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

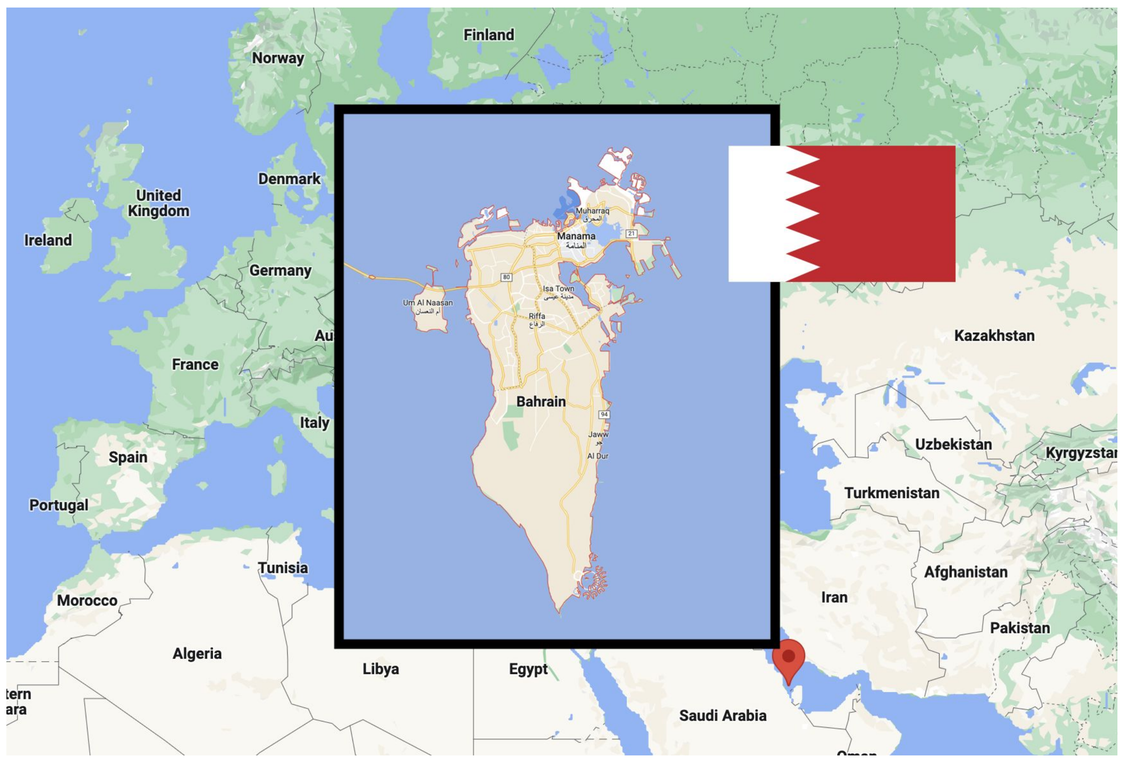
{darahim, lshamsan}@uob.edu.bh, {go.inoue, nizar.habash}@nyu.edu, salam.khalifa@stonybrook.edu

Introduction

- The Bahrain Corpus includes written texts as well as transcripts of audio files, belonging to a different genre (folktales, comedy shows, plays, cooking shows, etc.), comprising **620K words**.
- We provide automatic morphological annotations of the full corpus and validate its quality on a 7.6K word sample.

www.bahraincorpus.com

Kingdom of Bahrain



Corpus Contents

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	

Morphological Annotation

Accuracy of Automatic Annotation

	GLF	MSA	EGY	LEV
Lemma	**79.9%**	72.6%	76.6%	73.8%
POS	**90.6%**	74.4%	79.7%	87.9%
PAGN	**85.3%**	73.0%	75.6%	75.4%
Clitics	**93.1%**	57.1%	56.4%	60.9%

				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