Capturing Intra-Dialectal Variation in Qatari Arabic: A Corpus of Cultural and Gender Dimensions

Houda Bouamor[1], Sara Al-Emadi[2], Zeinab Ibrahim[1], Hany Fazzaa[3], Aisha Al-Sultan[4],

[1]Carnegie Mellon University in Qatar, [2] Hamad Bin Khalifa University,
[3] Georgetown University in Qatar, [4] Doha International Family Institute

hbouamor@cmu.edu, saal68500@hbku.edu.qa,
zmai22023@gmail.com, hf194@georgetown.edu

Abstract

We present the first publicly available, multi-dimensional corpus of Qatari Arabic, which captures intra-dialectal variation across Urban and Bedouin speakers. Although often grouped under the label "Gulf Arabic", Qatari Arabic exhibits rich phonological, lexical, and discourse-level differences shaped by gender, age, and sociocultural identity. Our dataset includes aligned speech and transcriptions from 255 speakers, stratified by gender and age, and collected through structured interviews on culturally important topics such as education, heritage, and social norms. The corpus reveals systematic variation in pronunciation, vocabulary, and narrative style, offering insights for both sociolinguistic analysis and computational modeling. We also demonstrate its utility through preliminary experiments in predicting dialects and genders. This work provides the first large-scale, demographically balanced corpus of Qatari Arabic, laying a foundation for both sociolinguistic research and the development of dialect-aware NLP systems.

1 Introduction

The linguistic landscape of Qatar has often been described in fragmented sources, with few comprehensive accounts capturing its internal diversity. Qatari Arabic is typically grouped under the broader "Gulf Arabic" category (Habash, 2010; Shockley, 2020), a generalization that overlooks meaningful intra-dialectal distinctions shaped by tribal, historical, and sociocultural factors. In practice, Qatari Arabic comprises a continuum of speech varieties, particularly those associated with Urban and Bedouin communities. These groups differ in pronunciation, vocabulary, and grammar, reflecting both inherited traditions and modern influences. Bedouin speakers tend to preserve conservative linguistic forms tied to tribal heritage, while Urban speakers, more exposed to education, media, and globalization, exhibit more borrowing and code-switching (al Sharekh and Freer, 2021; Theodoropoulou and Borresly, 2025). Qatari Arabic also diverges from neighboring Gulf dialects through distinct lexical items (e.g., صب /Sb/ "to pour" vs. MSA سكب) and includes borrowings from Turkish, Farsi, Hindi, and English, due to Qatar's trade history and migration patterns (Al-Mulla and Zaghouani, 2020a; Prochazka, 2021). Despite its sociolinguistic richness, Qatari Arabic remains underrepresented in linguistic and NLP research. Most Arabic corpora focus on Modern Standard Arabic (MSA) or broadly defined dialect groups such as Levantine or Gulf Arabic (Zaidan and Callison-Burch, 2014; Khalifa et al., 2016; Bouamor et al., 2018), limiting dialect-specific modeling and analysis in the Qatari context. To address this gap, we present the first publicly available, multimodal corpus capturing intra-dialectal variation in Qatari Arabic. Our corpus includes aligned audio and transcriptions from 255 native speakers, balanced by gender, age, and sociocultural group (Urban vs. Bedouin), who discuss culturally salient topics such as heritage, education, social norms, and national identity. The corpus supports sociolinguistic research, dialect-aware NLP applications, and broader cultural documentation efforts. We also present an analysis of lexical, phonological, and morphosyntactic patterns between groups, highlighting how language reflects gender and cultural identity. Finally, we demonstrate the computational utility of the corpus through two classification tasks: dialect identification and gender prediction, using different models trained on transcribed speech, showing its value for building inclusive Arabic NLP systems.

2 Related Work

The study of Arabic dialects has gained increasing attention, particularly through the development of large-scale corpora. Arabic dialects are often geographically grouped into Maghrebi, Egyptian, Levantine, Gulf, and Iraqi (Zaidan and Callison-Burch, 2011). Zaidan and Callison-Burch (2011) introduced the Arabic Online Commentary (AOC) corpus with texts from Gulf, Egyptian, and Levantine dialects. Similarly, the Shami Dialect Corpus (SDC) covers Jordanian, Palestinian, Syrian, and Lebanese dialects (Kwaik, 2018). Building on these early efforts, subsequent projects focused on larger-scale and more systematically designed resources. The Gumar Corpus (Khalifa et al., 2016) is a large Gulf Arabic dataset comprising over 100 million words from forum novels. DART (Alsarsour et al., 2018) offers a balanced collection of 25,000 tweets across five major dialect groups. The MADAR corpus (Bouamor et al., 2018) spans 25 cities and highlights the diversity within Arabic dialects, while Abdelali et al. (2020) provide a tweet-based dataset covering 18 MENA countries.

Alongside these broad regional corpora, more localized resources have been created to capture finer-grained variation. The Bahrain Corpus (Abdulrahim et al., 2022) features texts and audio transcripts from diverse genres, while Saudi dialect corpora such as SUAR (Al-Twairesh et al., 2018) and SDC (Tarmom et al., 2020) were designed to capture grammatical and morphological features of Saudi Arabic. There have also been efforts to document Algerian intra-dialectal variation (Bougrine et al., 2016, 2017). More recently, the Najdi Arabic Corpus has been introduced as a resource for another underrepresented Gulf variety, providing a systematically collected dataset for Najdi dialect research (Alhedayani, 2025). In contrast, the Qatari dialect has received relatively little attention. Existing resources include a Qatari idioms corpus (Al-Mulla and Zaghouani, 2020b), a corpus derived from television programs (Elmahdy et al., 2014), and oral history recordings related to the oil industry (AlNaama, 2012). Georgetown University in Qatar also developed a phrasebook app covering common Qatari expressions (Georgetown University in Qatar, 2017). Despite these efforts, Qatari Arabic remains underrepresented, with existing datasets limited in scope, genre, and demographic diversity. This lack hinders linguistic analysis, dialectal documentation, and NLP system development. To address these gaps, we present a new Qatari Arabic corpus built from semi-structured interviews, offering rich, culturally grounded, and demographically diverse spoken data.

3 Linguistic Background

Arabic in the Gulf region is far from monolithic. Instead, it encompasses a spectrum of dialects that reflect both deep historical roots and ongoing sociocultural change. Within this context, Gulf Arabic functions as the broader linguistic umbrella, under which more localized varieties, such as Qatari Arabic, develop and diverge.

3.1 Gulf Dialects

Gulf dialects represent a diverse cluster of Arabic varieties spoken across Bahrain, Saudi Arabia, Kuwait, Oman, Qatar, and the UAE, with eight major types identified: Coastal, Najdi, Baāārna, Kuwaiti Arabic, Eastern Arabian, Šawāwī (Omani S type), Gulf Pidgin Arabic, and Gulf koine (Holes, 2018; Skilliter, 1969). Their linguistic background is shaped by deep historical substrates from ancient Mesopotamian and South Arabian sources, alongside continuous contact with Modern South Arabian languages (Davey, 2016; Holes, 2018). Distinctive features include the retention of archaic phonemes such as interdentals and uvulars, complex feminine plural agreement in some varieties, and contact-induced simplification in others (Al-Bohnayyah, 2019; Bakir, 2010). While the region has cultural homogeneity, Gulf Arabic is far from linguistically uniform: dialects differ markedly in phonology, morphology, and lexicon, shaped by geography, social factors, and historical contact with other languages (Khalifa et al., 2016). Sociolinguistic factors, such as age, gender, sect, urbanization, and labor migration, play a major role in dialect variation, convergence, and divergence (al Qenaie, 2011; Holes, 1986). Urbanization has accelerated the development of homogenized varieties such as the Gulf koine, while multilingual labor migration has led to Arabic Gulf Pidgin (Bakir, 2010; Holes, 2018).

3.2 Qatari Dialect

The official language of Qatar is Arabic, and the variety predominantly spoken by Qatari nationals is commonly referred to as Qatari, a localized form of Gulf Arabic or Khaliji (El-Saba, 2016). While often grouped under the broader Gulf Arabic umbrella (Habash, 2010), the Qatari dialect exhibits notable internal variation shaped by historical, tribal, and sociocultural influences. The most salient division is between Urban and Bedouin varieties, which differ in pronunciation, vocabulary, and grammar and are readily recognized by Qatari speakers (Shoufan and Alameri, 2015). [1] Although the terms Urban and Bedouin carry cultural and historical associations, linguistic research employs them as analytical categories that simplify these complex social realities. Dialect affiliation depends not only on a family's tribal origin or historical settlement but also on patterns of migration, education, and social interaction. For instance, some families of Bedouin origin may speak Urban Qatari, reflecting the impact of demographic distribution, schooling, and intermarriage in modern contexts (Holes, 1990). Migration has long shaped the linguistic landscape of Qatar. Over time, numerous tribes, clans, and families established themselves in Qatar, leaving enduring linguistic and cultural imprints (see Appendix A.0.1).

4 Corpus Development Methodology

We followed the direct elicitation approach (Rickford, 2002) to collect data from native speakers of Qatari Arabic dialects. This method, widely used in sociolinguistics and dialectology, involves prompting participants with specific questions or topics to elicit particular types of language, such as lexical choices, speech patterns, or grammatical constructions within a structured or semi-structured setting. Unlike methods that are based solely on spontaneous conversation, this approach enabled us to engage directly with participants in a way that encouraged rich, culturally grounded responses, while maintaining consistency across all interviews. To support this process, we employed a single, systematically designed instrument: a structured, open-ended, qualitative questionnaire developed specifically for this study to elicit authentic spoken data. The questionnaire was tailored to reflect the linguistic diversity of Qatari society and ensure meaningful contributions from both Urban and Bedouin dialect speakers.

To account for the dialect variation, the questionnaire was deployed in two tailored versions, one for Urban dialect speakers and one for Bedouin dialect speakers, both administered to male and female participants across a range of age groups. These parallel versions ensured balanced data collection across Qatar's two major sociocultural groups while maintaining comparability in topic and structure. Each version included five broad, open-ended questions designed to prompt extended, naturalistic responses without infringing on participants' privacy or introducing personal, sensitive topics. The questions focused on the following culturally salient themes : (i) *social traditions*, including marriage practices, feasts, communal gatherings, and mourning rituals; (ii) *social perceptions* related to women's solo travel, employment, and access to education; (iii) *cultural heritage*, such as traditional crafts (e.g., shipbuilding, pearl diving), folk games, attire, oral traditions, chants, and musical instruments; (iv) *national identity and pride*, as expressed through participants' opinions on Qatar's hosting of international sports events, especially the FIFA World Cup 2022, and associated societal preparations; and (v) *inter-generational interests*, highlighting hobbies, values, and evolving preferences among contemporary Qatari youth. The full questionnaire is provided in Appendix A.0.2.

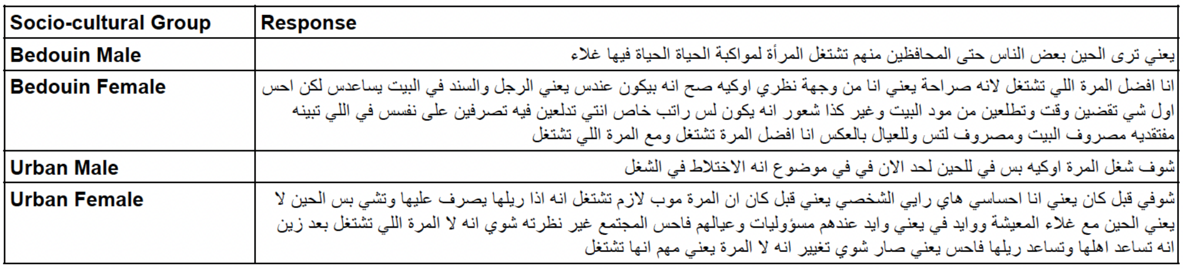
Figure 1 shows a small portion of the corpus theme, where it presents statements from different sociocultural groups regarding their perception and view of women's work in Qatar society. We chose each sample to show the general tone and point of view of each class: Bedouin male, Bedouin female, urban male and urban female. These quotes give us a look at how people of different backgrounds think about, expect and see women's roles in Qatar's workplace, and how this perception has affected over the years.

4.1 Interviewers and Participants

To construct our corpus, we employed a team of Qatari native speakers from both Urban and Bedouin backgrounds. All of them underwent structured training sessions to ensure consistency in conducting interviews and adhering to ethical and methodological protocols. The training focused on administering structured and semi-structured interviews, maintaining a natural yet culturally sensitive rapport with participants, and handling informed consent procedures. Special attention was given to strategies for eliciting spontaneous, culturally rich speech while minimizing interviewer bias.

The team was carefully balanced in terms of gender, with equal numbers of male and female interviewers, to facilitate comfortable and appropriate interactions with participants across gender lines, in accordance with social norms in Qatari

[1] We use the terms *Urban* and *Bedouin* to refer to dialect groupings in Qatari Arabic based on observable linguistic variation. While socially grounded, this classification reflects self-identified sociocultural affiliation and is used for analytical clarity.



Socio-cultural Group	Response
Bedouin Male	يعني ترى الحين بعض الناس حتى المحافظين منهم تشتغل المرأة لمواكبة الحياة الحياة فيها غلاء
Bedouin Female	انا افضل المرة اللي تشتغل لانه صراحة يعني انا من وجهة نظري اوكيه صح انه بيكون عندس يعني الرجل والسند في البيت يساعدس لكن احس اول شي تقضين وقت وتطلعين من مود البيت وغير كذا شعور انه يكون لس راتب خاص انتي تدلعين فيه تصرفين على نفسس في اللي تبينه مفتقديه مصروف البيت ومصروف لتس وللعيال بالعكس انا افضل المرة تشتغل ومع المرة اللي تشتغل
Urban Male	شوف شغل المرة اوكيه بس في للحين لحد الان في في موضوع انه الاختلاط في الشغل
Urban Female	شوفي قبل كان يعني انا احساسي هاي رايي الشخصي يعني قبل كان ان المرة موب لازم تشتغل انه اذا ريلها يصرف عليها وتشي بس الحين لا يعني الحين مع غلاء المعيشة ووايد في يعني وايد عندهم مسؤوليات وعيالهم فاحس المجتمع غير نظرته شوي انه لا المرة اللي تشتغل بعد زين انه تساعد اهلها وتساعد ريلها فاحس يعني صار شوي تغيير انه لا المرة يعني مهم انها تشتغل

Figure 1: Example of responses to the question:*"How does the society view and perceive the following: females' education, women's work, women's travel, family perceptions of boys and girls?"*

society. The interviewers were also selected to represent a range of tribal affiliations, age groups, and social backgrounds to enhance cultural relatability and participant trust—crucial factors in dialect-oriented sociolinguistic research.

Participant recruitment followed a mixed strategy combining purposeful and snowball sampling. Purposeful sampling was used to ensure representation across key demographic variables such as gender, age, region, and sociocultural identity (Urban vs. Bedouin), while snowball sampling helped reach speakers from less accessible or underrepresented communities by leveraging personal networks and community trust. This approach allowed us to build a linguistically and culturally representative corpus that captures the intra-dialectal diversity of Qatari Arabic. All participants were adults (18 years and older) and citizens of Qatar, drawn from major cities and regions across the country, including Al Shamal, Al Khor, Al Shahaniya, Umm Salal, Al Daayen, Doha, Al Rayyan, and Al Wakrah. Prior to the interviews, participants were required to complete and return signed informed consent forms, and confirm their consent verbally before the recording began.[2]

	Gender	18–30	31–45	46–60	Above 60	Total
Bedouin	Female	32	31	11	1	75
	Male	21	17	13	7	58
Urban	Female	32	33	18	10	93
	Male	19	6	2	2	29

Table 1: Distribution of Participants by Sociocultural Group, Gender, and Age

Table1 presents the demographic distribution of the Qatari interviewees in our corpus, categorized by sociocultural group (Urban vs. Bedouin), gender, and age group. The sampling aimed for balanced representation across key demographic variables to ensure diversity in speech patterns and cultural perspectives. First, the slightly higher proportion of Urban participants (52.2%) may reflect the demographic concentration of Qatar's population in urban areas such as Doha, where access to potential participants is more feasible. Urban residents are also more likely to be engaged with academic institutions and public initiatives, increasing their availability for structured interviews (Gardner, 2010).

The higher proportion of female participants in the Urban group (70% vs. 56.6% in Bedouin) likely reflects broader patterns of women's engagement in public and research-related activities within urbanized contexts. In Gulf countries, urban women, who tend to have greater access to education and public-sector employment, are more likely to participate in academic or institutional projects. In contrast, Bedouin communities often adhere to more conservative gender norms that limit women's visibility in such public domains (Krause, 2013).

The predominance of younger participants, with 40.8% aged 18–30 and 34.1% aged 31–45, likely reflects the practical constraints of participant recruitment. Younger individuals are more accessible through university networks and social media, and are generally more comfortable with the idea of being recorded. Older age groups (46–60 and above 60), who make up only 19.2% and 7.8% respectively, may be more reluctant to participate due to unfamiliarity with the research process or a preference for oral over documented interaction.

4.2 Data recording and Transcription

Each interview lasted between 45 to 60 minutes and was audio-recorded to ensure accuracy and fi-

[2] The study protocol, including recruitment and consent procedures, was reviewed and approved by the Institutional Review Board (IRB), ensuring compliance with ethical standards for research involving human subjects.

delity in data capture. Interviewers were equipped with high-quality recording devices and laptops to facilitate both the recording and subsequent transcription processes.[3] To ensure consistency and linguistic accuracy, all interviewers received training prior to data collection.

The transcription was handled by *Ramitechs* which was provided with the transcription guidelines to ensure consistency across all transcribed materials.[4] All transcriptions were reviewed for accuracy and adherence to conventions, with special attention to capturing sociolinguistic markers such as hesitations, code-switching, and phonetic variation. This rigorous process enabled the creation of a high-quality text corpus aligned with the audio recordings, supporting both linguistic and computational analyses.

Transcription Guidelines Summary: The transcription followed standardized conventions to preserve dialectal variation and ensure orthographic consistency. The main principles are as follows:

- **Phonological Variants**: Variants in pronunciation are represented using base letters with alternate forms in parentheses (e.g., ق(ج)لبي for /galbi/ pronounced as /jalbi/).
- **Orthographic Consistency:** Words must reflect the speaker's pronunciation (e.g., ض(ظ)ابط). When alternative spellings exist (e.g., برده / برضه), one consistent form should be used throughout.
- **Code-Switching:** English words are written in Latin script (e.g., `sorry`), while Arabicized terms like كمبيوتر are written in Arabic.
- **Overlaps and Noise:** Overlapping speech is only transcribed for the interviewer. Unintelligible speech is marked as (غير مسموع).
- **Exclusions:** Non-lexical utterances such as هه، آه، م are excluded. Diacritics are not used, except for tanwīn where pronounced (e.g., جداً، بتاتاً).
- **Orthographic Conventions:** Initial hamzated alifs (e.g., أمير) are written as امير. Prefixes like ما and يا, and suffix prepositions like ـل, are spaced from the verb (e.g., ما رحت، يا أخي، كتبت له).
- **Numerals and Scripts:** Numbers should be written in Arabic letters, not digits. Foreign words are written in their original scripts.
- **MSA Alignment:** Final letters such as ه، ة، ي، ى are written according to MSA conventions.

5 Corpus Analysis

To investigate sociolinguistic variation within Qatari Arabic, we conducted a detailed analysis of the corpus, focusing on distinguishing lexical patterns across Bedouin and Urban dialects. Our analysis aimed to uncover both cultural and gender-specific linguistic trends by examining the frequency and distribution of commonly used expressions. By comparing usage patterns across speaker groups, the corpus enabled the identification of lexemes that are characteristic of Bedouin speech versus those more prevalent in Urban settings. This comparative approach offers empirical insights into dialectal differentiation, particularly in the use of culturally salient and gender-marked terms.

5.1 Lexical and Phonological Variation Across Qatari Dialects

Expression	BM	UM	BF	UF
ايه/Ayh	32,365	23,325	26,462	29,157
ايوه/Aywh	12	0	0	2
نعم/ncm	11,266	1,405	904	421
امبلا/AmblA	0	6	92	193
صح/SH	2,950	2,468	8,336	4,597
امبلا/AmblA	0	6	92	193
انا اشهد/AnA A$hd	108	0	8	0
اكيد/Akyd	595	489	1,726	1,096
طبعا/TbçA	8	0	0	8
والله العظيم/wAllh Alcym	113	16	304	75
قسم بالله/qsm bAllh	8	0	136	6
ريال/ryAl	162	415	938	1,176
رجال/rjAl	1,853	127	2,004	107
رييل/ryAyl	0	0	4	4
رياييل/rjAyyl	2	4	58	2
برع/brc	2	111	10	392
برة/brh	389	94	1,278	561
برا/brA	0	0	4	2
بره/brh	12	24	64	62

Table 2: Frequency of Selected Expressions Across Gender and Dialect Groups

[3] It is important to note that the corpus is not segmented at the utterance or sentence level. Hence, corresponding timestamps are not provided.

[4] Ramitechs www.ramitechs.com is a company that creates and annotates several types of corpora and lexicons using expert linguists.

Our analysis reveals a range of salient linguistic phenomena that distinguish Bedouin and Urban speakers in Qatar. The list of features presented below was extracted from the corpus by a native Qatari speaker with sociolinguistic training, who systematically examined lexical, phonological, and morphosyntactic variation across speaker groups. This analysis focused on identifying patterns that reflect dialect-specific usage, with particular attention to forms that vary by gender, cultural register, or language contact. These include systematic phonological variation, lexical divergence influenced by borrowing from other dialects and languages, variation in demonstrative forms, and register-specific usage of culturally embedded expressions. The findings underscore the impact of sociolinguistic identity (Urban vs. Bedouin), gender, and patterns of language contact on dialectal variation within Qatari Arabic.

Phonological Shift: A clear phonological difference involves the realization of the /j/ sound as /y/ in Urban dialects. This is evident in words like رجال /rjAl/ ("men"), which is predominantly used by Bedouin speakers (BM:1,853; BF: 2,004), while Urban speakers favor the variant ريال /ryAl/, especially Urban females (UF: 1,176). Similarly, the morphological variant رياييل /ryAyil/ appears almost exclusively among Urban speakers, further emphasizing this sound shift.

Lexical Synonymy and Dialect Borrowing: The corpus shows several lexical items expressing the same meaning but differing by dialect. For instance, to say "yes," speakers use إيه أيوه نعم or إمبلا. The form إيه is dominant among Bedouin males (BM: 32,365), while نعم,the MSA form, also showsa notable presence among Bedouins (BM: 11,266). The Urban group, in contrast, favors إمبلا, a Levantine borrowing (UF: 193),reflecting dialect contact and media influence.

Code-Switching with English: The corpus also reveals systematic code-switching with English, as shown in Table 3. This practice is most frequent among Urban females, particularly in the younger cohorts (e.g., 5,511 tokens for ages 18–30), reflecting the influence of education and professional domains where English is dominant. Urban males display lower but still notable levels of English usage, while Bedouin speakers, especially older males, rarely code-switch. These findings indicate that English functions not merely as a source of lexical borrowing but as a resource for indexing modernity and cosmopolitan identity, contrasting with the more conservative, monolingual norms maintained in Bedouin speech.

Group	18–30	31–45	46–60	60+
Bedouin Female	243	112	35	9
Bedouin Male	58	21	12	**3**
Urban Female	**5,511**	3,291	804	212
Urban Male	1,027	462	187	66

Table 3: Frequency of English code-switching tokens across sociocultural groups and age cohorts.

Allophonic and Morphological Alternation in Spatial Terms: Lexical variation in Qatari Arabic frequently arises through allophonic and morphological alternation, where multiple surface forms convey the same semantic content. One such example is the word for "outside," which appears in the corpus with several variants: برع, برة, برا, and بره. The form برع, which is strongly preferred by Urban speakers (UM: 111; UF: 392), contrasts with the Bedouin-favored برة (BM: 389; BF: 1,278). These alternations reflect both regional lexical preferences and underlying allophonic variation, particularly in final vowel or consonant realizations. Meanwhile, the forms برا and بره appear less frequently and are more evenly distributed between groups, suggesting that they are neutral or transitional variants.

Discourse Markers and Epistemic Modality: Bedouin speakers frequently use epistemic markers such as انا اشهد (BM: 108), اكيد (BF = 1,726), and religious affirmations like والله العظيم and قسم بالله. These forms are related to the assertion of truth, politeness, or religious legitimacy. Urban speakers use these less frequently and prefer forms that index modernity or neutrality.

Standard Influence and Pragmatic Confirmation: The expression صح ("correct") is derived from MSA and is commonly used to confirm statements. It is especially prevalent among Bedouin women (BF: 8,336), which shows that MSA still influences spoken dialect in rural communities. Conversely, امبلا, borrowed from Levantine Arabic and used similarly to 'yes, indeed', is more prevalent in urban speech (UF: 193), indicating pragmatic convergence due to language contact.

Gendered Morphophonological Variation in Word-final Segments: A striking morphological distinction between the Bedouin and Urban varieties of Qatari lies in the gendered variation of the word-final segments for the feminine forms. In the Bedouin dialect, feminine nouns and adjectives frequently end with the affricates /s/ or /ts/, forming a characteristic lexical pattern. Examples include forms like بنتس/bnts and عاجبتس/Ajbts. However, Urban speakers tend to favor the palatal fricative // (rendered as ـتش), as seen in words such as اهدافتش/AhdAftš, ظروفتش/rwftš, and عاجبتش/Ajbtš. Interestingly, in both dialects, male speakers consistently use the masculine second-person possessive or descriptive suffix /k/, particularly in contexts involving possessive or descriptive constructions (e.g., بعطيك/bTyk, شكلك/klk).

5.1.1 Vocabulary Metrics

To complement the qualitative analysis of lexical and phonological variation, we also examined vocabulary diversity across groups in the corpus. Table 4 reports the total token counts, vocabulary size, and type-token ratio (TTR) for each demographic group. The results reveal clear sociolinguistic differences. Urban females contributed the largest volume of speech (over 1.29M tokens), yet their TTR is relatively low (0.0289), suggesting greater repetition and reliance on a stable lexicon. By contrast, Urban males contributed fewer tokens (422k) but show the highest TTR (0.0454), indicating proportionally richer lexical diversity. Bedouin speakers, particularly males, also demonstrate high lexical richness (TTR ≈ 0.040), reflecting broader use of culturally embedded vocabulary. Gender effects are also evident: while females overall produced nearly twice as many tokens as males (2.26M vs. 1.24M), males exhibit proportionally greater lexical variety (0.0337 vs. 0.0255). Finally, the entire corpus spans 3.5M tokens and over 78,000 unique word types, with an overall TTR of 0.0223, a value consistent with large-scale spoken corpora where lexical repetition increases with size.

Group	Total Tokens	Vocabulary Size	TTR
Urban Males (Total)	422,474	19,193	0.0454
Urban Females (Total)	1,299,825	37,526	0.0289
Bedouin Males (Total)	823,157	33,276	0.0404
Bedouin Females (Total)	969,089	37,622	0.0388
All Urban	1,722,299	45,481	0.0264
All Bedouin	1,792,246	56,688	0.0316
All Male	1,245,631	41,937	0.0337
All Female	2,268,914	57,799	0.0255
ENTIRE CORPUS	**3,514,545**	**78,418**	**0.0223**

Table 4: Vocabulary metrics across sociocultural groups, reporting total token counts, vocabulary size, and type-token ratio (TTR).

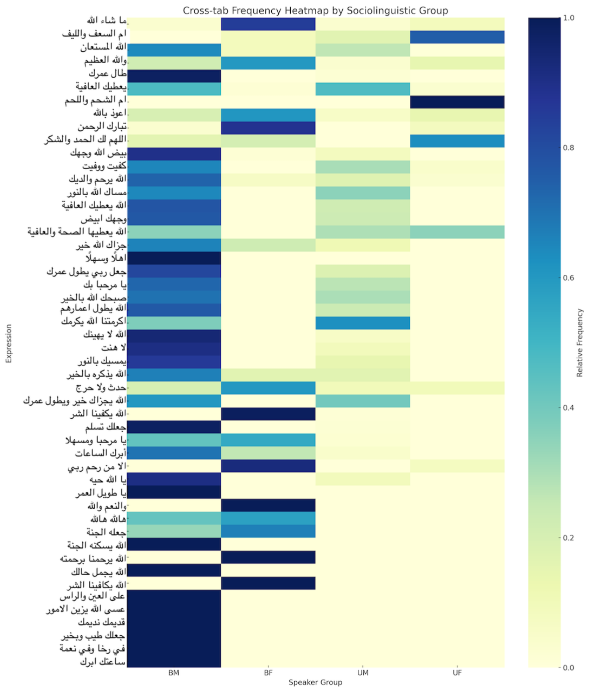


Figure 2: Normalized frequencies of selected culturally significant expressions across Qatari dialect groups: Bedouin Male (BM), Bedouin Female (BF), Urban Male (UM), and Urban Female (UF).

5.2 Sociolinguistic Patterns in Common Expressions

To explore the distribution of culturally significant expressions across Qatari dialectal groups, we conducted a cross-tab frequency analysis and visualized the results using a heatmap. The expressions selected for this analysis are among the most frequent formulaic phrases and cultural idioms found in the corpus. These include religious invocations, greetings, expressions of gratitude, and culturally embedded metaphors.

Figure 2 presents the normalized frequency of 50 expressions across four speaker categories: Bedouin Male (BM), Bedouin Female (BF), Urban Male (UM), and Urban Female (UF). The normalization accounts for unequal group sizes, enabling a more balanced comparison.

The heatmap reveals distinct sociolinguistic patterns. For example, the expression طال عمرك (*may your life be long*) occurs predominantly among

Bedouin male speakers, with negligible usage among other groups, reflecting its strong association with traditional Bedouin honorific discourse. In contrast, expressions like يعطيك العافية (*may God give you health*) and تبارك الرحمن (*blessed is the Merciful*) are more evenly distributed across groups, indicating their widespread use in both Urban and Bedouin settings.

Other expressions show clear gendered patterns. Urban females make frequent use of culturally rich metaphors such as أم السعف والليف and أم الشحم واللحم, both of which are almost absent among male speakers. Conversely, highly formulaic and religious expressions like والله العظيم and الله المستعان are more common among Bedouin males.

The heatmap also reveals that Urban speakers, especially females, use a broader range of metaphorical and heritage expressions, possibly due to greater exposure to cultural preservation discourse and social media usage. These findings point to the role of gender and cultural identity in shaping dialectal preferences and highlight the importance of capturing such intra-dialectal variation in computational modeling.

6 Initial Experiments on Dialect Identification and Gender Prediction

To explore the potential of the corpus for computational modeling and downstream NLP applications, we conducted two main experiments: (1) intra-dialectal dialect identification and (2) gender prediction based on linguistic features in transcribed speech.

6.1 Dialect Identification: Urban vs. Bedouin

Although dialect identification is a well-established task in Arabic NLP, this work focuses on intra-country linguistic variation, an underexplored but important dimension for building dialect-aware language technologies.

First, we trained a logistic regression model using TF-IDF representations of the transcribed interviews, with 80% of the data used for training and 20% for testing. The model achieved an overall accuracy of 77%, with detailed results shown in Table 5. The classifier performed well for the Bedouin class (F1: 0.83, recall: 0.91), but showed lower recall for Urban speakers (0.51), indicating that Urban speech is more lexically diverse or shares overlapping features with Bedouin speech, leading to misclassifications. This result aligns with the linguistic observations in Section 5, where Bedouin speakers consistently used more conservative or marked lexical and morphophonological forms (e.g., *-ts* suffixes, *rjAl*, *hAðy*), which may provide stronger cues for classification. In contrast, Urban speakers often exhibit greater borrowing and stylistic variation, which may blur dialectal boundaries from a feature-based modeling perspective. These results suggest that while dialect identity is strongly encoded in the corpus, especially for Bedouin speakers, future work should explore contextualized or multimodal representations to better capture Urban speech variation.

Dialect	**Precision**	**Recall**	**F1-Score**	**Support**
Bedouin	0.77	0.91	0.83	53,619
Urban	0.76	0.51	0.61	29,957
Accuracy		0.77		
Macro Avg	0.77	0.71	0.72	83,576
Weighted Avg	0.77	0.77	0.76	83,576

Table 5: Classification results for Urban vs. Bedouin dialect identification using logistic regression and TF-IDF

In addition to the logistic regression baseline, we experimented with transformer-based and feature-enriched models. Using AraBERT (Antoun et al., 2020)(bert-base-arabertv02), we obtained an accuracy of 71.7% and a macro-F1 of 0.65. As shown in Table 6, the model performs considerably better on the Bedouin class (F1: 0.80, recall: 0.89) than on the Urban class (F1: 0.51, recall: 0.41), confirming our earlier observation that Urban speakers exhibit greater lexical diversity and borrowing, making their speech more challenging to model reliably.

Class	**Precision**	**Recall**	**F1-Score**
Bedouin	0.72	0.89	0.80
Urban	0.67	0.40	0.50
Accuracy		0.7173	
Macro Avg	0.70	0.64	0.65
Weighted Avg	0.70	0.71	0.69

Table 6: Dialect identification results using AraBERT.

To improve performance, we extended the feature space with both lexical and morphological cues. The best-performing system combined word-level TF-IDF features (1–2 grams) with character-level TF-IDF features (3–5 grams), enabling the model to capture both lexical signals and morphological variation. Trained with a linear SVM classifier, this system achieved an accuracy of 83.8%, substantially outperforming both the logistic regression baseline (77%) and the AraBERT model. These findings demonstrate that intra-dialectal classification benefits from feature sets that jointly encode surface-level and morphological information, while contextual embeddings remain constrained by the heterogeneity of Urban speech.

6.2 Text Gender Prediction

To evaluate the degree to which gendered linguistic features in the corpus can be learned and predicted computationally, we conducted several binary classification experiments. First, we trained a logistic regression model to predict speaker gender (male vs. female) using TF-IDF representations of transcribed text segments. Data was split into 80% for training and 20% for testing, ensuring stratification by dialect and age to preserve demographic balance.

Gender	Precision	Recall	F1-Score	Support
Female	0.81	0.77	0.79	47,659
Male	0.72	0.77	0.74	35,917
Accuracy		0.77		
Macro Avg	0.77	0.77	0.77	83,576
Weighted Avg	0.77	0.77	0.77	83,576

Table 7: Classification results for gender prediction using logistic regression

The model achieved an overall accuracy of 77% on the held-out test set. As shown in Table 7, the classifier performs slightly better in identifying female speakers (F1: 0.79) than male speakers (F1: 0.74), with comparable recall scores for both groups (0.77). This suggests that certain lexical or morphophonological features characteristic of female speech in the corpus may be more distinctive or consistent across speakers. Overall, the macro-averaged F1 score is 0.77, indicating balanced performance across gender classes.

Next, we fine-tuned AraBERT on the corpus, and obtained an overall accuracy of 72% (Table 8). The model performed better on female speakers (F1: 0.76, recall: 0.77) than on male speakers (F1: 0.68, recall: 0.67), suggesting that lexical and stylistic markers of female speech are more consistent and thus more easily captured by contextual embeddings. In contrast, male speech exhibits greater heterogeneity, leading to lower classification performance. These results indicate that while AraBERT provides a strong baseline for gender prediction, there remain challenges in capturing intra-gender variation, which may require additional sociolinguistically informed features or multimodal cues.

Gender	Precision	Recall	F1-Score
Female	0.75	0.76	0.76
Male	0.68	0.66	0.67
Accuracy		0.72	
Macro Avg	0.71	0.71	0.71
Weighted Avg	0.72	0.72	0.72

Table 8: Gender classification results using AraBERT fine-tuned on the Qatari Arabic corpus. The model shows stronger performance for female speakers compared to male speakers.

Our findings provide empirical support for the sociolinguistic patterns observed in the corpus analysis. In particular, features such as morphophonological suffixes (e.g., *-ts* vs. *-š*), lexical preferences, and formulaic expressions appear to encode gender variation that can be effectively captured by relatively simple models.

7 Conclusion and Future Work

In this work, we presented the first publicly available, multimodal corpus of Qatari Arabic, capturing intra-dialectal variation across Urban and Bedouin speakers, balanced by gender and age. We detailed the data collection process, transcription conventions, and corpus analysis, including lexical diversity and code-switching patterns. We also reported baseline experiments on dialect and gender prediction, showing that surface-level lexical and morphological cues provide strong classification signals. These findings underscore the value of the corpus for both sociolinguistic inquiry and computational modeling. By filling a critical gap in Gulf Arabic resources, this work provides a foundation for inclusive language technologies and contributes to the documentation and preservation of Qatar's linguistic heritage.

Acknowledgments

We would like to express our sincere gratitude to our dedicated interviewers for their efforts in building the dataset. We also extend our thanks to the Qatari participants who generously agreed to be interviewed and whose contributions were essential to this project. Finally, we acknowledge Ramitechs for providing the transcription services that supported this work.

This work was supported by NPRP grant# NPRP12S-0301-190189 from the Qatar National Research Fund (a member of Qatar Foundation). The statements made herein are solely the responsibility of the authors.

References

Ahmed Abdelali, Hamdy Mubarak, Younes Samih, Salam Hassan, and Kareem Darwish. 2020. Arabic dialect identification in the wild. arXiv preprint arXiv:2005.06557. Cornell University.

Dana Abdulrahim, Go Inoue, Lama Shamsan, Salma Khalifa, and Nizar Habash. 2022. The bahrain corpus: A multi-genre corpus of bahraini arabic. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference (LREC 2022)*, pages 2345–2352, Marseille, France. European Language Resources Association (ELRA).

Moayyad Al-Bohnayyah. 2019. Dialect variation and change in eastern arabia: Al-ahsa dialect.

Mariam Al-Mulla and Wajdi Zaghouani. 2020a. An annotated corpus for qatari arabic. In *Proceedings of the LREC*.

Shaikha Al-Mulla and Wajdi Zaghouani. 2020b. Building a corpus of qatari arabic expressions. In *Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools (OSACT 2020)*, pages 23–30, Marseille, France. European Language Resources Association (ELRA).

Shamlan D. al Qenaie. 2011. *Kuwaiti Arabic: A Socio-Phonological Perspective*. Ph.D. thesis, University of Essex.

Alanoud al Sharekh and Courtney Freer. 2021. *Tribalism and Political Power in the Gulf State-Building and National Identity in Kuwait, Qatar and the UAE*. Bloomsbury, London.

Nora Al-Twairesh, Rasha N. Al-Matham, Nouf Madi, Nouf Almugren, Asma Al-Aljmi, Shahad Alshalan, Rawan Alshalan, Nouf Alrumayyan, Shatha Al-Manea, Shahad Bawazeer, Nouf Almutlaq, Najla Almanea, Wala B. Huwaymil, Dana Alqusair, Rawan Alotaibi, Shahad Al-Senaydi, and Areej Alfutamani. 2018. Suar: Towards building a corpus for the saudi dialect. *Procedia Computer Science*, 142:72–82.

Rukayah Alhedayani. 2025. The najdi arabic corpus: a new corpus for an underrepresented arabic dialect. *Language Resources and Evaluation*, 59(2):1593–1612.

Noor AlNaama. 2012. Torath al'ajdad. AlArab Newspaper. Accessed: 2025-07-05.

Ibrahim Alsarsour, Emad Mohamed, Rami Suwaileh, and Tamer Elsayed. 2018. Dart: A large dataset of dialectal arabic tweets. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Wissam Antoun, Fady Baly, and Hazem Hajj. 2020. Arabert: Transformer-based model for arabic language understanding. *arXiv preprint arXiv:2003.00104*.

M. Bakir. 2010. Notes on the verbal system of gulf pidgin arabic. *Journal of Pidgin and Creole Languages*, 25(2).

Houda Bouamor, Nizar Habash, and 1 others. 2018. The madar arabic dialect corpus and lexicon. In *Proceedings of LREC*.

Soumia Bougrine, Hadda Cherroun, Djelloul Ziadi, Abdallah Lakhdari, and Aicha Chorana. 2016. Toward a rich arabic speech parallel corpus for algerian subdialects. In *The 2nd workshop on arabic corpora and processing tools*, pages 2–10.

Soumia Bougrine, Aicha Chorana, Abdallah Lakhdari, and Hadda Cherroun. 2017. Toward a web-based speech corpus for algerian dialectal arabic varieties. In *Proceedings of the third arabic natural language processing workshop*, pages 138–146.

R. Davey. 2016. *Coastal Dhofari Arabic: A Sketch Grammar*.

A. M. El-Saba. 2016. Translating arabic speaking countries: Qatar. Globalization Partners International. Accessed: 2025-07-05.

Mahmoud Elmahdy, Mark Hasegawa-Johnson, and Eiman Mustafawi. 2014. Development of a tv broadcasts speech recognition system for qatari arabic. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC 2014)*, pages 3057–3061, Reykjavik, Iceland. European Language Resources Association (ELRA).

Andrew Gardner. 2010. *City of Strangers: Gulf Migration and the Indian Community in Bahrain*. Cornell University Press.

Georgetown University in Qatar. 2017. Gu-q launches qatari phrasebook app. Qatar Foundation. Accessed: 2025-07-05.

Nizar Y. Habash. 2010. *Introduction to Arabic Natural Language Processing*. Morgan & Claypool.

Clive Holes. 1986. The social motivation for phonological convergence in three arabic dialects. *International Journal of the Sociology of Language*, 61:33–51.

Clive Holes. 1990. *Gulf Arabic*. Psychology Press.

Clive Holes. 2018. *The Arabic Dialects of the Gulf*. Oxford Scholarship Online.

Salma Khalifa, Nizar Habash, Dana Abdulrahim, and Sara Hassan. 2016. A large scale corpus of gulf arabic. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, pages 4282–4289, Portorož, Slovenia. European Language Resources Association (ELRA).

Wanda Krause. 2013. Gender and participation in the arab gulf. In *The transformation of the Gulf*, pages 86–105. Routledge.

Khaled A. Kwaik. 2018. Shami: A corpus of levantine arabic dialects. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Stephan Prochazka. 2021. Loanwords in gulf arabic: A historical overview. *Journal of Arabic and Islamic Studies*, 21:125–146.

John Rickford. 2002. How linguists approach the study of language and dialect. Stanford University. Accessed: 2025-07-05.

Kristine Shockley. 2020. A sociophonetic study of gulf arabic dialects. In *Proceedings of the LSA*.

Abdelrahman Shoufan and Sultan Alameri. 2015. Natural language processing for dialectical arabic: A survey. In *Proceedings of the Second Workshop on Arabic Natural Language Processing (WANLP 2015)*, pages 36–48, Beirut, Lebanon. Association for Computational Linguistics.

S. A. Skilliter. 1969. Turkish grammar. by g. l. lewis, pp. xxiv, 304. oxford, clarendon press, 1967. 45s. *Journal of the Royal Asiatic Society of Great Britain & Ireland*.

Tarek Tarmom, W. J. Teahan, Eric Atwell, and M. A. Alsalka. 2020. Compression versus traditional machine learning classifiers to detect code-switching in varieties and dialects: Arabic as a case study. *Natural Language Engineering*, 26(6):663–676.

Irene Theodoropoulou and Dhyiaa Borresly. 2025. Stancing solidarity: Twitter communication in qatar during the blockade. *Humanities and Social Sciences Communications*, 12(1):1–12.

Omar Zaidan and Chris Callison-Burch. 2014. Arabic dialect identification. *Computational Linguistics*, 40(1):171–202.

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

A Appendix

A.0.1 Tribes and Families

Al Maadeed, Al Khulaifat, Al Sulata, Al Bin Ali, Bani Malik, Bani Hajer, Al Sudan, Al Mananaa, Al Bu Kuwara, Al Kibsa, Al Nuaim, AL Mazare', Al Emadiheya, Al Fakhroo, Al Gubaisat, Al Manaseer, Al Mahanda and Al Misnad, Al Dasim, Al Sada, Al Ibrahim (Kamal, 1901, p49), Al Muhazea, Al Attiyah, Bani khaled, Al Mesallam, Al Humaidat, Al Mutawaa, Al Nusairi, Al Zeyarah, Al Jubarah, Al Fudhalah, Al kaaban, Al Ruwashed, Al Mahandah, Al Haydos, Al Misnad, Al Muraikat, Al Mudahkah, Al Mutawaah, Al bu Rumai, Al Bu sumait, Al Duwaser, Qahtan, Al Ehbab, Al Namlan, Al khayareen, Al Shafi, Al shahwan, Al Salem, Al Khalifa, Al Sahlawi, Al Abdullah, Al Megalli, Al Hamad, Al Mohammad, Al Sultan, Al Jassim, Al Nubi, Al abdulrahman, Bani Tamim,, Al Saad, Al Hudaifi, Bu Rumaih, Al Naser, Al Buainain, Al khater, Al Muwalek, Al Derham, Al Mana, Al Shuraim, Al Jaber, Al Mahmoud, Al Muftah, Al Ibrahim, Al Abdulla, Al Yousef, Al Fakhroo, Al Derwish, Al Obaidan, Al khal, Al Nasser, Al Abadelah, Al Muhaizea, Al Rashid, Al Jassim, Al Burshaid, Al Fakhri, Al Sudan, Al Rabban, Al Mahmoud, Al Jusaiman, Subaea, Al Fayaheen, Al Sultan, Al Souailem, Al Suhol, Al Kulaifat, Al Ansar, Al Meslemani, Al Qubaisat, Otaibah, Al Shebani, Al Sheeb, Al Shehabi, Al Muthaffar, Al Abdulghani, Al Jaidah, Al Nemah, Al Jamali, Al Obaid, Al Eid, Al Jolo, Al Meer, Khafood, Al Awadhi, Al Khajah, Al Taher, Al Najjar, Al Najadah, Al Ghanem, Al Khathlan, Al Oolan, Al Dayel, Al Kharji, Al dulaimi, Al Jaber, Al Bahar, Al Nesef, Al bu Jallof, Al Khalaf, Al Sorour, Al Ahmad, Al Mohammed, Al Bu flasah, Bani Hashim, Al khori, Al Zaman, Al Saei, Al Manaseer, Al Theyab, Juhainah, Al Muwalek, Yam, Al Murrah, Al Ajman, Shahran, Bani Yafea, Al Saadi, Al Keldi, Al Suqatri, Al Salahi, Al Hajjaji, Al Rayashi, Al Ajji, Bani Hammad, Al Haram, Al Abadlah, Al Marazeeg, Al Ali, Al Aali, Al Aamri, Al Emadiah, Al Asmakh, Zainal, Al Meqbel, Al Humaid, Al Karani, Al Haydar, Al Fardan, Al Hayki, Al Makki, Al Haddad, Al bukeshisha, Al Sooj, Al dehniem, Al Sallat, Al Sayegh, Al Musawi, Al Sayed, Al Sharshani, Al Kunji, Al Derbesti, Nabina, Al Langawi, Al Janahi, Al sherawi, Shammar, Enizah, Al Qatami, Al Burdaini, Al Taweel, Al Zeydan, and more. It is worth noting that a number of families share the same name, yet they go back to different origins.

A.0.2 Interview Discussion Guide – Qatar Linguistic Map Project

Interviewer circles one response for each of the below: Age Group:

- 18–30 years
- 31–45 years
- 46–59 years
- 60 years and above

Gender:

- Male
- Female

Do you work?

- Yes
- No

Family/Tribe:

- Bedouin
- Urban

Education:

- Ph.D.
- Masters
- Undergraduate
- Associate
- Secondary

Interviewee Area of Residence in Qatar:

- Daayen
- Doha
- Khor
- Rayyan
- Salal
- Shahniya
- Shamal
- Wakra

Important Notice to Interviewer:

- Ask the participants not to say anything that is both identifiable and private in their responses to the open-ended questions.
- Also explain to them (in their dialect) that the questions below will be asked to stimulate a chat.

QUESTIONS

1. Have social norms and customs differed over time (from the past until the present) in terms of the following marriage rituals, social duties, social treats, solace and condolences, feasts? If yes, How?
2. How does the society view and perceive the following: females' education, women's work, women's travel, family perceptions of boys and girls?
3. Qatari heritage is full of elements such as: crafts (e.g. boat and ship building, hunting/fishing; pearl diving), folk games, traditional costumes, folk songs and chants, musical instruments, etc. Can you tell us something about all or any of them (as much as you know)?
4. What is your opinion of Qatar hosting of international sport and athletic championships? What's your opinion of Qatar hosting of World Football Cup 2022? What arrangements has Qatar done so far for hosting these events? Will you contribute to any of these arrangements? How? Will you attend some of the games? What are the values Qatari people need to adopt to ensure the success of these international events (e.g. accepting cultural differences, hospitality, etc.)?
5. What are your age group interests?