Mapping Linguistic Variations in Colloquial Arabic through Twitter

A Centroid-based Lexical Clustering Approach

Abdulfattah Omar[1]*
Department of English
College of Sciences and Humanities
Prince Sattam Bin Abdulaziz University
Department of English, Faculty of Arts
Port Said University

Hamza Ethleb[2]
Translation Department
Faculty of Languages
University of Tripoli
Tripoli, Libya

Mohamed Elarabawy Hashem[3]
Department of English
College of Science and Arts in Tabarjal,
Jouf University, KSA
Faculty of Languages and Translation
Cairo, Al-Azhar University, Egypt

***Abstract*—The recent years have witnessed the development of different computational approaches to the study of linguistic variations and regional dialectology in different languages including English, German, Spanish and Chinese. These approaches have proved effective in dealing with large corpora and making reliable generalizations about the data. In Arabic, however, much of the work on regional dialectology is so far based on traditional methods; therefore, it is difficult to provide a comprehensive mapping of the dialectal variations of all the colloquial dialects of Arabic. As thus, this study is concerned with proposing a computational statistical model for mapping the linguistic variation and regional dialectology in Colloquial Arabic through Twitter based on the lexical choices of speakers. The aim is to explore the lexical patterns for generating regional dialect maps as derived from Twitter users. The study is based on a corpus of 1597348 geolocated Twitter posts. Using principal component analysis (PCA), data were classified into distinct classes and the lexical features of each class were identified. Results indicate that lexical choices of Twitter users can be usefully used for mapping the regional dialect variation in Colloquial Arabic.**

Keywords*—Colloquial Arabic; computational statistical model; lexical patterns; linguistic mapping; principal component analysis (PCA)

I. Introduction

Sociolinguists have studied lexical variation and correlated the process through which speaker groups choose their vocabulary with a bundle of variables, such as gender, context, social status, topic [1-4]. More recently, researchers have focused on dialect geography in social media, due to the advances in technology and the unprecedented development of communication channels and networks [5-7]. It is true that these communication channels and networks provide good opportunities for researchers and sociolinguists to study and explore linguistic variation among different speaker groups. Interestingly, the study of linguistic variation through social media networks has been parallel to computational methods. These methods have the potential of dealing with big data and investigating linguistic variation on a larger scale which have positive implications to the generalizability and reliability issues [8-10]. In Arabic, however, much of the work on regional dialectology is so far based on traditional methods; therefore, it is difficult to provide a comprehensive mapping of the dialect variation of all the colloquial dialects of Arabic. As thus, this study is concerned with proposing a computational model for mapping the linguistic variation and regional dialectology in Colloquial Arabic through Twitter based on the lexical choices of speakers. The purpose of the study is to explore the lexical patterns for generating regional dialect maps as derived from Twitter users. In order to map the linguistic variation of Colloquial Arabic dialects, cluster analysis methods were used. This is a clustering method where each class or group has distinct features that make it different from other groups. In dialectology, speakers who share the same linguistic features should be grouped together. This should serve as a basis for exploring the distinctive features of each speaker group. The remainder of the article is organized as follows. Section 2 is a brief survey of the literature on linguistic mapping through social media networks. Section 3 describes the methods and procedures. Section 4 presents the lexical features of dialectal variations among Arab speakers. Section 5 is conclusion.

*Corresponding Author

II. Literature Review

Many advances have been made in the recent years in representing the world's linguistic diversity or what is referred to as language mapping, also referred to as linguistic cartography [11-13]. This is defined as "the visualization of linguistic and language-related data in geographic space and, hence, the representation of correlations between geographic and linguistic facts" [14]. Numerous research projects have been concerned with the regional classification of languages based on several parameters including phonetic and lexical variables. The premise is that there is a significant correlation between geographic location and the linguistic facts. It is argued that linguistic maps can be drawn or generated based on the lexical and phonetic variables as they still carry unique features that can distinguish speakers of the same language. Although the classification of languages and dialects is an established tradition in linguistic studies, the widespread of social media networks and platforms as well as electronic/digital databases has provided researchers with rich and untraditional resources to data. In fact, the social networks and platforms have become an integral part in people's daily

lives and created virtual speech communities which should not be ignored in sociolinguistic studies.

The unprecedented development of social media networks which have been parallel to the development of computational and statistical methodological frameworks have made it possible for researchers to investigate the issue of linguistic mapping on a larger scale. Today, computational approaches provide researchers with the potentials of processing big data in a fast and efficient way. In this regard, numerous studies have been developed using the potentials of computational systems in dealing with big data [15-17]. Studies of the kind are generally based on large corpora for investigating the correlation between lexical patterns and regional dialects [18]. The premise is that correlation between linguistics on the one hand and geography and population on the other hand can be best investigated through statistical and computational methods for their effectiveness in dealing with big corpora that are thought to have good implications to data representativeness and generalizability of the results. Another advantage of the use of computational models and systems in linguistic mapping is that they provide researchers with clear visualizations of the linguistic maps. Today, three-dimensional representation systems are used for the visualization of the geography of linguistic features [10, 19].

Moisl [20] argues that the integration of computational methods into linguistic mapping has significantly contributed to the literature. Traditionally, linguistic mapping of dialect variation was based on single linguistic (mainly phonetic and lexical) features. These maps were also normally limited to small bundles of dialects within the same language. Due to the capabilities of computational systems, maps of regional dialects can be based on multiple linguistic features. They can also be based on many dialects within the same language. An obvious example is the Atlas of German Dialects [21-23]. The Atlas provides a detailed classification of German dialects even beyond the political borders of Germany. The project records and documents the remaining German dialects which were spoken in Northern Moravia. This Atlas is different from traditional linguistic mapping projects of the German dialects which were based on partial explorations and lacked holistic view. The newly developed Atlas is determined and confirmed by multiple linguistic features and unified methods [24]. Computational approaches have also been used in mapping the dialect variation of different languages including English, German, and Chinese [14, 25-28].

Despite the recent advances in the classification of regional dialects using computational and statistical approaches, the studies of Arabic dialects have not been fully explored yet. Much of the work on the classification of the regional dialects of Arabic has been mainly based on comparing a small number of dialects using a limited set of linguistic variables. Although Mulki, et al. [29] suggested the use of recent clustering technologies and systems in the classification of social media language in Arabic, so far there is no holistic view of the regional dialects in Colloquial Arabic. This study seeks to address this gap in the literature through proposing a computational model for the classification of the regional dialects in Colloquial Arabic.

III. Methods and Procedures

For the purposes of the study, a corpus of 1597348 geolocated Twitter posts by 650847 users was designed. Selected tweets are limited to those written in Arabic. However, posts written in Arabizi or Phranco-Arabic are included. The rationale is that such alphabets are very popular today especially on social media networks and therefore should not be disregarded. Data were collected during December 2019 on the most important trends in the Arab world, according to the BBC News Arabic survey that included representatives from almost all the Arab countries. These topics included atheism, women's rights, refugees, honor killing, LGTB, and the Arab-Israeli conflict. Hashtags on these topics were selected and data were derived.

As an initial task, the tweets/posts were converted into what is known as bag of words. Tweets were represented as strings of lexical types. This is because the study is concerned with the lexical properties only. It asks whether lexical choices can map the linguistic variation of Colloquial Arabic dialects. This had the effect of having a corpus of 12778784 words. These were mathematically represented in a vector space matrix, henceforth referred to as colloquial_arabic_dial_ corpus. The matrix is built out of rows and vectors. The rows represent the number of speakers (650847 Twitter users) and the vectors represent all the lexical types included in this study (12778784 words).

One main problem with this corpus is that it is too big for any clustering system to handle. This problem is referred to as high-dimensionality of data. In cases of such kind, it is very challenging to identify the most distinctive lexical features within the corpus. In order to address the problem, (Term frequency-versus-document frequency) analysis TF-IDF was used. In term weighting applications, it is normally assumed that variables with the highest TF-IDF values are to be the most important. Given that hypothesis, variables 1-250 (representing the highest TF-IDF values) were only retained. This had the effect of reducing the matrix into just 250 lexical variables.

For validity purposes, principal component analysis (PCA) was used. PCA is one of the most reliable data reduction methods. It was revealed that the highest 217 lexical variables were the most important. In order to make sure that the corpus now includes only the most distinctive lexical variables, only the recurring or repeated variables in both TF-IDF and PCA tests were finally selected. This had the effect of reducing the matrix to only 113 lexical variables or words as shown in Table I.

Using Cluster Analysis, the 650847 speakers were classified into four main clusters, as shown in Fig. 1.

Referring to the personal information of users, it was found out that the clustering was not based on any geographic or regional grounds. The most distinctive lexical features of each cluster were thus investigated. It was clear then that clustering was based on thematic grounds. Accordingly, thematic words as well as proper names including murder, honor, killing, Israel, Palestine, and Trump were all deleted. This had the effect of reducing the corpus into just 68

variables. The assumption now is that any grouping of the tweets and users will not be based on thematic grounds.

Once again, cluster analysis was carried out for the Matrix colloquial_arabic_dial_corpus (650847, 68) where the former represents the number of users and the latter the number of lexical variables. Results are shown in Fig. 2.

TABLE I. EXTRACTED LEXICAL VARIABLES THROUGH THE EXECUTION OF PCA AND TF-IDF

عراقي	سوري	لاجئ	بلدنا	وظايفنا
Iraqi	Syrian	refugee	our country	Our jobs
عار	فلسطين	ترامب	اسرائيل	شرف
shame	Palestine	Trump	Israel	honor
دمار	قتل	مخنث	ميسي	شغل
destruction	murder	gay	Messi	job
مصاري	فلوس	بقد	حق	مطاعم
money	money	Really!		restaurants
بالحق	انتخابات	غزة	نتنياهو	عن جد
Really!	elections	Gaza	Netanyahu	Really!

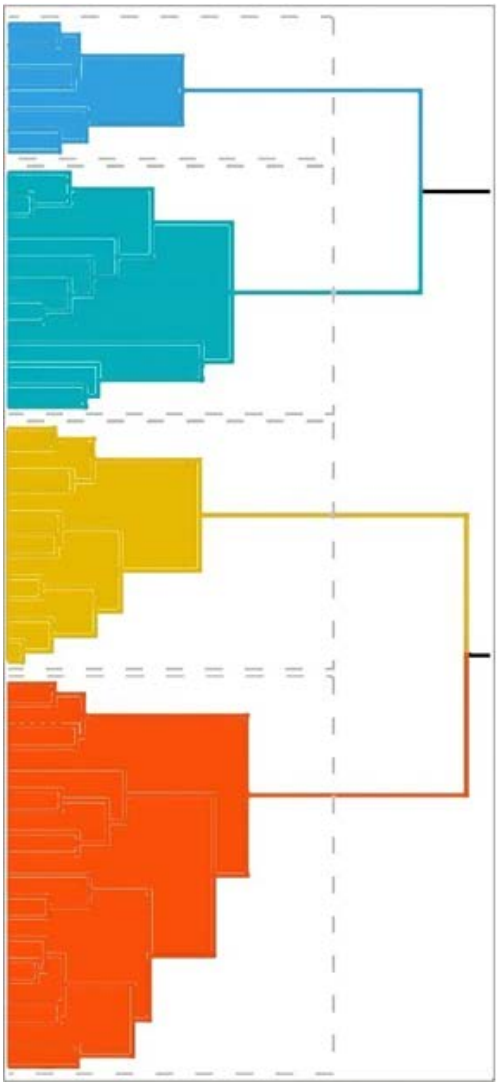

Fig. 1. A Cluster Analysis of the 650847, 113 Matrix.

The Matrix rows were assigned to nine clusters as shown in Fig. 2. Comparing the results of the clustering structures to the personal information available about the users, it was quite obvious that clustering was based on regional basis. These can be referred to as Groups 1-9. Thus, it can be claimed that clustering is based on regional basis.

Interestingly, more than 80% of the retained lexical variables which are considered the most distinctive features are best described as intensifiers and expressions of surprise. This may be due to the fact that such expressions are spontaneous in nature and frequently used in informal Arabic versions. In this regard, they (intensifiers and expressions of surprise) can be good indicators or predictors for mapping the linguistic variation in Colloquial Arabic. This will be the focus of the next section. The distinctive lexical features of each group are discussed.

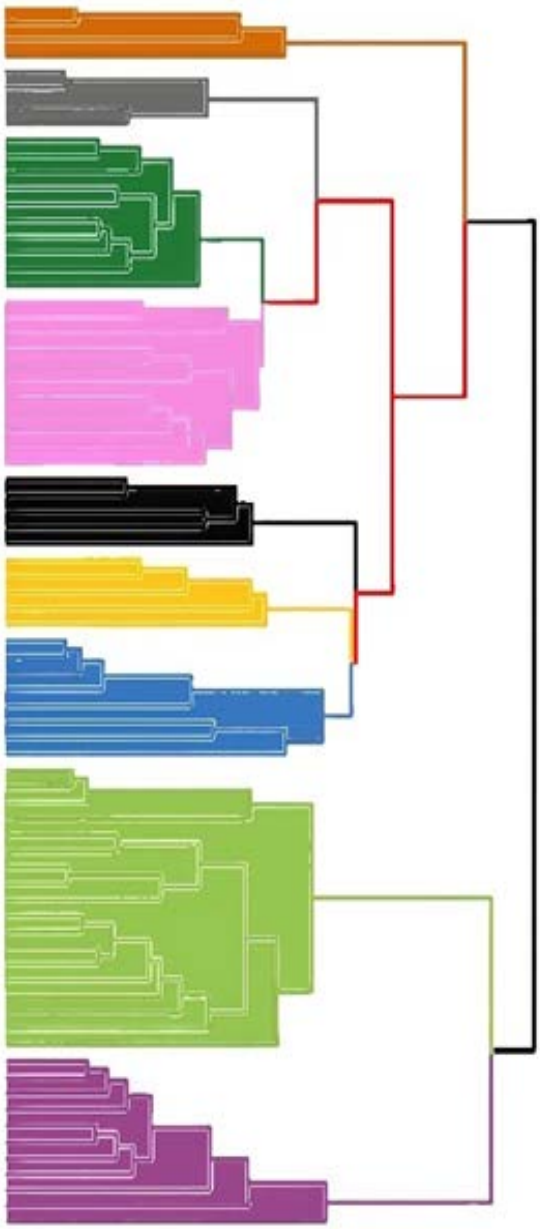

Fig. 2. A Cluster Analysis of the 650847, 68 Matrix.

IV. Analysis and Discussions

The clustering structure shown in Fig. 2 indicates that the datasets fall into nine distinct groups. It serves as a tool to gain insight into the distribution of data to observe characteristics of each cluster. In our case, each group is distinctive from other groups based on its lexical profile. Based on this clustering structure and the centroid analysis of the lexical features of each group or cluster, mapping the regional dialects of Colloquial Arabic can be useful.

It is obvious that intensifiers and expressions of surprise are the most distinctive lexical features of each cluster or speaker group. These expressions are normally known as degree words or intensifiers as they show a degree of intensity by the use of varied word classes [30-33]. According to Peters [34], intensifiers play a significant role in the social interaction and emotional expressions among language users. Based on these lexical features, the main regional dialects of Colloquial Arabic can be mapped through Iraq, the Arab Gulf, Levantine, Egypt, Tunisia, Morocco, Algeria and Libya, as shown in Fig. 3.

Interestingly, the four regional dialects of Algerian, Moroccan, Tunisian, and Libyan Arabic are traditionally classified under just one dialect known as Maghrebi Arabic. In our case, however, there are distinctive linguistic differences among these four dialects, as shown in Table II.

It is noticeable that Libyan and Tunisian dialects in the situation of expressing a surprise or joy are somehow close to each other. They both appear to include the use of 'حق' (hag), 'حقا' (haga) and 'بالحق' (balhag) interchangeably, depending on the mode of the speech. In fact, the word 'بالحق/حق/حقا' (haga/hag/balhag) is probably the most popular intensifying expressions that are employed by speakers of most Arabic dialects. This is due to its close derivational form from the Standard word 'حقيقة' (hagigatn) (truth; truly).

Algeria and Morocco primarily use different lexical choices in terms of showing joyful surprising news. The Algerians say 'منيتنك' (menitik) and the Moroccans use 'واش بصح' (wash'bishah) (what, is it true?) as first choice. These two different dialects sound arguably heavy to the ears of the Gulf and Levantine dialects speakers. Another choice that is popular in the Maghreb dialect is 'قول والله' (qul'walla) (swear by Allah), which received huge employment by most of the dialects speakers of the Arab region, but with huge difference in pronunciation, in terms of stress and pitch.

In Libyan Arabic, for example, a speaker would react to discomfort news as 'متقولهاش' (don't say it) or 'لا ياراجل' (No O man!), i.e. an intensifying phrase that vocally harmonizes with other Maghrebi and Egyptian dialects, and not peculiar to the ears of the Levantine and Mesopotamian dialects. Given the close geographical distance between Tunisia and Algeria, Morocco and Libya, the Tunisian expression of surprise 'يزي عاد' (that can be translated literally to 'stop it' and communicatively to 'really') has more tendency to be comprehended by speakers of those countries and easily identified as in the case of the Tunisian dialect.

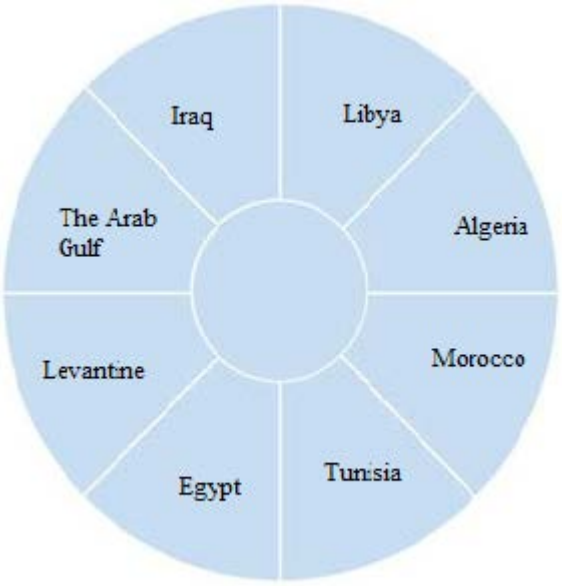

Fig. 3. Geographic Regions of Colloqual Arabic.

TABLE II. Maghreb's Distinctive Features of Intensifiers and Expressions of Surprise

Country	Libya	Algeria	Morocco	Tunisia
Expressions of surprise	Showing comfort			
	حق	منيتك	واش بصح	بالحق
Transliteration	haq	menitik	wash'bishah	balhag
Literal meaning	True	Kidding	What true	True
Communicative meaning	Really	Really	Really	Really
Expressions of surprise	Showing discomfort			
	متقلهاش	قول والله	بصح	يزي عاد
Transliteration	matqulhash	qul'walla	Besah	yazi'aad
Literal meaning	Don't say it	Swear by Allah	Correct	Stop it
Communicative meaning	Really	Really	Really	Really

Table III shows some of the distinctive features of dialectal intensifiers that are used by Maghrebi speakers as a form of intensives or downtoners– showing maximization and minimization. These of course are not the only ones used but as appeared in the methodology, they are more frequently used in the Maghreb region. Bolinger [33] refers to terminologies that are classified as 'adverbs' as amplifiers and downtoners. The latter are adverbs that usually reflect a small amount of quantity. The standard Arabic word for this is 'قليل' (qalil) which seems to have slight dialectal variations among Arabic dialects. However, the varieties of languages used in the Maghrebi dialect seem to influence its peripheries. The intensifier of maximization 'بزاف' is used in Algeria and Morocco. It is understood by the Tunisians and Libyans very well. This rings true to the argument of Harrat *et al.* [35] that dialects are morphologically and syntactically simplified, especially in the regions where one dialect coincides with one another.

TABLE III. MAGHREBI INTENSIFIERS

Country	Libya	Algeria	Morocco	Tunisia
Intensifiers	Showing maximization			
	هلبة	بزاف	بزاف	برشة
Transliteration	Halba	bezaf	bezaf	barsh
Meaning	A lot/too much/too many	A lot/too much/too many	A lot/too much/too many	A lot/too much/too many
Intensifiers	Showing minimization			
	شوي	شوية	شوية	شوية
Transliteration	Shwi	shwia	shwia	shwia
Meaning	A little/a few	A little/a few	A little/a few	A little/a few

This study explores a variety of distinctive lexical variations that are employed by dialect speakers of the Arab world. For instance, the Maghrebi and Egyptian dialect speakers appear to be having heated discussions on Twitter posts that handle political and cultural issues with respect to their individual countries. It is normal to see comments by different Maghrebi dialect users on a geopolitical topic with different lexical variations showing intensifiers and expressions of discomfort. This interaction on Twitter and other social media platforms has undoubtedly expanded the dialectological repertoire of speakers of Arabic dialects across the region.

It can be seen that Table IV shows a variety of lexical choices in expressing the concepts of showing surprise in positive and negative manners. It indicates that speakers' reflection of expression shows their positive attitude in one word at a time of the hearing of a particular piece of information that brings joy to their situation. In such case, the Egyptian speaker will say 'بقد' (beggad) (sure). This form can also be expressed by بجد (bejad) (sure). On the other hand, they would use 'ياخبر' to express discomfort of unwanted news, as shown in Table V.

As it has been previously mentioned, intensifiers are amplifiers and have the function of intensifying or maximizing a certain quantity [36]. The Egyptian dialect is widely spread among other Arabic dialects and this is probably due to its TV industry and famous civilization. The Egyptian people use 'أوي' (*awii*) when they intend to amplify certain situations in conversations. On the contrary, the use of the downtoners 'اليل' (alil) that is derived from 'قليل' (qalil) to express a small amount of quantity or not giving much importance to a quality. This is similar to the distinctive lexical item or 'حتة' (hita). It can be seen that the Egyptian dialect has some vocabularies that are inspired from other Arabic dialects, but with phonological alterations. In fact, there is an evident variation in the vocalization of most of the vocabulary in Egyptian dialect, especial where 'ق' is pronounced as 'أ' (a).

The Levantine group of dialects includes Lebanese, Syrian, Jordanian, and Palestinian dialects. They are quite similar in their ways of expressing intensification. For example, speakers from Lebanon use the expression 'عن جد' (aan'jed) or combined 'عنجد' (aanjed) – as in the case of Syrian and Jordan – to react to happy news. Similarly, in Palestine, they say 'بجد' (bejad), altering the first letter from 'ع' (a) into 'بـ' (b). Those expressions of surprise are close to the Egyptian expression with a slight alteration in the way letter 'ق' is uttered. Although the Levantine expressions of surprise are understood in the Maghreb region, they are rarely or never used by the speakers of the Maghreb region dialects.

The lexical items in Table VI are a result of mapping expressions of a speaker receiving sudden discomfort news of a given social phenomenon. It shows the differences among Arabic dialects, as shown elsewhere in this research. It has to be expressively evident that context is a determining factor in deciding the tone of the speaker; whether their reaction is positive or negative towards a particular intake. For example, the phrase 'قول والله' (swear by Allah), explained above, can be used in both contexts – expressions of happiness and sadness by most of speakers of Arabic dialects. The intensification use is, in fact, not restricted to certain dialects. It is used by dialects of Arabic in colloquial contextualized situations with different intonations and stresses on certain syllables. In this respect, Díaz-Campos and Navarro-Galisteo [37] suggest that the geographical factor in relation to dialects can play a significant role in recognizing lexical variations. This rings true with regard to the Levantine 'عن جد' (sure?) expression in relation to Lebanon, Syria, Jordan, and Palestine, as shown in Table VII.

TABLE IV. DISTINCTIVE FEATURES OF INTENSIFIERS AND EXPRESSIONS OF SURPRISE IN EGYPTIAN ARABIC

Country	Egypt
Expressions of surprise	Showing comfort
	بجد
Transliteration	beggad
Literal meaning	Sure
Communicative meaning	Really
Expressions of surprise	Showing discomfort
	ياخبر
Transliteration	ya'khber
Literal meaning	What a news
Communicative meaning	Really

TABLE V. EGYPTIAN INTENSIFIERS

Country	Egypt
Intensifiers	Showing maximization
	أوي
Transliteration	awii
Meaning	A lot/too much/too many
Intensifiers	Showing minimization
	شوي/اليل/حتى
Transliteration	Shwi/alil/hita
Meaning	A little/a few

TABLE VI. LEVANTINE'S DISTINCTIVE FEATURES OF INTENSIFIERS AND EXPRESSIONS OF SURPRISE

Country	Lebanon	Syria	Jordan	Palestine
Expressions of surprise	Showing comfort			
	عن جد	عنجد	عنجد	بجد
Transliteration	Aan'jed	Aanjed	Aanjed	bejad
Literal meaning	Sure	Sure	Sure	Sure
Communicative meaning	Really	Really	Really	Really
Expressions of surprise	Showing discomfort			
	والله عنجد	شو	والله	ايش
Transliteration	walla'anjed	shoo	walla	Ish
Literal meaning	Swear by Allah it is sure	what	Swear by Allah	what
Communicative meaning	Really	Really	Really	Really

TABLE VII. LEVANTINE INTENSIFIERS

Country	Lebanon	Syria	Jordan	Palestine
Intensifiers	Showing maximization			
	كثير	اكتير/خيرات الله	كوم/كوميات	كتير
Transliteration	kathir	Ekti/khirat allah	kom/komiat	ektir
Meaning	A lot/too much/too many	A lot/too much/too many	A lot/too much/too many	A lot/too much/too many
Intensifiers	Showing minimization			
	شوي	شوي	شوي	شوي
Transliteration	shwi	shwi	shwi	shwi
Meaning	A little/a few	A little/a few	A little/a few	A little/a few

The Levantine intensifiers are almost the same, especially those of downtoners. Data shows that to express a small amount of something, or to explain that certain quality or quantity is low, the minimizing term used is 'شوي' (shwi) or 'شوية' (shwia) depending on the contextual and other singular/plural, masculine/feminine factors. However, the utterance is sometimes colloquially-based and is not subject to syntactic structure. On the other hand, Levantine dialects, as seen in the data, employ the intensifier 'كتير' (ktir) to express maximization of quality or quantity. The morphological process in uttering the term involves stressing and stretching one sound more than the others, such as the case of Syria and Palestine, where the first syllable is stressed. Therefore, difference among Levantine dialects appears only in the surface and can hardly establish a distinction among its speakers. It can be argued that dialects in the Levantine region, and other regions in the Arab world, borrow many words from each other and use them interchangeably. However, the Jordanian intensifier 'كوم' or 'كوميات' (kom/komiat) is properly the most marked and distinctive one appeared in the data.

In the Gulf dialect, and Saudi Arabia in particular, speakers use 'أكيد' (akid). This is another form that is close to Standard Arabic and is usually used in most of the other mapped dialects. However, in the Gulf region, 'أكيد' (akid) appears to eclipse all the other existing intensifiers that express surprise. The data shows that it appears to be widely used in this geographical area. It is, in fact, a standardized form that can be understood by all Arab speakers. It is also used in Iraq in parallel with the intensifier 'صُدك' (sudek), the first syllable receives the primary stress. Expressions of surprise in Arabic dialects vary tremendously, but speakers of neighboring dialects may feel more comfortable where their lexical features coincide with other countries that have boundaries with. On the other hand, the expression 'ابد' (abad) (no/never?) is somehow a distinctive lexical item that expresses discomfort in disturbing situations. It enjoys a high frequency of usage in the Gulf region. It is important to stress here that these are not the only expressions of surprise that Gulf dialect speakers use, but these are the most frequent as per our data, as shown in Table VIII.

The intensifier 'واجد' (wajed) appears to be widely used in the Gulf Arabic, as shown in Table IX. Omar and Alotaibi [38] indicate that the term is used to state that something is provided in plenty in Saudi context. It is worth noting that this utterance has also been mapped in other countries in the regions, such as Libya, Tunisia and Sudan. To a certain degree, the lexical intensifier 'واجد' (wajed) is sometimes substituted by 'كتير' (katir), which is pronounced differently from the Levantine dialect. Contrary to maximization, the minimizer 'شوي' is also frequently used in the Gulf as is the case with almost all Arabic dialects.

According to Ito and Tagliamonte [39], the use of intensifiers is linked to colloquial usage and dialectal varieties. This argument is supported by the findings of this study, where the expressions collated from the data present a degree of difference cross the Arab countries surveyed in this study. The Iraqi dialects exhibit the most distinctive lexical features of the use of intensification, shown in Table X. In fact, some intensifiers still compete in occupying the first place or what really comes subconsciously to mind of speakers of a particular dialect. For instance, in Iraq, the use of intensifiers in expressing surprise in a positive manner is 'صدك' (sudek), but this could not be the case in other sub-regions in Iraq, where 'حقا' (haga) takes over. 'صدك' (sudek), with stressing the first syllable, is derived from the word 'صدق' (truth) and is used here to say that 'is this true?'; in other words, 'are you sure?'. On the contrary, the Iraqis generally react to discomfort news by 'جذب' (jedeb), meaning 'كذب' (lie) – 'you are lying', pronounced in a rising intonation to form a question, as shown in Table XI.

The intensifying utterance in Iraqi dialect is also a distinctive one that subscribes to the Standard Arabic word 'كثير' (kathir), but with a change of the first sound to 'j'. In terms of using minimization, the Iraqi downtoner is not different from the other Arabic dialects surveyed in this study. Quirk, et al. [40] state that downtoners are minimizing items that lessen the degree of intensity of something, they lower the efficacy to a degree of small extent. They always offer a downwards scale to things. Of course, they can be expressed

by using different words classes and structure, but as our data is limited to lexical items that are usually of one morpheme, rather than using 'chucks' or 'fixed phrases', the data show that the most frequent minimizer among Arabic dialects is 'شوي' (shwi), shown in Table XII.

TABLE VIII. GULF'S DISTINCTIVE FEATURES OF INTENSIFIERS AND EXPRESSIONS OF SURPRISE

Country	**The Arab Gulf**
Expressions of surprise	Showing comfort
	أكيد
Transliteration	Akid
Literal meaning	Sure
Communicative meaning	Really
Expressions of surprise	Showing discomfort
	ابد
Transliteration	abad
Literal meaning	No/Never
Communicative meaning	Really

TABLE IX. GULF INTENSIFIERS

Country	**Saudi Arabia**
Intensifiers	Showing maximization
	واجد/كتير
Transliteration	Wajed/katir
Meaning	A lot/too much/too many
Intensifiers	Showing minimization
	شوي
Transliteration	shwi
Meaning	A little/a few

TABLE X. IRAQ'S DISTINCTIVE FEATURES OF INTENSIFIERS AND EXPRESSIONS OF SURPRISE

Country	**Iraq**
Expressions of surprise	Showing comfort
	صدك
Transliteration	sudek
Literal meaning	Truth
Communicative meaning	Really
Expressions of surprise	Showing discomfort
	جذب
Transliteration	jedeb
Literal meaning	You lie
Communicative meaning	Really

TABLE XI. IRAQI INTENSIFIERS

Country	**Iraq**
Intensifiers	Showing maximization
	جثير
Transliteration	jethir
Meaning	A lot/too much/too many
Intensifiers	Showing minimization
	شوية
Transliteration	shwia
Meaning	A little/a few

TABLE XII. OTHER REGIONS' DISTINCTIVE FEATURES OF INTENSIFIERS AND EXPRESSIONS OF SURPRISE

Country	**Sudan**
Expressions of surprise	Showing comfort
	صحي
Transliteration	sahi
Literal meaning	Sure
Communicative meaning	Really
Expressions of surprise	Showing discomfort
	جدي
Transliteration	jedi
Literal meaning	No/Never
Communicative meaning	Really

Watson [41] deduces that the tendencies of using different lexical choices in colloquial interactions, in fact, appear to be a unification factor to the Arabic dialects. She attributes this to the regional tendencies of language usage among Arabs as well as to the "predictable phonological processes" that Standard Arabic goes through in deriving and uttering words [41]. The use of expressions of surprise and intensifiers travels across Arabic dialects and some of them give way to others in different regions. For example, the use of 'قول والله' (qul'walla), discussed above, or its sister variations "أحلف' (ahlef) (swearing as promising), 'والله' (walahi), and 'بالله عليك' (balhi alik) can be used both ways: to express welcomed and unwelcomed effects, depending on the context. It is worth noting that such intensifiers bear religious nuances in their conceptual structure. This expressive feature of showing surprise is more of idiosyncratic nature in colloquial usage of intensifiers.

Although the religious nuances appeared in the data at hand can be found across the Arabic dialects, Sudanese speakers would probably use both صحي (sahi) or جدي (jedi) to express surprise. The latter, which is used to express uninvited news, is closer to the Egyptian 'beggad'. Further to such distinctive lexical items, the Sudanese expression of discomfort 'ما تهظر' (ma'tehadher) (Don't joke) could pose difficulty to the Levantine users to comprehend, especially when uttered in decontextualized conversations. This is not due to the meaning of the expression as the words cause no challenge, but due to

its phonetic features as it is pronounced in a way that make it troublesome to the ears of other Arabic dialect users. It shows a change in the morphological structures of the consonants that are somehow close to the Egyptian 'ما تهزر' (ma tehazar) (Don't joke). Khrisat and Harthy [42] attribute such changes of morphological structure to the 'economy of effort' and 'ease of articulation' that speakers of certain dialects adopt.

TABLE XIII. SUDANESE INTENSIFIERS

Country	**Sudan**
Intensifiers	Showing maximization
	تُف
Transliteration	tuf
Meaning	A lot/too much/too many
Intensifiers	Showing minimization
	شوية/حبة
Transliteration	shwia/haba
Meaning	A little/a few

Unsurprisingly, Twitter provides exposure to all dialects of Arabic languages. It is a place where a variety of linguistic experience is gained by interacting with other dialect users. In fact, many Arabs tend to form friendships with people from other Arab countries who speak a different dialect. Such interaction has developed the linguistic reservoir among them in terms of comprehension. The minimizing downtoner conveys no difference to the above dialects. It is unmarked lexical item that seems to be rigorously used among Arab speakers. However, a more distinctive downtoner item is 'حبة' (haba). The Sudanese dialect lends similarity to the Egyptian dialect in expressing lexical items that feature minimization. According to [35], differences in lexical items among dialects are marked with variations in form. This indicates that even though dialects differ in their morphological structure, they still represent the same meaning. Further, the Sudanese dialect is marked with the distinctive use of 'تُف' (tuf) or 'فُل' (ful) to express intensity, as shown in Table XIII. Arguably, those phrases are challenging to people of other Arabic dialects, especially the Levantine and Iraqi.

V. CONCLUSION

In this paper, we have mapped the most frequent linguistic variations in most popular six dialects in Arabic language. For the purpose of limitation of such tremendous amount of data, this study proposed a computational statistical model for mapping regional dialectology in Colloquial Arabic through Twitter based on the lexical choices of speakers in relation to use of intensifiers and expressions of surprise. Using Twitter corpus of 12778784 words, we had to use a mathematically vector space matrix, henceforth referred to as colloquial_arabic_dial_corpus. The matrix is built out of rows and vectors. With such high-dimensionality of data, we had to adopt a term-frequency-versus-document frequency analysis (TF-IDF). This had the effect of reducing the matrix colloquial_arabic_dial_corpus into just 250 lexical variables. However, and for validity purposes, a reduction method of principal component analysis (PCA) was also employed. This had the effect of reducing the matrix to only 113 lexical variables or words. With this result of corpus reflecting the most frequently used distinctive dialects in colloquial Arabic in relation to intensifiers and expressions of surprise, we began to explore shapes of similarities and differences among Colloquial Arabic dialects.

Hence, twitter corpus was applied to Colloquial Arabic. It showed that Arab people use different lexical items in expressing their surprise. In fact, synonymous occurrences of every expression exist in bounty in every dialect in the study. This would permeate more linguistic variations and colloquial choices. Such variations could be in the morphological or phonological structures of the pattern. Most of the dialects of the Arab region exhibit huge difference in pronunciation in terms of intonation, stress and pitch. Further, the work is consistent with other studies claiming that speakers of neighboring dialects would have more tendencies to understand each other and can easily identify their dialects. Furthermore, it can be claimed that social media platforms such as Twitter is a reservoir of different dialects and a presentation mirror of lexical variations. This has been shown in the discussions of political and cultural issues with respect to their individual countries. Twitter users express geopolitical topics with lexical variations showing a degree of differences of intensifiers and expressions of discomfort. Those utterances sometimes depict a slight alteration in their orthographic structure.

Our findings show that speakers of different Colloquial Arabic dialectal varieties can distinguish speakers of other countries by the vocalization they use. More importantly, the data shows that the minimizer 'شوي' is almost used by all Arabic dialects with intonational differences. Our work could be expanded by examining other linguistic concepts rather than intensifiers and expressions of surprise as we come across the fact that regular contact with other dialects, especially through social media, contributes to the findings of aspects of similarities and differences among Arab speakers. We suggest further research in exploring emerging words and origin of terms that travel across regions in the Arab world. This would give a clearer pattern of their origin and more information on dialect variations in Arabic language.

ACKNOWLEDGMENTS

This publication was supported by the Deanship of Scientific Research at Prince Sattam bin Abdulaziz University, Alkharj, Saudi Arabia and the Research, Consulting and Training Center at the University of Tripoli, Tripoli, Libya. Authors would like also to thank Dr. Samira Farhat for her helpful comments and insightful suggestions.

REFERENCES

[1] P. Eckert, Meaning and Linguistic Variation: The Third Wave in Sociolinguistics. Cambridge: Cambridge University Press, 2018.

[2] J. K. Chambers and N. Schilling, The Handbook of Language Variation and Change. Wiley, 2018.

[3] D. Brouwer, Gender Variation in Dutch: A Sociolinguistic Study of Amsterdam Speech. Berlin: De Gruyter, 2011.

[4] J. Coates, Women, Men and Language: A Sociolinguistic Account of Gender Differences in Language. Taylor & Francis, 2015.

[5] J. King and S. Sessarego, Language Variation and Contact-Induced Change: Spanish across space and time. John Benjamins Publishing Company, 2018.

[6] A. Georgakopoulou and T. Spilioti, The Routledge Handbook of Language and Digital Communication. Taylor & Francis, 2015.

[7] E. Teich, Cross-Linguistic Variation in System and Text: A Methodology for the Investigation of Translations and Comparable Texts. Berlin:De Gruyter, 2012.

[8] M. Krug and J. Schlüter, Research Methods in Language Variation and Change. Cambridge: Cambridge University Press, 2013.

[9] G. Parodi, Working with Spanish Corpora. Bloomsbury Publishing, 2007.

[10] W. Maguire and A. McMahon, Analysing Variation in English. Cambridge Cambridge University Press, 2011.

[11] C. Boberg, J. A. Nerbonne, and D. Watt, The Handbook of Dialectology. Blackwell, 2018.

[12] Z. Cao, Linguistic Atlas of Chinese Dialects. Beijing: The Commercial Press, 2008.

[13] H. Goebl, "Dialectometry and quantitative mapping," in Language and Space. An International Handbook of Linguistic Variation, vol. 2, R. K. Lameli, & S. Rabanus Ed. (Language Mapping, Berlin/New York: De Gruyter Mouton, 2010, pp. 433–457.

[14] S. Rabanus, "Language Mapping Worldwide: Methods and Traditions," in Handbook of the Changing World Language Map, K. R. Brunn S., Ed. Cham: Springer, 2020.

[15] P. Rácz, Salience in Sociolinguistics: A Quantitative Approach. Berlin: De Gruyter, 2013.

[16] E. Benmamoun and R. Bassiouney, The Routledge Handbook of Arabic Linguistics. Taylor & Francis, 2017.

[17] J. M. Hernández-Campoy and J. C. Conde-Silvestre, The Handbook of Historical Sociolinguistics. Wiley, 2012.

[18] J. Grieve, A. Nini, and D. Guo, "Mapping Lexical Innovation on American Social Media," Journal of English Linguistics, vol. 46, no. 4, pp. 293 –319, 2018.

[19] H. Moisl and W. Maguire, "Identifying the Main Determinants of Phonetic Variation in the Newcastle Electronic Corpus of Tyneside English," Journal of Quantitative Linguistics, vol. 15, no. 1, pp. 46-69, 2008.

[20] H. Moisl, "Statistical corpus exploitation," in Handbook of Corpus Phonology, J. Durand, Gut, U., Kristofferson, G., Ed. Oxford: Oxford University Press, 2010.

[21] M. Muzikant, "Verkleinerungsformen in den deutschen Dialekten Mährens und Schlesiens - eine historische Reminiszenz (Diminutives in German Dialects in Moravia and Silesia - Historical Reminiscence)," in Sprachen verbinden Beiträge der 24. Linguistik- und Literaturtage, Brno/Tschechien, 2018.

[22] M. Wese and M. Muzikant, Atlas der deutschen Mundarten in Tschechien. Band III Lautlehre 2: Langvokale und Diphthonge (Atlas of german dialects in Czech republic. Volumen III Phonetics 2: Long vowels and diphthongs). Tübingen: Narr/Francke. 2016.

[23] K. Simet and M. muzikant, Atlas der deutschen Mundarten in Tschechien. Band IV Lautlehre 3: Konsonanten (Atlas of german dialects in Czech republic. Volumen IV Phonetics 3: Consonants). Tübingen: Narr/Francke, 2016.

[24] M. Rosenhammer, A. Dicklberger, D. Nuzel, and M. Muzikant, Atlas der deutschen Mundarten in Tschechien. Band II Lautlehre 1: Kurzvokale. (Atlas of german dialects in Czech Republic. Tuebingen: Narr/Francke, 2014.

[25] C. Xu and L. Mao, "The sociolinguistic meanings of syllable contraction in Chinese: A study using perceptual maps," Asia-Pacific Language Variation, vol. 3, no. 2, pp. 160-199, 2017.

[26] G. Roche and H. Suzuki, "Mapping the Minority Languages of the Eastern Tibetosphere," Studies in Asian Geolinguistics, vol. 6, pp. 28-42, 2017.

[27] S. Cullotta and G. Barbera, "Mapping traditional cultural landscapes in the Mediterranean area using a combined multidisciplinary approach: Method and application to Mount Etna (Sicily; Italy)," Landscape and urban planning, vol. 100, no. 1-2, pp. 98-108, 2011.

[28] M. Barni and G. Extra, Mapping linguistic diversity in multicultural contexts. Walter de Gruyter, 2008.

[29] H. Mulki, H. Haddad, M. Gridach, and I. Babaoglu, "Tw-StAR at SemEval-2017 Task 4: Sentiment Classification of Arabic Tweets," in Proceedings of the 11th International Workshop on Semantic Evaluations, Vancouver, Canada, 2017, pp. 664–669: Association for Computational Linguistics.

[30] O. Zaidan and C. Callison-Burch, "Arabic dialect identification," Computational Linguistics, vol. 52, no. 1, pp. 1-36, 2012.

[31] W. Dressler and L. M. Barbaresi, Morphopragmatics: diminutives and intensifiers in Italian, German, and other languages. Berlin, New York: Walter de Gruyter, 1994.

[32] C. Paradis, Degree Modifiers of Adjectives in Spoken British English. Lund: Lund University Press, 1997.

[33] D. Bolinger, Degree Words. The Hague & Paris: Mouton, 1972.

[34] H. Peters, "Degree adverbs in early modem English," in Studies in Early Modern English, D. Kastovsky, Ed. Berlin & New York: Walter de Gruyter, 1994, pp. 269-288.

[35] S. Harrat, K. Meftouh, M. Abbas, W.-K. Hidouci, and K. Smaïli, "An Algerian dialect: Study and Resources," International Journal of Advanced Computer Science and Applications, vol. 7, no. 3, pp. 384-396, 2016.

[36] R. Plo Alastrué and C. Pérez-Llantada, English as a Scientific and Research Language: Debates and Discourses, English in Europe. Berlin, Geramny; Boston, MA, USA: De Gruyter, 2015.

[37] M. Díaz-Campos and I. Navarro-Galisteo, "Perceptual Categorization of Dialect Variation in Spanish," in Selected Proceedings of the 11th Hispanic Linguistics Symposium, J. e. a. Collentine, Ed. Somerville, MA: Cascadilla Proceedings Project, 2009, pp. 179-195.

[38] A. Omar and M. Alotaibi, "Geographic Location and Linguistic Diversity: The Use of Intensifiers in Egyptian and Saudi Arabic," International Journal of English Linguistics, vol. 7, no. 4, pp. 220-229, 2017.

[39] R. Ito and S. Tagliamonte, "Well weird, right dodgy, very strange, really cool: Layering and recycling in English intensifiers," Language in Society, vol. 32, pp. 257-279, 2003.

[40] R. Quirk, S. , G. Greenbaum, G. Leech, and J. Svartvik, A Comprehensive Grammar of the English Language. London: Longman, 1987.

[41] J. Watson, "Arabic dialects " in The Semitic Languages: An internationalHandbook, S. Weninger, Khan, G, Streck, M and Watson, J, Ed. (Handbooks of Linguistics and Communication Science, Berlin Walter de Gruyter, 2011, pp. 851-896.

[42] A. A. Khrisat and Z. A. Harthy, " Arabic dialects and Classical Arabic Language," Advances in Social Sciences Research Journal, vol. 2, no. 3, pp. 254-260, 2015.