# A Morphologically Annotated Corpus of Emirati Arabic

**Salam Khalifa, Nizar Habash, Fadhl Eryani,**
**Ossama Obeid, Dana Abdulrahim,[†] Meera Al Kaabi[‡]**
Computational Approaches to Modeling Language Lab, New York University Abu Dhabi, UAE
University of Bahrain, Bahrain[†]
The United Arab Emirates University, UAE[‡]
{salamkhalifa@nyu.edu, nizar.habash@nyu.edu, darahim@uob.edu.bh}

**Abstract**

We present an ongoing effort on the first large-scale morphologically manually annotated corpus of Emirati Arabic. This corpus includes about 200,000 words selected from eight Gumar corpus novels in the Emirati Arabic variety. The selected texts are being annotated for tokenization, part-of-speech, lemmatization, English glosses and dialect identification. The orthography of the text is also adjusted for errors and inconsistencies. We discuss the guidelines for each part of the annotation components, and the annotation interface we use. We report on the quality of the annotation through an inter-annotator agreement measure.

**Keywords:** Gulf Arabic, Part-of-Speech Tagging, Morphology, Annotation

## 1. Introduction

There has been an increasing number of natural language processing (NLP) efforts focusing on dialectal Arabic, especially with the increasing amounts of written material on the web. However, resources for dialectal Arabic NLP tasks such as part-of-speech (POS) tagging, morphological analysis and disambiguation are still lacking compared to those for Modern Standard Arabic (MSA). MSA is the official language in more than 20 countries, where it is used in official communications, news, and education. Yet, it is not the commonly spoken variety of Arabic; the dialectal varieties of Arabic are what is used in the day-to-day communication. Dialectal Arabic is also commonly used in written form on social media platforms, forums and blogs.

Using available resources developed for MSA such as POS taggers and tokenizers gives limited performance when used on dialectal Arabic (Habash and Rambow, 2006; Jarrar et al., 2014; Khalifa et al., 2016a). Many researchers moved into the direction of creating tools and resources targeting the dialects specifically. Egyptian Arabic is one of the dialects that received earlier efforts for developing tools and resources. More resources are being developed for other dialects such as Levantine, Tunisian, Moroccan and Yemeni Arabic. Gulf Arabic, as we define it to be the native spoken variety in the Gulf Cooperation Council, is still lagging behind other Arabic dialects with respect to resource and tool creation, given the considerable amount of dialectal content online.

In this paper, we present an ongoing project for creating a manually annotated corpus of about 200,000 words of the Gulf Arabic of the United Arab Emirates – Emirati Arabic. The corpus is annotated for tokenization, POS, lemmas and English glosses in addition to spelling conventionalization and dialect identification. This resource will support the development of Arabic dialect enabling technologies, such as automatic POS tagging and morphological disambiguation, which in turn will facilitate efforts on different NLP tasks such as machine translation.

The rest of this paper is organized as follows. We discuss related work on dialectal corpora in Section 2. In Section 3. we describe the corpus used in this effort. We then present the annotation guidelines that are used to annotate the corpus in Section 4. We discuss the annotation process and the annotation quality results in Section 5.

## 2. Related Work

In this section we review a number of efforts on Arabic corpus creation, that significantly supported research and tool development for Arabic NLP.

### 2.1. Modern Standard Arabic Resources

The Penn Arabic Treebank (PATB) (Maamouri et al., 2004) has been a central resource for developing MSA resources. It was developed at the Linguistic Data Consortium (LDC), and it mainly consists of newswire text from different news sources. The PATB corpus is annotated for tokenization, segmentation, POS tagging, lemmatization, diacritization, English gloss and syntactic structure. The PATB has 12 parts of more than 1.3 million words. The annotated data has been a backbone of many state-of-the-art tools such as analyzers and disambiguators including MADAMIRA (Pasha et al., 2014) and its predecessor MADA (Habash et al., 2009), in addition to YAMAMA (Khalifa et al., 2016b), and most recently a neural morphological disambiguatior (Zalmout and Habash, 2017) and a fine grained POS tagger (Inoue et al., 2017). In addition, the PATB guidelines (Maamouri et al., 2009) have inspired the creation of similar guidelines for the dialects including our own.

### 2.2. Dialectal Arabic Resources

In the scope of dialectal Arabic, there have been many recent contributions to the development and creation of resources. Below, we discuss the highlights of those contributions.

**Egyptian Arabic Resources** Egyptian Arabic (EGY) was one of the first dialects that received the attention of the NLP community. The earliest effort, to the best of our knowledge, is the Egyptian Colloquial Arabic Lexicon (ECAL) (Kilany et al., 2002) which was developed as part of the CALLHOME Egypt corpus (Gadalla et al., 1997). The ECAL served as the seed to the EGY morphological analyzer (CALIMA) (Habash et al., 2012a). Later on, the Egyptian Arabic Treebank (ARZATB) (Maamouri et al., 2012a; Maamouri et al., 2014) was created by the LDC using CALIMA to provide analysis options for the annotation process. The ARZATB has currently 400,000 words in eight parts annotated in a similar fashion to the PATB. The annotation guidelines for the ARZATB (Maamouri et al., 2012b) followed that of the PATB with decisions specific to the dialect. Since the release, the ARZATB has been used extensively for developing EGY resources such as the EGY part of MADAMIRA, MADA and YAMAMA, in addition to a noise-robust morphological disambiguator for EGY (Zalmout et al., 2018). Other developed corpora and POS taggers for EGY include the work of Al-Sabbagh and Girju (2012) where they created their own POS tagset and corpus with the intention to facilitate certain NLP applications like subjectivity and sentiment analysis.

**Levantine Arabic and and Other Dialectal Arabic Resources** Levantine Arabic (LEV) received some notable efforts including the Levantine Arabic Treebank (LATB) of Jordanian Arabic (Maamouri et al., 2006) which contains around 27,000 annotated words in a similar fashion to ARZATB. A more recent resource is the annotated corpus of Palestinian Arabic (Curras) (Jarrar et al., 2014; Jarrar et al., 2016). ARZATB and Curras were used to create morphological analyzers and disambiguators (Eskander et al., 2016). Other dialects such as Yemeni and Moroccan Arabic followed the same approach (Al-Shargi et al., 2016). In addition to the dialects mentioned above, there were recent efforts on creating corpora for other dialects, namely Tunisian and Algerian (McNeil and Faiza, 2011; Masmoudi et al., 2014; Zribi et al., 2015; Smaïli et al., 2014). Other works targeted multi-dialect corpora (Diab et al., 2010; Zaidan and Callison-Burch, 2011; Diab et al., Forthcoming 2013; Bouamor et al., 2014; Cotterell and Callison-Burch, 2014), and, most recently, the ongoing Multi Arabic Dialect and Application Resources project (MADAR) (Bouamor et al., 2018) which includes corpora for 25 different city dialects.

**Gulf Arabic Resources** As far as Gulf Arabic (GLF) is concerned, the only existing annotated corpora include the Emirati Arabic Corpus (EAC) (Halefom et al., 2013) and the Emirati Arabic Language Acquisition Corpus (EMALAC) (Ntelitheos and Idrissi, 2017) that were created by linguists with emphasis on the phonological and morphosyntactic phenomena of Emirati Arabic. We recently collected a large-scale corpus of Gulf Arabic (Khalifa et al., 2016a) containing more than 100 million words covering six Gulf Arabic varieties. In regards to other tools and resources, we recently developed a morphological analyzer for Gulf Arabic verbs (CALIMA$_{\text{GLF}}$) (Khalifa et al., 2017). We are also aware of the previously developed rule-based stemmer for Arabic Gulf dialect (Abuata and Al-Omari, 2015).

In this work, we use about 200,000 words from the Emirati Arabic portion of the Gumar corpus to manually annotate for tokenization, POS tagging, lemma, English gloss and dialect identification. Additionally we conventionalize the spelling in accordance with the Conventional Orthography for Dialectal Arabic (CODA) rules (Habash et al., 2012b; Habash et al., 2018).

For recent surveys on Arabic resources for NLP, see Zaghouani (2014), Shoufan and Al-Ameri (2015) and Zeroual and Lakhouaja (2018).

## 3. Annotating the Gumar Corpus

We discuss next the Gumar Corpus and the portion of it we use to annotate in this effort.

### 3.1. Gumar Corpus

The Gumar corpus is a large-scale corpus of Gulf Arabic containing more than 100 million words. The corpus consists mainly of documents of long conversational novels also known as روايات النت 'Internet Novels'. This type of literature is very popular among female teenagers in the Gulf area. These novels are written mostly in dialectal Arabic, where the lengthy conversations between the characters of the story are in the dialect and the narration in between the conversations can sometimes be in MSA.

The writers of the novels remain anonymous and use *noms de plume*. The novels are publicly available online, where most of the writers ask for their pen name to be mentioned if the novel is to be published in a different platform than the original. The genre of the novels is mainly romantic, but also features tragedy and drama. The corpus can be browsed online,[1] it is currently annotated using MADAMIRA in EGY mode.

On the document level, Gulf Arabic text makes up more than 90% of the corpus, the rest of the corpus consists of other Arabic dialects in addition to MSA. Emirati Arabic text covers around 11% of the Gumar corpus.

### 3.2. The Annotated Gumar Corpus

We chose a set of 200,110 word tokens for the annotation task. The text consists of the first 25,000 words (rounded up to the nearest full sentence) from eight different novels by eight different authors. This allows us to cover different writing styles. The text is comprised of 15,277 sentences with an average of 13 words per sentence. Table 1 shows the list of the novels from which the text is selected. We name this subset of the corpus the *Annotated Gumar Corpus*. In the future we plan to continue adding annotations to it from other Gumar novels including different dialects.

Additionally, a total of about 12,000 words – 1,500 words from each of the eight parts rounded up to the nearest full sentence – are chosen to evaluate Inter-Annotator Agreement (IAA) throughout the annotation process. Thus, the total number of words to be annotated is about 212,000 words.

[1] Please visit `https://camel.abudhabi.nyu.edu/gumar/`

| Parts | Tokens | Sentences | العنوان | Title Transltation | المؤلف | Author |
|---|---|---|---|---|---|---|
| Part 1 | 25,022 | 1,387 | صوت دقات الخوافق لا طريته مثل حبات المطر فوق الخيام | 'The sound of the beating hearts when I remember it, is like the drops of rain on the tents' | b3thra had2a | 'Calm scatter' |
| Part 2 | 25,009 | 3,176 | والله أحبك لو على رقبتي سيف | 'I swear to God I love you even if there is a sword on my neck' | أم نشوان | 'Umm Nashwan' |
| Part 3 | 25,004 | 1,732 | بنيت لك في داخل فؤادي من اللّهفة قصر | 'I built you a castle in my heart out of longing' | أسماء | 'Asmaa' |
| Part 4 | 25,002 | 1,919 | مجنون ساره | 'Crazy about Sarah' | فاتنة دبي | 'The enchantress of Dubai' |
| Part 5 | 25,004 | 1,412 | وش عذرك يوم تخون وش ذنبي يوم صدقتك | 'What is your excuse when you betray what was my fault when I believed you' | غايتي رضا ربي | 'My Lord's satisfaction is my purpose' |
| Part 6 | 25,039 | 1,439 | ملامح وجهي القديم | 'Features of my old face' | حديقة الظلام | 'The garden of darkness' |
| Part 7 | 25,002 | 2,211 | له منزلن بالحيل مرموق ، وسط الحشا بأقصى الضمايا | 'Your love has become part of me' | ضحية حبيبي | 'The victim of my lover' |
| Part 8 | 25,028 | 2,001 | طحت طيحة في هواكم | 'I fell hard in your love' | السولعي | 'The Gazelle' |
| **Total** | **200,110** | **15,277** | | | | |

Table 1: The list of novels (parts) used for annotation and their raw word counts. The English titles and author names are approximate translations of the original Arabic ones. The author names are *noms de plume*.

## 4. Annotation Guidelines

In this section, we present the guidelines with examples for each of the different annotation tasks. The annotation contains six different tasks: spelling conventionalization according to CODA, tokenization, POS tagging, lemmatization, English glossing and dialect Identification on the word level.

### 4.1. CODA Spelling Guidelines

Emirati Arabic is similar to other dialects where there is no standard orthography. For example the word for 'hunger' may be spelled phonetically اليوع *Alywς*[2] or using the MSA cognate الجوع *Aljwς*. Hence, there will always be inconsistencies between different writers or even within the same writer (Habash et al., 2012b). In this annotation effort we follow the newly revised set of CODA* guidelines which include consonant mapping, vowel spelling and affixation and cliticization rules (Habash et al., 2018).

### 4.2. Tokenization Guidelines

Previous efforts used different tokenization and segmentation schemes depending on the goal of the task. In this annotation task we use the D3 tokenization (Habash, 2010), where we keep the baseword and separate all the clitics including the ال *Al* 'the' definite article. The clitics include all attachable prepositions, particles and pronouns. For example, the word بالجوع *bAljwς* 'with the hunger' is tokenized as [جوع]+ال+ب *b+Al+*[*jwς*], where the baseword in this case is [جوع] [*jwς*].

### 4.3. POS Guidelines

In this work, we opted to use a new POS tagset – CAMEL POS. CAMEL POS is inspired by the ARZATB tagset and guidelines (Maamouri et al., 2012b) which is based on the PATB guidelines (Maamouri et al., 2009). The CAMEL POS is designed as single tagset for both MSA and the dialects with the following goals in mind: (a) facilitating research on adaptation between MSA and the dialects, and among the dialects; (b) supporting backward compatibility with previously annotated resources; and (c) enforcing a functional morphology analysis that is deeper and more compatible with Arabic morphosyntactic rules than form-based analysis (Alkuhlani and Habash, 2011). The CAMEL POS tags and features are the union of those in MSA and the dialects. Features are available to use when needed. For example case and state features are used more often in MSA; but on the other hand, dialects tend to have many more clitics than MSA, including non-MSA ones.

One of the main differences between CAMEL POS and ARZATB is that the morphological features of both gender and number of nominals are annotated functionally (Alkuhlani and Habash, 2011; Smrž, 2007). This decision allows us to assign the features to the baseword without the need to specify the surface form affixes that mark form gender and number. This is not the case in ARZATB, where broken plural nouns are tagged singular because they do not use the sound plural affixes.

The other main difference is that we omit case and state features for nominals, and voice and mood for verbs as the dialects have almost lost them completely, except for some high frequency fossilized MSA forms, such as طبعاً *TabςAã* 'of course' which retains an indefinite ending.

The main part of the word, that is the baseword, is tagged in the following format: 'POS.features', where 'POS' is the core POS tag and 'features' is the possible feature combination that goes with the POS tag, a '.' separates the POS from the feature combination. Proclitics, however, get only a 'POS' tag since they have no features. However, pronom-

[2] Arabic transliteration is presented in the Habash-Soudi-Buckwalter scheme (Habash et al., 2007): (in alphabetical order)
أ ب ت ث ج ح خ د ذ ر ز س ش ص ض ط ظ ع غ ف ق ك ل م ن ه و ي
Â b t θ j H x d ð r z s š S D T Ď ς γ f q k l m n h w y
and the additional symbols: ' ء, Â أ, Ǎ إ, Ā آ, ŵ ؤ , ŷ ئ , ħ ة, ý ى.

inal enclitics get a similar tag format as the baseword (i.e. 'PRON.features').

CAMEL POS provides full array of features: (i) **A**spect with the values **P**erfective, **I**mperfective and **C**ommand; (ii) **P**erson with the values **1**st, **2**nd, **3**rd; (iii) **G**ender with values **M**asculine and **F**eminine; (iv) **N**umber with values **S**imgular, **D**ual and **P**lural and (v) **S**tate with values **D**efinite, **I**ndefinite and **C**onstruct; (vi) **C**ase with values **N**ominative, **G**enitive and **A**ccusative; (vii) **V**oice with values **A**ctive and **P**assive and (viii) **M**ood with values **S**ubjunctive, **I**ndicative and **J**ussive. Not all the features mentioned are necessarily relevant to the dialects. In the full POS tag, the specified values of the different features will appear in the following order:

**<POS>.<A><P><G><N>.<S><C><V><M>**
The second period is not necessary if none of the last four features is specified.

Table 2 shows the list of POS tagset used in this annotation effort compared with the ones used ARZATB. The tagset is divided into three categories according to the tokenization scheme we follow: *proclitics* (14 tags), *enclitics* (2 tags) and *baseword* (39 tags). Together with the features, CAMEL POS tagset maps to ARZATB and retains backward compatibility. It also offers an intuitive Arabic scheme that is suitable to use for annotation.

For a subset of POS tags in the baseword category, each POS tag has a limited number of possible feature combinations that is paired with it. Below is the list of the POS tags that take features and their possible ordered combination.

- **NOUN, NOUN_*, ADJ, ADJ_*** All nominals take the combination of **G**ender, **N**umber. For example جالس *jAls* 'sitting' is tagged ADJ.MS ; In the occasional uses of **S**tate, such as طبعًا *Tabς Aã* 'of course' the tag would be NOUN.MS.I
- **VERB** All verbs take the combination of **A**spect, **P**erson, **G**ender and **N**umber. For example يقطع *yqTς* 'cut' is tagged as VERB.I3MS
- **PRON** All pronouns take the combination of **P**erson, **G**ender and **N**umber. For example انتي *Anty* 'you [fs]' is tagged as PRON.2FS
- **PRON_DEM** All demonstrative pronouns take the combination of **G**ender and **N**umber. For example هَاذَا *hAðA* 'this' is tagged as PRON_DEM.3MS

In cases where a feature is not present, such as gender in verbs of first person inflections, the gender feature is simply dropped and does not require a placeholder since the possible feature values are ordered and unique. For example the imperfective 1st person verb اقول *Aqwl* 'I say' will be tagged as VERB.I1S

## 4.4. Lemma Guidelines

The lemma is the citation form of the the word. We follow the same guidelines of the lemma specification from Graff et al. (2009), where nominals are cited using the masculine singular form of the word or the feminine singular form if no masculine form exists. For example, the

| CAMEL POS Arabic | CAMEL POS | ARZATB POS |
|---|---|---|
| ***PROCLITIC* tags** | | |
| أداة_تعريف | PART_DET | DET |
| حرف_عطف | CONJ | CONJ |
| حرف_جر | PREP | PREP |
| أداة_نفي | PART_NEG | NEG_PART |
| أداة_استقبال | PART_FUT | FUT_PART |
| أداة_مضارعة | PART_PROG | PROG_PART |
| أداة_ربط | CONJ_SUB | SUB_CONJ |
| ضمير_إِشَارة | PRON_DEM | DEM_PRON |
| ضمير_استفهام | PRON_INTERROG | INTERROG_PRON |
| أداة | PART | PART |
| حرف_ربط | PART_CONNECT | CONNEC_PART |
| أداة_توكيد | PART_EMPHATIC | EMPHATIC_PART |
| جواب_شرط | PART_RC | RC_PART |
| أداة_نداء | PART_VOC | VOC_PART |
| ***ENCLITIC* tags** | | |
| أداة_نفي | PART_NEG | NEG_PART |
| ضمير | PRON | *SUFF_DO:[PGN] |
| ضمير | PRON | POSS_PRON_[PGN] |
| ضمير | PRON | PRON_[PGN] |
| ***BASEWORD* tags** | | |
| اسم | NOUN | NOUN |
| اسم_عدد | NOUN_NUM | NOUN_NUM |
| اسم_علم | NOUN_PROP | NOUN_PROP |
| اسم_كم | NOUN_QUANT | NOUN_QUANT |
| صفة | ADJ | ADJ |
| صفة_عدد | ADJ_NUM | ADJ_NUM |
| صفة_مقارنة | ADJ_COMP | ADJ_COMP |
| ظرف | ADV | ADV |
| ظرف_استفهام | ADV_INTERROG | INTERROG_ADV |
| ظرف_موصول | ADV_REL | REL_ADV |
| فعل | VERB | IV/PV/CV |
| شبه_فعل | VERB_PSEUDO | PSEUDO_VERB |
| اسم_فعل | VERB_NOM | VERB |
| ضمير | PRON | PRON_[PGN] |
| ضمير_إِشَارة | PRON_DEM | DEM_PRON_[GN] |
| ضمير_استفهام | PRON_INTERROG | INTERROG_PRON |
| ضمير_تعجب | PRON_EXCLAM | EXCLAM_PRON |
| ضمير_موصول | PRON_REL | REL_PRON |
| أداة | PART | PART |
| أداة_تعريف | PART_DET | DET |
| أداة_نفي | PART_NEG | NEG_PART |
| أداة_استقبال | PART_FUT | FUT_PART |
| أداة_مضارعة | PART_PROG | PROG_PART |
| أداة_فعل | PART_VERB | VERB_PART |
| أداة_نداء | PART_VOC | VOC_PART |
| أداة_استفهام | PART_INTERROG | INTERROG_PART |
| أداة_استثناء | PART_RESTRICT | RESTRIC_PART |
| أداة_تفصيل | PART_FOCUS | FOCUS_PART |
| أداة_توكيد | PART_EMPHATIC | EMPHATIC_PART |
| جواب_شرط | PART_RC | RC_PART |
| أداة_ربط | CONJ_SUB | SUB_CONJ |
| حرف_جر | PREP | PREP |
| حرف_عطف | CONJ | CONJ |
| حرف_ربط | PART_CONNECT | CONNEC_PART |
| رقم | DIGIT | NOUN_NUM |
| اختصار | ABBREV | ABBREV |
| تعجب | INTERJ | INTERJ |
| أجنبي | FORIEGN | FOREIGN |
| علامة_ترقيم | PUNC | PUNC |

Table 2: Table shows the CAMEL POS tagset used in the annotation of Annotated Gumar Corpus compared to the POS tagset in ARZATB. CAMEL POS Arabic shows the Arabic name of the tag.

lemma for the noun سيايير *syAyyr* ‘cars’ (NOUN.FP) is سَيَّارَة *say∼Arah̄* which is feminine singular since there is no masculine singular form of the word. The verbs are cited using the perfective 3rd person masculine singular form. For example, the lemma for the verb يشوفن *yšwfn* ‘they see [f.p]’ (VERB.I3FP) is شَاف *šAf* . For all other tags (i.e. particles, adverbs, ... etc) the lemma is the same form of the baseword. In this annotation effort, the lemma is the only form we require to be manually diacritized.

### 4.5. English Gloss Guidelines

The English gloss in this context refers to the semantic translation of the Arabic lemma. For nominals we use the singular form, and for verbs we use the infinitive form. An Arabic lemma could have multiple synonymous English glosses. For example كبير *kbyr* would have the following English glosses ‘large; great; important; major; senior’.

### 4.6. Word Level Dialect Identification

Dialect identification is the task of tagging a certain context with a given dialect tag. Deciding the dialect tag depends on the context of the sentence and/or the document. This can be challenging since many words in their written form may be shared by many dialects and MSA. Additionally, it is not uncommon to find dialect code switching between MSA and a dialect, and even a dialect with another dialect (less commonly) (Elfardy and Diab, 2012). Hence we tag per word, but rely on the context of the sentence and even the document to identify the dialect.

In Table 3 we show an example of a fully annotated sentence and the POS tag in ARZATB for comparison. For full description of each of the annotation tasks and examples, the full guidelines can be accessed online.

## 5. Annotation Process

In this section, we discuss the annotation process details, the tool we used, and some annotation quality evaluation results.

### 5.1. MADARi Interface

We used a newly developed interface for morphological annotation and spelling correction called MADARi (Obeid et al., 2018). MADARi is a web-based interface that supports joint morphological annotation (tokenization, POS tagging, lemmatization) and spelling correction at any point of the annotation process, which minimizes error propagation. English glossing and dialect identification are also supported in the interface. MADARi assigns initial answers to the new text using MADAMIRA in EGY mode, whose databases we extended with CALIMA$_{\text{GLF}}$ for more coverage. MADARi has many utilities to facilitate the annotation process that we utilize for more efficiency, of which examples are discussed in the next subsection. Figure 1 shows a screenshot of the annotation view in MADARi.

### 5.2. Manual Annotation

The annotator starts on an automatically pre-annotated document. They carefully examine the spelling of each word and all its analysis choices in context with reference to the raw text at all times. For each word the annotator faces one of the following scenarios:

- All annotation tasks are correct: the annotator has to only validate the answer.
- Correct analysis but wrong spelling: the annotator has to adjust the spelling and then validate the answer.
- Wrong analysis (wholly or partially) but correct spelling: the annotator can manually adjust the analysis or can use the ‘analysis search’ utility provided by MADARi to get an analysis for a word with similar structure and then they would only have to change the lemma and the gloss entries. Finally they validate the answer.
- Wrong analysis and spelling: the annotator has to adjust the spelling and follow the previous step.

At any point of the annotation process, the annotator is able to apply mass changes to spelling and/or analysis across the document they are working on. However, the annotator must insure that all the words affected by the change are in similar contexts. The annotator can also modify their answers any time during the annotation through feedback they get if they have any inquiries. This allows the annotator to skip over words they are not confident about and leave the answer unvalidated.

Once the annotation task is fully completed, the annotator may ‘submit’ the finished document to be later exported. This will allow all the analyses made by the annotator to be accessible to all the other annotators when they look up the analysis for similar words.

### 5.3. Inter Annotator Agreement

We evaluated the quality of the annotation using the Inter Annotator Agreement (IAA) measure between two annotators on a selected text of 1,500 words. We measured the agreement on: (i) word boundary, that is the agreement on whether word boundaries are the same (no splits/merges); (ii) CODA spelling; (iii) baseword form; (iv) baseword POS; (v) baseword features; (vi) clitic form (averaged across all clitic positions) and (vii) clitic POS (averaged across all clitic positions). To align the pair of annotations, we perform a word level alignment within the sentences. We use a weighted Levenshtein distance to maximize alignment, where insertions and deletions are weighted as 1 and substitutions are weighted as follows:

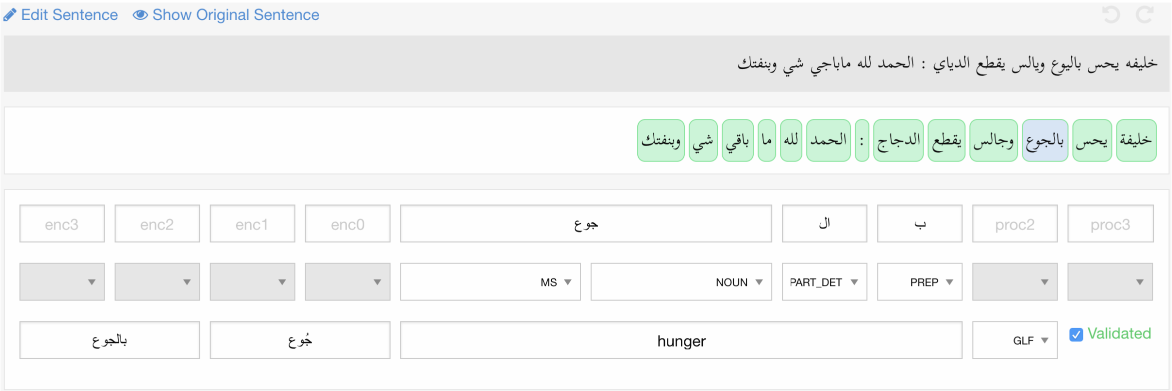
$$W_{edit}(t_1, t_2) = \frac{2Lev(t_1, t_2)}{max(|t_1|, |t_2|)} \quad (1)$$

Above, $t_1, t_2$ are the two word tokens, and $Lev$ is the Levenshtein distance at the *character* level. We employ this character-based weighing scheme to encourage the alignment of words with spelling changes. Using the same IAA measure, we measured the similarity between each annotator and the initial answers from the CALIMA$_{\text{GLF}}$-extended MADAMIRA.

The results are presented in Table 4 in terms of percent agreement. MADAMIRA provided a very helpful starting point. In at least 75% of the case, annotators agreed with

| **Raw sentence** | خليفه يحس باليوع ويالس يقطع الدياي : الحمد لله ماباجي شي وبنفتك | | | | | |
|---|---|---|---|---|---|---|
| **CODA sentence** | خليفة يحس بالجوع وجالس يقطع الدجاج : الحمد لله ما باقي شي وبنفتك | | | | | |
| **Transliteration** | xlyfħ yHs bAljwς wjAls yqtς AldjAj : AlHmd llh mA bAqy šy wbnftk | | | | | |
| **English Translation** | Khalifa feeling hungry and cutting chicken: Thank God it is almost over | | | | | |
| **Word** | **CODA** | **TOK/POS** | **Lemma** | **Gloss** | **Dialect** | **ARZATB analysis** |
| خليفه | خليفة | خليفة/NOUN_PROP.MS | خَلِيفَة | Khalifa | GLF | خليف/NOUN_PROP+ة/NSUFF_FEM_SG |
| يحس | يحس | يحس/VERB.I3MS | حَسّ | feel | GLF | ي/IV3MS+حس/IV |
| باليوع | بالجوع | ب/PREP+ال/PART_DET+جوع/NOUN.MS | جُوع | hunger | GLF | ب/PREP+ال/DET+جوع/NOUN |
| ويالس | وجالس | و/PREP + جالس/ADJ.MS | جَالِس | sitting | GLF | و/PREP+جالس/ADJ |
| يقطع | يقطع | يقطع/VERB.I3MS | قَطَّع | cut | GLF | ي/IV3MS+قطع/IV |
| الدياي | الدجاج | ال/PART_DET+دجاج/NOUN.MS | دِجَاج | chicken | GLF | ال/DET+دجاج/NOUN |
| : | : | :/PUNC | : | : | GLF | :/PUNC |
| الحمد | الحمد | ال/PART_DET+حمد/NOUN.MS | حَمد | gratitude | GLF | ال/DET+حمد/NOUN |
| لله | لله | ل/PREP+الله/NOUN_PROP.MS | اَللّه | God | GLF | ل/PREP+الله/NOUN_PROP |
| ماباجي | ما | ما/PART_NEG | مَا | not | GLF | ما/NEG_PART |
| | باقي | باقي/ADJ.MS | بَاقِي | remaining | GLF | باقي/ADJ |
| شي | شي | شي/NOUN.MS | شَيّ | something | GLF | شي/NOUN |
| وبنفتك | وبنفتك | و/CONJ+ب/PART_FUT+نفتك/VERB.I1P | اِفتَكّ | get rid of | GLF | و/CONJ+ب/FUT_PART+ن/IV1P+فتك/IV |

Table 3: An annotation example in the CAMEL POS scheme showing the different entries per word, in addition to the annotations in the ARZATB tagset for comparison. While Arabic is written from right to left, the tags above are displayed from left to right.



خليفه يحس باليوع ويالس يقطع الدياي : الحمد لله ماباجي شي وبنفتك

خليفة | يحس | بالجوع | وجالس | يقطع | الدجاج | : | الحمد | لله | ما | باقي | شي | وبنفتك

proc3: (empty) | proc2: (empty) | ب: PREP | ال: PART_DET | جوع: NOUN, MS | enc0: | enc1: | enc2: | enc3:

☑ Validated | GLF | hunger | جُوع | بالجوع

Figure 1: Example of the annotation step using the MADARi interface. The top gray box shows the raw sentence; next are the word tokens reflecting any spelling changes made. The section below shows all the fields required to annotate; they are initially populated using MADAMIRA. This example is of a manually annotated entry following the discussed guidelines.

| **Category** | **A1 vs M** | **A2 vs M** | **A1 vs A2** |
|---|---|---|---|
| Word Boundary | 89.7 | 89.1 | 98.9 |
| CODA | 78.8 | 78.1 | 94.7 |
| Baseword Form | 79.2 | 79.1 | 95.1 |
| Baseword POS | 80.2 | 80.4 | 96.1 |
| Baseword Features | 77.3 | 75.8 | 95.2 |
| Average Clitic Form | 96.0 | 95.9 | 99.4 |
| Average Clitic POS | 95.5 | 95.5 | 99.0 |

Table 4: Percentages of agreement between two annotators (i.e. A1 and A2) and between each annotator and the extended MADAMIRA (i.e. M) initial answers.

MADAMIRA's analysis choice. For each aligned pair of annotations, we compute the number of agreements for the considered categories (i–vii). The IAA score across the various categories ranges from 94.7% on CODA to over 99% on clitic annotations. Moreover, the measures between the annotators and MADAMIRA's answers show that both annotators changed many of the initial answers and their change was consistent to a large extent.[3]

## 6. Conclusion and Future Work

We presented an ongoing project for creating a manually annotated corpus of about 200,000 words of Emirati Arabic – the Annotated Gumar Corpus. We discussed the full guidelines for the different annotation components that include spelling adjustments, tokenization, POS tagging, lemmatization, English glossing and dialect identification. We used a newly developed interface for morphological annotation and spelling correction. We described the manual annotation process and finally measured the quality of the annotation through an IAA measure that found agreements

[3] At the time of writing this paper, the annotation of Parts 1, 2 and 3 had reached 75%, 65% and 66% of progress, respectively. The latest status of the annotation process can be viewed online along with all the guidelines mentioned in this paper. Please visit `http://resources.camel-lab.com`

ranging between 94.7% to more than 99% for different annotation tasks. In the future, we plan to expand the annotated text to include other genres and dialects. We are also interested in using the annotations to improve the quality of Arabic dialect POS tagging and morphological disambiguation.

## 7. Acknowledgements

This project is funded by a New York University Abu Dhabi Research Enhancement Fund. We would like to thank Ramy Eskander and the team of annotators at Ramitechs. We also thank the creators of MADARi from the MADAR project. Finally we thank Sondos Krouna for insightful discussions on POS decisions.

## Bibliographical References

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

Al-Sabbagh, R. and Girju, R. (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.

Al-Shargi, F., Kaplan, A., Eskander, R., Habash, N., and Rambow, O. (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.

Alkuhlani, S. and Habash, N. (2011). A Corpus for Modeling Morpho-Syntactic Agreement in Arabic: Gender, Number and Rationality. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics (ACL'11)*, Portland, Oregon, USA.

Bouamor, H., Habash, N., and Oflazer, K. (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).

Bouamor, H., Habash, N., Salameh, M., Zaghouani, W., Rambow, O., Abdulrahim, D., Obeid, O., Khalifa, S., Eryani, F., Erdmann, A., and Oflazer, K. (2018). The MADAR Arabic Dialect Corpus and Lexicon. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)*, May.

Cotterell, R. and Callison-Burch, C. (2014). A multi-dialect, multi-genre corpus of informal written Arabic. In *LREC*, pages 241–245.

Diab, M., Habash, N., Rambow, O., AlTantawy, M., and Benajiba, Y. (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*.

Diab, M., Hawwari, A., Elfardy, H., Dasigi, P., Al-Badrashiny, M., Eskander, R., and Habash, N. (Forthcoming – 2013). Tharwa: A Multi-Dialectal Multi-Lingual Machine Readable Dictionary.

Elfardy, H. and Diab, M. (2012). Token Level Identification of Linguistic Code Switching. *Proceedings of COLING 2012: Posters*, pages 287–296.

Eskander, R., Habash, N., Rambow, O., and Pasha, A. (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.

Gadalla, H., Kilany, H., Arram, H., Yacoub, A., El-Habashi, A., Shalaby, A., Karins, K., Rowson, E., MacIntyre, R., Kingsbury, P., Graff, D., and McLemore, C. (1997). CALLHOME Egyptian Arabic Transcripts. In *Linguistic Data Consortium, Philadelphia*.

Graff, D., Maamouri, M., Bouziri, B., Krouna, S., Kulick, S., and Buckwalter, T. (2009). Standard Arabic Morphological Analyzer (SAMA) Version 3.1. Linguistic Data Consortium LDC2009E73.

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

Habash, N., Soudi, A., and Buckwalter, T. (2007). On Arabic Transliteration. In A. van den Bosch et al., editors, *Arabic Computational Morphology: Knowledge-based and Empirical Methods*. Springer.

Habash, N., Rambow, O., and Roth, R. (2009). MADA+TOKAN: A toolkit for Arabic tokenization, diacritization, morphological disambiguation, POS tagging, stemming and lemmatization. In Khalid Choukri et al., editors, *Proceedings of the Second International Conference on Arabic Language Resources and Tools*. The MEDAR Consortium, April.

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

Habash, N., Diab, M. T., and Rambow, O. (2012b). Conventional Orthography for Dialectal Arabic. In *LREC*, pages 711–718.

Habash, N., Khalifa, S., Eryani, F., Rambow, O., Abdulrahim, D., Erdmann, A., Faraj, R., Zaghouani, W., Bouamor, H., Zalmout, N., Hassan, S., shargi, F. A., Alkhereyf, S., Abdulkareem, B., Eskander, R., Salameh, M., and Saddiki, H. (2018). Unified Guidelines and Resources for Arabic Dialect Orthography. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)*, May.

Habash, N. (2010). *Introduction to Arabic Natural Language Processing*. Morgan & Claypool Publishers.

Halefom, G., Leung, T., and Ntelitheos, D. (2013). A corpus of Emirati Arabic. Technical Report NRF Grant (31 H001), United Arab Emirates University.

Inoue, G., Shindo, H., and Matsumoto, Y. (2017). Joint Prediction of Morphosyntactic Categories for Fine-Grained Arabic Part-of-Speech Tagging Exploiting Tag Dictionary Information. In *Proceedings of the 21st Conference on Computational Natural Language Learning*

*(CoNLL 2017)*, pages 421–431, Vancouver, Canada, August. Association for Computational Linguistics.

Jarrar, M., Habash, N., Akra, D., and Zalmout, N. (2014). Building a Corpus for Palestinian Arabic: a Preliminary Study. *ANLP 2014*, page 18.

Jarrar, M., Habash, N., Alrimawi, F., Akra, D., and Zalmout, N. (2016). Curras: An Annotated Corpus for the Palestinian Arabic dialect. *Language Resources and Evaluation*, pages 1–31.

Khalifa, S., Habash, N., Abdulrahim, D., and Hassan, S. (2016a). A Large Scale Corpus of Gulf Arabic. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*, Portorož, Slovenia.

Khalifa, S., Zalmout, N., and Habash, N. (2016b). YAMAMA: Yet Another Multi-Dialect Arabic Morphological Analyzer. In *Proceedings of the International Conference on Computational Linguistics (COLING): System Demonstrations*, pages 223–227.

Khalifa, S., Hassan, S., and Habash, N. (2017). A Morphological Analyzer for Gulf Arabic Verbs. *WANLP 2017 (co-located with EACL 2017)*, page 35.

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

Maamouri, M., Bies, A., Buckwalter, T., and Mekki, W. (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.

Maamouri, M., Bies, A., Buckwalter, T., Diab, M., Habash, N., Rambow, O., and Tabessi, D. (2006). Developing and Using a Pilot Dialectal Arabic Treebank. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation, LREC 2006*.

Maamouri, M., Bies, A., Krouna, S., Gaddeche, F., and Bouziri, B., (2009). *Penn Arabic Treebank Guidelines*. Linguistic Data Consortium.

Maamouri, M., Bies, A., Kulick, S., Tabessi, D., and Krouna, S. (2012a). Egyptian Arabic Treebank DF Parts 1-8 V2.0 - LDC catalog numbers LDC2012E93, LDC2012E98, LDC2012E89, LDC2012E99, LDC2012E107, LDC2012E125, LDC2013E12, LDC2013E21.

Maamouri, M., Krouna, S., Tabessi, D., Hamrouni, N., and Habash, N. (2012b). Egyptian Arabic Morphological Annotation Guidelines.

Maamouri, M., Bies, A., Kulick, S., Ciul, M., Habash, N., and Eskander, R. (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).

Masmoudi, A., Ellouze Khmekhem, M., Esteve, Y., Hadrich Belguith, L., and Habash, N. (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).

McNeil, K. and Faiza, M. (2011). Tunisian Arabic Corpus : Creating a Written Corpus of an "Unwritten" Language.

Ntelitheos, D. and Idrissi, A. (2017). Language Growth in Child Emirati Arabic. *Hamid Ouali (ed.) Perspectives on Arabic Linguistics 29*, pages 229–248.

Obeid, O., Khalifa, S., Habash, N., Bouamor, H., Zaghouani, W., and Oflazer, K. (2018). MADARi: A Web Interface for Joint Arabic Morphological Annotation and Spelling Correction. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)*, May.

Pasha, A., Al-Badrashiny, M., Kholy, A. E., Eskander, R., Diab, M., Habash, N., Pooleery, M., Rambow, O., and Roth, R. (2014). MADAMIRA: A Fast, Comprehensive Tool for Morphological Analysis and Disambiguation of Arabic. In *In Proceedings of LREC*, Reykjavik, Iceland.

Shoufan, A. and Al-Ameri, S. (2015). Natural Language Processing for Dialectical Arabic: A Survey. In *ANLP Workshop 2015*, page 36.

Smaïli, K., Abbas, M., Meftouh, K., and Harrat, S. (2014). Building resources for Algerian Arabic dialects. In *15th Annual Conference of the International Communication Association Interspeech.*

Smrž, O. (2007). ElixirFM — Implementation of Functional Arabic Morphology. In *Proceedings of the 2007 Workshop on Computational Approaches to Semitic Languages: Common Issues and Resources*, pages 1–8, Prague, Czech Republic, June. ACL.

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

Zaidan, O. F. and Callison-Burch, C. (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.

Zalmout, N. and Habash, N. (2017). Don't Throw Those Morphological Analyzers Away Just Yet: Neural Morphological Disambiguation for Arabic. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 704–713.

Zalmout, N., Erdmann, A., and Habash, N. (2018). Noise-Robust Morphological Disambiguation for Dialectal Arabic. In *Proceedings of the 26th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies Conference (HLT-NAACL2018).*

Zeroual, I. and Lakhouaja, A., (2018). *Arabic Corpus Linguistics: Major Progress, But Still A Long Way to Go*, pages 613–636. Springer International Publishing, Cham.

Zribi, I., Ellouze, M., Belguith, L. H., and Blache, P. (2015). Spoken Tunisian Arabic Corpus" STAC": Transcription and Annotation. *Research in computing science*, 90:123–135.