A Large Scale Corpus of Gulf Arabic

Salam Khalifa, Nizar Habash, Dana Abdulrahim†, Sara Hassan
Computational Approaches to Modeling Language Lab, New York University Abu Dhabi, UAE
†University of Bahrain, Bahrain
{salamkhalifa,nizar.habash,sah650}@nyu.edu,darahim@uob.edu.bh

Abstract

Most Arabic natural language processing tools and resources are developed to serve Modern Standard Arabic (MSA), which is the official written language in the Arab World. Some Dialectal Arabic varieties, notably Egyptian Arabic, have received some attention lately and have a growing collection of resources that include annotated corpora and morphological analyzers and taggers. Gulf Arabic, however, lags behind in that respect. In this paper, we present the Gumar Corpus, a large-scale corpus of Gulf Arabic consisting of 110 million words from 1,200 forum novels. We annotate the corpus for sub-dialect information at the document level. We also present results of a preliminary study in the morphological annotation of Gulf Arabic which includes developing guidelines for a conventional orthography. The text of the corpus is publicly browsable through a web interface we developed for it.

Keywords: Arabic Dialects, Corpus, Large-Scale, Gulf Arabic

1. Introduction

Most Arabic natural language processing (NLP) tools and resources are developed to serve Modern Standard Arabic (MSA), the official written language in the Arab World. Using such tools to understand and process Dialectal Arabic (DA) is a challenging task because of the phonological and morphological differences between DA and MSA. In addition, there is no standard orthography for DA, which only complicates matters more. Some DA varieties, notably Egyptian Arabic, have received some attention lately and have a growing collection of resources that include annotated corpora and morphological analyzers and taggers. Gulf Arabic (GA), broadly defined as the variety of Arabic spoken in the countries of the Gulf Cooperation Council (Bahrain, Kuwait, Oman, Qatar, Saudi Arabia, and the United Arab Emirates), however, lags behind in that respect.

In this paper, we present the Gumar Corpus,[1] a large-scale corpus of GA that includes a number of sub-dialects. We also present preliminary results on GA morphological annotation. Building a morphologically annotated GA corpus is a first step towards developing NLP applications, for searching, retrieving, machine-translating, and spell-checking GA text among other applications. The importance of processing and understanding GA text (as with all DA text) is increasing due to the exponential growth of socially generated dialectal content in social media and printed works (Sarnākh, 2014), in addition to existing materials such as folklore and local proverbs that are found scattered on the web.

The rest of this paper is structured as follows. We present some related work in Dialectal Arabic NLP in Section 2. This is followed by a background discussion on GA in Section 3. We then discuss the collection of the corpus and describe its genre in Section 4. We present our preliminary annotation study and evaluate it in Section 5. Finally, we present the Gumar Corpus web interface in Section 6.

[1]Gumar قُمَر /gumɛr/ is the word for 'moon' in Gulf Arabic.

2. Related Work

2.1. Dialectal Corpora

There have been many notable efforts on the development of annotated Arabic language corpora (Maamouri and Cieri, 2002; Maamouri et al., 2004; Smrž and Hajič, 2006; Habash and Roth, 2009; Zaghouani et al., 2014). Most contributions however targeted MSA, developing annotation guidelines and producing large-scale Arabic Treebanks. These resources were instrumental in pushing the state-of-the-art of Arabic NLP.

Contributions that are specific to DA are limited in size, more scattered and more recent. Some of the earliest and relatively largest efforts have targeted Egyptian Arabic (EGY). They include CALLHOME Egyptian Arabic (CHE) corpus (Gadalla et al., 1997) and its associated Egyptian Colloquial Arabic Lexicon (ECAL) (Kilany et al., 2002). In addition, there is the YADAC corpus (Al-Sabbagh and Girju, 2012), which was based on dialectal content identification and web harvesting of blogs, micro blogs, and forums of EGY content. And most recently, the Linguistic Data Consortium collected and annotated a sizable EGY corpus (Maamouri et al., 2012b; Maamouri et al., 2012a; Maamouri et al., 2014). Levantine Arabic received less attention, with notable efforts including the Levantine Arabic Treebank (LATB) of Jordanian Arabic (Maamouri et al., 2006) and the Curras corpus of Palestinian Arabic (Jarrar et al., 2014). Efforts on other dialects include corpora for Tunisian Arabic (Masmoudi et al., 2014) and Algerian Arabic (Smaïli et al., 2014). There are also some efforts that targeted multiple dialects such as the COLABA project (Diab et al., 2010) which annotated dialectal content resources for Egyptian, Iraqi, Levantine, and Moroccan dialects from online weblogs, the Tharwa multi-dialectal lexicon (Diab et al., 2014), the multidialectal parallel Arabic corpus (Bouamor et al., 2014), and the highly dialectal online commentary corpus (Zaidan and Callison-Burch, 2011). Most recently, in this conference proceedings, Al-Shargi et al. (2016) present two morphologically annotated corpora for Moroccan Arabic and Sanaani Yemeni Arabic.

As far as Gulf Arabic is concerned, Halefom et al. (2013) created an Emirati Arabic Corpus (EAC) consisting of 2 million words of transcribed Emirati TV and radio shows. The corpus was transcribed in broad IPA and translated to English. Morphological and lexical annotations as well as Arabic script annotation was manually done for a small portion of the corpus (around 15,000 words). Furthermore, Ntelitheos and Idrissi (Forthcoming) created an Emirati Arabic Language Acquisition Corpus (EMALAC) consisting of 78,000 words following the style of the widely studied CHILDES collection of corpora (MacWhinney, 2000). Both EAC and EMALAC were created by linguists with the primary purpose of studying the grammatical system of the Emirati Arabic dialect and its development. There is a lot of emphasis in the annotations of these corpora on the phonological and morphosyntactic phenomena of Emirati Arabic. Our Gumar Corpus is differently oriented and designed: text as opposed to speech is the starting point. And computational models of GA is our target. Our corpus only includes language created by adult speakers (unlike EMALAC) and that is a slightly conventionalized novel-like form. Finally the Gumar Corpus includes texts from a number of the Gulf countries and is not limited to the UAE. For recent surveys of Arabic resources for NLP, see Zaghouani (2014) and Shoufan and Al-Ameri (2015).

2.2. Dialectal Orthography

Due to the lack of standardized orthography guidelines for DA, along with the phonological differences from MSA, and dialectal variations within the dialects themselves, there are many orthographic variations for written DA content. Writers in DA, regardless of the context, are often inconsistent with others and even with themselves when it comes to the written form of a dialect, writing with MSA driven orthography, or phonologically driven orthography in Arabic script or even Latin script (Darwish, 2013; Al-Badrashiny et al., 2014). These orthographic variations make it difficult for computational models to properly identify and reason about the words of a given dialect (Habash et al., 2012b), hence, a conventional form for the orthographic notations is important. Habash et al. (2012b) proposed a Conventional Orthography for Dialectal Arabic (CODA). CODA is designed for the purpose of developing conventional computational models of Arabic dialects in general which makes it easy to be extended to other dialects. Initially, the guidelines of CODA were mainly specific to EGY. Jarrar et al. (2014) extended the existing CODA to cover Palestinian Arabic. Recent work on Tunisian (Zribi et al., 2014), Algerian (Saadane and Habash, 2015) and Maghrebi Arabic (Turki et al., 2016) extended the original version of CODA. We extend CODA to cover Gulf Arabic.

2.3. Arabic Dialect Morphological Modeling

Most of the work that explored morphology in Arabic focused on MSA (Al-Sughaiyer and Al-Kharashi, 2004; Buckwalter, 2004; Graff et al., 2009; Pasha et al., 2014). Contributions to DA morphology analysis are usually based on either extending available MSA tools to cover DA characteristics, as in the work of Abo Bakr et al. (2008) and Salloum and Habash (2011), or modeling DAs directly, without relying on existing MSA contributions (Habash and Rambow, 2006). One of the notable recent contributions for Egyptian Arabic morphological analysis is CALIMA (Habash et al., 2012a). The CALIMA analyzer for EGY and the commonly used SAMA analyzer for MSA (Graff et al., 2009) are central in the functioning of the EGY morphological tagger MADA-ARZ (Habash et al., 2013), and its successor MADAMIRA (Pasha et al., 2014), which supports both MSA and EGY. Eskander et al. (2013) describe a technique for automatic extraction of morphological lexicons from morphologically annotated corpora and demonstrate it on EGY. Al-Shargi et al. (2016) apply the technique of Eskander et al. (2013) and build two morphological analyzers for Moroccan Arabic and Sanaani Yemeni Arabic. As for GA, we are aware of a single effort on a rule based stemmer (Abuata and Al-Omari, 2015) that works on sets of words collected online; they compare their results to other well known MSA stemmers. In this paper we use MADAMIRA-EGY as a starting point for the GA morphological annotation following the approach taken by Jarrar et al. (2014).

3. Gulf Arabic Dialect

Strictly speaking, Gulf Arabic refers to the linguistic varieties spoken on the western coast of the Arabian Gulf, in Bahrain, Qatar, and the seven Emirates of the UAE (Qafisheh, 1977), as well as in Kuwait and in Al-Hasa – the eastern region of Saudi Arabia (Holes, 1990). Omani, Hijazi, Najdi, and Baharna Arabic, among other additional dialects spoken in the Arabian Peninsula, are usually not included in grammars of Gulf Arabic due to the fact that they considerably vary in their linguistic features from the set of dialects listed above. In this current project, we extend the use of the term 'Gulf Arabic' to include any Arabic variety spoken by the indigenous populations residing the six countries of the Gulf Cooperation Council: Bahrain, Kuwait (KW), Oman (OM), United Arab Emirates (AE), Qatar (QA), and Kingdom of Saudi Arabia (SA).

The cultural homogeneity of the Gulf region does not necessarily entail linguistic homogeneity. Indeed, GA dialects extensively differ in their morpho-phonological and lexical features, reflecting a number of geographical and social factors (Holes, 1990) in addition to being influenced by different contact languages at different time periods. A number of linguistic features set GA dialects apart from other dialects spoken in the Arab world. One of the distinguishing phonological features of most GA dialects includes maintaining the pharyngealized fricative /$z^{\text{ʕ}}$/,[2] as well as the interdental /θ/ and /ð/, unlike what happens in other Arabic dialects. Among the most prominent phonological features in GA are the variant pronunciations of the sounds /q/, /dʒ/, /ʃ/, and /k/. /q/ may be realized in certain dialects as /g/ as in /gaːl/ 'he said', or as /dʒ/ as in /dʒɪdɪr/ 'pot'. /dʒ/, on the other hand, may be realized in some varieties as /j/, as in /jɪmɛl/ 'camel'. The palatal /tʃ/ and the velar /k/ may both turn into the alveopalatal /tʃ/ as in /tʃaːj/ 'tea', and /tʃɛf/ 'palm'. Moreover, The 2nd singular feminine possessive and object pronoun /kɪ/ retains its phonological form

[2]Phonetic transcription is presented in IPA.

Gumar Corpus	
Words	112,410,688
Sentences	9,335,224
Documents	1,236

Table 1: Statistics on the Gumar Corpus

in certain dialects (e.g. some Saudi dialects) but is realized in some other dialects as /tʃ/, /ʃ/, or /ts/.

In terms of the morphological features, and as in the case with most spoken Arabic dialects, GA dialects have also lost case inflection (with the exception of some Bedouin dialects, e.g. /bɪntɪn ʔəsˤiːlɛ/ 'respectable girl'). Possession may also be marked by clitics such as /maːl/ and /ħag/ (Holes, 1990), e.g. /lɪ-ktaːb maːlɛl mʌrʌh/ 'The book of the lady'. Negation is also marked by the particles /muː/ (and its variants /mʊb/, /mʊhʊb/, and /hʊb/) (Holes, 1990). The plural and the dual masculine and feminine forms of verbs and nouns are collapsed into one form in most dialects, but some distinctions are still maintained in certain others. For instance, in some varieties of Saudi, Emirati, and Omani Arabic, verbs and possessive pronouns inflected for the 3rd plural feminine are quite distinct from the masculine forms:

- /gaːmɛt/ 'She stood up' - /gaːmɛn/ 'They[2FP] stood up'.
- /wɛdʒhɪk/ 'Your[2FS] face' - /wədʒuːhkɪm/ 'Your[2FP] faces'.

Additional morphological features include the cliticized /ba/ for future (instead of the standard /sa/ prefix). Alternatively, in some varieties of GA, the motion verb /raːħ/ has grammaticalized into a future marker (e.g. Kuwaiti, Bahraini, and some varieties of Saudi Arabic). Less prominently, yet still a distinctive feature in some GA dialects, is the the epenthetic /n/ found in the active participle in some varieties of Emarati and Baharna Arabic: /maːʕtˤɪnhəm/ 'I've/'d given them'. Finally, the lexicon of GA consists of standard Arabic cognates that may or may not follow the phonotactics of the respective dialects. Unsurprisingly, many cognate expressions that are highly frequent in a given dialect have reduced in form, such as

- /liʔay ʃayʔ/ 'For which thing'→/leːʃ/ 'Why'.

As for lexical borrowings in GA dialects, there is an undoubtedly substantial amount of lexical items that have been borrowed from various contact languages throughout different historical periods, e.g.,

- /ʕɛmbɛluːsˤ/ 'Ambulance' from English.
- /hɛst/ 'There is' from Persian.
- /ʔaːluː/ 'Potatoes' from Hindi.

4. Corpus Description

Corpus Collection Gulf Arabic, just like any other Arabic dialect has no written convention nor is it used as a formal mean of communication in the media, education or official documents. Hence there are no known go-to resources. A unique genre of written materials that is specifically known to GA is online anonymous publicly published long conversational novels. We have found a huge collection of these novels online in one place.[3] We automatically downloaded about 1,200 MS Word documents. Usually, such novels are written in lengthy threads that can be found in online forums. The data we got was collected by volunteering forum members into MS Word documents and then published by another member in an organized matter.[4]

Corpus Genre The main theme of most of the novels is romantic; but they also include drama and tragedy. The structure of a novel is simple. It starts with a brief introduction that contains the title of the novel, the writer's pen name (no real names are used) and the country of the novel. The introduction is then followed by a prologue that usually contains a small piece of dialectal poetry or a small piece of literary writing usually in MSA. It also contains a brief description of the novel characters, though some writers prefer to introduce the characters as their role appears. Then comes the main body of the novel, which is often a dialogue between the characters. There are also some pieces of narration between conversations in either the dialect or MSA. The last part of the novel usually has some "moral" lessons narrated by the writer. Writers tend to ask the audience for positive criticism and opinions and whether they should continue writing more novels or not.

The novels are entirely written in DA except for the parts mentioned above. The dialect of the novel is not necessarily the same as the dialect of the writer. Most of the time the writers remain anonymous under nicknames, though they ask to be credited if the novel is transferred to another forum. Hence some writers are quite famous among the audience. The targeted audience is mainly female teenagers, the nature of publishing the novels is highly interactive and dependent on the activity of the audience. The writer usually ends each "part" in the novel with a teaser and demands participation and encouragement from the audience. Table 1 shows statistics on all the collected text. Words are whitespace tokenized and the counts include punctuation. The number of sentences represents the number of lines. Most of the time, each document represents a single novel; but in few cases a novel may be split into more than one document.

Corpus Dialects and Dialect Annotation We have annotated the corpus on the level of documents for the dialect, novel name and writer name for each. The dialect of the written text was the most challenging to know. In some documents, the dialect or the country of the writer was explicitly stated; in others, names of cities clearly indicated the origin country. However, in many cases, further investigation was needed. The GA dialects are closely intertwined, yet when thoroughly observed show evident differences. These differences were observed through common trends in relation to each GA dialect. It is important to point that the names given to the characters in each story have shown a trend with the dialect used, for example: SA

[3] www.graaam.com

[4] There are no copyright claims by the anonymous writers or organizers; and we do not claim any copyrights to the text. We will make the cleaned up and extended versions of the data fully publicly available.

dialect novels have repeatedly used the names فيصل fySl[5] /feːsˤɛl/ 'Faisal', عبد العزيز ʕbd Alʕzyz /ʕɛbdɪlʕɛziːz/ 'Abdulaziz' and تركي trky /tɪrkej/ 'Turky' for male characters and لما lmA /lɛmɛ/ 'Lama', وسن wsn /wɛsɛn/ 'Wasan' and نجود njwd /ndʒuːd/ 'Njood' for female characters. On the other hand the AE dialect novels have used the names هزّاع hz∼Aʕ /hɛzzaːʕ/ 'Hazza'', زايد zAyd /zaːjɪd/ 'Zayed' and راشد rAšd /raːʃɪd/ 'Rashid' as male names and حصّة HS∼ħ /ħɪsˤsˤɛ/ 'Hessa', ميثاء myθA' /meːθɛ/'Maytha' and شمّة šm∼ħ /ʃɛmmɛ/ 'Shamma' for female characters. It is worth mentioning that in both SA and AE dialect novels the male names could be related to the leaders of each country and hence this may be a cultural influence that could be related to the author's intuition when selecting character names. The KW dialect novels used the names ضاري DAry /dˤaːrej/ 'Dhari', مشاري mšAry /mʃaːrej/ 'Mshari' and صباح SbAH /sˤʊbaːħ/ 'Subah' for male characters while the names سبيجة sbyjħ /sɪbiːtʃɛ/ 'Sebichah' and فوزيّة fwzy∼ħ /fozɪjjɛ/ 'Fawziyya' were noticed to be prominent for female characters. The second noticed trend is the use of Hijri dates in SA dialect novels, which is explained by the country's official calendar use of the Hijri date. This trend was only noticed in SA dialect novels. Thirdly, the use of commonly spoken words that could be directly traced to a particular dialect such as هالحزّة hAlHzzħ /hɛlhɛzzɛ/ 'Now' and يا معود yA mʕw∼d /jaː mʕɛwwɛd/ 'O man' in KW dialect, بزران bzrAn /bɪzraːn/ 'Little kids', وراك wrak /wɛraːk/ as in وراك ما تجاوب wrAk mA tjAwb /wɛraːk maː tdʒaːwɪb/ 'Why are you[2MS] not answering?' which is traced to SA dialect and AE words such as سيدا sydA /siːdɛ/ 'Straight forward', اقرب Aqrb /ɪgrəb/ 'Come in, please!' and خاشوقة xAšwqħ /xaːʃuːgɛ/ 'Spoon'. The above are trends noticed when annotating over a thousand GA novels that helped in adding efficiency to the task. These trends were noticed as the process progressed and selecting a dialect was only completed once parts of the story were read, alongside facts provided by the author, which include providing an insight to the readers that the story will be written in a particular dialect and/or constantly providing details of where events took place (i.e. city and/or country names).

Following on the annotation effort, we present the distribution of the dialects across the corpus, see Table 2. We have observed that 92% of the entire corpus is actually written in GA with SA being the most dominant and BH the least. There is also around 10% that is identified as GA (other) which are the cases of a novel containing a combination of several GA dialects that is due to multiple writers with different dialects or due to the existence of different characters in the novel. It was sometimes hard to differentiate through the text between the three dialects of OM, QA and AE even with a native speaker annotating, hence these cases we marked as GA (other) also. The rest of the corpus (7.93%) is mostly MSA (original text or translation attempts of existing non Arabic text) and other DA such as Egyptian, Iraqi, Levantine, ... etc.

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

Dialect	Percentage
SA	60.52
AE	13.35
KW	5.91
OM	1.13
QA	0.65
BH	0.49
GA (other)	10.03
Arabic (other)	7.93

Table 2: Distribution of Dialects across the Gumar Corpus

5. Preliminary Investigation into GA Annotation

We describe next a pilot study in semi-automatic annotation of GA. We use the MADAMIRA (Pasha et al., 2014) morphological tagger (which works in two modes: MSA and EGY) and manually change its output in accordance to the orthography and morphology guidelines that are discussed next in this section.

Orthography Guidelines GA speakers who write in the dialect produce spontaneous inconsistent spellings that sometimes reflect the phonology of the GA, and other times the word's cognate relationship with MSA. We follow the work of Habash et al. (2012b) for CODA in order to overcome these inconsistencies. There are the general CODA rules that apply for every dialect, among which is affix spelling. Affix spelling includes the spelling of the Ta Marbuta; if the Ta Marbuta is at the end of the word it will always be ة ħ and not ه h regardless of the pronunciation. When inside a word (before a clitic) Ta Marbuta will become تـ, (e.g. سيّارة حصّة جديدة syArħ HSħ jdydħ /sɛjjaːrɛt ħisˤsˤah jɪdiːdɛh/ 'Hissa's car is new', سيّارتها syArthA /sajjaːretha:/ 'her car'). Following on the discussion about GA dialect properties in section 3., we extend several aspects of the original CODA. The root consonant mapping rules are extended to cover the GA pronunciations that are unseen in other dialects, see Table 3. Another aspect is the spelling of the 2nd person singular feminine pronominal clitic; if spelled differently than the ك k /kɪ/ equivalent in MSA, it is mapped to ج j /dʒ/, (e.g. كتابش ktAbš /ktaːbɪʃ/, كتابتس ktAbts /ktaːbɪts/, and كتابج ktAbj /ktaːbɪtʃ/ 'Your[FS] book', becomes كتابج ktAbj /ktaːbɪtʃ/ but not كتابك ktAbik /ktaːbɪk/) in CODA. As with the original CODA for EGY, we also maintain a list of exceptional spellings for uniquely dialectal words. One example in GA is the spelling of the perfective verb كان kAn /kaːn/ 'was' and its variant جان jAn /tʃaːn/: the perfective verb form is used as in MSA, however the other variant is considered a modal auxiliary (Brustad, 2000). Both spellings are kept due to the difference in their usage despite the fact that they share the same origin. Another example is the the negation particles مب mb /mʊb/

CODA	Pron. variation	Example
ق *q*	/q/ or /g/ or /dʒ/	قدام *qid∼Am* /dʒɪddaːm/ ‘Front’ قلم *elm* /gɛlˁɛm/ ‘pen(cil)’
ك *k*	/k/ or /tʃ/ or /ts/	كبد *kbd* /tʃɛbd/ ‘Liver’
ج *j*	/dʒ/ or /j/	جلس *jls* /jɪlas/ ‘He sat’
ش *š*	/ʃ/ or /tʃ/	شاي *šAy* /tʃaːj/ ‘Tea’

Table 3: GA root consonants mapping rules

‘not’ and مانيب mAnyb /mɛniːb/ ‘I’m not’, both of which have a number of non-CODA variants such as موب mwb and منيب mnyb, respectively. The complete guidelines for the GA CODA will be available separately as a technical report. An example of the application of several CODA rules is presented in Table 4

Morphology Guidelines For every input word MADAMIRA produces a list of analyses specifying every possible morphological interpretation of that word, covering all morphological features of the word (diacritization, part-of-speech (POS), lemma, and 13 inflectional and clitic features). MADAMIRA then applies a set of models (support vector machines and N-gram language models) to produce a prediction, per word in-context, for different morphological features, such as POS, lemma, gender, number or person. A ranking component scores the analyses produced by the morphological analyzer using a tuned weighted sum of matches with the predicted features. The top-scoring analysis is chosen as the predicted interpretation for that word in context (Pasha et al., 2014).
We follow a similar approach that was used to morphologically annotate both the EGY and LEV corpora. We select the following set of features with their initial values from the output of MADAMIRA-EGY on a given GA text to annotate:

- **Word orthography** We follow the previously discussed orthography guidelines.
- **Morphemic tokenization** A word is split into its morphemes and stem.
- **Part of Speech** We use the MADAMIRA POS tag set
- **CATiB 6 POS** Except the tag for passive verbs (Habash and Roth, 2009).
- **Lemma** Diacritized form of the lemma.
- **English gloss** The English translation of the lemma

Beside the above guidelines, there exist cases of erroneous merging and splitting of words that gets no analysis from the automatic annotation. For merged words, we place a ‘#’ symbol where a split should happen, this fix aggregate through all the other annotated features. In the case where there is a split, we place the ‘#’ symbol at the end of the first split part to indicate the merging position.
Table 5 shows an annotation example following the above guidelines. Where green and red shaded cells indicates a change.

Evaluation We conduct an evaluation of the quality of automatic morphological annotation tools (taggers) on this corpus to assess the amount of effort needed to manually annotate it. Following the annotation guidelines discussed above, we manually annotated around 4K words from four different novels with a goal to capture different dialects, styles of writing, . . . etc. An example of an annotated sentence is shown in Table 5. As a preliminary experiment, we investigated the frequency of out-of-vocabulary (OOV) words in both systems of MADAMIRA: MSA and EGY. The Egyptian model OOV (5.6%) was almost half that of MSA (9.3%), suggesting it is better to work with Egyptian as the base system for manual annotation.

Feature	MADAMIRA-MSA	MADAMIRA-EGY
CODA	83.81	88.34
Morph	76.16	83.62
POS	72.37	80.39
CATiB	76.28	81.51
Lemma	64.03	77.02

Table 6: Results on evaluation of our Gold annotation against the output of MADAMIRA in both modes: MSA and EGY

We evaluated using the accuracy measure for word orthography, morphemic tokenization, POS, CATiB POS and lemma against the output of both models of MADAMIRA: MSA and EGY. Table 6 shows the results of the the evaluation for all words. These numbers allow us to assess the basic quality of the these tools on GA. As expected, MADAMIRA-EGY outperforms MADAMIRA-MSA between 4 and 13% absolute on different metrics, confirming that it is better to use it as a baseline. This is similar to results reported by Jarrar et al. (2014) on Palestinian Arabic.

Error Analysis We manually investigated the four sets of 100 words from different parts of the MADAMIRA-EGY annotated sub-corpus (total 400 words, with an average of 30 words containing at least one error). 52.1% of the errors are likely due the wrong assignment of POS especially in proper nouns that look like nouns or adjectives. Another major source of error is the lemma 18.2% and out of vocabulary related errors are 16.5% and this happens for two main reasons, either the word is never seen before or because of a typo. Finally, errors that come from a mistake in merging or splitting word tokens, typos and tokenization words combine around 13%.

6. Gumar Interface

Following the collection of the corpus, we created a simple online interface that is specific for searching the corpus.[6] The entire text of the corpus is stored in a relational database in an optimized manner. The lookup of the data is a simple search query that matches the user input to either a word token, lemma or stem form. Through the website interface, the rows of results are displayed to the user including the full context, word analysis that includes the POS,

[6] http://camel.abudhabi.nyu.edu/gumar/

Example 1	Raw	...اسمع هالحتسي منتس ياويلتس ...Asmς hAlHtsy mnts yAwylts ...
	CODA	...اسمع هالحكي منج ياويلج ...Asmς hAlHky mnj yAwylj ...
	English	[If] I hear this talk from you [again][,] you will suffer
Example 2	Raw	...عسى الغدى جاهز؟ ...ςsý Alγdý jAhz?
	CODA	...عسى الغدا جاهز؟ ...ςsý AlγdA jAhz?
	English	...Is lunch ready?
Example 3	Raw	ساره: منيب صغيرونه انا اللحين في الجامعه sArh: mnyb Sγyrrwnh AnA AllHyn fy AljAmςh
	CODA	سارة: مانيب صغيرونة انا الحين في الجامعة sArħ: mAnyb Sγyrwnħ AnA AlHyn fy AljAmςħ
	English	Sarah: I'm not a child I'm now in university.

Table 4: Example of application of CODA rules

MADAMIRA-EGY							
Raw		**CODA**	**Morph**	**POS**	**CATiB 6**	**Lemma**	**Gloss**
زياد	zyAd	zyAd	zyAd	noun_prop	PROP	ziyAd	Ziad
:	:	:	:	punc	PNX	:	:
ليساء	lysA’	lysA’	lysA’	noun	NOM	Âaloyas	valiant
ميساء	mysA’	mAysA’	mA+y+sA’	verb	VRB	ÂasA’	be_harmed
انتولازم	AntwlAzm	NOAN	NOAN	NOAN	NOAN	NOAN	NOAN
تجابون	tjAbwn	tjAbwn	t+jAb+wn	verb	VRB	ÂajAb	be_answered
ع	ς	ςlY	ςlY	prep	PRT	ςalaY	on
كل	kl	kl	kl	noun_quant	NOM	kul∼	all
الي	Aly	Ally	Ally	pron_rel	NOM	All∼iy	which
يقولكم	yqwlkm	yqwlkm	y+qwl+km	verb	VRB	qAl	said

Manual Annotation							
Raw		**CODA**	**Morph**	**POS**	**CATiB 6**	**Lemma**	**Gloss**
زياد	zyAd	zyAd	zyAd	noun_prop	PROP	ziyAd	Ziad
:	:	:	:	punc	PNX	:	:
ليساء	lysA’	lysA’	lysA’	noun_prop	PROP	laysaA’	Laysaa
ميساء	mysA’	mysA’	mysA’	noun_prop	PROP	MayosaA’	Maysaa
انتولازم	AntwlAzm	Antw#lAzm	Antw#lAzm	pron#noun	NOM#NOM	Antw#lAzim	you#necessary
تجابون	tjAbwn	tjAwbwn	t+jAwb+wn	verb	VRB	jAwab	comply
ع	ς	ςlY	ςlY	prep	PRT	ςalaY	on
كل	kl	kl	kl	noun_quant	NOM	kul∼	all
الي	Aly	Ally	Ally	pron_rel	NOM	All∼iy	which
يقولكم	yqwlkm	yqwlh#lkm	y+qwl+h#l+km	verb#prep	VRB#PRT	qAl#la	said#to

Table 5: Example of manual annotation following the orthography and morphology guidelines. Columns represent features to be annotated and rows represent words. NOAN means that no analysis was given automatically.

lemma, stem and gloss entries in addition to the information about the novel the word belongs to. See Figure 1.

7. Conclusion and Future Work

We collected the Gumar Corpus that consists of 100 million words from 1200 forum novels. We annotated the corpus for sub-dialect information at the document level of the novels in addition to the informations about the name and the writer's name of the novel. We also performed a preliminary investigation on the annotation of GA text. As an initial experiment we annotated around 4K words from four different novels using proposed orthography and morphology guidelines that followed previous efforts. We compared our gold annotations to the automatic annotations provided by MADAMIRA on its both MSA and EGY modes. The evaluation of the accuracy suggests that using MADAMIRA-EGY automatic annotations as a starting point for manual annotation of GA speeds up the process. We plan to semi-automatically annotate the corpus and include a careful manual check at a large portion of it (1M words). We are also looking forward to building a morphological analyzer for GA. We also plan to use the Gumar Corpus dialect annotations for some NLP tasks such as dialect identification. We will make this corpus and its annotations publicly available.

8. Acknowledgments

We would like to thank Dimitrios Ntelitheos for helpful discussions. We would also like to thank Mustafa Jarrar and

Figure 1: Online web interface for browsing Gumar

Rami Asia for sharing their set up of the Curras database browsing website.

9. Bibliographical References

Abo Bakr, H., Shaalan, K., and Ziedan, I. (2008). A Hybrid Approach for Converting Written Egyptian Colloquial Dialect into Diacritized Arabic. In *The 6th International Conference on Informatics and Systems, INFOS2008*. Cairo University.

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-Badrashiny, M., Eskander, R., Habash, N., and Rambow, O. (2014). Automatic Transliteration of Romanized Dialectal Arabic. *CoNLL-2014*, page 30.

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.

Al-Sughaiyer, I. A. and Al-Kharashi, I. A. (2004). Arabic morphological analysis techniques: A comprehensive survey. *Journal of the American Society for Information Science and Technology*, 55(3):189–213.

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).

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

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

Darwish, K. (2013). Arabizi Detection and Conversion to Arabic. *CoRR*.

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. T., Al-Badrashiny, M., Aminian, M., Attia, M., Elfardy, H., Habash, N., Hawwari, A., Salloum, W., Dasigi, P., and Eskander, R. (2014). Tharwa: A large scale dialectal arabic-standard arabic-english lexicon. In *LREC*, pages 3782–3789.

Eskander, R., Habash, N., and Rambow, O. (2013). Automatic Extraction of Morphological Lexicons from Morphologically Annotated Corpora. In *Proceedings of tenth Conference on Empirical Methods in Natural Language Processing*.

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 the 44th Meeting of the Association for Computational Linguistics (ACL'06)*, Sydney, Australia.

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

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., 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). Con-

ventional Orthography for Dialectal Arabic. In *LREC*, pages 711–718.

Habash, N., Roth, R., Rambow, O., Eskander, R., and Tomeh, N. (2013). Morphological Analysis and Disambiguation for Dialectal Arabic. In *Proceedings of the 2013 Conference of NAACL-HLT*, Atlanta, GA.

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

Holes, C. (1990). *Gulf Arabic*. Psychology Press.

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

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. and Cieri, C. (2002). Resources for Arabic Natural Language Processing. In *International Symposium on Processing Arabic*, volume 1.

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, LRECâĂŹ06*.

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).

MacWhinney, B. (2000). The childes project: Tools for analyzing talk: Volume i: Transcription format and programs, volume ii: The database. *Computational Linguistics*, 26(4):657–657.

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).

Ntelitheos, D. and Idrissi, A. (Forthcoming). Language Growth in Child Emirati Arabic. *Hamid Ouali (ed.) Perspectives on Arabic Linguistics 29*.

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.

Qafisheh, H. A. (1977). A Short Reference Grammar of Gulf Arabic.

Saadane, H. and Habash, N. (2015). A Conventional Orthography for Algerian Arabic. In *ANLP Workshop 2015*, page 69.

Salloum, W. and Habash, N. (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.

Sarnākh. (2014). *Sukayk 1: riwāyah emārātiyya maḥaliyyah sākhirah*. Midad Publishing & Distribution.

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

Turki, H., Adel, E., Daouda, T., and Regragui, N. (2016). A Conventional Orthography for Maghrebi Arabic. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*, Portorož, Slovenia.

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

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. Association for Computational Linguistics.

Zribi, I., Boujelbane, R., Masmoudi, A., Ellouze, M., Belguith, L., and Habash, N. (2014). A Conventional Orthography for Tunisian Arabic. In *Proceedings of the Language Resources and Evaluation Conference (LREC), Reykjavik, Iceland*.