Generating a Lexicon for the Hijazi dialect in Arabic

Fatimah Abdullah Alqahtani[1,2][0000-0002-4164-721X] and Mark Sanderson[1][0000-0003-0487-9609]

[1] Computer Science, School of Science, RMIT University, Melbourne, Australia
[2] College of Computer Science & Information Systems, Jazan University, Jazan, Kingdom of Saudi Arabia
fatimah.alqahtani,mark.sanderson@rmit.edu.au

Abstract. We present a methodology for creating a lexicon for a low-resource Arabic dialect in Saudi Arabia: Hijazi. We show the differences between the Hijazi dialect and Modern Standard Arabic. We annotate articles and tweets using recruited native speakers. We create a lexicon of Hijazi adapted from two resources: Sebawai and Quranic Arabic Corpus. The lexicon is created both manually and automatically by using Hijazi morphology. We detail the methodology to build this lexicon and present results of an evaluation of the corpus formation process.

Keywords: Hijazi Dialect, Lexicon Generation.

1 Introduction

Arabic dialects are a set of linguistic characteristics that belong to a particular environment [1], and are often used in informal daily communication. An increased awareness of the existence and functioning of these dialects has emerged due to the influence of social media, where these dialects are now written.

The Egyptian, Levantine, and Moroccan dialects as well as MSA are considered high-resource; however, others, including the Hijazi Dialect are low-resource [2]. The lack of resources has created an obstacle for researching Hijazi. A dialect of western cities (Makkah, Madinah, Jeddah and Taif) in Saudi Arabia, Hijazi is spoken in the second most populous region[1]. Hijazi has two varieties: urban and rural, and this study focuses on the urban variety. To the best of our knowledge, no one has built resources for Hijazi.

We applied a methodology to create a Hijazi lexicon in which potential Hijazi words were annotated through a comparison with an MSA lexicon. Then, the Hijazi words were analyzed using an approach employed by Darwish, Sajjad and Mubarak [3] for Egyptian dialects to automatically generate an expanded Hijazi word list. We also annotated 3,000 tweets to identify Hijazi content. Our work addresses the following research question: Can a methodology used to create a High-resource Egyptian lexicon be adapted to create a Low-resource Hijazi lexicon?

Section 2 of the research reviews relevant work, Section 3 shows the approach use to build Hijazi lexicon. In section 4 Standard and Hijazi Arabic are compared. The approach to generating the Hijazi dialect is presented in Section 5. Section 6 describes our evaluation. The paper concludes and gives insight into future work in Section 7.

[1] 10,090,256 people in 2015 - http://www.cdsi.gov.sa

2 Related work

2.1 Creating resources in Arabic Dialects

Prior Arabic Dialect corpus building work focused on building monolingual or parallel corpora. Methods vary in the building with most using recruited native speakers.

The COLABA project collected resources from Arabic blogs for four dialects: Egyptian, Iraqi, Levantine, and Moroccan [4]. For harvesting, Diab, Habash, Rambow, Altantawy and Benajiba [4] asked 25 native speakers to generate 40 dialectal queries containing words with multiple orthographies that cover social issues, religion, and politics. The authors asked annotators to translate the queries to MSA and English. The queries were used to extract matching blog data from the web. The researchers created a tool to process and manage the data harvested from the blog.

Harrat, Meftouh and Smaili [5] created a parallel corpus for MSA, Algiers, Annaba, Tunisian, Palestinian, and Syrian. They collected around 2.5K Algiers dialect sentences from transcribed films and TV shows, which were then translated to MSA and Annaba by a native speaker. In the same way, the researchers collected the corpus of Annaba dialect for approximately 3.9K sentences from the transcribed recordings of the daily life of some people of Annaba, which were then translated to MSA and Algiers. Finally, they translated a whole collection of MSA around 6.4K sentences to Tunisian, Palestinian and Syrian by native translators. The Dialect and MSA translation were conducted by twenty-five persons in total for free.

The Gumar Corpus is a large-scale collection of Gulf Arabic consisting of 100 million words from over 1,200 novels published online [6]. The researchers annotated the corpus manually into a sub-dialect of Gulf Arabic, which includes the Saudi, UAE, Bahraini, Qatari, and Omani dialects. Annotations were at the document level. They found that names given to the characters, cities, and event names in the novel helped determine the dialect. The researchers extracted features and rules for these dialects to understand the morphology and to build tools.

Most recently, there is the Curras corpus for the Palestinian dialect [7] that was manually annotated by two annotators for one year. The researchers of this study identified 56,700 tokens in 190 documents compiled from resources such as Facebook, Twitter, blogs, forums, Palestinian stories, Palestinian terms, and scripts from Palestinian TV shows. The researchers used the DIWAN Dialect Word Annotation tool [8] and the MADAMIRA [9] morphological analyzer tool for MSA and Egyptian. A quantitative evaluation was performed for three of the documents, which consist 1,529 tokens by two annotators, who met to review and discuss their annotations. Their agreement was measured by Kappa, and the outcome was almost perfect desirable for different tags (e.g., POS, stem, prefix, etc.).

Darwish, Sajjad and Mubarak [3] employed manual and automatic approaches to collect three lexicons of Egyptian Dialect. In the manual approach, they asked a linguist to extract 1,300 high frequency Egyptian words (MAN) from the Egyptian side of the LDC2012T09 corpus [10] while in the automatic approach, they applied Egyptian morphology rules to generate verbs from Sebawai Arabic roots [11]. The rules added prefixes and suffixes such as pronouns and negation that are compatible with the Egyptian

dialect. The rules also substituted letters to change a word to the Egyptian dialect. Filters were applied to the verb and letter substitutions to remove words that were MSA. An MSA word list was drawn from 63 million Arabic tweets and Aljazirah articles.

Mubarak and Darwish [12] presented a multi-dialect corpus from Twitter for Saudi, Egyptian, Iraqi, Lebanese, Syrian, and Algerian. They collected 92 million tweets, which had a user location. The user location of tweets was assigned to one of the Arab countries in GeoNames, which indicate the location in each country. Then, the researchers manually reviewed the location, which they mapped with GeoNames. Also, they manually tried to map locations, which were non-matching with GeoNames. Also, they collected all n-gram words that occurred at least three times in AOC, Aljazeera interview articles, and the GigaWord corpus, which is a text archive of Arabic news sources by Linguistic Data Consortium LDC. These n-gram words were manually labelled by a native Arabic speaker, knowledgeable in different dialects to specify if it was a dialect word and to which dialect it belonged. This resulted in around 2,500 dialect words. The researchers filtered the tweets as dialect tweets by the n-gram dialect words and they got 6.5 million dialect tweets based on the following assumption *"if a sentence contained one of these n-grams, then the sentence is dialectal"*. They used crowdsourcing to evaluate 100 randomly extracted tweets per dialect. They asked crowdsourcing workers, who were from the same countries from which the tweet was issued, to judge whether the tweet dialect coincides in their country. They were not able to get judges for Qatar and Bahrain.

Overall, the reviewed literature shows that many contributions have been made to Egyptian, Levantine and North Africa dialects, also, sub-dialects of Levantine such as Syrian, Palestinian, Jordanian and Lebanese.

2.2 Creating resources in Low-Resource languages

Different means of collecting and generating data for low-resource languages have been tried. In 2011, Outahajala, Zenkouar, & Rosso [13] manually built a corpus for the Amazighe language: a low-resource language in Morocco, Algeria, Tunisia, Libya, and areas of Egypt. They extracted text from various sources such as the Royal Institute for Amazighe Culture's newsletter and website as well as three primary school textbooks. The corpus was manually annotated by a team of four annotators. It consisted of different POS features to the tokenized Amazighe texts. Three linguists chose random texts and evaluated them. The annotator agreement was 94.89%.

By drawing on the concept of manual tagging, Ramrakhiyani and Majumder [14] provided a corpus of temporal expression recognition in Hindi called ILTIMEX2012. There were three temporal expression classes: data time, a time or duration expression; a frequency, which is a date or time expression; or period, which is a frequency expression. The corpus is composed of 300 documents of a set of articles from the Hindi newspaper, The FIRE 2011 Hindi corpus [15]. Each document has more than 500 words. ILTIMEX2012 was manually labelled by using the General Architecture for Text Engineering (GATE) tool annotation module. 514 periods, 110 frequency, and 1295 date-time for temporal expressions were included in the corpus. This corpus was used for Hindi temporal expressions identification and classification.

Bird, Gawne, Gelbart and McAlister [16] applied a Android application, which supports recording of speech directly [17]. When finishing a recording, users were asked to add metadata such as name, language, and image. Users could segment the audio and write a transcription and translation. The researchers used the application for collecting audio from Brazil and Nepal. They collected 100k words from 10 hours of audio. A challenge with this approach is lack of the participation and scarcity of electricity in the village which is important to charge the device. It was noted in the study that the collection of data for these low-resources languages appeared firstly by collecting speech, which shows that there is no written source for these languages. Also, the crowdsourcing and human approaches consider the conventional method for collecting and building the data for low-resource languages.

For Hijazi, Alahmadi [18] collected more than 30 Hijazi words by asking native Hijazi speakers to give the correct dialect word for an image.

3 Approach

In defining the methodology for our problem of building a Hijazi corpus, we considered our situation. There are no Hijazi language resources available and only a limited amount of edited Hijazi text is available online. However, there is a notable amount of social media content written in Hijazi. Therefore, we examined an approach to building a corpus using a combination of manual and automatic techniques that is adapted from an approach by Darwish, Sajjad and Mubarak [3]. The approach requires access to an initial Hijazi word list, a set of morphological rules, and a list of Arabic word roots. There are two phases of the methodology: corpus creation and evaluation.

We used different resource to build the Hijazi list: articles, dictionaries and building from roots. We found a collection of Hijazi articles in Okaz news[2] which is read in western Saudi Arabia; 156 Hijazi articles were collected manually from February 2011 to September 2014. The set contained 59,225 tokens. An analysis of the most frequent words found most were Hijazi pronouns, as shown in **Error! Reference source not found.**.

The overview of our process is to pre-process the list of Hijazi words (see **Fig. 1** (a)), which involves removing MSA stop words. Next, we check if each remaining word matches to a set of Hijazi verbs rules (section 4.2), if there is a match, the word is considered Hijazi and is added to the MANHijazi list. If there is not a match, the words are compared with three MSA dictionaries: Maajim[3], Alwased,[4] and Alsahah[5], if the word is not found in these dictionaries, then the words is also considered Hijazi and is added to the MANHijazi list.

[2] http://www.okaz.com.sa
[3] http://www.maajim.com/dictionary/
[4] http://shamela.ws/browse.php/book-7028
[5] http://www.almaany.com

Table 1. The 10 most frequent words

Words	Frequency	Source
Ally اللي which	1168	Hijazi
yA يا	597	MSA
All~h الله Allah	435	MSA
E$An عشان because	410	Hijazi
hAdy هادي This	299	Hijazi
hdA هدا This is	258	Hijazi
kdh كده like this	283	Hijazi
Ay$ ايش What	229	Hijazi
wAlly واللي And that	188	Hijazi

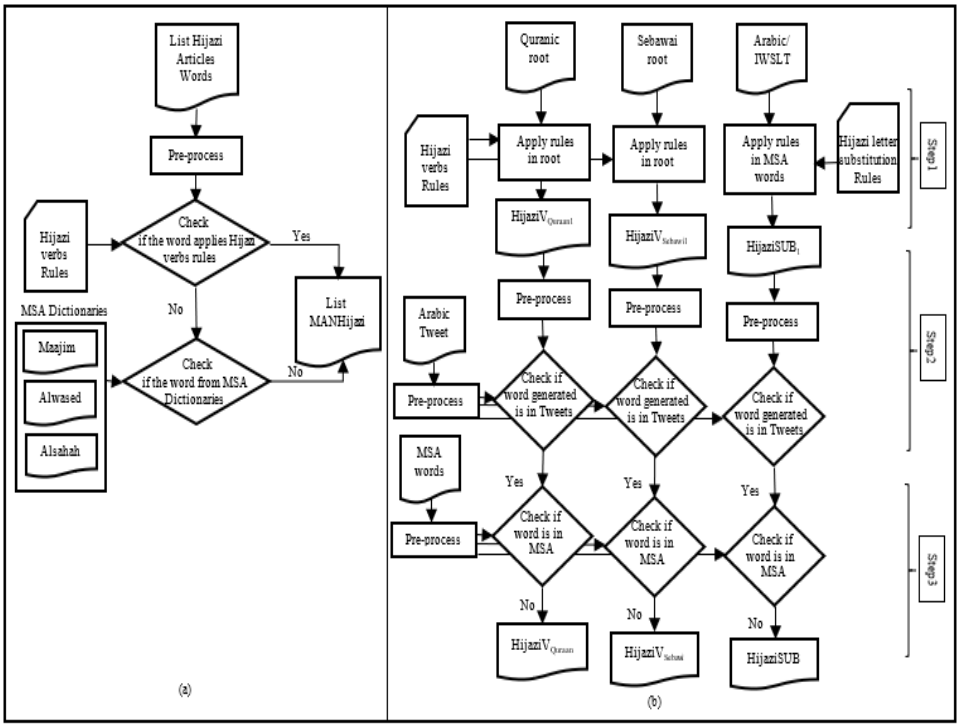

Fig. 1. (a) Creating a manually formed Hijazi list (MANHijazi) from articles by comparing with MSA dictionaries and Hijazi rules, (b) Creating three different list of Hijazi then filter

Fig. 1 (b) illustrates the steps of the automatic approach to adding to the list in MAN-Hijazi. First, a set of roots drawn from the two sources: 10,406 from the Sebawai system [11] and 943 verb roots from the Quran. Each root was transformed into multiple verbs by applying a set of Hijazi verbs rules to form two lists of words: HijaziV$_{Quraan}$, and HijaziV$_{Sebawai}$. We used the Quran root because it has a real verb root while Sebawai has a large number of roots that are automatically generated. The coverage of this latter

set is greater, but it is noisy because of the roots are extracted from different words such as the verbs, nouns and adjectives. In addition, MSA to Hijazi letter substitution rules were applied to a list of MSA words from the Arabic side of the English/Arabic parallel corpus from the International Workshop on Spoken Language Translation (Arabic/IWSLT) 2013 to generate the list $\text{Hijazi}_{\text{SUB1}}$.

In the next step, the words in the three lists were compared to a list of three million words drawn from Arabic tweets, which were pre-processed using the AraNLP tool [19]. Words that matched were assumed to be some form of Arabic. In the final step, the remaining words in the three lists were compared against a corpus of MSA words OSAC6 [20], which contains 18,183,511 tokens. Those words that did not match against the corpus were assumed to be Hijazi. We evaluated the lists of words intrinsically by manually investigating a random sample of words drawn from each of the lists.

4 Hijazi Dialect

We describe the basic syntax of the Urban Hijazi Dialect by contrasting it with MSA. We focus on verbs and letter substitutions because these are where the main differences between MSA and Hijazi are found. We examine irregular structures, such as lexical and independent pronouns, and regular structures such as verb and letter substitution.

4.1 Irregular

Lexical. Three past papers – [21-23] – described three differences between the lexicon words in MSA and Hijazi.

1. Lexicon words in Hijazi which have the same meaning in MSA but with different letters. For example, (Bridge, كبري "kbry" in Hijazi, جسر "jsr" in MSA), (Maybe, بلكن"blkn" in Hijazi, احتمال "AHtmAl" in MSA) and (Good, كويس "kwys" in Hijazi, طيب "Tyb" in MSA).
2. Distinguishing dialectal terms in Hijazi which come from the combination of two words, which are called a blend in linguistics. For example, دحين 'dHyn' (now), comes from the words ذا الحين '*A AlHyn'; and كمان 'kmAn' (also), it comes from the words إن كما 'kmA <n'
3. Words in Hijazi and MSA which are written the same word but have different meanings. For example, دول 'dwl', which means 'those' in Hijazi and 'countries' in MSA.

Independent Pronouns. Independent pronouns in Hijazi are distinct from those in MSA. The pronoun for "we" has more than one form. For example, 'AHnA' احنا or 'nHnA' نحنا in Hijazi, but it is 'nHn' نحن in MSA. Also, the plural person pronoun is 'AntW'انتو in Hijazi, not antum أنتم as in MSA. The first person 'Anti' انتِ can be written in another form 'Anty' انتي. However, dual pronouns 'ntmA' أنتما and 'hmA' هما and the feminine plural pronouns 'Antn' انتن and 'hn' هن are not used in Hijazi.

[6] https://sites.google.com/site/motazsite/arabic/osac

The demonstrative pronouns in Hijazi are different from MSA. For the singular, “this” is 'dA' دا for masculine and 'dy' دي for feminine. While in MSA, 'h*A' هذا and 'h*h' هذه are used for masculine and feminine respectively. Also, Hijazi speakers use 'hAdA' هادا for masculine, while 'hAdy' هادي for feminine, feminine plural nouns and for inanimate masculine plural nouns. For plural "these", they used 'hdWl' هدول or 'hdWlA هدولا or 'dWl' دول or 'dWlA' دولا instead of 'h&lA'' هؤلاء in MSA.

4.2 Regular

We present the main features of the Hijazi in negation, verb, and letter substitution.

Negation. The 'mA-' ما prefix is the only particle used for negation in all tenses of Hijazi verbs. However, in MSA, 'lA' لا, 'lm' لم, 'ln' لن, 'lys' ليس are used to negate a verb in addition to 'mA-' ما. Also, the 'mA' ما prefix is used for the negative pronouns in Hijazi instead of 'lys' ليس in MSA. In negating adjectives and nouns, the word ‘mw’ مو can be used while in MSA 'lA' لا, and 'lys' ليس are used [21].

The Verb. We consider verbs from two points: tense and structure. There are three main tenses in MSA: past, present, and future. However, [21-23]showed that some verb forms differ between MSA and Hijazi:

- Hijazi adds a verbal particle as a prefix /bi-/ب + verb (e.g. /bi-yiktub/ بيكتب ‘byktb’ he writes) in the expression of present tense, but MSA does not have this verbal particle.
- Hijazi adds a verbal particle as a prefix /in-/ إنـ or انـ + verb (e.g. /inktub/انكتب ‘Anktb’ was written), although some speakers add the prefix /At/ اتـ ‘At’ + verb (e.g. /Atktab/ اتكتب ‘Atktb’ was written) in the expression of passive past tense, but MSA does not have this verbal particle.
- Hijazi adds a verbal particle as a prefix /ħa-/ ح ‘H’ + verb (e.g. /ħa-yiguul/حيقول ‘Hyqwl’ he’ll say) in the expression of future tense. Or Add the word /raħ/ راح ‘rAH’ + verb (e.g. /raħ-yiguul/ راح يقول ‘rAH yqwl’ he’ll say). Instead of the word سوف ‘swf’ ‘will’, which is add before the verb in MSA, or the letter س‘s’, which is add in prefix of verb in MSA.

There are two types of verbs across MSA and Hijazi: sound (صحيح SHyH) or weak (معتل mEtI):

- **Sound verbs** are those verbs that do not include (w)و or (y) ي in the root letters. In the past verb of singular feminine, Hijazi adds the letter ‘ي’؛y’ to the end of a word such as the word كتبت ‘ktbti’ (she wrote) to be كتبتي ‘ktbty’ in Hijazi. Furthermore, there are double-sound verbs (الفعل المضعف AlmuDEf), where the second and third letter of the root are the same, such as دقّ daqqa -يدقّ yadiqqu (to knock). In the past verb of singular, Hijazi removes the third letter to be دقيت ‘dqyt’ (knocked) and add the suffix يت ‘yt’ in You (masculine), while دققت ‘dqqtu’ in MSA with adding suffix ت‘t’. Also, there are (الفعل المهموز mahmuuz) Hamzated verbs, where ء is one of the

consonants, such as أكل ‘>kal’ - يأكل ‘y>kl’ (to eat). In general, there is no difference between masculine and feminine in Hijazi for all tenses of verbs. In Hijazi, the formulas for the masculine and feminine are the same in the plural form.

- **Weak verbs** are those verbs that contain و (w) or ي (y), as one or two of the root letters. There are three types of weak verbs, Hijazi weak verbs have different patterns among themselves and with MSA:

  — Assimilated verbs (الفعل المثال), which begin with و ‘w’ or ي ‘y’ such as وقف ‘wqf’ يقف ‘yqf’, ‘to stand up’. In the present form, MSA removes the letter و ‘w’ and replaces it with ا ‘A’ like اقف ‘Aqf’, while Hijazi keeps the letter و ‘w’ and adds the prefix ب ‘b’ like باوقف ‘bAwqf’ (will stand).

  — Hollow verbs (الفعل الأجوف), which are the second letter in the root is ا ‘A’, و ‘w’ or ي ‘y’, such as قام ‘qAm’ يقوم ‘yqwm’, ‘to get up’. ا ‘A’ is replaced with و ‘w’ in the present tense. Also, ا ‘A’ is replaced with ي ‘y’ in the present tense, such as باع ‘bAE’ يبيع ‘ybyE’, ‘to sell’.

  — Defective verbs (الفعل الناقص), which end the root of the verb with و ‘w’ or ي ‘y’ such as رمى ‘rmY’ يرمي ‘yrmy’, ‘to throw’. ‘Y’ is replaced by ‘y’ or ا ‘A’ is replaced by ‘w’, as in the example نما ‘nmA’ ينمو ‘ynmw’ ‘to grow’.

Letter Substitution. Hijazi people write in a way that reflects their pronunciation. This leads to some letter substitutions between MSA and Hijazi [21-23]. Example substitutions as shown in **Table 2**

Table 2. The letter substitutions between MSA and Hijazi

Letter in MSA	Letter substitution in Hijazi	Example in MSA	Example in Hijazi
‘ذ’ ‘*’	‘ز’ z / ‘د’ d	‘*kryAt’ (memories)/ ‘k*Ab’ (liar)	‘zkryAt’/ ‘kdab’
‘ث’ ‘v’	‘ت’ t / ‘س’ s	‘vqyl’ (heavy)/ ‘mvAl’ (example)	‘tqyl’/ ‘msAl’
‘ض’ ‘D’	‘ز’ z	‘bAlDbT’ (exactly)	‘bAlzbT’
‘أ’ ‘>’	‘ي’ y	‘bd>t’ (I started)	‘bdyt’
‘ئ’ ‘}’	‘ي’ y	‘EwA}l’ (families)	‘EwAyl’
‘ه’ ‘h’	‘و’ w	‘klh’ (All of)	‘klw’

5 Experiment

This section firstly presents the process used to collect our Hijazi dialect corpus. We collected data from two different domains: articles and tweets. Each domain required different approaches to annotation, as described below. Also, this section shows the method used to generate the lexicon of Hijazi dialect verbs automatically.

5.1 Articles

To obtain a word list, we split the gathered articles into tokens by using whitespace and other punctuation characters (".,?!") as delimiters. In informal text, words are not always split by whitespace, so the letter و "w" "and", was also used as a delimiter. Next,

we remove Arabic stop-words by using a list from the Ranks NL [24]. The remaining words were confirmed as Hijazi by:

1. Checking manually if the word is a verb, then checking if it fits the Hijazi verb structure described in section 4. If yes, then it is a Hijazi verb.
2. If the word is not a verb or the name of a person or a place, then a search for the word in three extensive MSA lexicons was conducted: Maajim, Alwased and Alsahah. If the word was in one of the three, then it was considered an MSA word, otherwise, it was considered an Hijazi word.

This leaves us with 1,363 MANHijazi words. A manual examination of the words revealed that the distribution of POS were 904 verbs (simple present - passive past - future), which common use the verb form "فعل" and "انفعل", 62 pronouns, 21 prepositions, 56 adverbs, 82 adjectives, 18 question, 12 interjections, 6 phrases and 202 nouns. The highest number of verbs demonstrates that the main difference between Hijazi and MSA is in verbs.

5.2 Tweets

To label tweets, we used Hijazi native speakers. Three thousand tweets, which were geo-located in the western cities of Saudi Arabia, Jeddah, Makkah, Taif, Medina and Yanbu were collected from March to May 2014. The tweets were pre-processed through manual inspection. All URLs, embedded images and user-related information (i.e. display name, avatar/ display image and user mention) were removed from the tweets and their content was checked to ensure none were offensive.

Native speakers. We obtained 3000 tweets from Mourad, Scholer and Sanderson [25] that had locations in the western cities of Saudi Arabia: Jeddah, Makkah, Taif, Medina, and Yanbu. Three native Hijazi speakers were recruited to annotate the 3000 tweets. We used a questionnaire, which had the list of tweets, to ask the speakers to label the tweets. The speakers were asked to record the tweet's dialect type (Hijazi or non-Hijazi) and details of which Hijazi words indicated that the tweet belonged to the dialect. The speakers were given one week for the task.

Fleiss` Kappa measured an annotator agreement of 0.89. From the annotation, we found that there were 372 Hijazi tweets from the original 3000, Error! Reference source not found. shows the number of tweets from each city. There were 666 words in the Hijazi tweets and 311 unique word forms. Statistical comparisons between the Hijazi articles and tweets are shown in Error! Reference source not found. We can see that this approach to labeling Hijazi tweets generated a limited lexicon, and therefore an automatic approach is also needed to extend the lexicon Hijazi.

Table 3. Identifying Hijazi and non-Hijazi dialect tweets by cities

City	Hijazi	Non-Hijazi	Total
Jeddah	200	1277	1477
Makkah	69	520	589
Taif	45	331	376
Medina	42	377	419
Yanbu	16	123	139
Total	372	2628	3000

Table 4. Description of Hijazi articles and tweets

Features	Articles	Tweets
Number of	156	372
words	59223	4672
Unique words	16270	2872
Average sentence length	96.3	10
Short words (<=3 characters)	17537	1555
Long words (>=7 characters))	9602	424
Unique Hijazi words	1367	35

5.3 Automatic Generation of Hijazi Dialect Words

The methodology used to expand the Hijazi lexicon verbs, follows a two-stage approach from Darwish, Sajjad and Mubarak [3]:

1. Generating a lexicon by using morphological rules of the dialect combined with roots (Quranic verbs root and Sebawai roots). The rules will generate multiple verbs from a single root.
2. Filtering the lexicon.

Automatic generating word. To generate Hijazi verbs from each root, prefixes were add to the root to set a tense (present, future, present/future passive). In Hijazi, object pronouns are attached to the verb as suffixes. A suffix set of Hijazi dialect are shown in **Table 5** . Also, there are ten subject pronouns: I "أنا", you "أنتَ", you "أنتِ", you "أنتم", you "أنتنَّ", we "نحن", they "هم", they "هن", he "هو", she "هي", and each pronoun has associate suffixes from the suffixes set.

Table 5. Suffixes set in Hijazi Dialect

Suffix in Hijazi	Example
Suffix length 1	ت, ه, ي, ك, و
Suffix length 2	وك, كي, ها, هم, كم, وا, وه, ني, نا
Suffix length 3	وكي, وكم, وها, وهم, وني, ولي, ونا

The rules have the following form: Sign, Root_Length, Condition, Prefix, Suffix

- Sign: contains (=, >) to compare with the root length.
- Root_Length: the length of the root in the rule to apply the rule condition. Note that the first letter begins with index 0.
- Condition(s): it has three main sections:
  - index: contains the index number to verify, it can provide a specific index, a range of index or all indexes in the root.
  - verify: contains a letter or collection of letters to check if the letter is in the given index.
  - action: if the condition of verifying achieved then apply the action, which can be R to remove a specific letter in the index or change to a letter or collection of letters, which leads to generate multi-root.
- Prefix: the prefixes are added at the beginning of the root. If multiple prefixes are listed, then from single root multi verbs will be generated.
- Suffix: the suffixes are added at the end of the verb, which generated from the Prefix stage, by attaching appropriate suffixes as appropriate for the pronouns to generate Hijazi, ($HijaziV_{Quraan}$) and ($HijaziV_{Sebawai}$).

After applying the rules, the total number of Hijazi words generated from the two roots sets, Sebawai and Quraan, are 1,585,461 and 184,227 respectively. Example of the generation process are shown in **Table 6**.

For letter substitution, we used an MSA list of words from an English/Arabic parallel corpus[7], which consists of a dataset of around 150k sentences. We apply substitution rules (from section 4.2) if there is a letter in any word of MSA list match in Hijazi letter substitution then changed it to the appropriate letter in the Hijazi. We have obtained in 3,800 letter substitutions in Hijazi Arabic ($Hijazi_{SUB}$).

Filtering the generating word. The lists of Hijazi words, ($HijaziV_{Quraan}$, $HijaziV_{Sebawai}$, and $Hijazi_{SUB}$) were filtered by using steps 2 and 3 as shown in **Fig. 1**.(b). The purpose of step 2 was to remove ambiguous or error words produced by one of the automatic generation methods. Also, this technique had a disadvantage; there might have been deletions of unused Hijazi words in tweets. We obtained 30,389, 25,599 and 1,630 words for $HijaziV_{Sebawai}$, $HijaziV_{Quraan}$, and $Hijazi_{SUB}$ respectively from this step. Then, we applied the step 3 to ensure there were no MSA words in the lists. In the end, the total number of $HijaziV_{Sebawai}$, $HijaziV_{Quraan}$, and $Hijazi_{SUB}$ from step 3 were 24,413, 12,428 and 1,074 respectively. The verbs intersection in $HijaziV_{Sebawai}$ and $HijaziV_{Quraan}$ are around 10,595 verb words, while there is no intersection between the list of verbs ($HijaziV_{Sebawai}$ and $HijaziV_{Quraan}$) and $Hijazi_{SUB}$.

[7] From the International Workshop on Arabic Language Translation.

Table 6. An example illustrating the automatic generating Hijazi word

<table>
<tr><td colspan="6">Phase 1</td></tr>
<tr><td rowspan="3">Context</td><td colspan="5">=, 3, (1,ا , (ا, ي, و)), (ب,با), (-,كم, كي, ك, هم, ها, ه)</td></tr>
<tr><td>Sign</td><td>Root_Length</td><td colspan="3">Condition</td><td>Prefix</td><td>Suffix</td></tr>
<tr><td>=</td><td>3</td><td colspan="3">(1,ا , (ا, ي, و))
Index | verify | action
1 | ا | و, ي, ا</td><td>با, ب</td><td>ه, ها, هم, ك, كي, كم</td></tr>
<tr><td colspan="6">Phase 2</td></tr>
<tr><td>Root</td><td colspan="5">نام “nAm” sleep , length of root (نام)= 3</td></tr>
<tr><td>Changing the root by using action</td><td colspan="5">Check if the index 1 in the root نام “nAm” has a letter “ا”, then apply the action by changing the letter “ا” to “ا”, ”ي”, ”و”.
word1: نام</td></tr>
<tr><td></td><td>word1: نام</td><td colspan="2">word2:نيم</td><td>word3:نوم</td></tr>
<tr><td>Add prefix</td><td>word11: بنام, word12: بانام</td><td colspan="2">word21:بنيم , word22: بانيم</td><td>word31:بنوم , word32</td></tr>
<tr><td>Add suffix</td><td>word111: بنام, word112: بنامه, word113:بنامها , word114: بنامهم, word115: بنامك, word116: بنامكي, word117: بنامكم word121: بانام, word122: بانامه, word123:بانامها , word124: بانامهم, word125: بانامك, word126: بانامكي, word127: بانامكم</td><td colspan="2">word211: بنيم, word212: بنيمه, word213:بنيمها , word214: بنيمهم, word215: بنيمك, word216:بنيمكي , word217: بنيمكم word221: بانيم, word222: بانيمه, word223:بانيمها , word224: بانيمهم, word225: بانيمك, word226:بانيمكي word227:بانيمكم</td><td>word311: بنوم, word312: بنومه, word313:بنومها , word314: بنومهم, word315: بنومك, word316: بنومكي, word317: بنومكم word321: بانوم, word322: بانومه, word323:بانومها , word324: بانومهم, word325: بانومك, word326:بانومكي , word327:بانومكم</td></tr>
</table>

6 Evaluation

Comparative evaluation with past work is not available since to the best of our knowledge this is the first study to generate a Hijazi lexicon. Instead, we used intrinsic evaluation to show the accuracies of the three Hijazi lexicons. We manually investigated randomly sampling a 2% from $\text{HijaziV}_{\text{Sebawai}}$, $\text{HijaziV}_{\text{Quraan}}$, and $\text{Hijazi}_{\text{SUB}}$: 490, 250, and 22 words, respectively, to estimate the coverage of the accuracy of lists. The error rate (ER) was calculated from the proportion of error words. As shown in the **Table 7**, $\text{HijaziV}_{\text{Quraan}}$ has the highest number of Hijazi dialect words (227) with an error rate of 0.09. In contrast, $\text{Hijazi}_{\text{SUB}}$ has the highest error rate of 0.27 in 22 words while the $\text{HijaziV}_{\text{Sebawai}}$ was 0.16. An analysis of errors was conducted, see **Table 7**. Three error types were found:

- Alternative root: the generated verb looks like a Hijazi word, but a Hijazi speaker would use a different root.
  - In HijaziV$_{Sebawai}$, alternative Hijazi root has the highest error rate of 69 error words. For example, the word اتعمش 'AtEm$', 'weak eyesight' from the word 'Em$', is used as the adjective and it is اعمى 'AEmY', 'blind', ضعف نظره 'DEf nZrh' as an adjective in Hijazi, or انعمى 'AnEmY' 'to blind' as the verb.
  - In HijaziV$_{Quraan}$, the highest number of such errors was 14. Examples include: حيميد 'Hymyd', "to shake from the root" ميد where 'myd' is replaced with هز 'hz' to be حيهز 'Hyhz'.
- Incorrect root: the generated verb looks like a Hijazi word, but the root is not an MSA verb root. The incorrect root was due to errors in the automatically created Sebawai list. An example of this is in Sebawai [11] , ملم 'mlm', 'has knowledge', the root is a noun according to the dictionary Maajim not a verb root, has been applied the Hijazi rule to become حتملم 'Atmlm' in HijaziV$_{Sebawai}$, which is not Hijazi verb.
- Error a rule: the rule should not have been applied to all words. In HijaziSUB, the errors are due to the mistaken assignment for rules in word or root. For example, in the word الرضا 'AlrDa' "satisfaction", the letter ض "D" changed to ز 'z' to become 'Alrza', which is not Hijazi words.

Table 7. The Distribution of Hijazi and non-Hijazi words in the evaluation

Type of list	No. words	Sample words 2%	Hijazi words in 2%	Non-Hijazi words in 2%	Error rate	Distribution error rate		
						Alternative Hijazi root	Incorrect root	Error rule
HijaziV$_{Sebawai}$	24,413	490	409	81	0.16	69	6	6
HijaziV$_{Quraan}$	12,428	250	227	23	0.09	14	0	8
HijaziSUB	1,074	22	16	6	0.27	0	0	6

This technique of using the morphological rule of Hijazi is considered a good starting point to generate the Hijazi automatically, where we started from 1,633 and now we have 24,413, 12,428 and 1,074 for HijaziV$_{Sebawai}$, HijaziV$_{Quraan}$, and Hijazi$_{SUB}$ respectively. The intrinsic manual evaluation for HijaziV$_{Sebawai}$ and HijaziV$_{Quraan}$ give a different type of answer: once has more words with high error rate while the other has few words with low error rate.

7 Conclusion

In this paper we asked the following research question: can a methodology, used to create a High-resource Egyptian, lexicon be adapted to create a Low-resource Hijazi lexicon?

We explained a method for building a lexicon of a low-resource Arabic dialect in Saudi Arabia: Hijazi, using human experts. We expanded a lexicon of Hijazi words by

covering different Hijazi morphologies rules in a manual and in automatic approaches that can be used in this research and the future as a benchmark. Morphological rules were applied to two different root sets: Quranic and Sebawai roots. The Hijazi lexicon generated can be used for computational linguistic research and NLP tools. This lexicon would help in some applications such as electronic translation.

We evaluated Hijazi lexicon manually. Based on the result of our experiments, we found that the methodology can be applied to create a Hijazi lexicon. Also, we found that the linguistic phenomena in the Hijazi are the first step that allows us to build a Hijazi lexicon. For future work, we plan to expand the size of our corpus to maximize the coverage of the domain of the Hijazi dialect lexicon and mapping with MSA. We also plan to develop a morphological analyzer for the Hijazi dialect, and then build an automatic classification to annotate the Hijazi dialect.

References

1.Anis, I.: On The Arabic Dialects. Maktabat al-Anglo al-Misriyya, Cairo (1952).
2.Biadsy, F., Hirschberg, J., and Habash, N.: 'Spoken Arabic dialect identification using phonotactic modeling. In In Proceedings of the eacl 2009 workshop on computational approaches to semitic languages2009, pp. 53-61. Association for Computational Linguistics, (2009).
3.Darwish, K., Sajjad, H., and Mubarak, H.: Verifiably Effective Arabic Dialect Identification. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 1465-1468.(2014).
4.Diab, M., Habash, N., Rambow, O., Altantawy, M., and Benajiba, Y.: COLABA: Arabic dialect annotation and processing. In Lrec workshop on semitic language processing , pp. 66-74. (2010).
5.Harrat, S., Meftouh, K., and Smaili, K.: Creating Parallel Arabic Dialect Corpus: Pitfalls to Avoid. In 18th International Conference on Computational Linguistics and Intelligent Text Processing (CICLING) (2017).
6.Khalifa, S., Habash, N., Abdulrahim, D., and Hassan, S.: A large scale corpus of Gulf Arabic. arXiv preprint arXiv:1609.02960, 2016
7.Jarrar, M., Habash, N., Alrimawi, F., Akra, D., and Zalmout, N.: Curras: an annotated corpus for the Palestinian Arabic dialect. Language Resources and Evaluation, 51, (3), 745-775 (2017).
8.Al-Shargi, F., and Rambow, O.: Diwan: A dialectal word annotation tool for Arabic. In Proceedings of the Second Workshop on Arabic Natural Language Processing, pp. 49-58. (2015).
9.Pasha, A., Al-Badrashiny, M., Diab, M.T., El Kholy, A., Eskander, R., Habash, N., Pooleery, M., Rambow, O., and Roth, R.: MADAMIRA: A Fast, Comprehensive Tool for Morphological Analysis and Disambiguation of Arabic. In: LREC, vol. 14, pp. 1094-1101. (2014).
10.Zbib, R., Malchiodi, E., Devlin, J., Stallard, D., Matsoukas, S., Schwartz, R., Makhoul, J., Zaidan, O.F., and Callison-Burch, C.: Machine translation of Arabic dialects.
11.Darwish, K.: Building a shallow Arabic morphological analyzer in one day. In: Proceedings of the ACL-02 workshop on Computational approaches to semitic languages (2002).
12.Mubarak, H., and Darwish, K.: Using Twitter to collect a multi-dialectal corpus of Arabic. In: Proceedings of the EMNLP 2014 Workshop on Arabic Natural Language Processing (ANLP), pp. 1-7. (2014).

13.Outahajala, M., Zenkouar, L., and Rosso, P.: Building an annotated corpus for Amazighe. InWill appear In Proc. of 4th International Conference on Amazigh and ICT (2011).
14.Ramrakhiyani, N., and Majumder, P.: Approaches to temporal expression recognition in Hindi. ACM Transactions on Asian and Low-Resource Language Information Processing, 14 (1), pp. 2 (2015).
15.Palchowdhury, S., Majumder, P., Pal, D., Bandyopadhyay, A., and Mitra, M.: Overview of FIRE 2011. In Multilingual Information Access in South Asian Languages, pp. 1-12. Springer, Berlin, Heidelberg (2013).
16.Bird, S., Gawne, L., Gelbart, K., and McAlister, I.: Collecting bilingual audio in remote indigenous communities. In Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers, pp. 1015-1024. (2014).
17.Hanke, F.R., and Bird, S.: Large-scale text collection for unwritten languages. In Proceedings of the Sixth International Joint Conference on Natural Language Processing pp. 1134-1138. (2013).
18.Alahmadi, S.D.: Loanwords in the urban Meccan Hijazi dialect: An analysis of lexical variation according to speakers' sex, age and education. International Journal of English Linguistics, 5, (6), pp. 34 (2015).
19.Althobaiti, M., Kruschwitz, U., and Poesio, M.: AraNLP: a Java-based Library for the Processing of Arabic Text. In: Proceedings of the 9th language resources and evaluation conference (LREC), pp. 4134-413. Reykjavik. (2014).
20.Saad, M.K., and Ashour, W.: Osac: Open source arabic corpora. In: Proceedings of the EEECS'10 the 6th International Symposium on Electrical and Electronics Engineering and Computer Science, pp. 118-123. European University of Lefke, Cyprus (2010).
21.Sieny, M.E.: The Syntax of Urban Hijazi Arabic. Librairie du Liban, Beirut (1973).
22.Omar, M.K.: Saudi Arabic, Urban Hijazi Dialect: Basic Course. Washington, D.C.: Foreign Service Institute (1975).
23.Arabic Variant Identification Aid, http://terpconnect.umd.edu:80/~nlynn/AVIA/Level3/index.htm, last accessed 2015/06/06
24.RANKS NL - Arabic stopword list.:https://www.ranks.nl/stopwords/arabic. last accessed 2015/02/02
25.Mourad, A., Scholer, F., and Sanderson, M.: Language influences on tweeter geolocation. In European Conference on Information Retrieval, pp. 331-342. Springer, Cham (2017).