# A Morphological Analyzer for Gulf Arabic Verbs

**Salam Khalifa, Sara Hassan and Nizar Habash**
Computational Approaches to Modeling Language (CAMeL) Lab
New York University Abu Dhabi
{salamkhalifa,sah650,nizar.habash}@nyu.edu

## Abstract

We present CALIMA$_{\text{GLF}}$, a Gulf Arabic morphological analyzer currently covering over 2,600 verbal lemmas. We describe in detail the process of building the analyzer starting from phonetic dictionary entries to fully inflected orthographic paradigms and associated lexicon and orthographic variants. We evaluate the coverage of CALIMA$_{\text{GLF}}$ against Modern Standard Arabic and Egyptian Arabic analyzers on part of a Gulf Arabic novel. CALIMA$_{\text{GLF}}$ verb analysis token recall for identifying correct POS tag outperforms both the Modern Standard Arabic and Egyptian Arabic analyzers by over 27.4% and 16.9% absolute, respectively.

## 1 Introduction

Until recently, Dialectal Arabic (DA) was mainly spoken with little to no publicly available written content. Modern Standard Arabic (MSA) on the other hand is the official language in more than 20 countries, where most written documents from news articles, to educational materials and entertainment magazines, are written in MSA. Hence, most of the tools that are available for Natural Language Processing (NLP) tasks are focused on MSA. With the introduction of social media platforms online, dialectal written content is being produced abundantly. Using existing tools that were developed for MSA on DA proved to have limited performance (Habash and Rambow, 2006; Khalifa et al., 2016). Having resources specific to DA, such as morphological lexicons is important for Arabic NLP tasks, such as part-of-speech (POS) tagging and morphological disambiguation. Recently, dialects such as Egyptian (EGY) and Levantine (LEV) Arabic have been receiving increasing attention. Morphological analyzers for EGY and LEV proved to perform well when used for morphological tagging (Eskander et al., 2016). To our knowledge, there exist no full morphological analyzers for Gulf Arabic (GLF) that produce segmentation, POS analysis and lemmas. Although we note the work of Abuata and Al-Omari (2015) on developing a Gulf Arabic stemmer. In this paper, we present CALIMA$_{\text{GLF}}$,[1] a morphological analyzer for GLF. In the current work, we present the effort focusing on GLF verbs only. We utilize a combination of computational techniques in addition to explicit linguistic knowledge to create this resource. We also evaluate it against wide coverage tools for MSA and EGY. CALIMA$_{\text{GLF}}$ verb analysis token recall in terms of identifying correct POS tagging outperforms on both MSA and EGY by over 27.4% and 16.9% absolute, respectively. CALIMA$_{\text{GLF}}$ will be made publicly available to researchers working on Arabic and Arabic dialect NLP.[2]

The rest of this paper is organized as follows. In Section 2 we review related literature, then we briefly describe the main characteristics of GLF in Section 3. In Section 4 we describe the approach and the resources involved and evaluate in Section 5. We conclude and discuss future work in Section 6.

## 2 Related Work

### 2.1 Arabic Morphological Modeling

Much work has been done on Arabic morphological modeling, covering a wide range of different system designs. Earlier systems such as BAMA, SAMA and MAGEAD (Buckwalter, 2004; Graff

[1] In Arabic كلمة *kalimaħ* means 'Word'. We follow the naming convention from (Habash et al., 2012a) who developed CALIMA$_{\text{EGY}}$ since we are using the same format and analysis engine for the databases we create.

[2] CALIMA$_{\text{GLF}}$ can be obtained from http://camel.abudhabi.nyu.edu/resources/.



*Proceedings of The Third Arabic Natural Language Processing Workshop (WANLP)*, pages 35–45,
Valencia, Spain, April 3, 2017. 

et al., 2009; Habash and Rambow, 2006) were entirely manually designed. Similarly, Habash et al. (2012a) developed CALIMA, a morphological analyzer for Egyptian Arabic (hence $\text{CALIMA}_{\text{EGY}}$). $\text{CALIMA}_{\text{EGY}}$ was developed based on a lexicon of morphologically annotated data using several methods and then manually verified. Furthermore, Salloum and Habash (2011) extended existing SAMA and $\text{CALIMA}_{\text{EGY}}$ resources using hand crafted rules which extended affixes and clitics based on matching on existing ones. Recently, Eskander et al. (2013) developed a technique that generates a morphological analyzer based on an annotated corpus. They describe a technique in which they define inflectional classes for lexemes that represents morphosyntactic features in addition to inflected stems. They automatically 'complete' these classes in a process called paradigm completion. They also show that using manually annotated iconic inflectional classes helps in the overall performance. Using the aforementioned paradigm completion technique, a Moroccan Arabic and a Sanaani Yemeni Arabic morphological analyzers were created (Al-Shargi et al., 2016). And very recently Eskander et al. (2016) presented a single pipeline to produce a morphological analyzer and tagger from a single annotation of a corpus; they produced resources for EGY and LEV. Other works that involve DA morphological modeling include the work of Abuata and Al-Omari (2015). Who developed a rule-based system to segment affixes and clitics in GLF text. They compare their results to other well known MSA stemmers.

In this paper, we create morphological paradigms similar to the iconic inflectional classes discussed by Eskander et al. (2013). Our paradigms map from morphological features to fully inflected orthographic forms. The paradigms abstract over templatic roots; and lexical entries are specified in a lexicon as root-paradigm pairs, in a manner similar to the work of Habash and Rambow (2006). We convert the paradigms to the database representation used in MADAMIRA (Pasha et al., 2014) and $\text{CALIMA}_{\text{EGY}}$ (Habash et al., 2012a).

### 2.2 Dialectal Orthography

Due to the lack of standardized orthography guidelines for DA, and given the major differences from MSA, dialects are usually written in ways that reflects the words' pronunciation or etymological relation to MSA cognates (Habash et al., 2012b), and even then with a lot of inconsistency. Furthermore, as with MSA, Arabic orthography ignores the spelling of short vowel diacritics, thus increasing the ambiguity of the written forms. As a result, it is rather challenging to computationally process *raw* DA text directly from the source, or even agree on a common normalization. Habash et al. (2012b) proposed a Conventional Orthography for Dialectal Arabic (CODA) as part of a solution allowing different researchers to agree on a set of DA orthographic conventions for computational purposes. CODA was first defined for EGY, but has been extended to Palestinian, Tunisian, Algerian, Maghrebi and Gulf Arabic (Jarrar et al., 2014; Zribi et al., 2014; Saadane and Habash, 2015; Turki et al., 2016; Khalifa et al., 2016). We follow the conventions defined by Khalifa et al. (2016) for CODA GLF.

### 2.3 Dialectal Arabic Resources

In addition to the above mentioned morphological analyzers, there exist other resources such as dictionaries and corpora for both DA and MSA. For annotated MSA corpora, several developed such as (Maamouri and Cieri, 2002; Maamouri et al., 2004; Smrž and Hajič, 2006; Habash and Roth, 2009; Zaghouani et al., 2014).

Many efforts targeted DA, notably, EGY (Gadalla et al., 1997; Kilany et al., 2002; Al-Sabbagh and Girju, 2012; Maamouri et al., 2012b; Maamouri et al., 2012a; Maamouri et al., 2014). As for LEV, there exist morphologically annotated corpora and a treebank (Jarrar et al., 2014; Jarrar et al., 2016; Maamouri et al., 2006). Newly developed corpora for other dialects include (Masmoudi et al., 2014; Smaïli et al., 2014; Al-Shargi et al., 2016; Khalifa et al., 2016) for Tunisian, Algerian, Moroccan, Yemeni and Gulf Arabic respectively. Other notable efforts targeted multiple dialects such as the COLABA project, and the Tharwa dictionary (Diab et al., 2010; Diab et al., 2014). Parallel dialectal corpora by Bouamor et al. (2014) and Meftouh et al. (2015), in addition to the highly dialectal online commentary corpus by Zaidan and Callison-Burch (2011).

Specifically for GLF, we use the Qafisheh Gulf Arabic Dictionary (Qafisheh, 1997) as well as the Gumar Corpus (Khalifa et al., 2016) in developing our analyzer.

## 3 Gulf Arabic

### 3.1 Background

From a linguistic point of view, Gulf Arabic refers to the linguistic varieties spoken on the western coast of the Arabian Gulf, that is Bahrain, Qatar, and the seven Emirates of the United Arab Emirates, as well as in Kuwait and the eastern region of Saudi Arabia (Holes, 1990; Qafisheh, 1977). We extend the use of the term 'Gulf Arabic' (GLF) to include any Arabic variety spoken by the indigenous populations residing the six countries of the Gulf Cooperation Council. In this paper, we focus specifically on Emirati Arabic.

### 3.2 Orthography

Similar to other dialects, GLF has no standard orthography (Habash et al., 2012b). As such, words may be written in a manner reflecting their pronunciation or their etymological relationship to MSA cognates. For example the word for 'dawn' /al-fayr/ may be written as الفير *Alfyr*[3](reflecting pronunciation) or as الفجر *Alfjr* (reflecting its MSA cognate). In this work we follow the same CODA standards for GLF that were introduced by the authors in (Khalifa et al., 2016) extending the original CODA in (Habash et al., 2012b). We use CODA in developing the morphological databases; but we also add support for non-CODA variants and evaluate on raw non-CODA input. Another challenge caused by Arabic orthography in general (for MSA and other dialects including GLF) is that Arabic orthography does not require writing short vowel diacritics, which adds a lot of ambiguity.

### 3.3 Morphology

GLF shares many of the same morphological complexities of MSA and other Arabic dialects. Arabic rich morphology is represented templatically and affixationally with a number of attachable clitics. This representation in addition to the fact that short vowel diacritics are usually dropped in text add to the text's ambiguity. In comparison to MSA, EGY and LEV, GLF shares and differs in several aspects:

- Like MSA, but unlike EGY and LEV, GLF has no negation enclitic marker, namely the ش *iš* '[negation]' ending such as ما قلتش *mA qultiš* in EGY and LEV as opposed to ما قلت *mA qilt* in GLF 'I did not say'.

- Unlike MSA, but like EGY and LEV, GLF has an indirect object enclitic which is written separately in CODA (but not necessarily in raw form), e.g., قلتلك *qultlik* (CODA قلت لك *qult lik*) in LEV and قلتلج *qiltlij* (CODA قلت لج *qilt lij*) in GLF 'I told you[FS]'.

- GLF has different imperfect verb subject suffixes for second and third person plural and second person feminine singular from EGY and LEV, e.g. تقولوا *tquwlwA* in EGY and LEV and تقولون *tquwluwn* in GLF for 'you[P] say'; and تقولي *tquwliy* in EGY and LEV and تقولين *tquwliyn* in GLF for 'you[FS] say'. It is interesting to note that both forms exist in MSA where they indicate different moods.

- GLF shares with EGY and LEV the absence of the dual forms of the verb and imperfective moods, both of which are present in MSA.

- Unlike MSA, GLF shares with EGY and LEV the ambiguous forms of second masculine singular and first person perfective verbs, e.g., كتبت *katabt* 'I wrote or you wrote' in EGY, GLF and LEV; while MSA has *katabtu* 'I wrote' and *katabta* 'you wrote'.

- GLF has different second person singular direct object enclitics from EGY, LEV and MSA. The second masculine singular form in GLF ك *ik*, sounds like the second feminine singular form in EGY and LEV, and is differnt from MSA's ك *ka*; and the second feminine singular form in GLF ج *ij* (pronounced /itš/), is altogether different. For example, LEV شفتك *šuftik* maps to GLF شفتج *šiftij* 'I saw you[FS]'.

- The future verbal particle in GLF is ب *b* which is different from the MSA equivalent (س *sa*), and can be easily confused with the present progressive particle ب *b* in both EGY and LEV in. GLF does not have a progressive particle.

[3]All Arabic transliterations are provided in the Habash-Soudi-Buckwalter transliteration scheme (Habash et al., 2007).

## 4 Building CALIMA$_{GLF}$

### 4.1 General Approach

Our goal is to build a morphological analysis and generation model for GLF. We focus on verb forms in this paper, but plan to extend the work to other POS in the future. We employ two databases that capture the full morphological inflection space from lemmas and morphological features to fully inflected surface forms and in reverse. The two databases are (1) a collection of root-abstracted paradigms which map from features to root-abstracted stems, prefixes and suffixes; and (2) a lexicon specifying verbal entries in terms of roots and paradigm IDs. These two structures together define for any verb all the possible analyses allowed within GLF morphology. The two databases are then merged to create a full model. The merging can be done as a finite state machine. However, the implementation we chose is a variant based on the BAMA/SAMA databases following the representation used in MADAMIRA (Pasha et al., 2014) and CALIMA$_{EGY}$ (Habash et al., 2012a).

Next we discuss step by step the process we took to build CALIMA$_{GLF}$, starting with a phonetic dictionary all the way to building a fully functional morphological analyzer that even models non-CODA spelling variants.

### 4.2 The Qafisheh Gulf Arabic Verb Lexicon

Our starting point is the Qafisheh Gulf Arabic Verb Lexicon (QGAVL), which is a portion of the Qafisheh (1997) dictionary. Each entry in the lexicon includes a root, perfective and imperfective verb inflections, Verb Form (as in form II or VII) and English gloss. See Table 1 for some example entries. The Arabic entries are in a phonetic representation and not in Arabic script. The verb forms are only in third person masculine singular inflection (PV3MS and IV3MS, for perfective and imperfective aspect, respectively); and no clitics are attached. In total, there are 2,648 verb entries.

### 4.3 Orthographic Mapping

The first step we took was to create the orthographic spelling of all the verb entries. This included mapping to the appropriate vowel spelling as well as following the CODA spelling rules for stem consonants and morphemes. This step was first done automatically and then checked manually for every entry. See Table 2 for an example of the result of mapping the entries in Table 1. We mapped the roots in two ways, one following CODA and one reflecting a phonological spelling. This information will be used later to make the analyzer robust to non-CODA spellings.

### 4.4 PV-IV Pattern Extraction

Next, we identified for each verb its orthographic inflected templatic pattern, i.e., the pattern that would directly produce the surface form once the root radicals are inserted. This approach to pattern definition is most like the work of Eskander et al. (2013) in it being a one shot application of root-template merging to generate surface orthography. The approach differs from the work of Habash and Rambow (2006), who use a large number of rewrite rules for phonology, morphology and orthography after inserting the roots into the templates.

The pattern extraction was done automatically and then manually checked. It was only done to the forms available in the lexicon so far (PV3MS and IV3MS). The PV-IV Pattern (perfective-imperfective pattern) uses digits (e.g., 1,2,3,4,5) to represent root radicals. In this pattern, all vowels and glottal stop (Hamza) forms are explicitly spelled because they tend to vary within single paradigms. For example, the first entry in Table 5 specifies the PV-IV Pattern 1A3-y1uw3, which when merged with the root radicals *qwl* generates the perfective and imperfective forms *qAl* and *yquwl*.

### 4.5 Basic Paradigm Construction

We identified 72 unique PV-IV patterns in the lexicon, which represent 72 different paradigms. Arabic Verb Forms (I, II, III, etc.) are too general to capture the different variations within the paradigms. That is due to the different root classes (i.e. hamzated, hollow, defective, geminate and sound); and other root-pattern interactions, such as the different forms of Form VIII (اقترب/افتعل, ازدهر/افدعل, اضطرب/افطعل, etc.). All of these phenomena can be handled with orthographic, phonological and morphological rules as was done by Habash and Rambow (2006). However, here we embedded the result of such rule application in the paradigm directly. See Table 3 for counts of PV-IV patterns per Verb Form.

| Root | Perfective 3MS | Imperfective 3MS | Form | English Gloss |
|---|---|---|---|---|
| gwl | gaal | yguul | I | to say, tell |
| syr | saar | ysiir | I | to leave, go |
| ṭrš | ṭarraš | yṭarriš | II | to send, forward s.th. |

Table 1: Example of a Qafisheh Gulf Arabic Verb Lexicon Entry.

| Phono Root | CODA Root | PV3MS | IV3MS | Form | English Gloss |
|---|---|---|---|---|---|
| Gwl گول | qwl قول | qAl قال | yquwl يقول | I | to say, tell |
| syr سير | syr سير | sAr سار | ysiyr يسير | I | to leave, go |
| Trš طرش | Trš طرش | Tar∼aš طرّش | yTar∼iš يطرّش | II | to send, forward s.th. |

Table 2: Orthographic mapping of the entries in Qafisheh Gulf Arabic Verb Lexicon. The Root is orthographically spelled in two ways reflecting phonology and etymology (CODA style); PV3MS and IV3MS refer to the perfective and imperfective third masculine singular verb forms.

| Verb Form | (وَزن) | BP Count |
|---|---|---|
| I | (فَعَل) | 21 |
| II | (فَعَّل) | 6 |
| III | (فَاعَل) | 3 |
| IV | (أَفعَل) | 3 |
| V | (تفَعَّل) | 5 |
| VI | (تفَاعَل) | 6 |
| VII | (اِنفَعَل) | 4 |
| VIII | (اِفتَعَل) | 10 |
| IX | (اِفعَلَّل) | 1 |
| X | (اِستَفعَل) | 7 |
| Q | (فَعلَل) | 3 |
| Qt | (تفَعلَل) | 3 |

Table 3: Basic Paradigm counts for every Verb Form Class (وزن *wazn*).

We use the PV-IV patterns as keys (indices) for the paradigms. We then proceed to build a database of Basic Paradigms (BP). A BP is defined as the complete set of possible morphological features (except for clitic features) along with the corresponding stem. The features included are Aspect (perfective **PV**, imperfective **IV**, command **CV**), Person (1, 2, 3), Gender (masculine **M**, feminine **F**, unspecified **U**), and Number (singular **S**, plural **P**). The total number of allowable feature combinations is 19. The BP is defined in a similar fashion to the iconic inflectional classes that was defined by Eskander et al. (2013). Each form of the BP is divided into prefix, stem template, and suffix. See Table 4 for examples of two BPs.

### 4.6 Affixational Orthographic Rules

While we covered most of the orthographic, phonological and morphological rules by embedding them in the BPs, there are still a small number of additional orthographic rules that apply to specific stem-suffix combinations. Specifically, suffixes beginning with ت *t* and ن *n* that attach to stems ending with the same letter are modified as a result of the orthographic gemination rule (الشّدّة Shadda). For example the verb نحت+ت *naHat+t* 'I sculpted' should be written as نحتّ *naHat∼*; and the verb ضمن+نا *Daman+nA* 'we guaranteed' should be written as ضمّنا *Daman∼A*. We automatically identified all root-paradigm pairs that cause the above rules to apply, and we created new paradigms from them. For example, the root نحت *nHt* is linked with the paradigm 1a2a3-yi12a3-**t** and the root ضمن *Dmn* is linked with the paradigm 1a2a3-yi12a3-**n**. This resulted in 32 additional paradigms, bringing the total to 104 paradigms.

### 4.7 Lexicon Construction

From the set of PV-IV patterns, which we used as paradigm keys, and the lexical entries converted from QGAVL, we constructed our lexicon automatically and then manually validated all the entries. The lexicon consists of 2,648 entries that are linked to the paradigms. See Table 5 for examples of the lexical entries in previous tables. Each entry specifies the root (in phonological spelling and CODA) as well as the paradigm key and gloss.

| | Paradigm 1A3-y1uw3 | | | | Paradigm 1a2∼a3-y1a2∼i3 | | | |
|---|---|---|---|---|---|---|---|---|
| **Morph.Feat.** | **Prefix** | **Stem** | **Suffix** | **Example** | **Prefix** | **Stem** | **Suffix** | **Example** |
| PV1US | | 1i3 | t | قلت | | 1a2∼a3 | t | طرّشت |
| PV1UP | | 1i3 | nA | قلنا | | 1a2∼a3 | nA | طرّشنا |
| PV2MS | | 1i3 | t | قلت | | 1a2∼a3 | t | طرّشت |
| PV2FS | | 1i3 | tiy | قلتي | | 1a2∼a3 | tiy | طرّشتي |
| PV2UP | | 1i3 | tawA | قلتوا | | 1a2∼a3 | tawA | طرّشتوا |
| **PV3MS** | | **1A3** | | قال | | **1a2∼a3** | | طرّش |
| PV3FS | | 1A3 | at | قالت | | 1a2∼a3 | at | طرّشت |
| PV3UP | | 1A3 | awA | قالوا | | 1a2∼a3 | awA | طرّشوا |
| IV1US | Aa | 1uw3 | | اقول | Aa | 1a2∼i3 | | اطرّش |
| IV1UP | n | 1uw3 | | نقول | n | 1a2∼i3 | | نطرّش |
| IV2MS | t | 1uw3 | | تقول | t | 1a2∼i3 | | تطرّش |
| **IV3MS** | **y** | **1uw3** | | يقول | **y** | **1a2∼i3** | | يطرّش |
| IV3FS | t | 1uw3 | | تقول | t | 1a2∼i3 | | تطرّش |
| IV2FS | t | 1uw3 | iyn | تقولين | t | 1a2∼3 | iyn | تطرّشين |
| IV2UP | t | 1uw3 | uwn | تقولون | t | 1a2∼3 | uwn | تطرّشون |
| IV3UP | y | 1uw3 | uwn | يقولون | y | 1a2∼3 | uwn | يطرّشون |
| CV2MS | | 1uw3 | | قول | | 1a2∼i3 | | طرّش |
| CV2FS | | 1uw3 | iy | قولي | | 1a2∼3 | iy | طرّشي |
| CV2UP | | 1uw3 | awA | قولوا | | 1a2∼3 | awA | طرّشوا |

Table 4: Example of BP for a paradigm of Form I and another of Form II for the roots قول *qwl* and طرش Trš respectively. The verb قال *qAl* means 'he said' and the verb طرّش *Tar∼aš* means 'he sent'.

| Phono Root | CODA Root | PV3MS | IV3MS | Paradigm Key | Form | English Gloss |
|---|---|---|---|---|---|---|
| گول Gwl | قول qwl | قال qAl | يقول yquwl | 1A3-y1uw3 | I | to say, tell |
| سير syr | سير syr | سار sAr | يسير ysiyr | 1A3-y1iy3 | I | to leave, go |
| طرش Trš | طرش Trš | طرّش Tar∼aš | يطرّش yTar∼iš | 1a2∼a3-y1a2∼i3 | II | to send, forward s.th. |

Table 5: Example of lexicon entries. For each entry there is: (a) a phonological root, which will be used to model possible non-CODA variations, (b) a CODA root, (c) two verbal forms (PV3MS and IV3MS), (d) the paradigm key, (e) Verb Form, and (f) English gloss.

### 4.8 Clitic Extension of the Basic Paradigms

At this point, we have a complete inflectional model of GLF verbs except that they do not include any of the numerous clitics written attached in Arabic. We define a set of rules for extending the paradigms to include the clitics. Our extensions include two types of resources.

**Clitic Locations and Forms** First is the list of clitics with their morpheme POS (a la Buckwalter tag) and their relative location around the basic inflected verb, and any conditions for their application. For example, the future particle proclitic ب *b* appears immediately before the basic verb form, but can only occur with imperfective verbs; the conjunction proclitics و *wi* 'and' and ف *fa* 'so' can appear as the first clitics in any series of clitics; and so on. All possible clitic combinations are then applied to each form in the paradigm along with the necessary spelling changes. The negative proclitic ما *mA* and the indirect pronominal enclitics introduced with the preposition ل are introduced as attached at this point (which is non-CODA compliant). With this information, we are able to model the verb ومابيكتبهالهم *w+mA+b+***y-ktb**+*hA+l+hm* 'and+not+will+**he-write**+it+for+them' (the bolded substring is the only element from the BP).

We extended the paradigms with a total of 25 clitics, including five proclitics which are و *wi* 'and', ف *fa* 'so', the future particle ب *b* 'will' and the two negation particles ما *mA* and لا *lA*.

For the enclitics, we extended with all possible 10 direct object enclitics which are: ني *ny* 'me', نا *nA* 'us' , ج *ij* 'you[FS]', ك *ik* 'you[MS]' , ها *hA* 'she' ,هم *hum* 'them', هن *hun* 'them[FP]', كم *kum* 'you[P]', كن *kun* 'you[FP]' and their respective 10 indirect objects enclitics by adding the preposition ل *li* 'for'. With all of the additional clitics and their features, the total number of allowable feature combinations (or rows in the paradigms) increases from 19 to 24,321 per paradigm.

**Clitic Rewrite Rules** We apply a number of clitic rewrite rules which are mandated by CODA spelling conventions. One example is the change of the stem Alif Maqsura to Alif when it is not word final. For example the basic verb اشترى+ها *Aštrý+hA* 'he bought + it' is rewritten as اشتراها *AštrAhA* (ى *ý* → ا *A*). Another example is the drop of the Alif of the plural suffix pronouns وا *wA* when it is not word final. For example, اشتروا+ها *Aštrwa+hA* 'they boaught + it' is rewritten as اشتروها *AštrwhA* (وا *wA* → و *w*).

### 4.9 Database Generation

To generate the database, we used the same toolkit used in (Al-Shargi et al., 2016; Eskander et al., 2016) which generates a morphological analyzer database in the representation used in MADAMIRA (Pasha et al., 2014) and CALIMA$_{\text{EGY}}$ (Habash et al., 2012a). The conversion was straightforward once we converted our paradigm and lexicon database to the forms expected by the database generation tool. This conversion included providing a POS tag for every prefix, stem and suffix. We use the Buckwalter POS tag style used by many other databases for Arabic morphology (Graff et al., 2009; Habash et al., 2012a).

### 4.10 Extending to Non-CODA Variants

The generated database at this point expects only CODA input, which is not realistic for dealing with *raw* dialectal text. We extended the database for the set of complex prefixes (pronoun prefixes and proclitics), complex suffixes (pronoun suffixes and enclitics) and stems. For the complex affixes we used the same extensions used in (Habash et al., 2012a) as we don't have enough annotated data to learn from. As for the stems, we inflected the phonological roots that correspond to the CODA roots in the lexicon to their respective stems, which are mapped to the CODA stems in the database. With these extensions we will be able to correctly model a non-CODA input like يابو *yAbw* 'they brought' as correct CODA form جابوا *jAbwA*.

## 5 Evaluation

### 5.1 Experimental Setup

**Dataset** We used a part of an Emirati novel in raw text from the Gumar corpus. We contextually annotated all the verbs appearing in first 4,000 words of the novel – a total of 620 verbs. The annotation includes identifying the CODA spelling, full Buckwalter tag and the morphemic segmentation. Table 6 shows an annotation example of one sentence from the data.

In this work we only use one dataset for the evaluation as we didn't use any feedback from the evaluation in the current state of work, i.e., this was a blind test.

**Metrics** We report token recall on verbs only. We report in terms of CODA spelling, segmentation and POS. We report in two modes of input: *raw* input and CODA compliant input of the same text. Token recall counts the percentage of the time one of the analyses returned by the morphological analyzer given a particular input word matches the gold analysis of the input word in the aspect evaluated (e.g., CODA, segmentation or POS). This is similar to the evaluation carried by Habash and Rambow (2006).

**Systems** We used six different analyzers for our experiments.

- SAMA analyzer for MSA (Graff et al., 2009).
- CALIMA$_{\text{EGY}}$ for EGY, which includes MSA (Habash et al., 2012a).
- CALIMA$_{\text{GLF}}$ for GLF.
- CALIMA$_{\text{GLF-CODA}}$ is CALIMA$_{\text{GLF}}$ without the extensions discussed in 4.10.
- CALIMA$_{\text{GLF}}$ extended with SAMA.
- CALIMA$_{\text{GLF}}$ extended with CALIMA$_{\text{EGY}}$.

### 5.2 Results

SAMA performs the least amongst all systems in all aspects which is consistent with results reported by Habash and Rambow (2006) and Khalifa et al. (2016). CALIMA$_{\text{EGY}}$ performs much

*Original Gulf Arabic*

:: يوم الخميس :: على صوت اذان الفير [[كانت]] فطامي ناشه عليه ،، وعقب [[سارت]] [[تاخذ]]
شاور و [[وعت]] ابوها و مرت ابوها و عقب [[سارت]] [[اتصلي]] و [[تقرالها]]
كمن آيه ،، ويوم [[خلصت]] [[حضرت]] الريوق و [[دخنت]] البيت
و [[عدلته]] و [[نظفته]] ،، كل يوم على هالحاله ،، وهيه صابره ،،

*English literal translation*

Thursday, upon the call for the dawn prayer, Fattami [[was]] awakened; then she [[went]] and [[took]] a shower and [[woke]] her father and step mother up; and then she [[went]] to [[pray]] and [[read]] few verses; and when she [[finished]], she [[prepared]] breakfast and [[scented]] the house with incense and [[fixed]] it and [[cleaned]] it; every day is the same and she is always patient.

| Raw | | CODA | | Segmentation | Full POS tag | English Gloss |
|---|---|---|---|---|---|---|
| كانت | kAnt | كانت | kAnt | kAn+t | PV+PVSUFF_SUBJ:3FS | she was |
| سارت | sArt | سارت | sArt | sAr+t | PV+PVSUFF_SUBJ:3FS | she went |
| تاخذ | tAx* | تاخذ | tAx* | t+Ax* | IV3FS+IV | to take [3FS] |
| وعت | wEt | وعت | wEt | wE+t | PV+PVSUFF_SUBJ:3FS | she woke someone up |
| سارت | sArt | سارت | sArt | sAr+t | PV+PVSUFF_SUBJ:3FS | she went |
| اتصلي | AtSly | تصلي | tSly | t+Sly | IV3FS+IV | to pray [3FS] |
| تقرالها | tqrAlhA | تقرا لها | tqrAl hA | t+qrA+l+hA | IV3FS+IV+PREP+PRON_3FS | to read for herself |
| خلصت | xlSt | خلصت | xlSt | xlS+t | PV+PVSUFF_SUBJ:3FS | she finished |
| حضرت | HDrt | حضرت | HDrt | HDr+t | PV+PVSUFF_SUBJ:3FS | she prepared |
| دخنت | dxnt | دخنت | dxnt | dxn+t | PV+PVSUFF_SUBJ:3FS | she scented |
| عدلته | Edlth | عدلته | Edlth | Edl+t+h | PV+PVSUFF_SUBJ:3FS+PVSUFF_DO:3MS | she fixed it |
| نظفته | nZfth | نظفته | nZfth | nZf+t+h | PV+PVSUFF_SUBJ:3FS+PVSUFF_DO:3MS | she cleaned it |

Table 6: Annotation example. In this sentence, there are total of 12 verbs marked with [[ ]]. For each verb we provide the CODA spelling, morphemic segmentation and the full Buckwalter POS tag.

better than SAMA which is also consistent with previous results (Khalifa et al., 2016; Jarrar et al., 2014). $\text{CALIMA}_{\text{GLF}}$ outperforms both SAMA and $\text{CALIMA}_{\text{EGY}}$ on all measured conditions. The merged forms of $\text{CALIMA}_{\text{GLF}}$ (with SAMA and $\text{CALIMA}_{\text{EGY}}$) outperform $\text{CALIMA}_{\text{GLF}}$. The best system we have is the result of merging $\text{CALIMA}_{\text{GLF}}$ and $\text{CALIMA}_{\text{EGY}}$, which effectively includes GLF, EGY and MSA. The evaluation of $\text{CALIMA}_{\text{GLF-CODA}}$ highlights the added value of our non-CODA modeling, which contributed to over 11% absolute increase in recall (from $\text{CALIMA}_{\text{GLF-CODA}}$ to $\text{CALIMA}_{\text{GLF}}$) for raw input on all evaluated conditions.

### 5.3 Error Analysis

We conducted an error analysis on the analyzed verbs for $\text{CALIMA}_{\text{GLF}}$. We identified three main sources of errors. First are *typos* in the raw text which lead to no possible analysis. Examples include ابلس *Abls* instead of البس *Albs* 'I wear' and ادخلوو *Adxlww* instead of ادخلوا *AdxlwA* 'come in'. These kinds of errors are around 19%. Second are non-CODA-compliant input words that lead to different segmentations and POS, e.g., the word اتصلي *AtSly* (CODA تصلي *t+Sly* 'she prays') is analyzed as *AtSl+y* 'call! [FS]'. These make up around 18% of errors. Third are the out-of-vocabulary (OOV) cases, which for us include words with lemmas not in our lexicon, or words with affixes not modeled in our paradigms. For example, we encountered some EGY-like verbal constructions that we did not expect to see in GLF: تقوليلي *tqwlyly* 'you[FS] tell me' instead of تقولينلي *tqw-lynly*, تاخذولي 'you [P] take for me' instead of تاخذونلي *tAx*wnly*. These cases are about 63% of the errors. When we compare the performance of our best system ($\text{CALIMA}_{\text{GLF}}$+$\text{CALIMA}_{\text{EGY}}$) to $\text{CALIMA}_{\text{GLF}}$, we note that the errors of the first two types do not change, but there is a drop of 13% absolute in the OOV error cases.

## 6 Conclusion and Future Work

We presented $\text{CALIMA}_{\text{GLF}}$, a morphological analyzer for GLF currently covering over 2,600 verbal lemmas. $\text{CALIMA}_{\text{GLF}}$ verb analysis token recall with CODA input outperforms both SAMA and

| | Raw Input | | | CODA Input | |
|---|---|---|---|---|---|
| **Analyzer** | **CODA** | **Segmentation** | **BW POS tag** | **Segmentation** | **BW POS tag** |
| ***CALIMA$_{GLF}$ + CALIMA$_{EGY}$*** | **90.7** | **85.5** | **87.3** | **92.7** | **92.3** |
| ***CALIMA$_{GLF}$ + SAMA*** | 89.7 | 83.9 | 84.4 | 91.1 | 90.7 |
| ***CALIMA$_{GLF}$*** | **89.7** | **81.8** | **81.5** | **88.7** | **87.7** |
| ***CALIMA$_{GLF\text{-}CODA}$*** | 78.4 | 70.5 | 68.7 | 88.7 | 86.0 |
| *CALIMA$_{EGY}$* | 83.7 | 70.8 | 65.7 | 78.9 | 70.8 |
| *SAMA* | 71.6 | 52.7 | 51.8 | 64.4 | 60.3 |

Table 7: Token recall evaluation on CODA matching, Buckwalter POS tag and morphemic segmentation. Evaluation is on verbs only. The evaluated analyzers are (1) SAMA for MSA, (2) CALIMA$_{\text{EGY}}$ for EGY, which includes MSA, (3) CALIMA$_{\text{GLF}}$ for GLF, and (4) CALIMA$_{\text{GLF-CODA}}$, which is CALIMA$_{\text{GLF}}$ without the extension discussed in 4.10.

an CALIMA$_{\text{EGY}}$ by over 27.4% and 16.9% absolute, respectively, in terms of identifying correct POS tag. We plan to morphologically annotate a large portion of the Gumar corpus to learn different spelling variations and grow the coverage of lemmas. We also plan to extend CALIMA$_{\text{GLF}}$ beyond verbs using those annotations. We also plan to use a similar building process to create morphological analyzers and lexicons for other dialects given the availability of resources.

## Acknowledgments

We would like to thank Kevin Schluter and Meera Al Kaabi for helpful discussions and for providing us with transcribed entries of the Qafisheh dictionary used to build CALIMA$_{\text{GLF}}$. We also would like to thank Ramy Eskander for help in providing some of the tools we used to create the analyzer databases. We also are thankful to Maverick Alzate, who helped in the early stages of the conversion from the Qafisheh dictionary.

## References

Belal Abuata and Asma Al-Omari. 2015. A rule-based stemmer for Arabic Gulf Dialect. *Journal of King Saud University - Computer and Information Sciences*, 27(2):104 – 112.

Rania Al-Sabbagh and Roxana Girju. 2012. A supervised POS tagger for written Arabic social networking corpora. In Jeremy Jancsary, editor, *Proceedings of KONVENS 2012*, pages 39–52. ÖGAI, September. Main track: oral presentations.

Faisal Al-Shargi, Aidan Kaplan, Ramy Eskander, Nizar Habash, and Owen Rambow. 2016. A Morphologically Annotated Corpus and a Morphological Analyzer for Moroccan and Sanaani Yemeni Arabic. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*, Portorož, Slovenia.

Houda Bouamor, Nizar Habash, and Kemal Oflazer. 2014. A multidialectal parallel corpus of arabic. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014)*. European Language Resources Association (ELRA).

Tim Buckwalter. 2004. Buckwalter Arabic Morphological Analyzer Version 2.0. LDC catalog number LDC2004L02, ISBN 1-58563-324-0.

Mona Diab, Nizar Habash, Owen Rambow, Mohamed AlTantawy, and Yassine Benajiba. 2010. Colaba: Arabic dialect annotation and processing. In *Proceedings of the LREC Workshop for Language Resources (LRs) and Human Language Technologies (HLT) for Semitic Languages: Status, Updates, and Prospects*.

Mona T Diab, Mohamed Al-Badrashiny, Maryam Aminian, Mohammed Attia, Heba Elfardy, Nizar Habash, Abdelati Hawwari, Wael Salloum, Pradeep Dasigi, and Ramy Eskander. 2014. Tharwa: A large scale dialectal arabic-standard arabic-english lexicon. In *LREC*, pages 3782–3789.

Ramy Eskander, Nizar Habash, and Owen Rambow. 2013. Automatic Extraction of Morphological Lexicons from Morphologically Annotated Corpora. In *Proceedings of tenth Conference on Empirical Methods in Natural Language Processing*.

Ramy Eskander, Nizar Habash, Owen Rambow, and Arfath Pasha. 2016. 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*, pages 3455–3465, Osaka, Japan, December.

Hassan Gadalla, Hanaa Kilany, Howaida Arram, Ashraf Yacoub, Alaa El-Habashi, Amr Shalaby, Krisjanis Karins, Everett Rowson, Robert MacIntyre, Paul Kingsbury, David Graff, and Cynthia McLemore. 1997. CALLHOME Egyptian Arabic Transcripts. In *Linguistic Data Consortium, Philadelphia*.

David Graff, Mohamed Maamouri, Basma Bouziri, Sondos Krouna, Seth Kulick, and Tim Buckwalter. 2009. Standard Arabic Morphological Analyzer (SAMA) Version 3.1. Linguistic Data Consortium LDC2009E73.

Nizar Habash and Owen Rambow. 2006. MAGEAD: A Morphological Analyzer and Generator for the Arabic Dialects. In *Proceedings of ACL*, pages 681–688, Sydney, Australia.

Nizar Habash and Ryan Roth. 2009. CATiB: The Columbia Arabic Treebank. In *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*, pages 221–224, Suntec, Singapore.

Nizar Habash, Abdelhadi Soudi, and Tim Buckwalter. 2007. On Arabic Transliteration. In A. van den Bosch and A. Soudi, editors, *Arabic Computational Morphology: Knowledge-based and Empirical Methods*. Springer.

N. Habash, R. Eskander, and A. Hawwari. 2012a. A Morphological Analyzer for Egyptian Arabic. In *NAACL-HLT 2012 Workshop on Computational Morphology and Phonology (SIGMORPHON2012)*, pages 1–9.

Nizar Habash, Mona T Diab, and Owen Rambow. 2012b. Conventional Orthography for Dialectal Arabic. In *LREC*, pages 711–718.

Clive Holes. 1990. *Gulf Arabic*. Croom Helm Descriptive Grammars. Routledge, London / New York.

Mustafa Jarrar, Nizar Habash, Diyam Akra, and Nasser Zalmout. 2014. Building a Corpus for Palestinian Arabic: a Preliminary Study. *ANLP 2014*, page 18.

Mustafa Jarrar, Nizar Habash, Faeq Alrimawi, Diyam Akra, and Nasser Zalmout. 2016. Curras: an annotated corpus for the Palestinian Arabic dialect. *Language Resources and Evaluation*, pages 1–31.

Salam Khalifa, Nizar Habash, Dana Abdulrahim, and Sara Hassan. 2016. A Large Scale Corpus of Gulf Arabic. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*, Portorož, Slovenia.

H. Kilany, H. Gadalla, H. Arram, A. Yacoub, A. El-Habashi, and C. McLemore. 2002. Egyptian Colloquial Arabic Lexicon. LDC catalog number LDC99L22.

Mohamed Maamouri and Christopher Cieri. 2002. Resources for Arabic Natural Language Processing. In *International Symposium on Processing Arabic*, volume 1.

Mohamed Maamouri, Ann Bies, Tim Buckwalter, and Wigdan Mekki. 2004. The Penn Arabic Treebank: Building a Large-Scale Annotated Arabic Corpus. In *NEMLAR Conference on Arabic Language Resources and Tools*, pages 102–109, Cairo, Egypt.

Mohamed Maamouri, Ann Bies, Tim Buckwalter, Mona Diab, Nizar Habash, Owen Rambow, and Dalila Tabessi. 2006. Developing and Using a Pilot Dialectal Arabic Treebank. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation, LREC06*.

Mohamed Maamouri, Ann Bies, Seth Kulick, Dalila Tabessi, and Sondos Krouna. 2012a. Egyptian Arabic Treebank DF Parts 1-8 V2.0 - LDC catalog numbers LDC2012E93, LDC2012E98, LDC2012E89, LDC2012E99, LDC2012E107, LDC2012E125, LDC2013E12, LDC2013E21.

Mohamed Maamouri, Sondos Krouna, Dalila Tabessi, Nadia Hamrouni, and Nizar Habash. 2012b. Egyptian Arabic Morphological Annotation Guidelines.

Mohamed Maamouri, Ann Bies, Seth Kulick, Michael Ciul, Nizar Habash, and Ramy Eskander. 2014. Developing an Egyptian Arabic Treebank: Impact of Dialectal Morphology on Annotation and Tool Development. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014)*. European Language Resources Association (ELRA).

Abir Masmoudi, Mariem Ellouze Khmekhem, Yannick Esteve, Lamia Hadrich Belguith, and Nizar Habash. 2014. A corpus and phonetic dictionary for tunisian arabic speech recognition. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014)*. European Language Resources Association (ELRA).

Karima Meftouh, Salima Harrat, Salma Jamoussi, Mourad Abbas, and Kamel Smaili. 2015. Machine translation experiments on padic: A parallel arabic dialect corpus. In *The 29th Pacific Asia conference on language, information and computation*.

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

Hamdi A Qafisheh. 1977. A Short Reference Grammar of Gulf Arabic.

H.A. Qafisheh. 1997. *NTC's Gulf Arabic-English dictionary*. NTC Pub. Group.

Houda Saadane and Nizar Habash. 2015. A Conventional Orthography for Algerian Arabic. In *ANLP Workshop 2015*, page 69.

Wael Salloum and Nizar Habash. 2011. Dialectal to Standard Arabic Paraphrasing to Improve Arabic-English Statistical Machine Translation. In *Proceedings of the First Workshop on Algorithms and Resources for Modelling of Dialects and Language Varieties*, pages 10–21, Edinburgh, Scotland.

Kamel Smaïli, Mourad Abbas, Karima Meftouh, and Salima Harrat. 2014. Building resources for Algerian Arabic dialects. In *15th Annual Conference of the International Communication Association Interspeech*.

Otakar Smrž and Jan Hajič. 2006. The Other Arabic Treebank: Prague Dependencies and Functions. In Ali Farghaly, editor, *Arabic Computational Linguistics: Current Implementations*. CSLI Publications.

Houcemeddine Turki, Emad Adel, Tariq Daouda, and Nassim Regragui. 2016. A Conventional Orthography for Maghrebi Arabic. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*, Portorož, Slovenia.

Wajdi Zaghouani, Behrang Mohit, Nizar Habash, Ossama Obeid, Nadi Tomeh, Alla Rozovskaya, Noura Farra, Sarah Alkuhlani, and Kemal Oflazer. 2014. Large scale Arabic error annotation: Guidelines and framework. In *International Conference on Language Resources and Evaluation (LREC 2014)*.

Omar F Zaidan and Chris Callison-Burch. 2011. The Arabic online commentary dataset: an annotated dataset of informal arabic with high dialectal content. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers-Volume 2*, pages 37–41. Association for Computational Linguistics.

Ines Zribi, Rahma Boujelbane, Abir Masmoudi, Mariem Ellouze, Lamia Belguith, and Nizar Habash. 2014. A Conventional Orthography for Tunisian Arabic. In *Proceedings of the Language Resources and Evaluation Conference (LREC), Reykjavik, Iceland*.