔ.t	l-mudarris.i:n	/ l-mudarris.e:n	[Accusative]
heard.1S	the-teacher.P	/ the-teacher.D	
'I heard the teachers'		/ 'I heard the two teachers'	

Also, in the genitive construct state, the consonant < ن > from the case markers *-i:n* < ين > (for the plural noun) and *ajn* < يْن > (for the dual noun) is deleted in MSA. However, in KA, this process does not apply, and the consonant < ن > appears even in construct states. Compare the following examples:

(3) MSA

muslim-**i:**	ʔu:rubba:
Muslims-P.GEN	Europe
'Europe's Muslims'	

(4) KA

muslim-**i:n**	ʔu:rubba:
Muslims-P	Europe
'Europe's Muslims'	

More specifically, the morpheme -i:n no longer shows case distinctions in KA but is simply a marker for the masculine plural noun.

Similarly, mood markers, that are also short vowels *-u* -ó and *-a* ó- suffixed to the imperfective verb to mark indicative and subjunctive mood respectively, are commonly dropped in pronunciation. However, complex mood markers change depending on the context. Indicative verb jadrusu:n < يدرسون > 'they study' is changed into yadrusu: < يدرسوا > in the subjunctive sentences; the consonant < ن > is deleted from the mood agreement suffix in MSA. In KA, the plural imperfective verb agreement morpheme *-u:n* < ون > does not exhibit any change related to mood.

Some subject agreement morphemes are shortened in KA compared to MSA. This is clear with the perfective verb 2nd person plural masculine agreement morpheme *-tum* < تُم > in MSA, which is shortened to *-taw* < تَو > in KA, where the consonant < م > is dropped, for example *darastum* < درستم > 'you all studied,' which is *darstaw* < درستو > in KA. Also, using the dual subject agreement is limited to MSA in general, and usually indicates an instance of code-switching when used in the written context.

Another critical morphological feature of KA is the use of prefix *b* < بـ > with the imperfective verb to indicate future tense. It is believed to be contracted from the verb *jabi*: < يبي > 'want'(15) A similar prefix is used in other Arabic dialects to indicate progressive aspects such as Jordanian Arabic and Egyptian Arabic; however, that function – for the prefix – is not attested in KA(16). Therefore, this prefix in KA functions similar to the future tense prefix in MSA, as shown in the following examples:

(5) MSA	vs.	KA
sa-jadrusu		**b**-jadris
FUT-study.3SM		FUT-study.3SM
'he will study'		'he will study'

Another important morpheme common to Arabic dialects and not attested in MSA is the use *of* < ش > as an interrogative particle clitic for example: *ʃ-ga:l?* < شقال؟ > 'what did he say?' *or ʃ- ħagga?* < شحقّه؟ > 'what for?'. The same clitic is also used to express exclamation, especially when added to degree words such as *ʃ-kubrah* < شكبَرَه! > 'how big!' and *ʃ-ħala:tah* < شحلاتَه! > 'how lovely!'

2.3. Distinctive Syntactic and Grammatical Features of KA

Many words and phrases are highly frequent in KA. They are mostly grammatical elements and have specific syntactic functions. They are included in this description either because they are distinctively written from MSA or because they are specifically unique to KA.

One example is free pronouns, which are pronounced in a special way in KA. They tend to be pronounced with an epenthetic vowel that could be related to the same process discussed in (2.1). Table I shows that many of these 3rd person pronouns are preceded by a connective hamza. The 1st person plural pronoun iħna < احنا > is one of the most frequent words in our data (see Table III).

Table I: Free pronouns of KA

Pronoun	Syllable structure
1st person, singular	ʔa:na آنا
1st person, plural	iħna احنا
2nd person, plural	intawانتَو
3rd person, singular, masculine	uhwa اهو
3rd person, singular, feminine	ihja اهي
3rd person, plural	uhum اهم

Another example is demonstratives. They are generally like MSA with minor differences. There are two forms of demonstratives; one indicates proximity, whereas the other indicates distance. For singular male and female demonstratives, there is no considerable difference except in the use of (-i) to refer to the female, as shown in the table below (Table II). The form used for the plural is slightly different from that used in MSA. In MSA, plural is referred to by *ha:ʔula:ʔ* < هؤلاء > 'this.P,' whereas in KA, it is *haðawl* < هذول > 'this.P' or *haðawlak* < هذولاك > 'that.P.'

Table II: Demonstratives in Kuwaiti Arabic

Demonstrative	Proximity (close = this)	Proximity (distant = that)
Singular, male	*haða* هذا	*haða:k* هذاك
Singular, female	*haði:*هذي	*haði:ʧ*هذيچ
Plural (male & female)	*haðawl* هذول	*haðawlak* هذولاك

In addition to these forms usually used to refer to animate or inanimate objects with some gender reference, there is the shorter form ha < هَـ >, which is used to indicate deictic reference without reference to gender, as shown in the following example:

(6)

ba-sa:fir	ʔasba:nja	ha-ʃʃahar
FUT-1s.travel	Spain	this-the-month

'I will travel to Spain this month'

Closely related to the demonstrative construction is the presentative construction that also uses a clitic ha- < هَـ >. The difference between the two is that the presentative is directed to the second person only and has three main forms ha:k < هاك > for the singular male addressee, ha:ʧ < هاچ > for the singular feminine, and ha:kum < هاكُم > for the plural, which is then followed by the object being presented, as shown in the following example:

(7)

ha:-kum	l-kita:b
PRST-2P	DEF-book

'Here you have the book' or 'Here! Take the book.'

Presentative *ha-* is also found in MSA. What is different is the use of *kA-* < كا > in a presentative construction in KA. This construction may be related to the existential *ʔaku* < أكو > discussed in the next paragraph. Additionally, *ka:-* is not limited to the 2nd person but may be used with the 1st and 3rd person:

(8)

ka:-hu:	l-kta:b
PRS-3SM	DEF-book

'Here is the book' or 'Here you have the book'

(9)

ka:-ni:	ji:.t
PRS-1S	came.1S

'Here! I came'

The existential construction in Arabic is usually headed by a locative preposition *fi:* < في > 'in' with third-person singular agreement *fi:h* < فيه > 'there is'. In KA, another form may be used which is *ʔaku:* < أكو > 'there is…'. For example:

(10)

ʔaku:	fa:r	taħt	l-siri:r
EXT	mouse	under	the-bed

'There is a mouse under the bed'

The form *ʔaku*: < أكو > is commonly used with the negative particle *ma*: < ما > as in ma:-*ku*: *ʃay* < ماكو شَيَ > 'There is nothing' or with the interrogative particle ʃ- as in *ʃa-ku*:? < شكو؟ > 'What's there?' or 'What's wrong?'

Also, distinctive of Kuwaiti Arabic is the use of ma:l to indicate possession. It is usually used as an adjective following the word expressed as possessed by the subject and showing gender, number, and person agreement with the subject. Al-Qenaie[17] observes that when the possessed object is singular masculine, the form *ma:l* < مال > is used, and when it is singular feminine, the form *ma:lat-* < مالت > is used and a plural form *malu:t* < ملوت > is used when the possessed object is plural regardless of its gender, as shown in the following examples:

(11) Possessed singular

a. l-qalam	ma:l-ha:
the-pen.SM	POS.SM-her
'Her pen'	
b. l-liʔbah ma:lt-ah	
the-toy.SF	POS.SF-his
'His toy'	

(12) Possessed Plural

l-ʔalʔa:b	malu:t-ah
the-toys.P	POS.P-his
'His toys'	

Finally, one of the main features that differentiates an Arabic variety from another regionally and from MSA is the choice of functional verbs that indicates aspectual or modal functions in the sentence. These verbs include those expressing will and desire or verbs such as do and make. For example, in MSA *juri:d* < يريد > , *jawaddu* < يَوَد >,

and *jabɣi*: < يَبغي > can all be used to mean 'he wants' as a verb of will and desire, but in KA, the verb used is always jabi: < يَبي > which is shortened from *jabɣi*: by omitting the consonant (ɣ).

In other closely related Gulf dialects, the same verb is used with slight phonological changes that help listeners to differentiate one dialect from another, for example; in the Emirates, it is pronounced as *jiba*: < يبا >[18], in Bahraini Arabic *jabbi* < يبّي > with geminated /b/, in Hijazi Arabic, it may be jiba < يبا > or *jabɣa* < يبغى >. In Omani Arabic, they use the participle form of *jabɣi*:, which is *baːɣi*: < باغي > or its alternative *baːja*[19] 'wanting'. Other dialects would use different forms altogether, such as *widd* < ود > and its variant *bidd* < بد > that are used in the Levantine dialects or j*riːd* < يريد > used in some Iraqi and Gulf dialects.

The functional verb do in MSA is either derived from the root √FʕL < فعل >, √ṢNʕ < صنع > or √ʕML < عمل >. In KA, the verb used to indicate the functional sense of DO is derived from a completely different root √SWJ < سوى >, which means 'fix' or 'align' in Arabic. Other dialects use different roots, such as *ʕmil* < عمل > 'work' as in Egyptian or *daːr* < دار > 'turn or go round' as in Moroccan with perfective and participle forms. These verbs are essential indicators of ADs.

2.4. Distinctive Kuwaiti Vocabulary and Lexical Items

Several lexical items are distinctive to KA or shared by KA with other Gulf Arabic dialects. These words are highly frequent as evident from the data collected in this study (discussed in Section 4). These words are either adverbs, intensifiers, or answer particles. They can be simple words or even phrases. We will discuss those that are highly frequent and clearly distinctive of KA.

The first group are adverbs, which are either temporal adverbs such as *ʔalħiːn* < الحين > 'now', *ħazzah* < حزّة > 'moment,' and *baːʧir* < باچر > 'tomorrow,' locatives *ħadir* < حَدِر > 'under' and *siːdah* < سيدة > 'straight ahead' and adverbs of manner, such as *zain* < زين > 'well' and *killiʃ* < كلِّش > 'at all,' which are very frequent in KA data. Second, KA uses a small number of distinct intensifiers, such as *ħadd* < حدّ > 'extremely,' *waːjid* < وايد > 'a lot,' and ħeil < حيل > 'intensively.' Out of these three words, only *ħadd* shows agreement with the subject:

(13) **ħad**-ha: ʕadʒiːbah
extreme-2SF amazing.F
'She's so amazing'

(14)	**wa:jid**	ħilwah	ha-l-iɣnija
	a lot	beautiful	this-the-song
	'This song is so beautiful'		
(15)	nafnu:f-ha:	**ħeil**	ðˤajjidʒ
	dress-her	intensive	tight
	'Her dress is so tight'		

Second, answering in affirmation in Arabic is achieved using particles such as *ʔi*: < إي >, *naʕam* < نعم >, *bla*: < بلى > and *ʔadʒal* < أجل > meaning 'yes.' In KA, these forms are used in addition to another distinctive form, which is *ʔimbala*: < امبلا > . Furthermore, ʔadʒal has another form as shown in the following example:

(16)

- killi-na:	bi-nru:ħ	l- ħadi:qa
All-us	FUT-1P.go	the-garden
'we are all going to the garden'		
- ʕayal	b-aru:ħ	maʕa:-kum
AFF	FUT-1S.GO	with-you

'Then I will go with you'

Finally, some of the most frequent words found in the dataset include ham < هم > 'also,' *xwʃ* < خوش > 'good' – which is a borrowed word – and *killiʃ* < كِلِّش > 'not at all,' which is a phrase used either as negation or intensifier.

These are the main phonological, morphological, syntactic, and lexical features and elements that help distinguish KA from MSA or from other AD. These features are attested in the written data and should be incorporated into any morphological analyzer that aims at analyzing written KA. These features were collected to design the KA-specific extension for MADAMIRA[20] morphological analyzer, as shall be explained in the following methodology section. **3. Methodology**

3.1. Designing the morphological analyzer for KA

In this research, we aimed at developing a morphological analyzer that can account for written KA data. Considering that KA data is not clearly distinguishable from MSA,

especially in the written form, we aimed at expanding a morphological analyzer that was built for MSA to cover KA. The main motivation behind working on expanding an existing morphological analyzer, rather than creating a new one from scratch, is the assumption that first, KA shares many vocabulary and morphological and orthographic features with MSA, and second, most users tend to code-switch between MSA and KA which is a diglossic situation.(21) The choice was set on the Morphological Analysis and Disambiguation tool of Arabic (MADAMIRA) to accommodate entries of KA text and provide a linguistic analysis for them. MADAMIRA is an Arabic morphological analyser that uses natural language processing (NLP) techniques to analyse, and segment given Arabic text into its constituent morphemes. The main goal of MADAMIRA is to provide an accurate analysis of Arabic text, including its diacritics, stem, and affixes. This information is crucial for many NLP tasks as it can help to disambiguate the meaning of words and improve the accuracy of tasks like machine translation, part-of-speech tagging, and named entity recognition.

The following is an explanation of how MADAMIRA works:

1. Tokenization: The input text is tokenized, which means that it is divided into individual words or tokens.

2. Diacritisation: The tokens are diacritised by adding the unwritten short vowels. This step is necessary because Arabic text is typically written without short vowels, making it difficult to analyse the text accurately.

3. Lemmatisation: Each token is lemmatised, which means that it is reduced to its root form. For example, the word *ktb* < كتب > (books) is reduced to its root *ktAb* < كتاب > (book).

4. Morphological analysis: The lemmatised tokens are analysed morphologically to determine their stem, root, and affixes. This involves applying various morphological rules and patterns to the tokens to identify their parts of speech, verb conjugation, and noun declension.

5. Disambiguation: Finally, the output of the morphological analysis is disambiguated to resolve any ambiguity that may arise from the complex nature of the Arabic language. This involves selecting the most likely interpretation of each token based on its context.

Morphological analysis is a crucial stage for most text processing applications. It provides syntactic tags, missing diacritics, tokens, and other elements, for the input text preparing it for further process. MADAMIRA is a toolkit designed to provide such linguistic information. What sets MADAMIRA apart from similar tools is that it takes word context into account, which makes the generated analysis more accurate.

For the purpose of this study, approximately 100,000 location-specific tweets were collected using Twitter API. Only tweets originating from Kuwait were collected. To limit the search to Kuwaiti users, a set of common hashtags was used in the search, specifically those related to Kuwaiti parliament elections such as < #مرزوق-ارحل > `marzouq_leave`. Those we believe are not of interest to non-Kuwaiti people in Kuwait.[22] The period of the search was January 2019. The following steps followed the data collection process.

3.2. Pre-processing Data

Pre-processing the data is an essential step that needs to be carried out before the application of any morphological analysis. Textual data – especially that in social media – includes unnecessary information, such as emojis, URLs, contact details, and so forth. Those non-linguistic instances need to be removed initially.

3.3. Dealing with Variation

This involves converting the Arabic text into standard form with the least variation. This is done by unifying characters that have similar shapes but different Unicode value depending on their position in the word (such as hamza letter, ya:, and ta: marbuta). Moreover, it is observed that social media users typically use repeated letters to express exaggeration about something. Take the examples: < حدددده > 'extremely' and < واااايد > 'a lot.' These words should be restored to their original form by deleting repeated letters. In order to pre-process the text and unify the orthographic representation, a program was written in Python.

KA does not have a standard orthography system. The lack of orthographic guidelines results in variation in the written text. Also, we do not expect all writers to master Arabic writing rules, especially the hamza rules known to be complex. To control these types of variations, we added some rules that allow certain variations in letters. Examples of these rules are:

(1) Gliding hamza: e.g., [qA}l][23] < قائل > → [gAyl] < قايل > / [fA}dp] < فائدة > → [fAydp] < فايدة >

(2) Deleting the word final hamza e.g., [$ay'] < شَيء > → [$ay] < شَي >

(3) Substitutions between different forms of hamza e.g., [ymtl}] < يمتلئ > → [ymtl >] < يمتلأ >

3.4. Expanding MADAMIRA

The expansion involved enriching the analyzer with the necessary dictionary of orthographic, morphological, syntactic, and lexical items representative of KA. These were generalizations based first on our linguistic judgments as specialists of the Arabic

language and natural speakers of KA, and second, on a thorough examination of the errors from running the original MADAMIRA on the tweets. An initial re-examination of errors from the first run showed a number of unprocessed data that repeatedly appeared within KA tweets; these errors were not mere spelling mistakes but appear to be written commonalities that may raise to the level of conventions amongst KA Twitter users, which we discuss in detail in the results (Section 4).

Furthermore, MADAMIRA requires a SAMA style set of prefix, stem, and suffix dictionaries. SAMA[24] (Standard Arabic Morphological Analyzer) is a software tool for the morphological analysis of Standard Arabic. In our current project, we worked on expanding the coverage of SAMA dictionaries to include KA nominal and verbal prefixes and suffixes. That is because KA hosts some extra prefixes and suffixes, as discussed earlier in Section (3).

Moreover, a total of 3600 KA lemmas were added to the dictionary along with their MSA equivalents.[25] The KA lemmas were collected from two sources. The first source is the encyclopedia of Kuwaiti dialect[26], from which we extracted 3400 words. The second resource is Twitter[27]; we collected many tweets that were posted from Kuwait, and then we created a Python script that works on extracting words and listing them according to the number of times they appear in the extracted tweets. After removing functional words and MSA words, 200 Kuwaiti words with high occurrence rates were extracted and added to the SAMA stem dictionary[28] (see Table III below for the most frequent words). In addition to the KA lemmas, we added some compound words to the lexicon.

Table III: Most frequent KA words from Twitter.

item	count	item	count	item	count	item	count	item	count	item	count
اي	2091	حيل	1710	حده	1580	صجك	1376	اللي	1251	حسافة	1040
شنو	2070	شلون	1703	تخلينا	1568	اشوه	1367	هذي	1250	شريت	1006
احنا	1987	كيفك	1699	وايد	1567	شدراك	1349	خلصت	1247	مستانس	986
هم	1877	ليش	1698	ماكو	1561	عشان	1337	ينقال	1230	شدعوه	985
مافي	1876	مسوين	1656	بس	1558	اوكي	1323	مالتي	1187	متوهق	981
جذي	1861	علشان	1650	ليا	1534	شوي	1315	يبونها	1179	سوي	978
نروح	1740	نبي	1627	هذيل	1524	دق	1309	خلاص	1168	السالفة	977
نشوف	1738	جدام	1618	وينج	1496	حطي	1297	انتي	1103	عفيه	965
خوش	1738	منو	1608	توهقت	1465	بغيت	1254	كلش	1101	يسكر	943
مو	1719	وينك	1598	مبجر	1425	بنطر	1251	يازينه	1045	قفلوا	937

Multiple levels of quality checks were performed on the output of each step in the creation process to improve the coverage of the extended analyzer. The steps of the methodology are summarized in the following figure:

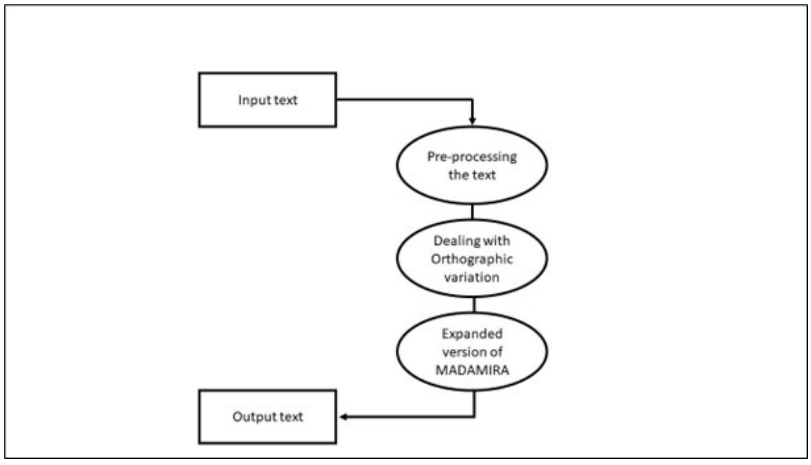

Figure 1: Methodology in steps

4. Results and Discussion

The expanded version of MADAMIRA – which we call MADAMIRA-KA – was tested by analyzing a total of 100,000 tweets written in KA. The raw data of the tweets were retrieved using Twitter Premium API in JSON format. The query used to retrieve the data from the API was based on keywords, hashtags, and account name mentions.(29)

The original version of MADAMIRA failed to produce an analysis for 29.9% of the words. We tested the expanded version MADAMIRA-KA with the same data and followed this with many rounds of quality checks and error analysis to determine the gaps in the system and the areas of weakness. After several modifications to the analyzer, with several reruns, words with no-analysis dropped to only 11.3% of total words. This is a substantial improvement in data analysis.

Following the initial test of the original MADAMIRA on KA data, many modifications had to be added to the tool, which resulted in the modified version. These modifications included the addition of several phonological, morphological, and lexical elements distinctive of KA, which we discussed earlier in Section 2. In the written data, we needed to address the following cases by adding them as rules to the analyzer:

(1) Words including the grapheme [k] < ك > may be substituted by the following graphemes [J] < چ > or [j] < ج >, especially if it was the feminine singular 3rd person pronoun such as in the following examples: [klb] < كلب > = [Jlb] < چلب > 'dog,' [qlmk] < قلمك > = [qlmJ] < قلمچ > 'your pen,' and [Endk] < عندك > = [EndJ] < عندچ > 'you have.' (ʧ)???

(2) Words that include the grapheme [y] < ي > instead of [j] < ج > such as [yAkm] < ياكم > 'he came to you' instead of [jAkm] < جاكم > or [yry$] < يريش > 'groats' instead of [jry$] < جريش >.

(3) Words that are pronounced with [g] instead of [q] are written with the letter [q] < ق > such as [qlb] < قلب > 'heart' or [sAq] < ساق > 'leg', whereas words that are pronounced with [j] are found written with the letter [j] < ج > instead of [q] < ق > such as [jdAm] < جدام > 'in front of' or [jAbl] < جابل > 'faced.'

(4) Words that have the grapheme [D] < ض > in MSA are usually substituted by [Z] < ظ > instead, such as [ZAbT] < ظابط > 'officer' and [ZfdE] < ظفدع > 'frog.'

(5) Instances of vowel epenthesis in KA tend to appear with an additional grapheme [A] < ا > at the beginning of the word, such as with the following examples from our data: [ASxlh] < اصخله > 'goat' and [AHTbh] < احطبه > 'brick.'

(6) Instances of hamza deletion include deletion of hamza that is part of the definitive particle [Al] < ال > such as in the following example: [$HlAt lbywt] < شحلات لبيوت > 'what lovely houses!' They also happen with imperatives such as [drswA] < درسوا > 'study you all' [lEbwA] < لعبوا > 'play you all' instead of how it is written in MSA < ادرسوا – العبوا >.

(7) Pronouns that have a different written form when compared to MSA such as indicated in the following table:

Table IV: Personal pronouns and demonstratives as spelled in written KA

Pronoun	Syllable structure	Spelling
1st person, singular	/ ʔa:-na/	[AnA] آنا
1st person, plural	/ʔiħ-na/	[AhnA] احنا
2nd person, plural	/ʔin-taw/	[Antw] انتو
3rd person, singular, masculine	/ʔu-hu/	[Ahw] اهو
3rd person, singular, feminine	/ʔi-hi/	[Ahy] اهي
3rd person, plural	/ʔu-hum/	[Ahm] اهم
Demonstrative, singular, male, distant	/ha-ða:k/	[h*Ak] هذاك
Demonstrative, singular, female, close	/ha:-ði/	[h*y] هذي
Demonstrative, singular, female, distant	/ha-ði:ʧ/	[h*yJ] هذيج
Demonstrative, plural (male & female), close	/ha-ðu:l/	[h*wl] هذول
Demonstrative, plural (male & female), distant	/ha-ðu:-la:k/ /ha-ði:-la:k/	[h*wlAk] هذولاك [h*ylAk] هذيلاك

The final step was to manually inspect the analyzer's output to find out the main factors for not producing an analysis for the remaining words. We can categorize the non-analyzed words to the following points:

1- Non-Arabic words, such as: [AwnlAyn] < اونلاين > 'on line,' [lAyk] < لايك > 'like,' [lwk] < لوك > 'look,' and [Awky] < اوكي > 'OK')

2- Noncovered KA words, such as: [xrbwTp] < خربوطة >, [mzhbaP] < مزهبة >

3- Pause fillers: [AAAAh] < اااااه >, [hmmmm] < هممم >.

4- Named entities such as proper names: [saEdyap mfrH] < سعدية مفرح > and places [Hwly] < حولي >

5- Words from other dialects such as the Egyptian word [buS] < بُص > 'look.'

6- Mis-spelt words (typo or KA writing system) < شفتلك > < قلتلك >

As for the first point, these words can be added separately to the analyzer as borrowed words into MSA and not just for KA. The second point can instantly be improved by adding more Kuwaiti lemmas to the KA dictionary that have not yet been included. Pause-fillers need to be introduced as a separate linguistic class of words that are not included in the typical MSA or KA dictionaries. Furthermore, the problem of the proper names can be solved by adding a dictionary of named entities. For this work, Arabic Named Entity Gazetteer[30] was used to extract proper names and introduce them as Nprop within the SAMA dictionary. The final point is a positive output because we intended for the analyzer to analyze only KA data, alongside MSA.

The expanded version of MADAMIRA has shown substantial results in processing Kuwaiti tweets.[31] Furthermore, the results have shown that KA has written conventions 'unconsciously' standardized amongst KA users. It relies heavily on the conventions of written MSA with some additional features discussed above. Furthermore, enriching the morphological analyzer with 3600 KA lemmas and 200 of the most frequent KA vocabulary items has proven to be important in the successful function of the analyzer. Finally, in relation to other morphological analyzers for dialectal Arabic (DA), this analyzer certainly fills a gap in the field since it is the first of its kind dedicated to KA, an Arabic dialect notably distinct from other dialects.

5. Conclusion

The current study presented a detailed linguistic description of written Kuwaiti Arabic. The characteristics of KA were extracted from examining more than 100,000 Kuwaiti tweets and finding consistencies in the way KA users reflected their dialect. This showed that there are many characteristics on every linguistic level that can set KA

apart from other Arabic dialects. Despite that the users appear to adhere to MSA orthographic conventions, there are some areas where KA stands out, especially in the use of the connective hamza, choice of consonants, and the vocabulary used.

Another critical contribution of this study is the extension and improvement of a morphological analyzer dedicated to KA texts. MADAMIRA-KA achieved excellent results in the analysis of KA data. The improvement is owed greatly to the incorporation between linguistic description and computational programming. Without the linguistically described input, many of the results would have come across as unanalyzed errors. We anticipate that the expanded analyzer will be a useful tool in developing most NLP applications for KA. For example, a morphological analyzer is necessary for machine translation, especially one that can translate different varieties of Arabic in addition to MSA. Other applications include part of speech tagging, sentiment analysis, information retrieval, and speech recognition systems. All these applications can be significantly improved once they are capable of analysing KA as well as MSA.

6. Notes and References:

(1) Holes, *Modern Arabic: Structures, Functions, and Varieties* (2004), p. 2.

(2) Habash, Diab and Rambow, "Conventional Orthography for Dialectal Arabic", *LREC,* 2012, pp. 711–8.

(3) A note on transcription and transliteration: The examples provided in this paper are transcribed in IPA – taking the recommendations of our anonymous reviewers - followed by the way they are written in Arabic graphemes, to clearly show how they are pronounced. In some instances, especially in the methodology section, a different transliteration system is used to show how the Arabic examples are written in the morphological analyser which applies the Buckwalter transliteration scheme [Habash, Soudi, and Buckwalter, "On Arabic Transliteration." *In Arabic Computational Morphology* (2007), pp. 15-22]. The Buckwalter transliteration scheme substitutes the arabic grapheme for a Latin grapheme, hence when a short vowel (diacritic) is not written in the Arabic example, the system will not compensate for it and vice versa. Using this transliteration shows directly one of the difficulties that are faced in developing NLP systems to deal with written Arabic. Finally, the IJMES transliteration system is used for the Arabic references in the endnotes and bibliography following the journals requirements.

(4) Matar, *Khaṣa'iṣ Al-lahja Al-kuwaitia Dirasa Lughawia Maidania* [Charactaristics of Kuwaiti Dialect: A Field Linguistic Study] (1969); Brustad, *The Syntax of Spoken Arabic: A Comparative Study of Moroccan, Egyptian, Syrian, and Kuwaiti Dialects* (2000); Alghunaim, Yaqoub, *'alfāẓ al-lahja al-kuwaitia fi lisān al-'arab* [Vocabulary of Kuwaiti Dialect in lisanu' Al Arab Dictionary] (2004); Al-Rashed*, Encyclopedia of Kuwaiti Dialect*, (2011); Al-Bahri, *A Grammar of Hadari Arabic: A Contrastive-typological Perspective*. PhD Diss. (2014); AlBader, *Semantic Innovation and Change in Kuwaiti Arabic: A Study of the Polysemy of Verbs*. PhD diss. (2015); Al-Fhaid, *al-lahja al-kuwaitia fi al-rub' al-thālith min al-qarn al-'ishrīn: dirasa ṣawtiyya ṣarfiyya* [Kuwaiti Dialect in the Third Quarter of the Twentieth Century: a Morpho-Phonological Study], MA diss. (2015).

(5) Alsharhan and Ramsay. "The Development of a Speech Corpus Annotated for the Main Arabic Dialects". *Arab Journal for the Humanities,* 2020, p. 155.

(6) Beesley. "Computer Analysis of Arabic Morphology: A Two-level Approach with Detours", in Third *Annual Symposium on Arabic Linguistics* 1989, pp. 155-172; Buckwalter, "Buckwalter Arabic Morphological Analyser (BAMA) version 2.0", *Linguistic Data Consortium LDC*, 2004; Roth et al.,"Arabic Morphological Tagging, Diacritization, and Lemmatization Using Lexeme Models and Feature Ranking". *In Proceedings of ACL-08: HLT short papers*, 2008, pp. 117-120; Graff et al.,"Standard Arabic Morphological Analyser (SAMA) version 3.1", *Linguistic Data Consortium* 2009, pp. 53-56; Pasha et al.,"MADAMIRA: A Fast, Comprehensive Tool for Morphological Analysis and Disambiguation of Arabic", *LREC* 14, 2014, pp. 1094-1101; Abdelali et al., "Farasa: A Fast and Furious Segmenter for Arabic", *in Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations,* 2016, pp. 11-16; and Boudchiche et al.,"Al-Khalil Morpho Sys 2: A Robust Arabic Morpho-syntactic Analyser", *Journal of King Saud University-Computer and Information Sciences,* 29.2 (2017), pp. 141-146.

(7) Habash and Rambow. "MAGEAD: A Morphological Analyser and Generator for the Arabic Dialects". *The 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics,* 2006, pp. 681-688; and Pasha et al.,"MADAMIRA".

(8) Examples of efforts toward building resources for these two dialects are: [Al-Sabbagh and Girju, "A Supervised POS Tagger for Written Arabic Social Networking Corpora", *KONVENS* (2012), pp. 39-52], [Habash, Eskander, and Hawwari, "A morphological Analyser for Egyptian Arabic", The *Twelfth Meeting of the Special Interest Group on Computational Morphology and Phonology* (2012), pp. 1-9], [Maamouri et al.,*Egyptian Arabic Treebank DF Parts 1-8,V2.0,* 2012]; [Habash et al.,"Morphological analysis and disambiguation for dialectal Arabic", *The 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (2013), pp. 426-432]; and [Eskander et al., "Creating Resources for Dialectal Arabic from a Single Annotation: A Case Study on Egyptian and Levantine", *in the 26th International Conference on Computational Linguistics*: Technical Papers (2016), pp. 3455-3465]. However, other dialects have received less attention ([Al-Shargi et al.,"Morphologically Annotated Corpora and Morphological Analysers for Moroccan and Sanāni Yemeni Arabic", in *10th Language Resources and Evaluation Conference*, 2016, pp. 1300 – 1306; Harrat et al.,"Building Resources for Algerian Arabic Dialects", *15th Annual Conference of the International Communication Association Interspeech* (2014); and Al-Twairesh et al.,"Suar: Towards Building a Corpus for the Saudi Dialect". *Procedia Computer Science* 142, 2018, pp. 72-82]).

(9) Bouamor, Houda, Nizar Habash, Mohammad Salameh, Wajdi Zaghouani, Owen Rambow, Dana Abdulrahim, Ossama Obeid et al. "The MADAR Arabic Dialect Corpus and Lexicon." In *LREC*. 2018.

(10) Alotaibi, *Event Phrase and the Syntax of TMA Verbs in Kuwaiti Arabic.* PhD diss. 2019] for the different functions of */kān/ and /chān/.*

(11) Al-Qenaie, *Kuwaiti Arabic: A Socio-phonological Perspective*. PhD diss, 2011.

(12) Ingham. *Northeast Arabian dialects* 1982, p. 37; and De Jong. "Gahawa-Syndrome". *Encyclopedia of Arabic Language and Linguistics 2* (2006), p. 151.

(13) Holes, "Kuwaiti Arabic". *Encyclopaedia of Arabic Language and Linguistics pp,608-620*.

(14)

(15) Al-Bahri, Khaled, A *Grammar of Hadari Arabic: A Contrastive-typological Perspective*, PhD Dissertation (2014), p. 72.

(16) As one of our reviewers indicated, there is a relatively extensive discussion dating to 1900 concerning the derivation of preverbal b-. In fact, it has been called one of the most widely debated topics in Arabic dialectology. Examples of such references are [Kampffmeyer, Georg, Die arabische Verbalpartikel b (m). Beitrage zur Dialektologie des Arabischen II. *Mitteilungen des Seminars fiir Orientalische Sprachen zu Berlin, zweite Abteilung*: (1900), p. 50.], [Stewart, Devin J." Clitic Reduction in the Formation of Modal Prefixes in the Post- Classical Arabic Dialects and Classical Arabic Sa-/Sawfa". Arabica 45(1), 1998, p. 109.], [Eksell, Kerstin, "The origin and development of the cursive b-imperfect in Syrian". In Arabic. K. Eksell and T. Vinther (eds.), Change in Verbal Systems: Issues on Explanation. (2006), p. 75.], [Persson, Maria, "The Role of the b-prefix in Gulf Arabic Dialects as a Marker of Future, Intent and/or Irrealis". *Journal of Arabic and Islamic Studies* 8, (2008), p. 26.], [Retsö, Jan, "The bi-imperfect once again: Typological and diachronic perspectives". In Lutz Edzard and John Huehnergard (eds.) Proceedings of the Oslo-Austin Workshop in Semitic Linguistics. (2014), pp. 64], [Davey, Richard J, Coastal Dhofārī Arabic: A sketch grammar. (2016), p250], [Jarad, Nabil Ismail, "Grammaticalization in Emirati Arabic". Arabica, 64, (2017), p. 750], [Lentin, Jérôme, "The Levant". In Clive Holes (ed.) Arabic Historical Dialectology. (2018), p. 170], [Owens, Jonathan, "Dialects (speech communities), the apparent past, and grammaticalization: Towards an understanding of the history of Arabic". In Clive Holes, ed., Arabic Historical Dialectology. (2018), pp. 206] and [Bettega, Simone. *Tense, Modality and Aspect in Omani Arabic*. (2019). P. 139].

(17) Al-Qenai. Kuwaiti Arabic, p.106.

(18) As noted by an anonymous reviewer, In Emirati Arabic it may either by yibā, yabi, or yibġā, see [Leung, T, D. Ntelitheos, and M. Al Kaabi. *Emirati Arabic: A comprehensive grammar.* (2020), p227]. Also see [Qafisheh, H, A *Short Reference Grammar of Gulf Arabic*. (1977)] for examples of other Gulf Arabic dialects.

(19) From personal communication with Assistant professor of Linguistics and a native speaker of Omani Dr Suaad Ambu-Saidi.

(20) MADAMIRA is a state-of-the-art tool that produces a rich output. The tool produces a list of analyses for each word in each sentence. The analysis ranking component then scores each word analysis list based on how well each analysis agrees with the model predictions and then sorts the analyses based on that score. A non-commercial license of MADAMIRA is freely available at: www.innovation.columbia.edu/technologies/CU14012.

(21) Alruwayeh. *Diglossic Code-switching in Kuwaiti Newspapers*. PhD diss. (2016).

(22) The set of hashtags used can be shared publicly upon request from the authors.

(23) As indicated in endnote 3, in this methodology section, we use the Buckwalter transliteration system to show how exactly these words are added to the analyser. The Latin letters are put

between square brackets [ ], and Arabic equivalent between less and grater than signs < >.

(24) SAMA is an updated version of Buckwalter Arabic Morphological Analyzer (BAMA). SAMA analyzes each Arabic word token by providing all possible prefix-stem-suffix segmentations and lists all possible annotation solutions, with the assignment of all diacritic marks, morpheme boundaries, and all Part-of-Speech (POS) tags. The choice is then left to users to select the most appropriate annotation among the generated output. Accessing this tool is exclusively available to LDC members through this link: https://catalog.ldc.upenn.edu/LDC2010L01.

(25) The set of lemmas extracted for the purpose of this study can be publicly provided by the authors upon request.

(26) Al-Rashed. Encyclopaedia of KA.

(27) We thank Dr. Salah Alnajim for providing us with the data (tweets) needed to evaluate the developed tool.

(28) Here is an example for mapping verb شيل، يشيل 'carry'

```
;--- $Al
;; $Al_1
$Al $Al      PV          carried
$yl $yl      IV_no-Pref-l      carry
$yl $yl      IV_need-Pref-l    carry
An$Al        Ain$Al   PV_Pass  be carried
```

(29) https://developer.twitter.com/en/premium-apis

(30) Arabic Named Entity Gazetteer is an Arabic "fine-grained" gazetteer that was automatically compiled from the Arabic Wikipedia [Alotaibi and Lee, "Automatically Developing a Fine-grained Arabic Named Entity Corpus and Gazetteer by Utilizing Wikipedia", in *Proceedings of the Sixth International Joint Conference on Natural Language Processing* (2013), pp. 392-400].

(31) MADAMIRA-KA was applied to other projects conducted by the authors on other sets of KA tweets with great results in processing KA data (see [Alsharhan and Alotaibi, "The Development of Efficient Transcription System for Kuwaiti Broadcast news and conversational speech", *Arab Journal for the Humanities*, 2021, p 333]).

Bibliography

Abdelali, Ahmed; Kareem Darwish; Nadir Durrani; and Hamdy Mubarak, "*Farasa: A Fast and Furious Segmenter for Arabic*", presented at *2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations* in San Diego, California, Jun. 2016.

AlBader, Yousuf, *Semantic Innovation and Change in Kuwaiti Arabic: A Study of the Polysemy of Verbs*, PhD dissertation (University of Sheffield, 2015).

Al-Bahri, Khaled, *A Grammar of Hadari Arabic: A Contrastive-typological Perspective*, PhD Dissertation (University of Sussex, 2014).

Al-Fhaid, Abdullah, *Al-lahja al-kuwaitia fi al-rubʿ al-thālith min al-qarn al-ʿishrīn: dirasa ṣawtiyya ṣarfiłyya* [Kuwaiti Dialect in the Third Quarter of the Twntieth Century: a Morpho-Phonological Study], MA Dissertation (Alshariqa University, 2015)

Alghunaim, Yaqoub, *ʾalfāẓ al-lahja al-kuwaitia fī lisān al-ʿarab* [Vocabulary of Kuwaiti Dialect in Lisanu' Al Arab Dictionary] (Kuwait: Centre of research and Kuwaiti studies, 2004).

Alotaibi, Bashayer, *Event Phrase and the Syntax of TMA Verbs in Kuwaiti Arabic*, PhD dissertation (Newcastle University(2019 ,.

Alotaibi, Fahd, and Mark Lee, "Automatically Developing a Fine-grained Arabic Named Entity Corpus and Gazetteer by Utilizing Wikipedia", presented at *Sixth International Joint Conference on Natural Language Processing* in Nagoya, Japan, Oct. 2013.

Al-Qenaie, Shamlan, *Kuwaiti Arabic: A Socio-phonological Perspective*, PhD dissertation (Durham University, 2011).

Al-Rashed, Khaled, *Mawsūʿat al-lahja al-kuwaitiyya* [Encyclopedia of Kuwaiti Dialect], (2011).

Alruwayeh, Marwah, *Diglossic Code-switching in Kuwaiti Newspapers*, PhD dissertation (Newcastle University, 2016).

Al-Sabbagh, Rania, and Roxana Girju. "A Supervised POS Tagger for Written Arabic Social Networking Corpora", presented at *11th Conference on Natural Language Processing KONVENS 2012* in Vienna, 19 Sept. 2012.

Al-Shargi, Faisal; Aidan Kaplan; Ramy Eskander; Nizar Habash; and Owen Rambow, "Morphologically Annotated Corpora and Morphological Analysers for Moroccan and Sanāni Yemeni Arabic", presented in *10th Language Resources and Evaluation Conference,* Portorož (Slovenia): 23-28 May 20166.

Alsharhan, Eiman, and Allan Ramsay. "The Development of a Speech Corpus Annotated for the Main Arabic Dialects", *Arab Journal for the Humanities*, 2020, pp 155.

Alsharhan, Eiman, and Bashayer Alotaibi. "The Development of Efficient Transcription System for Kuwaiti Broadcast news and conversational speech", *Arab Journal for the Humanities*, 2021, pp 329-348.

Al-Twairesh, Nora; 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. "Suar: Towards Building a Corpus for the Saudi Dialect". *Procedia Computer Science* 142 (2018): pp. 72–82.

Beesley, Kenneth. "Computer Analysis of Arabic Morphology: A Two-level Approach with Detours", presented in *Third Annual Symposium on Arabic Linguistics* in Salt Lake City, University of Utah, 1989.

Bettega, Simone. *Tense*, *Modality and Aspect in Omani Arabic*. (Naples: Università degli Studi di Napoli "L'Orientale," Dipartimento di Asia, Africa e Mediterraneo. Series Minor XCI. 2019).

Bouamor, Houda, Nizar Habash, Mohammad Salameh, Wajdi Zaghouani, Owen Rambow, Dana Abdulrahim, Ossama Obeid et al. "The MADAR Arabic Dialect Corpus and Lexicon." *LREC*. 2018.

Boudchiche, Mohamed; Azzeddine Mazroui; Mohamed Ould Abdallahi Ould Bebah; Abdelhak Lakhouaja; and Abderrahim Boudlal. "Al-Khalil Morpho Sys 2: A Robust Arabic Morpho-syntactic Analyser." *Journal of King Saud University-Computer and Information Sciences* 29 (2017): pp. 141-146.

Brustad, Kristen. *The Syntax of Spoken Arabic: A Comparative Study of Moroccan, Egyptian, Syrian, and Kuwaiti Dialects* (Georgetown University Press, 2000:

Buckwalter, Tim. *Buckwalter Arabic Morphological Analyser (BAMA) version 2.0*, Linguistic Data Consortium LDC*,* University of Pennsylvania, Philadelphia, 2004).

Davey, Richard J. *Coastal Dhofārī Arabic*: *A sketch grammar*. (Leiden: Brill. 2016).

De Jong, Rudolph. "Gahawa-Syndrome". *Encyclopedia of Arabic Language and Linguistics 2* (2006), pp. 151–3.

Eksell, Kerstin. "The origin and development of the cursive b-imperfect in Syrian." In Arabic. K. Eksell and T. Vinther (eds.), *Change in Verbal Systems*: *Issues on Explanation*. (Peter Lang, Frankfurt am Main, 2006), pp. 73-98.

Eskander, Ramy; Nizar Habash; Owen Rambow; and Arfath Pasha, "Creating Resources for Dialectal Arabic from a Single Annotation: A Case Study on Egyptian and Levantine", presented in *the 26th International Conference on Computational Linguistics: Technical Papers,* 2016.

Graff, David; Mohamed Maamouri; Basma Bouziri; Sondos Krouna; Seth Kulick; and Tim Buckwalter. "Standard Arabic Morphological Analyser (SAMA) version 3.1", *Linguistic Data Consortium LDC2009E73* (2009): pp. 53-56.

Habash, Nizar and Owen Rambow. "MAGEAD: A Morphological Analyser and Generator for the Arabic Dialects", presented in *The 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics,* 2006.

Habash, Nizar; Abdelhadi Soudi; and Timothy Buckwalter. "On Arabic Transliteration", In *Arabic Computational Morphology* (Springer, Dordrecht, 2007): pp. 15–22.

Habash, Nizar; Mona T. Diab; and Owen Rambow. "Conventional Orthography for Dialectal Arabic", *Language Resources and Evaluation Conference,* 2012.

Habash, Nizar; Ramy Eskander; and Abdelati Hawwari. "A morphological Analyser for Egyptian Arabic",presented in *The Twelfth Meeting of the Special Interest Group on Computational Morphology and Phonology,* 2012.

Habash, Nizar; Ryan Roth; Owen Rambow; Ramy Eskander; and Nadi Tomeh. "Morphological analysis and disambiguation for dialectal Arabic", *The 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies,* 2013.

Harrat, Salima; Karima Meftouh; Mourad Abbas; and Kamel Smaïli. "Building Resources for Algerian Arabic Dialects", presented in *15th Annual Conference of the International Communication Association Interspeech* in Singapore, 2014.

Holes, Clive. *Modern Arabic. Structures, functions, and varieties* (Georgetown University Press, 2004).

Holes, Clive. "Kuwaiti Arabic". in *Encyclopedia of Arabic Language and Linguistics 2* (2006).

Ingham, Bruce. *Northeast Arabian dialects* (London and Boston: Kegan Paul International, 1982).

Jarad, Nabil Ismail. "Grammaticalization in Emirati Arabic". *Arabica*, 64, 2017: pp: 742–760.

Kampffmeyer, Georg. Die arabische Verbalpartikel *b* (*m*). Beitrage zur Dialektologie des Arabischen II. *Mitteilungen des Seminars fiir Orientalische Sprachen zu Berlin, zweite Abteilung*: (Westasiatische Sprachen, Berlin/Stuttgart: W. Spemann, 1900), pp 48-101.

Lentin, Jérôme. "The Levant". In Clive Holes (ed.) *Arabic Historical Dialectology*. (Oxford: Oxford Uni-

versity Press, 2018), pp. 170-205.

Leung, Tommi Tsz-Cheung, Dimitrios Ntelitheos, and Meera Al Kaabi. *Emirati Arabic: A comprehensive grammar*. (New York: Routledge, 2020).

Matar, Abdulaziz, *Khasʿaʾisʿ Allahja Alkuwaitia Dirasa Lughawia Maidania* [Charactaristics of Kuwaiti Dialect: A Field Linguistic Study] (Kuwait: Alrisala Publishing, 1969).

Maamouri, Mohamed; Ann Bies; Seth Kulick; Dalila Tabessi; and Sondos Krouna. *Egyptian Arabic Treebank DF Parts 1-8 V2.0* (2012).

Owens, Jonathan. "Dialects (speech communities), the apparent past, and grammaticalization: Towards an understanding of the history of Arabic". In Clive Holes, ed., *Arabic Historical Dialectology*. (Oxford: Oxford University Press, 2018), pp. 206-256.

Pasha, Arfath; Mohamed Al-Badrashiny; Mona T. Diab; Ahmed El Kholy; Ramy Eskander; Nizar Habash; Manoj Pooleery; Owen Rambow; and Ryan Roth. "MADAMIRA: A Fast, Comprehensive Tool for Morphological Analysis and Disambiguation of Arabic", *Language Resources and Evaluation Conference,* 2014.

Persson, Maria. "The Role of the b-prefix in Gulf Arabic Dialects as a Marker of Future, Intent and/or Irrealis". *Journal of Arabic and Islamic Studies* 8, 2008. pp. 26–52.

Qafisheh, H. *A Short Reference Grammar of Gulf Arabic.* (Tucson, University of Arizona Press, 1977).

Retsö, Jan. "The *bi*-imperfect once again: Typological and diachronic perspectives". In Lutz Edzard and John Huehnergard (eds.) *Proceedings of the Oslo-Austin Workshop in Semitic Linguistics*. (Wiesbaden: Harrassowitz, 2014), pp. 64–72.

Roth, Ryan; Owen Rambow; Nizar Habash; Mona Diab; and Cynthia Rudin. "Arabic Morphological Tagging, Diacritization, and Lemmatization Using Lexeme Models and Feature Ranking", presented in *Proceedings of ACL-08: HLT short papers,* 2008. pp. 11.

Stewart, Devin J,. "Clitic Reduction in the Formation of Modal Prefixes in the Post- Classical Arabic Dialects and Classical Arabic Sa-/Sawfa". *Arabica* 45(1), 1998: pp. 104–128.