Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 2345–2352
Marseille, 20-25 June 2022

The Bahrain Corpus: A Multi-genre Corpus of Bahraini Arabic

Dana Abdulrahim, Go Inoue[†], Latifa Shamsan, Salam Khalifa[‡], Nizar Habash[†]
University of Bahrain, [†]New York University Abu Dhabi, [‡]Stony Brook University
{darahim,lshamsan}@uob.edu.bh
{go.inoue,nizar.habash}@nyu.edu
salam.khalifa@stonybrook.edu

Abstract

In recent years, the focus on developing natural language processing (NLP) tools for Arabic has shifted from Modern Standard Arabic to various Arabic dialects. Various corpora of various sizes and representing different genres, have been created for a number of Arabic dialects. As far as Gulf Arabic is concerned, Gumar Corpus (Khalifa et al., 2016) is the largest corpus, to date, that includes data representing the dialectal Arabic of the six Gulf Cooperation Council countries (Bahrain, Kuwait, Saudi Arabia, Qatar, United Arab Emirates, and Oman), particularly in the genre of "online forum novels". In this paper, we present the Bahrain Corpus. Our objective is to create a specialized corpus of the Bahraini Arabic dialect, which includes written texts as well as transcripts of audio files, belonging to a different genre (folktales, comedy shows, plays, cooking shows, etc.). The corpus comprises 620K words, carefully curated. We provide automatic morphological annotations of the full corpus using state-of-the-art morphosyntactic disambiguation for Gulf Arabic. We validate the quality of the annotations on a 7.6K word sample. We plan to make the annotated sample as well as the full corpus publicly available to support researchers interested in Arabic NLP.

Keywords: Arabic, Gulf Arabic, dialect, morphology, corpus

1. Introduction

There is an abundance of corpora of various sizes, types, and representing various genres, that cater primarily to Modern Standard Arabic (MSA). Dialectal Arabic (DA), on the other hand, has only recently received attention in the field of corpus linguistics, in particular, and the field of natural language processing (NLP), in general.

According to Habash (2021), there are numerous challenges that face NLP researchers of DA. Most noteworthy is the large number of forms of Arabic words, resulting from the rich morphology of Arabic language (Holes, 2004). Lexical items in Arabic can be marked for different inflectional features such as number, person, gender, aspect, mood, case, voice, etc. All of that in addition to pronouns and particles (e.g. prepositions, conjunctions, the definite article, etc.) that can attach to the lexical item. Dialects of Arabic do not all share the same exact set of inflectional affixes or clitics (neither with each other, nor with MSA). Due to the significant differences that exist between MSA and DA, MSA processing tools that have been previously developed have proved to be quite insufficient in the attempt of processing DA data (e.g. Khalifa et al. (2016)). Additionally, lack of a standard orthography for DA results in inconsistent spelling of lexical items in DA (Habash, 2021). For instance, speakers of Bahraini Arabic would provide the following different spellings for the phrase مسوية لك *mswyp lk*[1] 'I (fem) have made for you (masc)': مسويت لك *mswyt lk*, مسويه لك *mswyh lk*, مسويتلك *mswytlk*, or even امسوية لك *Amswyp lk*. These orthographic inconsistencies have always posed a great challenge for corpus developers interested in Arabic dialects. One of the measures taken by NLP researches of DA is to propose a Conventional Orthography for Dialectal Arabic, or CODA (Habash et al., 2012) as a means of providing researchers in the field with a common convention for writing in DA.

These given challenges have not entirely dissuaded Arabic NLP researchers from creating various DA corpora of different sizes (see §2). As far as Gulf Arabic (GLF) is concerned, there are few corpora available to date that represent some GLF dialect. One of the largest GLF corpora is the Gumar Corpus (Khalifa et al., 2016), which is a large-scale corpus containing texts from various dialects used in the Arabian Gulf (Bahraini, Emarati, Kuwaiti, Omani, Qatari, and Saudi). The Gumar Corpus exclusively consists of anonymously published online forum novels, that are highly conversational in style, which limits the genre of texts available at the corpus to written novels only.

It is without a doubt that a huge corpus, like the Gumar Corpus, offers innumerable benefits for quantitative and computational linguistic research. However, the lack of balance with respect to the different genres and the different registers of dialectal Arabic is still an issue that needs to be addressed. In the Bahrain Corpus, we aimed to create a more balanced corpus, depicting Bahraini Arabic as present in both spoken and written mediums. More importantly, this independent corpus of Bahraini Arabic aspires to represent not only the linguistic diversity found in Bahrain (see §3), but also the cultural diversity that distinguishes this country. The

[1] Arabic transliteration is in the Buckwalter transliteration scheme (Buckwalter, 2002).

texts that comprise this corpus (whether originally written, or transcripts of TV shows) vary from a retelling of folktales, songs, recipes of traditional Bahraini dishes, TV series reenacting stories from older times and older ways of speaking, interviews with prominent Bahraini personalities, as well as texts that include social commentary. At the time of this publication, the corpus comprises 620K words, carefully curated. In sum, this project is an attempt at documenting the Bahraini dialect, thought, and culture.

We enrich the Bahrain Corpus text with automatic morphological annotations using state-of-the-art morphosyntactic disambiguation for Gulf Arabic (Inoue et al., 2022). We validate the quality of the annotations on a 7.6K word sample. We make the full corpus as well as the annotated sample publicly available to support researchers interested in Arabic NLP.[2]

The rest of the paper is organized as follows. In §2, we present related work, followed by a description of Bahraini Arabic in §3. We describe the details of our corpus creation in §4. In §5, we discuss the annotation process and evaluation. We conclude and discuss future directions in §6.

2. Related Work

There have been many efforts on the development of Standard Arabic text corpora, whether raw or annotated, across different genres, and different variants (Modern and Classical) (Maamouri and Cieri, 2002; Maamouri et al., 2004; Smrž and Hajič, 2006; Habash and Roth, 2009; Belinkov et al., 2016; Taji et al., 2017; Altammami et al., 2019; Al-Ghamdi et al., 2021; Habash and Palfreyman, 2022; Habash et al., 2022, among others).

Dialectal Arabic, however, did not receive as much attention until recently with the increased use of social media. Even though there is an abundance of dialectal text, text corpora are relatively scarce when compared to MSA. There are several (some large-scale) dialectal corpora but their annotation are usually minimal. For example, Gumar (Khalifa et al., 2016) is a large-scale corpus of Gulf Arabic that contains raw text from six different Gulf dialects with a minority of other dialects. Gumar is manually annotated for dialectal information on the document level. Shami (Abu Kwaik et al., 2018) is a corpus of Levantine Arabic dialects. NADI (Abdul-Mageed et al., 2020) covers more dialects, but it is only automatically annotated for dialectal information. Such corpora usually target multiple dialects at a time while providing minimal dialectal information.

In contrast, there are corpora that focus on parallelism across the dialects (Bouamor et al., 2014; Meftouh et al., 2015; Bouamor et al., 2018). MADAR (Bouamor et al., 2018) for instance was carefully curated to covers 25 major Arab cities' dialects but it is small in size when compared to Gumar. For an extensive listing and comparison of these and similar corpora, see Baimukan et al. (2022).

Finally, there is a number of manually annotated corpora that target specific dialects in terms of morphosyntactic tagging. Early efforts included the Levantine Arabic Treebank (specifically Jordanian) (Maamouri et al., 2006), the Egyptian Arabic Treebank (Maamouri et al., 2014), and Curras, the Palestinian Arabic annotated corpus (Jarrar et al., 2014). More recently, Al-Shargi et al. (2016), Alshargi et al. (2019), and Darwish et al. (2018) provided morphologically annotated corpora for a number of dialects with varying levels of annotation. Gulf Arabic in particular has very few curated and annotated corpora. Khalifa et al. (2018) presented a morphologically annotated corpus of Emirati Arabic. The corpus contains about 200,000 words from eight novels extracted from the Gumar corpus. Furthermore, Al-Twairesh et al. (2018) presented SUAR, a Saudi Arabic corpus that has been partially annotated for morphology.

To the best of our knowledge, the Bahrain Corpus is the first of its kind to target Bahraini Arabic in particular. The corpus is carefully curated to cover multiple genres. Additionally, a portion of the corpus is manually annotated for morphological features.

3. Bahraini Arabic

The Kingdom of Bahrain is a small archipelago consisting of 33 islands, with an area size of around 780 square kilometers. It is situated on the northeastern coast of Saudi Arabia, and is connected to this part of Saudi Arabia by the 25-kilometer King Fahad Causeway. The population of Bahrain is estimated at 1.5 million, and it comprises 45% of Bahraini citizens (from various ethnic backgrounds), and 55% of non-Bahraini residents.

This diversity of the population in Bahrain, both local and non-local, is clearly reflected in its rich historical and cultural heritage. Naturally, the Bahraini speech community also reflects this diversity, which is apparent in the existence of two main linguistic varieties in Bahrain. The first unmarked variety of Bahraini Arabic shares features with other Gulf Arabic dialects spoken in some eastern parts of Saudi Arabia, as well as Kuwait and Qatar; while the second variety (Baḥrāni Arabic) is also spoken in Al-Hasa, and Al-Qatif in Saudi Arabia, as well as some parts of Oman.

Various studies have previously addressed either or both varieties (Al-Tajir, 1982; Holes, 1983; Holes, 1987; Holes, 2001; Holes, 2003; Holes, 2006; Prochazka, 1981, among others). And many have noted the differences between them with respect to their phonology, morphology, as well as lexical characteristics (Holes, 2006, among others).

It is worth noting that due to the multiple ethnicities that comprise the Bahraini speech community, there do exist sub-varieties (or accents) even within each main variety. For instance, a speaker of the first variety of

[2] http://www.bahraincorpus.com

Bahraini Arabic, can usually identify individuals who come from the city of Muharraq, as opposed to speakers who come from the capital city Manama. On the other hand, speakers of Baḥrāni Arabic can identify individuals from Sitra, as opposed to, also, Baḥrāni speakers living in Manama.

Phonological features are the primary distinguishing factors of Bahraini Arabic from the rest of Gulf Arabic dialects. One example is the use of the emphatic, open back vowel /ɑ/ as a variant of the open front vowel /a/, which is more generally used in other dialects of the Gulf, particularly in Saudi Arabia and Kuwait. An example would be saying /tɛʕbɑn/, instead of /tɛʕban/ for the word تعبان *tEbAn* 'tired'. This vowel appears to be strongly marked, and is a way of identifying a Bahraini speaker of Arabic by other GCC citizens. Admittedly, in certain speech communities, especially in the island of Muharraq, /ɑ/ can show up as the open-mid back rounded vowel /ɔ/ (e.g. tɛʕbɔn/).

In this paper, we use the term 'Bahraini Arabic' to refer to all Arabic varieties used in Bahrain, and to refer to all the spoken and written data that was collected for this corpus (see §4). Sub-dialect annotations and phonological representations are outside of the scope of the presented work in this paper.

4. Corpus Description

In this section we discuss the corpus creation and text collection efforts as well as the automatic morphological annotations we used.

4.1. Corpus Creation

As mentioned earlier, our aim is to create a corpus that is somewhat balanced. No specific variety of Bahraini Arabic was exclusively targeted in the process of data collection, since the main concern was to get whatever written texts and transcripts of spoken data that were available to us. Transcriptions of spoken data accounts for 66.0% of the entire corpus, while the remaining consists of originally written data. In Table 1, we show the overall statistics of our corpus. Words are based on white space tokenization. The number of sentences represents the number of lines.

Spoken The *Spoken* portion of the corpus comprises mainly of transcriptions of YouTube videos. These videos correspond to numerous local TV shows, such as comedy shows, drama, interviews, talk shows, cooking shows, plays, etc., some of which are scripted while the others are unscripted. Table 2 shows the different genres of these videos with representative examples.

The process of transcribing spoken data (into written texts) started in the year 2016. These transcriptions were produced by Bahraini college students either on a volunteer basis, or as part of the requirements for a linguistics course. In addition, the students were made aware of the importance of building a dialectal corpus of Bahraini Arabic as a means of documenting the language, and providing indispensable resources for language researchers, among other objectives.

The Bahrain Corpus	
#Word	620,301
#Unique Word	59,846
#Sentence	46,030
#Document	112

Table 1: Statistics on the Bahrain Corpus.

Transcribers were instructed to simply convert what they hear on the videos to plain text, without providing any further annotations, including false starts, hesitations, and fillers. They were also introduced to CODA (Habash et al., 2012; Habash et al., 2018) and encouraged to follow them while transcribing the videos. Transcribers were most importantly instructed to only use the Arabic script letters used in MSA, and avoid using Arabic script letters used for other languages such as پ /p/ or گ /g/, as per the CODA guidelines.[3]

Written Table 2 shows the different sources from which the written corpus data was collected. 92.6% of the text was obtained from the portions of Gumar corpus tagged as Bahraini Arabic (Khalifa et al., 2016). As such, genre-wise, they are online forum novels. Also, 4.7% of the written data comes from a folktale collection, narrated in Bahraini Arabic, and collected by the Bahraini author, Dr. Anisa Fakhroo (Fakhroo, 2019). The remaining part of the written data comprises short texts that span different genres (proverbs, parables, stories, jokes, cooking recipes, advice, etc.), which we collected from over 180 Bahraini individuals.

4.2. Automatic Morphological Annotation

To automatically annotate our corpus, we use the BERT-based morphosyntactic tagger used in Inoue et al. (2022). Their system for GLF is trained on the annotated portion of the Gumar corpus (Khalifa et al., 2018), where they fine-tune the CAMeLBERT-Mix pre-trained language model (Inoue et al., 2021) for the morphosyntactic tagging task. Following their recommendation, we use the GLF unfactored model without the morphological analyzer to obtain morphological feature predictions. For lemmas, we use the model with the analyzer.

We present below all of the annotations that we produce automatically. In the next section, we discuss validating the results using a representative sample from the corpus.

Orthography The current system from Inoue et al. (2022) assumes the input to be in CODA. This is clearly a limitation whose effect we evaluate in §5.4 by comparing the performance on raw input.

Morphology The current system from Inoue et al. (2022) provides the following features automatically.

[3] http://coda.camel-lab.com

Spoken (66.0% of the words in the corpus)		
Genre	**#Word (%)**	**Example Document**
drama	128,439 (31.4%)	مسلسل البيت العود، حلقة ١٢ The Big House TV show, episode 12
interview	103,889 (25.4%)	برنامج وطني، مقابلة مع الممثل محمد ياسين Watani TV show, interview with the actor Mohammed Yasin
comedy	62,995 (15.4%)	مسلسل سوالف طفاش، الجزء الثاني حلقة ١٣ Tafash Stories TV show, season 2 episode 13
play	60,150 (14.7%)	مسرحية بيت خاص جدا Very Special House play
monologue	29,095 (7.1%)	برنامج فنيال قهوة مع سناء السعد، حلقة الاشاعات Cup of Coffee TV show with Sanaa Alsaad, Rumors episode
cooking	16,271 (4.0%)	برنامج المطبخ، حلقة برياني الدجاج The Kitchen TV show, chicken biryani episode
reality	5,811 (1.4%)	برنامج حياة خوات، الحلقة الأولى Sisters' lives Reality TV show, episode 1
cartoon	2,862 (0.7%)	برنامج بو جليع، موسم ٢٠١٥ حلقة ٢٢ Bu Jlee'a TV show, season 2015 episode 22
Total	409,512	

Written (34.0% of the words in the corpus)		
Genre	**#Word (%)**	**Example Document**
forum novels	195,132 (92.6%)	قصة مريم عيون سلمان Maryam in the Eyes of Salman novel
folktales	9,897 (4.7%)	حزاوي أمي العودة ـ د. أنيسة فخرو Bahraini Folktales - Dr Anisa Fakhroo
mix	5,760 (2.7%)	قصص وحكم ونكات ووصفات شعبية stories, parables, jokes, traditional recipes
Total	210,789	

Table 2: The genre distribution of the Bahrain Corpus.

- **Lemma**: The lemma is an abstraction over the various inflectional forms of a particular lexical item with a specific derivation. Inoue et al. (2022)'s system follows the common conventions used for Arabic, such as using the 3rd person singular perfective for verbs, and the singular masculine or feminine (if no masculine form exists) for nouns and adjectives.

- **Core POS**: The part-of-speech tagset consists of the following 35 tags:[4] abbrev, adj, adj_comp, adj_num, adv, adv_interrog, adv_rel, conj, conj_sub, digit, interj, latin, noun, noun_num, noun_prop, noun_quant, part, part_det, part_focus, part_fut, part_interrog, part_neg, part_restrict, part_verb, part_voc, prep, pron, pron_dem, pron_exclam, pron_interrog, pron_rel, punc, verb, verb_nom, and verb_pseudo.

- **PAGN (Person, Aspect, Gender, Number)**: Inflectional features consist of person (*per*:{1, 2, 3}), aspect (*asp*:{**i**mperfective, **p**erfective, **c**ommand}), gender (*gen*:{**f**eminine, **m**asculine}), and number (*num*:{**s**ingular, **d**ual, **plural**}).

- **Clitics**: All proclitics and enclitics are specified using their form and also their own clitic POS. There are in total 38 proclitics (*prc3, prc2, prc1, prc0*) and 34 enclitics (*enc0*).[5] Examples of proclitics include **l_prep** (the ل *l* preposition 'for/to') and **w_conj** (the و *w* conjunction 'and'). Examples of enclitics include direct object pronouns and possessive pronouns, e.g. **2fs_dobj** for the $2nd$ feminine singular direct objective pronoun clitic.

Table 3 presents an annotated example.

[4]For more details on the POS guidelines, see `http://guidelines.camel-lab.com`

[5]The number following the clitic feature indicates its relative distance from the base word.

					Clitics				PAGN				Clitics	
Word		**CODA**		**Lemma**	**POS**	**prc3**	**prc2**	**prc1**	**prc0**	**per**	**asp**	**gen**	**num**	**enc0**	***English***
yA	يا	yA	يا	يا	part_voc	0	0	0	0	na	na	na	na	0	O!
bnyty	بنيتي	bnyty	بنيتي	بنية	noun	0	0	0	0	na	na	f	s	1s_poss	my girl
:	:	:	:	:	punc	0	0	0	0	na	na	na	na	0	:
<*A	إذا	A*A	اذا	اذا	conj_sub	0	0	0	0	na	na	na	na	0	if
ywEtj	يوعتج	jwEtj	جوعتج	جوع	verb	0	0	0	0	3	i	f	s	2fs_dobj	she makes you hungry
mrt	مرت	mrp	مرة	مرة	noun	0	0	0	0	na	na	f	s	0	wife of
>bwj	أبوج	Abwj	ابوج	اب	noun	0	0	0	0	na	na	m	s	2fs_poss	your father
،	،	،	،	،	punc	0	0	0	0	na	na	na	na	0	,
|nh	آنه	AnA	انا	انا	pron	0	0	0	0	1	na	u	s	0	I
b|gnyj	بآغنيج	bAgnyj	باغنيج	اغنى	verb	0	0	b_fut	0	1	i	u	s	2fs_dobj	will make you rich
wb|ETyj	وبآعطيج	wbAETyj	وباعطيج	عطى	verb	0	w_conj	b_fut	0	1	i	u	s	2fs_dobj	and will give you
>kl	أكل	Akl	اكل	اكل	noun	0	0	0	0	na	na	m	s	0	food
wAyd	وايد	wAjd	واجد	واجد	adj	0	0	0	0	na	na	m	s	0	a lot

Table 3: Example of manual annotation. Columns represent features to be annotated and rows represent words.

Manual Annotation	
#Word	7,609
#Word (CODA)	7,680
#Sentence	1,141

Table 4: Statistics of the manually annotated portion of the Bahrain Corpus.

5. Evaluation of Morphological Annotation

In this section, we present a small study where we manually annotate a portion of the corpus. This annotation allows us to investigate the performance of the state-of-the-art morphological disambiguation systems on Bahraini Arabic and whether there is a need to develop corpora and tools that are specific to Bahraini Arabic.

5.1. Data Selection

We extracted a representative sample of 7,609 words from six texts: transcripts of two monologues, a comedy show, and a drama show (spoken); as well as extracts from the published Bahraini folktales (Fakhroo, 2019), and the body of mixed texts collected by the authors (written). We did not include forum novels in the sample because the tool from Inoue et al. (2022) was trained on data from the same genre (Khalifa et al., 2018), and we were concerned with its performance on the texts from the other genres, which were curated specifically as part of this corpus. As part of preparing the sample texts for morphological annotation, we split paragraphs into complete sentences, or independent standalone phrases and utterances.
Statistics on the annotated sample are in Table 4.

5.2. Annotation Guidelines

Orthography For orthographic annotation, we followed the CODA Star guidelines (Habash et al., 2018). This task includes spelling changes as well as word splits and merges. We report on the degree of spelling change due to CODA and evaluate morphological annotation accuracy with raw input and CODA input in §5.4. See Table 3 for an example sentence with raw and CODA annotations

Morphology For morphological annotation, we followed the annotation guidelines used in the Gumar Annotated Corpus (Khalifa et al., 2018), and the LDC's Egyptian Arabic guidelines (Maamouri et al., 2012) which Khalifa et al. (2018) also referred to. However, we chose to represent the annotations in terms of feature-value pairs as used in a number of Arabic NLP systems (Pasha et al., 2014; Obeid et al., 2020) including the system by Inoue et al. (2022), which we used for automatic annotation. See Table 3 for an example sentence with full morphological annotation.

5.3. Annotation Process

The automatic annotation was subsequently and jointly evaluated by the two Bahraini authors of the paper (1st and 3rd authors). All automatically-generated annotations were inspected and corrected, if needed. In addition, a CODA compliant spelling was supplied for lexical items with alternative spellings.

5.4. Evaluation

We evaluate a number of automatic morphosyntactic taggers (Inoue et al., 2022) on the manually annotated portion of our corpus.

Experimental Settings We report on using the state-of-the-art BERT-based morphosyntactic taggers (Inoue et al., 2022) for GLF, as well as other variants of Arabic, i.e., MSA, Egyptian (EGY), and Levantine (LEV). We use the best setup for each variant: For GLF, we use the unfactored model without a morphological analyzer to obtain morphological feature predictions, and the model with an analyzer for lemma predictions. For MSA and EGY, we use the factored model with a morphological analyzer. For LEV, we use the factored

	Raw				CODA			
	GLF	**MSA**	**EGY**	**LEV**	**GLF**	**MSA**	**EGY**	**LEV**
Lemma	**79.9%**	72.6%	76.6%	73.8%	**83.1%**	76.3%	79.7%	77.1%
POS	**90.6%**	74.4%	79.7%	87.9%	**92.2%**	76.0%	81.8%	89.6%
PAGN	**85.3%**	73.0%	75.6%	75.4%	**87.2%**	74.7%	77.2%	76.3%
Clitics	**93.1%**	89.3%	89.5%	90.8%	**94.9%**	90.9%	91.0%	92.2%

Table 5: Results of the CAMeLBERT morphological disambiguator model in GLF, MSA, EGY, and LEV on the manually annotated portion of our corpus.

model without an analyzer for morphological features, and the one with an analyzer for lemmas. For preprocessing, we remove diacritics and Tatweel/Kashida using CAMeL Tools (Obeid et al., 2020).

Evaluation Metric We report the accuracy in terms of the following metrics:

- **Lemma**: The accuracy of the dediacritized lemma choice.
- **POS**: The accuracy of the core POS.
- **PAGN**: The accuracy of the core inflectional features: person (*per*), aspect (*asp*), gender (*gen*), and number (*num*).
- **Clitics**: The accuracy of all proclitics (*prc3, prc2, prc1, prc0*) and enclitics (*enc0*).[6]

We also evaluate using two versions of the text: raw and CODA to measure the effect of orthographic changes. In the raw evaluation, we consider predictions for the raw words involved in merge or split in CODA as *incorrect*, since our morphological annotation is done on the CODA space.

Results Table 5 shows the morphological annotation results for the different models on the manually annotated corpus in both raw and CODA.

Table 6 presents the number and types of changes done in CODAfying the raw text. Over 91.3% of the words are left unchanged; and about 1% of the words went through a split/merge operation.

In terms of overall performance, the GLF model consistently performs the best in all the metrics for both raw and CODA text. We observe that lemma identification is the most challenging task, followed by PAGN, POS, and Clitics. On average, the GLF model performs better than the other three models by 5.6% in lemma prediction, 10.0% in POS, 10.7% in PAGN, and 3.2% in Clitics in the raw text evaluation. This validates our choice of the GLF model over the other models trained on different variants.

To measure the effect of orthographic normalization on the performance of the morphosyntactic tagger, we compare the results on the raw and CODAfied text. We observe that the improvement in performance due to CODAfication is 3.2% in lemma, 1.6% in POS, 1.9% in PAGN, and 1.8% in Clitics.

When we look at the performance difference of the GLF unfactored model on the Gumar corpus (Khalifa et al., 2018) and our dataset, the same model achieves 97.8% accuracy in POS tagging on the Gumar corpus (Inoue et al., 2022), while it achieves 92.2% on our dataset (with CODA input). While these results allow a more robust evaluation on the performance of automatic taggers for GLF Arabic, they also suggest that our dataset is more challenging than the Gumar corpus dataset. This provides further motivation for NLP researchers to fine-tune the current models with more dialect-specific uses from other variants of GLF Arabic.

Orthographic Change	**#Word (%)**
Split	73 (1.0%)
Merge	2 (0.0%)
Substitute	665 (8.7%)
Unchanged	6,944 (91.3%)

Table 6: The number of non-CODA words involved in orthographic changes.

6. Conclusion and Future Work

We presented a carefully curated 620K word corpus of Bahraini Arabic that includes written texts, as well as transcripts of spoken data, belonging to a different genre (folktales, comedy shows, plays, cooking shows, etc.). We provided automatic morphological annotations of the full corpus using state-of-the-art morphosyntactic disambiguation for Gulf Arabic, and validated the quality of these annotations. We make the annotated sample as well as the full corpus publicly available to support researchers interested in Arabic NLP.

In the future, we plan to continue to expand the corpus with more materials, more metadata, and richer annotations. We will continue to update its automatic annotations as better systems for Gulf morphosyntactic disambiguation become available. We also plan to study the distributional vocabulary differences among texts from different genres in the corpus. We will explore the use of the Bahrain Corpus to assist natural language processing tasks such as speech recognition and dialect identification.

[6]We remove the diacritics from the lexical parts of the proclitic features to increase matchability with the simplifying conventions used in the GLF annotation. For example, the conjunction + و *w+* 'and' is specified as *wa_conj* in MSA and *wi_conj* in EGY; we map both to *w_conj*.

Acknowledgments

We would like to thank all of the students and volunteers who helped us create this resource over the last few years. We would like to especially acknowledge the generous contribution of Dr. Anisa Fakhroo who provided the authors with one of her latest publications on Bahraini folktales to be included in the Bahrain Corpus. Part of this work was carried out on the High Performance Computing resources at New York University Abu Dhabi. We thank the anonymous reviewers for their insightful suggestions and comments.

7. Bibliographical References

Abdul-Mageed, M., Zhang, C., Bouamor, H., and Habash, N. (2020). NADI 2020: The first nuanced Arabic dialect identification shared task. In *Proceedings of the Fifth Arabic Natural Language Processing Workshop*, pages 97–110, Barcelona, Spain (Online), December. Association for Computational Linguistics.

Abu Kwaik, K., Saad, M., Chatzikyriakidis, S., and Dobnik, S. (2018). Shami: A corpus of Levantine Arabic dialects. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Al-Ghamdi, S., Al-Khalifa, H., and Al-Salman, A. (2021). A dependency treebank for classical Arabic poetry. In *Proceedings of the Sixth International Conference on Dependency Linguistics (Depling, SyntaxFest 2021)*, pages 1–9, Sofia, Bulgaria, December. Association for Computational Linguistics.

Al-Shargi, F., Kaplan, A., Eskander, R., Habash, N., and Rambow, O. (2016). Morphologically Annotated Corpus and a Morphological Analyzer for Moroccan and San'ani Yemeni Arabic. In *Proceedings of the Language Resources and Evaluation Conference (LREC)*, Portorož, Slovenia.

Al-Tajir, M. (1982). *Language and linguistic origins in Baḥrain: the Baḥārna dialect of Arabic*. Kegan Paul International.

Al-Twairesh, N., Al-Matham, R., Madi, N., Almugren, N., Al-Aljmi, A.-H., Alshalan, S., Alshalan, R., Alrumayyan, N., Al-Manea, S., Bawazeer, S., Al-Mutlaq, N., Almanea, N., Huwaymil, W. B., Alqusair, D., Alotaibi, R., Al-Senaydi, S., and Alfutamani, A. (2018). SUAR: Towards building a corpus for the Saudi dialect. In *Proceedings of the International Conference on Arabic Computational Linguistics (ACLing)*.

Alshargi, F., Dibas, S., Alkhereyf, S., Faraj, R., Abdulkareem, B., Yagi, S., Kacha, O., Habash, N., and Rambow, O. (2019). Morphologically annotated corpora for seven Arabic dialects: Taizi, Sanaani, Najdi, Jordanian, Syrian, Iraqi and Moroccan. In *Proceedings of the Fourth Arabic Natural Language Processing Workshop*, pages 137–147, Florence, Italy, August. Association for Computational Linguistics.

Altammami, S., Atwell, E., and Alsalka, A. (2019). The Arabic–English parallel corpus of authentic hadith. *International Journal on Islamic Applications in Computer Science And Technology-IJASAT*.

Baimukan, N., Habash, N., and Bouamor, H. (2022). Hierarchical aggregation of dialectal data for Arabic dialect identification. In *Proceedings of the Language Resources and Evaluation Conference (LREC)*, Marseille, France. The European Language Resources Association.

Belinkov, Y., Magidow, A., Romanov, M., Shmidman, A., and Koppel, M. (2016). Shamela: A large-scale historical Arabic corpus. In *Proceedings of the Workshop on Language Technology Resources and Tools for Digital Humanities (LT4DH)*, pages 45–53, Osaka, Japan, December. The COLING 2016 Organizing Committee.

Bouamor, H., Habash, N., and Oflazer, K. (2014). A multidialectal parallel corpus of Arabic. In *Proceedings of the Language Resources and Evaluation Conference (LREC)*, Reykjavik, Iceland.

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 Language Resources and Evaluation Conference (LREC)*, Miyazaki, Japan.

Buckwalter, T. (2002). Buckwalter Arabic morphological analyzer version 1.0. Linguistic Data Consortium (LDC) catalog number LDC2002L49, ISBN 1-58563-257-0.

Darwish, K., Mubarak, H., Abdelali, A., Eldesouki, M., Samih, Y., Alharbi, R., Attia, M., Magdy, W., and Kallmeyer, L. (2018). Multi-dialect Arabic pos tagging: A CRF approach. In *Proceedings of the Language Resources and Evaluation Conference (LREC)*, Miyazaki, Japan.

Fakhroo, A. (2019). *Hazawi Ummi Al'ooda*. That Al-Salaasel.

Habash, N. and Palfreyman, D. (2022). ZAEBUC: An annotated Arabic-English bilingual writer corpus. In *Proceedings of the Language Resources and Evaluation Conference (LREC)*, Marseille, France.

Habash, N. and Roth, R. (2009). CATiB: The Columbia Arabic Treebank. In *Proceedings of the Joint Conference of the Association for Computational Linguistics and the International Joint Conference on Natural Language Processing (ACL-IJCNLP)*, pages 221–224, Suntec, Singapore.

Habash, N., Diab, M., and Rambow, O. (2012). Conventional Orthography for Dialectal Arabic. In *Proceedings of the Language Resources and Evaluation Conference (LREC)*, pages 711–718, Istanbul, Turkey.

Habash, N., Eryani, F., Khalifa, S., 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 Language Resources and Evaluation Conference (LREC)*, Miyazaki, Japan.

Habash, N., AbuOdeh, M., Taji, D., Faraj, R., Gizuli, J. E., and Kallas, O. (2022). Camel Treebank: An open multi-genre Arabic dependency treebank. In *Proceedings of the Language Resources and Evaluation Conference (LREC)*, Marseille, France.

Habash, N. (2021). Arabic computational linguistics. In Karin Ryding et al., editors, *The Cambridge Handbook of Arabic Linguistics*, chapter 18, pages 427–445. Cambridge University Press, Cambridge.

Holes, C. (1983). Bahraini dialects: sectarian differences and the sedentary/nomadic split. *Zeitschrift für arabische Linguistik*, pages 7–38.

Holes, C. (1987). *Language variation and change in a modernising Arab state: The case of Bahrain*, volume 7. Taylor & Francis.

Holes, C. (2001). *Dialect, culture, and society in Eastern Arabia, Glossary*. Brill (Leiden [ua]).

Holes, C. (2003). *Colloquial Arabic of the Gulf*. Routledge.

Holes, C. (2004). *Modern Arabic: Structures, Functions, and Varieties*. Georgetown Classics in Arabic Language and Linguistics. Georgetown University Press.

Holes, C. (2006). Bahraini Arabic. In Kees Versteegh, et al., editors, *Encyclopedia of Arabic Language and Linguistics*, pages 241–255. Brill, Leiden.

Inoue, G., Alhafni, B., Baimukan, N., Bouamor, H., and Habash, N. (2021). The interplay of variant, size, and task type in Arabic pre-trained language models. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, pages 92–104, Kyiv, Ukraine (Virtual), April. Association for Computational Linguistics.

Inoue, G., Khalifa, S., and Habash, N. (2022). Morphosyntactic tagging with pre-trained language models for Arabic and its dialects. In *Proceedings of the Findings of the Association for Computational Linguistics: ACL2022*, Dublin, Ireland, May. Association for Computational Linguistics.

Jarrar, M., Habash, N., Akra, D., and Zalmout, N. (2014). Building a Corpus for Palestinian Arabic: A Preliminary Study. In *Proceedings of the Workshop for Arabic Natural Language Processing (WANLP)*, pages 18–27, Doha, Qatar.

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

Khalifa, S., Habash, N., Eryani, F., Obeid, O., Abdulrahim, D., and Kaabi, M. A. (2018). A morphologically annotated corpus of Emirati Arabic. In *Proceedings of the Language Resources and Evaluation Conference (LREC)*, Miyazaki, Japan.

Maamouri, M. and Cieri, C. (2002). Resources for Natural Language Processing at the Linguistic Data Consortium. In *Proceedings of the International Symposium on Processing of Arabic*, pages 125–146, Manouba, Tunisia.

Maamouri, M., Bies, A., Buckwalter, T., and Mekki, W. (2004). The Penn Arabic Treebank: Building a Large-Scale Annotated Arabic Corpus. In *Proceedings of the International 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 Language Resources and Evaluation Conference (LREC)*, Genoa, Italy.

Maamouri, M., Krouna, S., Tabessi, D., Hamrouni, N., and Habash, N. (2012). 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 Language Resources and Evaluation Conference (LREC)*, Reykjavik, Iceland.

Meftouh, K., Harrat, S., Jamoussi, S., Abbas, M., and Smaili, K. (2015). Machine translation experiments on PADIC: A parallel Arabic dialect corpus. In *Proceedings of the Pacific Asia Conference on Language, Information and Computation*.

Obeid, O., Zalmout, N., Khalifa, S., Taji, D., Oudah, M., Alhafni, B., Inoue, G., Eryani, F., Erdmann, A., and Habash, N. (2020). CAMeL tools: An open source python toolkit for Arabic natural language processing. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 7022–7032, Marseille, France, May. European Language Resources Association.

Pasha, A., Al-Badrashiny, M., Diab, M., Kholy, A. E., Eskander, R., Habash, N., Pooleery, M., Rambow, O., and Roth, R. (2014). MADAMIRA: A fast, comprehensive tool for morphological analysis and disambiguation of Arabic. In *Proceedings of the Language Resources and Evaluation Conference (LREC)*, pages 1094–1101, Reykjavik, Iceland.

Prochazka, T. (1981). The Shī'ī dialects of Bahrain and their relationship to the Eastern Arabian dialect of Muḥarraq and the Omani dialect of Al-Ristāq. *Zeitschrift für arabische Linguistik*, pages 16–55.

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.

Taji, D., Habash, N., and Zeman, D. (2017). Universal dependencies for Arabic. In *Proceedings of the Workshop for Arabic Natural Language Processing (WANLP)*, Valencia, Spain.