An Exploratory Study of The Development of a Speech Corpus Annotated for the Main Arabic Dialects

*** Eiman Tawfeeq Alsharhan**
**** Allan Ramsay**

* Assistant Professor, Dept. of Arabic Language & Literature, College of Arts, Kuwait University, Kuwait

** Professor, School of Computer Science, Faculty of Engineering and Physical Sciences, University of Manchester, UK

Abstract

Arabic varieties differ substantially in all aspects of linguistics. These differences call for dialect specific modeling when building Arabic automatic speech recognition systems. The paper introduces the development of a multi-dialect annotated corpus of dialectal Arabic with data obtained from Linguistic Data Consortium (LDC). The annotation process is applied to GALE (phase 3) broadcast news and broadcast conversational speech. The annotation process resulted in assigning a dialect label for about 2900 speakers who contributed to this substantial Arabic resource. The final evaluation of the annotations shows that it achieved a substantial level of agreement. The annotations are fully available online for searching and downloading along with a set of access tools to help extract specific information from the database. The researchers' goal is for this dataset to be used for the development of NLP applications, which pay attention to issues that arise because of the wide range of Arabic accents.

1 Introduction and Objectives

The availability of comprehensive corpora is a major factor in the recent advance in natural language processing development and evaluation. Arabic, however, can still be considered a relatively poorly resourced language when compared to other languages such as English (Alsharhan and Ramsay). This lack has negative consequences in performing Arabic NLP tasks; researchers have to start from scratch and spend considerable time in constructing their own resources (Shoufan and Alameri). This will consequently lead to the reluctance of carrying on research in this area. This obstacle is more serious when discussing dialect resources which are particularly poor and limited.

There are two main forms of Arabic used at the present time: Modern Standard Arabic (MSA), which is widely taught at schools and universities, used in the media, formal speeches, courtrooms, and any kind of formal communication. MSA is considered to be the official language in all Arabic speaking countries. In addition, people generally speak in their own dialects in daily communication. These dialects are neither taught at schools nor even have any standardised written form. Arabic dialects differ substantially from MSA in terms of phonology, morphology, vocabulary, and syntax, as will be seen in section 3. Dialectal Arabic (DA), also known as Colloquial Arabic, is the natural spoken language in everyday life. It varies from one country to another and sometimes more than one dialect can be found within a country. MSA is, therefore, not a native language of any Arab person. Speakers learn it from schools while using their own dialect in everyday communication.

The stark dissimilarity between different forms of Arabic poses a problem for developing NLP applications. For instance, a speech recogniser trained with Gulf dialect is unlikely to work effectively with speech produced with an Egyptian dialect. The differences in sounds, vocabularies, and syntax between MSA and the dialects and between the dialects themselves make it necessary to use dialect specific corpora instead of MSA speech or text-based corpora or even a mixture of dialects corpora when building dialects applications.

This paper presents the first phase of an attempt at building an efficient Arabic speech recognition system (ASR). This phase aims at providing information about the speakers' regional origin, which is a great source of variation in speech. This information can be employed to develop a dialect detection tool to help classify speakers during training and testing the ASR system.

The research introduces an annotated version of the GALE (phase 3) broadcast news and broadcast conversation data to include information about the speakers' dialects. The annotation process resulted in assigning a dialect label of about 2900 speakers. Each speaker was assigned to an accent group by three

annotators. Annotators did not always agree; hence, the final database records all three judgments, so that developers who wish to use this data can choose whether to use examples where the annotators were unanimous or whether examples where two out of the three agreed are acceptable.

The paper describes the annotation effort made to identify the speakers' dialect of this substantial Arabic resource, including the online annotation interface, our methodology in choosing speakers from GALE data, and a description of the annotators who participated in the development of this dataset.

The annotators were assessed to ensure their reliability and only reliable annotators were chosen. Annotations were eventually evaluated by measuring inter-annotator agreement. The final assessments show that the refined annotations are highly consistent.

The annotations are fully available online for searching and downloading(1). A set of access tools is also shared to help extracting specific information from the database. However, due to licensing conditions of the Linguistic Data Consortium (LDC), the associated wave and text files cannot be shared. The annotations developed in this paper are, hence, an enrichment of the GALE data, but using them does require access to this data.

In addition to the aforementioned topics, the paper gives a detailed discussion of Arabic varieties and linguistic differences between MSA and dialects and between dialects themselves. The linguistic study covers all language aspects as found in the GALE (phase 3) speech data. At the end of this paper, a thorough analysis is given in regards to the collected annotations. The analysis covers the following topics: distribution of dialectal label in the GALE data, sources of low agreement as found in some cases, and interdialectal confusability measurement.

2 Relation to Previous Work

This section introduces recent efforts in constructing linguistics resources to assess problems of dealing with dialectal Arabic for NLP applications. Most of the work in the literature is focused on the written form of the language ((Cotterell and Callison-Burch), (Zaidan and Callison-Burch), (Elfardy and Diab, Al-Sabbagh and Girju) among others). However, the discussion in this section focuses on works that deal with developing multi-dialect spoken Arabic corpora.

Masmoudi et al. and Selouani and Boudraa have introduced a corpus for Tunisian Arabic and Algerian Arabic, respectively. The Tunisian corpus contains audio recordings and transcriptions extracted from dialogues in the Tunisian Railway Transport Network. The Algerian corpus is composed of MSA

speech pronounced by 300 Algerian native speakers from different regions. The developed corpora can be used to build dialect-specific Arabic ASR systems. Almeman et al. developed Arabic parallel texts and speech corpora that cover three main Arabic dialects: Gulf, Egyptian and Levantine as well as MSA. The 32 hours of recordings were collected with the aid of 52 participants. The researchers chose a specific linguistic domain to work with, namely travel and tourism.

Researchers in Biadsy et al. constructed the largest dialectal Arabic speech corpus with the aid of more than 125 million people in Egypt, Jordan, Lebanon, Saudi Arabia, and the United Arab Emirates. This multi-dialect corpus is used to build cross dialects speech recognition systems to support voice search, dictation, and voice control for the general Arabic speaking public.

The work described in this paper differs from previously constructed corpora in three ways:

- The researchers work on annotating a speech dataset that is well-known and publicly available to researchers in the field.
- The researchers' annotation is based on the classification of Arabic dialects into five groups, namely: Gulf (GLF), Iraqi (IRQ), Egyptian (EGY), Levantine (LEV), and Magharebi (MGR). This is the most complete coverage of any dialectal corpus known to the authors.
- Speakers are labeled with their dialectal origin even when they speak in MSA. This is important since the effect of the speakers' original dialect can be seen even when speaking in MSA, as shown in the example in section 3.

3 Linguistic Background

3.1 The Dialectal Varieties of Arabic

Arabic dialects can be classified based on different aspects, in terms of geography and social class. Regional dialects can be divided into the following five main groups: Gulf (GLF), Iraqi (IRQ), Egyptian (EGY), Levantine (LEV), and Magharebi (MGR). Figure 1 shows the geographical areas for different Arabic varieties in the Arab world.

3.2 Differences between MSA and DA

Despite the fact that MSA and DA have large overlaps, substantial differences can be observed between MSA and DA and between dialects themselves in terms of their syntax, morphology, phonology, and lexicon.

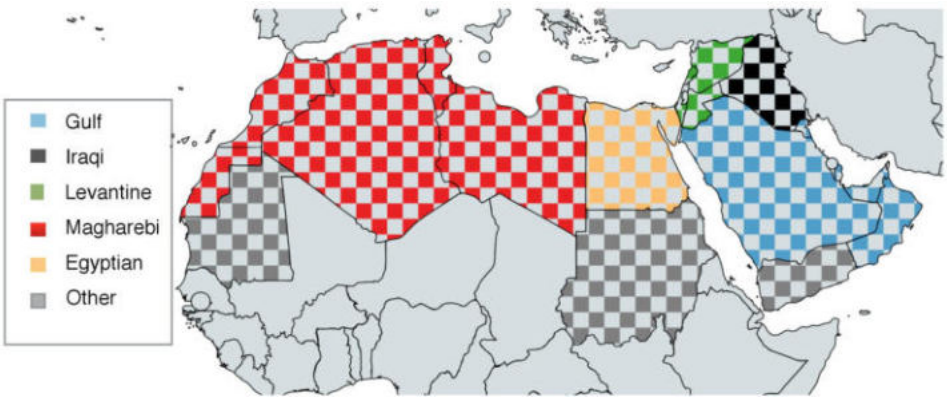


Figure 1: Division of Arabic dialects based on region.

MSA is regulated with strict linguistic rules established centuries ago describing the standards of forming words and sentences. Contrary to MSA, DA does not have any explicitly written set of grammar rules. In addition, DA has no standard orthography system, and there are plenty common words used in each specific region that are not part of MSA. Moreover, the speakers' regional dialects have a noticeable impact on the way they speak MSA. Take the example of the MSA word رجل /*rajul* (meaing: man) which could be pronounced /radʒul/ in Gulf countries and Iraq as the sound /dʒ/ in the English word *jungle*; /ragul/ in Egypt as the sound /g/ in the English word *get*; or /raʒul/ in Levant and Maghareb as the sound /ʒ/ in the English word *leisure*.

Below is a summary of the differences between MSA and DA in all linguistic aspects as found in the GALE (phase 3) speech data:

3.2.1 Morpho-syntactic Features

MSA and DA differ significantly in terms of syntax and morphology. The following is a short review of the phenomena that are generally observed from the data available to the researchers:

i. All Arabic dialects lack grammatical case, whereas MSA has a complex case system. Case endings are short vowels that are attached to the ends of words to indicate the word's grammatical function. Case endings are rarely explicitly written; however, they are supposed to be pronounced when speaking in MSA. Dialects tend to drop the grammatical case in all aspects of speech. A previous research found that speakers tend to drop the case markers even when speaking in MSA.

ii. Arabic varieties differ noticeably in style and sentence composition as well. For instance, because of its rich inflectional morphology, MSA allows a great deal of freedom in word order. For example, VSO, VOS, SVO, OVS are all allowable patterns in MSA. However, two prominent word orders are the most widely used in all Arabic varieties, namely: VSO and SVO. It is notable that MSA tends to prefer the use of word order VSO over the use of SVO, while most dialects prefer the alternative word order SVO. This observation is confirmed by Aoun and Sportiche study which found a higher incidence of VSO sentences in MSA compared to dialects.

iii. In MSA, the grammatical person and number as well as the mood are designated by a variety of prefixes and suffixes. Dialects use only a limited number of affixes. For instance, a present tense imperfect verb may be found in one of three grammatical moods: the indicative, the subjunctive and the jussive. For the indicative mood, the suffixes "wn" and "yn" are used for 2nd and 3rd person masculine plural and 2nd person singular feminine, respectively, while for the subjunctive and the jussive mood the letter "n" is deleted, and the suffixes used are "wA"[(2)] and "y". Some Arabic dialects, for example, Egyptian and Levantine speakers, make use of only "wA" and "y" even when the verb is in indicative mood. Take (1):

(1) *MSA*: /ʔanti taktubi:na Aldars/ أنت تكتبين الدرس (meaning: You write the lesson)

Egyptian: /ʔinti tuktubi: Aldars/ إنت تكتبي الدرس.

Similarly, dialects use only one form of noun suffixes such as "yn" instead of "wn" as in (2):

(2) *MSA*: /ʔalmudarisu:n mumta:zu:n/ المدرسون ممتازون (meaning: teachers are excellent).

Gulf: /ʔilmudarisi:n mumta:zi:n/ المدرسين ممتازين.

Also, the so-called "five nouns" are used in only one form in most Arabic dialects (e.g. ">bw" أبو (meaning: father of) instead of ">bA" أبا or ">by" أبي).

Moreover, MSA has a dual form in addition to the singular and plural forms, whereas the majority of dialects lack the dual form in both nouns and verbs. MSA has also two plural forms, one masculine and one feminine, whereas many (though not all) dialects prefer not to make such gendered distinction and use only masculine plural.

iv. Cliticisation system: Dialects make use of particular morphological patterns that do not exist in MSA and often make alternations to some morphological constructs. The following are some examples:

- Dialects have a more complex cliticisation system than MSA, allowing the use of circumfix negation and for indirect object forms, which appear in MSA as separate words (e.g. ly "to me", lahu "to him") that are attached onto the verb. For example, the MSA sentence: *lam taktubwhA lahu/* لم تكتبوها له (meaning: you did not write it to him) is merged into one word in Egyptian dialect *makatabtwhAlw$/* ما كتبتاهالوش . This can lead to complicated agglutinative constructs.
- Many Arabic dialects have specific affixation systems, allowing the use of developed prefixes and suffixes. For example, the prefix *sa-* -س (meaning: will) is converted to *Ha-* -ح in Egyptian. In Magharebi Arabic first person singular verbs begin with a *n-* -ن while in Egyptian the prefix *b-* is added in front of the verb in present tense. An example of the changes in suffixes is the plural suffix *–wA* -وا which is converted to *-w$* -وش in Magharebi Arabic.
- Various forms of Arabic demonstrative and relative pronouns appear shorter in many dialects compared to their classical form. For example, Gulf dialect uses *ha* ها instead of *hA*A* هذا (meaning: this). Magharebi dialect uses *dak- dik- duk/* داك- ديك- دوك (masculine/feminine/plural) instead of **Alk* (meaning: that). In Egyptian Arabic, the shortened demonstrative pronouns follow the noun. For example, *AlkitAb dah/* الكتاب ده (meaning: this book), *Albinti di/* البنت دي (meaning: this girl). MSA uses six varieties of relative pronouns. All dialect groups use only the shortened form of the relative pronouns (*illy* اللي) regardless of gender/number.

v. Some dialects are known for dropping specific articles and prepositions in some syntactic constructs. For example, the preposition *<lY* إلى (meaning: to) in *>nA rHt <lY Almadrasa* أنا رحت إلى المدرسة (meaning: I went to school) is typically dropped in most dialects. In addition, the particle *>n* أن (meaning: to) is dropped in the sentence *>nA mHtAj >n > nAm* أنا محتاج أن أنام (meaning: I need to sleep).

vi. Possessives (Idafa construction): Possessive pronouns take the form of suffixes in MSA. In addition to the use of these suffixes, different Arabic varieties replace the possessive suffixes with additional separate prepositions, as in (3):

(3) *MSA*: *mudarisyny* مدرسيني (meaning: my teachers)

Egyptian: *ilmudarisyn butwEy* المدرسين بتوعي

Magharebi: *ilmudarisyn dyali* المدرسين ديالي

3.2.2 Phonetics and Phonology

MSA has 36 phonemes, of which six are vowels, two diphthongs, and 28 are consonants. Arabic varieties differ significantly in terms of their phonemic system. Each variety poses major changes to the standard MSA phonemic system. Those changes can

appear by introducing new phonemes, deleting some phonemes, changing phonemes' duration, and alternating them. Those changes are more apparent in vowels than in consonants. Going through a comprehensive study of the phonetic system of each variety is rather complicated and out of the scope of this paper. However, the researchers of the paper go through the main differences that they found by analysing the spoken language in their data. The following substitutions in consonants are common:

i. The uvular stop ق /q/ is frequently voiced to [g], as the word /qa:l/ which is pronounced /gal/ by Gulf speakers, debuccalised to /ʔ/ as the pronunciation /ʔa:l/ by Egyptian speakers, or fronted to /k/ as the pronunciation /ka:l/ by rural Jordanian speakers. The sound /q/ can also be seen palatalised sometimes by Gulf speakers, as in the name of the city of *Sharjah*.

ii. The velar stop consonant ك /k/ is frequently palatalised to /tʃ/ in Iraq and some Arabian Gulf regions.

iii. The consonant ج /dʒ/ has many pronunciations. It is pronounced as /g/ in most of Egypt and some regions in Sudan, Yemen, and southwestern Oman. It is frequently softened to /ʒ/ in most North Africa and most of the Levant. In certain regions of the Arabian Gulf, it is palatalised to /j/.

iv. The dental fricatives ث /θ/ and ذ /ð/ have different realisations in Arabic dialects. For instance, the consonant /θ/ is pronounced as /s/ or /t/ in Egypt and Levant, and /t/ in the Magharebi dialect. Meanwhile, /ð/ consonant is pronounced as /z/ in Egypt and the Levant.

v. The emphatic consonants ض /dˤ/ and ظ /ðˤ/ were found to be merged in most Arabic varieties, where both letters are pronounced /ðˤ/. Egyptian speakers were found to make this distinction between both letters. However, they are inconsistent in the way they pronounce both letters. For instance, /dˤ/ was found to be pronounced /zˤ/ in the word ضابط (meaning: officer) while the /ðˤ/ was found to be pronounced /dˤ/ as in the word /ʔildˤuhr/ (meaning: afternoon). In addition, the consonant /ðˤ/ is pronounced as /zˤ/ in the Levant and Egypt as in the word عظيم /ʔazˤ i:m/ (meaning: great).

In MSA there are six vowel phonemes forming three pairs of corresponding short and long vowels, namely: /a, i, u, a:, i:, u:/. In his artical, "the length of Stem-Final Vowels in Colloquial Arabic", John McCarthy investigated whether Arabic vowels are the same at the phonetic level when spoken by speakers from different Arabic regions, including Saudi Arabia, Sudan, and Egypt. The author found that the phonetic implementation of the standard Arabic vowel system differs significantly according to dialects. The following is a brief description of the changes made to vowels in Arabic spoken varieties that are found in GALE speech data:

i. New vowels were introduced in addition to the basic three vowels and two diphthongs, such as /o/ and /e/ along with their long counterparts (Rosenhouse et al.) and (Alghamdi). Many examples of the extra types of vowels were found in our data. Take the examples: موت/mawt/ (meaning: death), which is pronounced /mo:t/, the word زيت/zajt/ (meaning: oil), which is pronounced /ze:t/, and the word مدن/mudun/ (meaning: cities), which is pronounced /mudon/ in many Arabic dialects.

ii. In some regions of the Levant, the vowel /a/ is raised to /e/. For example, the word /sitta/ستة (meaning: six) is pronounced /sitte/. Also, the word-medial /a:/ is raised to /e:/ such as in the word شباب /ʃaba:b/, which is pronounced /ʃebe:b/.

iii. Many Arabic dialects pose significant vowels shifting. For example, the pronouns of the second person are usually pronounced /ʔinta/ and /ʔinti/ (meaning: you (masc./fem.)) instead of the standard pronunciation /ʔanta/ أنتَ and /ʔanti/أنتِ. Another example is the pronunciation of the word كتب as /kitab/ instead of /katab/ by some Gulf accented speakers.

iv. In many Arabic varieties, collapse of short vowels can be observed. For instance, short /i/ and /u/ have collapsed to schwa /ə/ in many dialects and exhibit very little distinction. This collapse of the two closed vowels is either total or partial. For instance, some dialects (such as Magharebi) have only two short vowels, /a/ and /ə/, while other dialects (such as some Levantine dialects) show only partial collapse of the two closed vowels /i/ and /u/ where it is restricted to specific contexts. Meanwhile, a number of dialects still allow three short vowels (/a/, /i/, /u/) in all positions, such as Egyptian Arabic.

v. Many Arabic varieties have a tendency to delete short vowels (especially closed vowels) in many phonological contexts. This phenomenon is more obvious in word finals where short vowels have grammatical functions. Speakers prefer to delete final short vowels rather than assign the wrong ones. Deleting short vowels can also be observed in word-medial positions. For example, the word كتاب/kita:b/ is pronounced /kta:b/ in Iraqi Arabic.

vi. Insertion of a vowel into final two-consonant sequences is quite common in Arabic varieties. For example, the word حلم /ħilm/ (meaning: dream) is pronounced /ħilim/, the word بنت/bint/ (meaning: daughter) is pronounced /binit/, and the word بحر/baħr/ (meaning: sea) is pronounced /baħar/ in Levantine, Iraqi, and Gulf Arabic, respectively.

3.2.3 Rhythm and Intonation

Many annotators in this study found speech rhythm cues to be helpful to

distinguish speakers from different Arabic regions, especially between speakers from North Africa (Magharebi) and those from the Middle East. By comparing utterances from different speakers (mainly from the western and eastern regions), the researchers of this paper found that the reduced vowels duration and the complex syllable formation in Magharebi dialect are the main source for differences found in rhythmic structures. The significant difference in the duration of vowels between Western and Eastern speakers is found to be very obvious even when speaking in MSA. As a result, this rhythm metric could be used as a distinctive parameter within MSA to classify speakers according to their geographic region.

3.2.4 Lexicon

Arabic dialects differ significantly in the lexical choice. This difference results partly from the changes made to the structure of some words. These changes affect the properties of Arabic words and the lexical inventory. For example, the following particles can be found in different Arabic varieties:

i. Demonstrative and relative pronouns: ده/dah, دي/di, دول/dwl, هداك/hdAk, هذي/hA*y, هذول/ha*wl, اللي/Ally.

ii. Negative particles: مش/mi$, مو/mw, ماهو/mahw (masc.), ماهي/mahy (fem.).

iii. Interrogative particles: شو /$w, شنو/$nw, ايش/Ai$, ايه/Ayh.

Another source of variation is the presence of a substantial amount of vocabulary with no common roots with its MSA synonyms. Those vocabulary items are either loanwords such as ساندويش/sandwich, or new words that have evolved within the dialect. Examples for the new words are given in Table 1.

Table 1: Examples of Words That Have Evolved in Different Arabic Varieties

English Gloss	MSA	Arabic Varieties
I want	Auryd اريد	بدي/Bdy, عايز/EAyiz, ابي/Aby, بغيت/bgyt
Now	Alln الآن	دلوقتي/dilwa'ti, الحين/AlHyn, هلأ/halla, هسه/hassa
Leave me	Autrukny اتركني	هدني/Hidny, سيبني/sybny, عوفني/Ewfny
man	rajul رجل	رجال/Rajjal, راجل/ra:gil, زلمة/zalameh, زول/zol

An interesting experiment conducted by Kirchhoff and Vergyri (Kirchhoff and Vergyri) calculated vocabulary overlaps between different Arabic varieties as a percentage of shared unigrams, bigrams and trigrams. The authors then compared these numbers with equivalent statistics for two English varieties. It was found that the inventory of words (unigrams) in Arabic only overlaps by 10.3%. This can be compared to English; however, the percentage of shared unigrams was found to be 44.5%.

In both cases (changes in words' structure and the creation of new words), there have been significant transformations between the original vocabularies found in MSA and the current expression. The vocabulary choice provides with important information to help identify the speaker's dialect. However, since most Arabic speakers from different regions share much of the vocabulary, it is not always trivial to identify the speakers' dialect by solely looking for dialectal words in a given utterance. The linguistic differences and characteristics discussed above require a special approach to treat different varieties of Arabic speech.

4 Annotation Methodology

4.1 GALE Dataset

The annotation process is done to the GALE (phase 3) Arabic broadcast news and broadcast conversational speech dataset released by Linguistic Data Consortium (LDC). This dataset consists mainly of two parts: the first one contains approximately 132 hours of Arabic broadcast news speech (BN) collected from 13 Arabic channels. The second part contains approximately 129 hours of Arabic broadcast conversation speech (BC) collected from 17 channels. The dataset contains a combination of spontaneous and scripted speech. It consists of mainly MSA speech and a smaller but still significant amount of DA speech, especially in conversational recordings. This dataset is available exclusively for LDC membership holders.

4.2 Annotation Interface

In order to start the collection of annotations, a web-based annotation tool was created. Annotators are first asked to log in with their names. Afterwards, the annotation interface will display a number of speech files, randomly selected among different speakers from the GALE data. For each recording, the annotator is instructed to listen carefully to it and make judgment about the regional dialect of the speaker. The dialect labels were Egyptian, Gulf, Iraqi, Levantine, Magharebi, Unknown, and nonspeech. The "unknown" label is chosen when a speaker's dialect cannot be definitively identified. The "nonspeech" label is chosen when the recording file is empty or contains nonspeech materials such as music or white noise.

A label for MSA was not included, seeing as the speaker's regional background can have a noticeable influence on the speech even when speaking in MSA. Since the main application for this data source is developing a dialect identification tool and improving the performance of ASR systems, the researchers thought that the information of speakers' dialect is important even when speaking in MSA. In addition, Arabic speakers do not think about MSA and their dialect as separate languages. They code switch between the two language forms regularly. Almost no native Arabic speaker is willing to sustain continuous and spontaneous production of spoken MSA.

In order to help the annotators make the right judgment, annotators have the option to listen to the recordings as many times as they want. In addition, three different recordings for each speaker are introduced in case the annotator was unsure about the dialect of the speaker by listening to only one recording. Therefore, only speakers with at least three utterances of no less than three seconds in duration are introduced.

4.3 Speakers

The total number of speakers was found to be 3685 speakers for the BN dataset, and 832 speakers for the BC dataset. This makes a total of 4517 speakers in the whole dataset who have at least one non-empty recording. However, some speakers only produced a very small amount of data, and for such speakers the annotators often had difficulty in assigning an accent. Consequently, the researchers decided that only speakers with at least three recordings of no less than three seconds should be chosen to be annotated. This step reduced the total number of speakers to 2865. Each speaker is annotated by three different annotators to make a judgment of their regional dialect. Speakers were chosen in a random order to minimise the chance of a given set of annotators all being given the same speakers to annotate.

4.4 Annotators

Overall, 83 annotators participated in the task, 23 of whom annotated at least 200 speakers. Figure 2 provides a line chart to show the distribution of participation for the annotators.

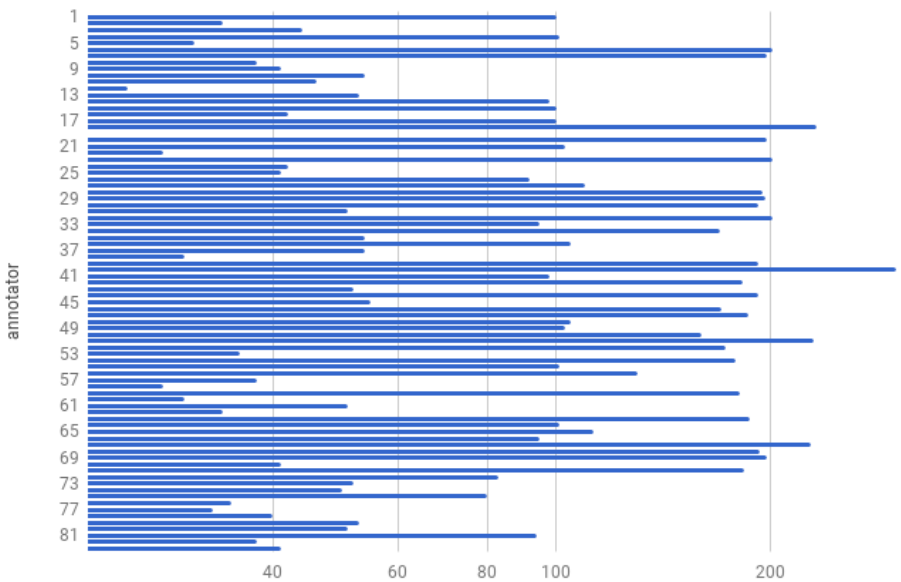


Figure 2: Distribution of participation among annotators.

The participation in this task is restricted to native Arabic speakers. All annotators are students at the department of Arabic language and linguistics, Kuwait University.

The collected annotations are only useful if a high level of consistency is achieved across annotators. Therefore, all annotators have taken part in a short tutorial and agreed to a set of annotation protocols prior to any annotation process. During the tutorial, special instructions were given to help to distinguish between different Arabic varieties by means of many sources of information extracted from the speech information. Annotators were reminded of some of the phonotactic, prosodic, lexical, and morpho-syntactic features of different Arabic varieties to help them make the right judgment. All annotators watched a demonstration to understand how to use the website properly. The annotators were free to do any number of annotations (preferably 200) and they were instructed to calculate the time they spent in the annotation process.

There is a reasonable level of mutual intelligibility across Arabic dialects. Some annotators, however, indicated that they had difficulties in assigning some speakers' dialect. The extent to which a particular annotator is able to recognise other dialects depends mainly on his/her exposure to Arab culture and literature from outside of their own country. For instance, most Arabic speakers have little trouble identifying and even understanding Egyptian dialect, owing to the Egyptians' popularity across the Arab world and the film industry and TV shows supplied to the Arabic speaking world. Yet, Eastern speakers find the Magharebi dialect difficult to recognise. The annotators indicate that they identify local Magharebi speakers from their special tone variations without understanding their speech. The following summarises the main sources of difficulties in recognising the speakers' dialect, as stated by the annotators:

- Many speakers in the dataset are news presenters who are trained to use the standard form of the language and to be linguistically neutral. Those professional speakers adopt the use of MSA speaking rules, hiding their regional dialect as far as possible.
- Some recordings are very short, which makes it hard to make a decision.
- Some of the recordings are extremely noisy, which makes it hard to hear the speaker's voice clearly.
- Annotators found it tricky to categorise Sudanese speakers' dialect. Some of the speakers in Sudan have similar accents to the ones in Saudi Arabia, while some others have accent similarities with Egyptian Arabic.

5 Annotation Evaluation

Delivering annotated data for speech and language applications requires careful quality control to ensure that the results are based on accurate information. For this reason, the reliability of the collected annotations was examined by first analysing the annotators' behaviour. This enables the researchers to detect any spamming behaviour and exclude unreliable annotators. In addition, several inter-annotators metrics were proposed for inter-annotator agreement evaluation. This section provides with sums based on the current collection of annotations. The researchers calculated the percentage of agreement, Fleiss' Kappa scores, and the distributions of agreement among annotators for the entire dataset

5.1 Annotators' Behaviour

By doing this assessment, the researchers aim to identify unreliable annotators and reject their annotations. The reliability of an annotator is defined as the degree of agreement of the annotations done by this annotator with the annotations done by the co-annotators. This assessment is delivered by calculating the Fleiss' Kappa for each individual annotator to measure the extent they agree with their co-annotators. This assessment reveals that annotators' Kappa scores vary from 0.84 (top score) and -0.125 (lowest score), with an average of 0.54 Kappa score. Details are given in Figure 3.

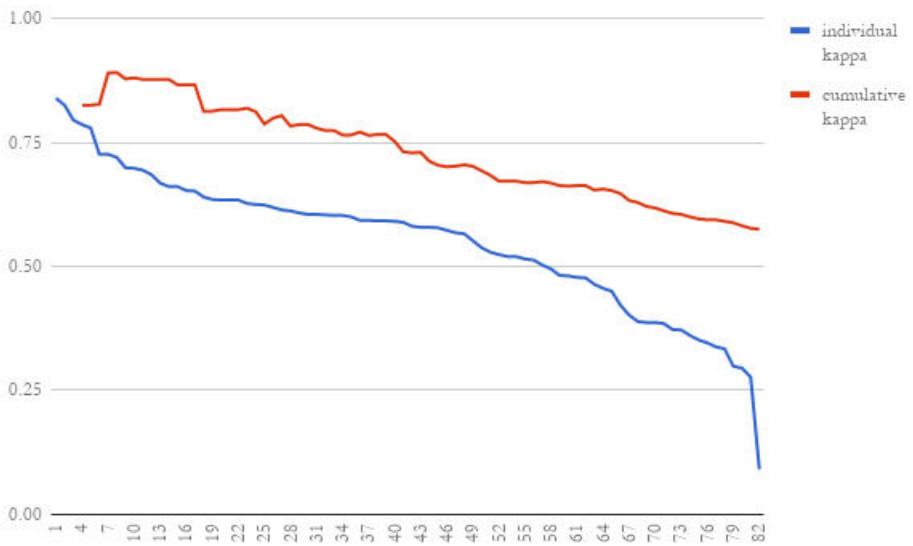


Figure 3: Individual and cumulative Kappa scores for the whole data.

5.2 Inter-annotator Agreement

The goal of this evaluation step is to measure the degree of agreement among annotators. Inter-annotator agreement is often evaluated by measuring Kappa coefficient. This method gives a score of how much homogeneity there is in the given annotations. Like most correlation coefficients, Kappa ranges from 0 to 1, where 0 is no agreement (even made by chance), 1 is perfect agreement. When an annotator scores less than 0, that indicates poorer than chance agreement (Fleiss and Cohen).

The total agreement between the annotators was found to be 39%. This corresponds to a kappa value of 0.43. The distribution of agreement is as following: all annotators agree: 1113 cases, two annotators agree: 1235 cases, all annotators disagree: 517 cases.

The agreement percentage increases after taking out the cases where a speaker is annotated as "unknown". The key here is that "unknown" is not really an annotation. If annotator (A) says "Gulf" and annotator (B) says "Egyptian" then they are disagreeing; if annotator (A) says "Gulf" and annotator (B) says "unknown" then they are not. The percentage of agreement among annotators after removing the "unknown" label comes to 52.9%, which makes a Kappa score of 0.57, with the following distribution: all annotators agree: 1100, two agree: 790, all disagree:189. In their article, "An Application" of Hierardnical Kappa-type Statistics, Landis and Koch) gave the following table for interpreting kappa values:

Table 2: Interpretations of Kappa scores as provided by Landis and Koch

Kappa	Interpretation
< 0	Poor agreement
0.01 – 0.20	Slight agreement
0.21 – 0.40	Fair agreement
0.41 – 0.60	Moderate agreement
0.61 – 0.80	Substantial agreement
0.81 – 1.00	Almost perfect agreement

According to the information given in this table, the annotators exhibit a moderate level of agreement. If various annotators do not agree, either the judgment on some data is hard, or the annotators are not doing their job consistently. In particular, it seems likely that annotators for whom the kappa score was significantly below the norm were probably making mistakes (an obvious reason for an annotator to disagree with their co-annotators to an unusually high degree). For this reason, another round of annotation was carried out after removing annotators who scored low Kappa scores and came at the bottom of the list. Details are given in the next section. Figure 3 provides information about the Kappa score for each individual annotator (sorted from best to worst) along with the cumulative Kappa score for the whole data.

6 Annotation Refinement

The accuracy of any machine learning task primarily depends on the quality of the data. The collected annotations will be distributed and used by researchers in the field of processing Arabic speech. Using unreliable data may lead to the degradation of performances of learning algorithms that use this data. It is, therefore, crucial to provide the most reliable information to guarantee the usability of the annotated corpus.

For optimising the Kappa score and to ensure delivering the most reliable information, the researchers excluded all annotations provided by annotators who achieved a low Kappa score. Those annotators were believed to be "unreliable". The cutoff point was decided to be at the point where the cumulative Kappa comes to 0.7. In total, 39% of the assignments were rejected on this basis. This leaves the researchers with the annotations of 49 annotators, which form around 61% of the total data.

Subsequently, "reliable" annotators who were willing to perform another round of annotation were asked to redo the unreliable annotations, while ensuring they complete their job faithfully. The revised annotations were evaluated again. The percentage of total agreement was found to be 61%. The achieved Kappa score is 0.62. The distribution of agreement is as following: all annotators agree: 1282, two agree: 710, all disagree: 107. Given the information in Table 2, the extent of agreement can be now qualified as substantial agreement.

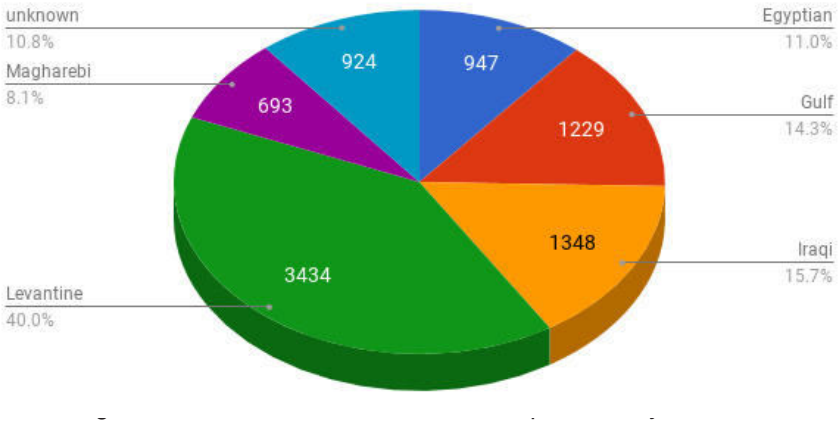
6.1 Label Distribution

Upon examination of the provided dialectal labels for the GALE (phase 3) speakers and their distribution, the researchers conclude that:

- The most chosen dialect label is Levantine. This might be explained by the dominance of Levantine broadcasters working in the Arabic broadcast media in general. For instance, in *Al Jazeera* broadcast news channel, which is operated in Doha, Qatar, approximately 60% of the broadcasters are from Levantine origins. Similarly, about 72% of broadcasters are from Levantine origins in *Al Arabiya* broadcast news, which is operated from Dubai[3]. Another possible explanation is the large geographical area of the Levant region with a high population.
- Despite its wide geographical spread, the Magharebi dialect was the least chosen dialect label. Looking at the sources of data in the GALE dataset, the researchers find that it contains only one Magharebi TV channel, "Tunisia TV". This will definitely lead to the lack of Magharebi speakers. Another reason is the fact that Magharebi dialect cannot be considered to be mutually intelligible with other dialect groups, unlike other dialects. This fact leads to having the Magharebi speakers semiconsciously trying to standardise their language as much as possible. The process of language standardisation resulted in having

a new form of the language called *plain* or *white dialect*. When using a white dialect, speakers try to detach themselves from their native dialect as much as possible in order to be mutually understood. Consequently, many Magharebi speakers who choose to neutralise their language will not be identified. More details of this phenomena will be discussed in 7.3.

Figure 4 provides a pie chart to show the distribution of the main five Arabic varieties labels in the collected annotations, in addition to the "unknown" label.



7 Results Analysis

In this section, the researchers highlight and analyse the main observations from the annotated data that were delivered from the final assignments. This is done particularly to help the researchers of the paper have a better understanding of the features of the dataset available to them, and to highlight some dialect related issues for future treatment.

7.1 Reasons Behind Low Agreement in Some Cases

In this investigation, the researchers try to find the source for low agreement in the first place, then they try to see if there is a pattern in the cases of disagreement. This investigation is carried out in two steps. The first step is by listening to the recordings where all three annotators disagree on their judgment about the dialect of the speaker.

There are 107 cases where all three annotators disagree. By listening to these cases, the researchers produced findings that confirm observations made by annotators about difficulties they had, namely:

- Sudanese speakers who cannot be categorised into a specific dialect group.
- Professional MSA speakers.
- Recordings with noisy backgrounds like those recorded in streets, or recordings with instant interpretations where two speakers speak at the same time.

From a computational point of view, the disagreement in those cases could be regarded as useful information. As a matter of fact, when the speaker's dialect cannot be identified by a human annotator, computers will not be expected to identify it. It is, therefore, useful to provide a special treatment for those cases when building dialect-specific ASR systems.

7.2 Interdialectal Confusability

The second step towards finding the sources of difficulties is by trying to find the rates of confusability among different dialects. The aim of this step is to understand the amount of ambiguity that each dialect hides, causing a high competition between different dialect judgments for one speaker. Here, the researchers try to find answers to the following questions:

i. Which language seems to be easy to get right? And which language seems to be hard?
ii. What languages get mixed together causing a high ambiguity rate?

The dialect label is assigned to each speaker in the dataset based on judgments provided by three different annotators. In order to measure the rates of confusability between dialects, each judgment provided to each dialectal label was counted, then the ratio between each dialect cluster was calculated. Information about the provided annotations is shown in Figure 5, where the confusion matrices were computed and used as a way to find answers to the researchers' questions.

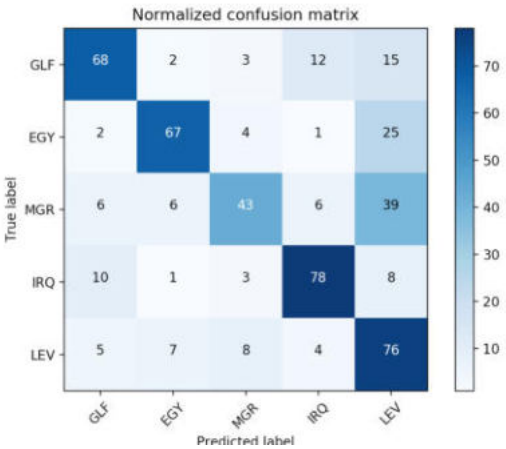


Figure 5: Confusion matrices for dialect labels.

The researchers infer from the confusion matrix in Figure 5 that the Iraqi dialect was the easiest dialect to be distinguished from other varieties. Iraqi speakers reveal themselves clearly by their way of intonation and their distinctive use of complex tonal patterns. The employment of intonation is essential in Iraqi speech as it marks word prominence, connects words in a sentence, draws the boundaries, and adds attitudinal meaning to the utterance.

The Iraqi dialect is also marked by the way the emphasis spreads over the neighboring syllables making adjacent vowels more prominent compared to other Arabic varieties. Additionally, final vowels are known to be fairly long in duration (vowel stretching). These aspects of Iraqi dialect add special characteristics to an utterance, which makes its recognition straightforward even when an Iraqi speaker is talking in MSA.

It can be also noted from the confusion matrices that Magharebi dialect is the hardest dialect to recognise, and in 39% of the cases it gets confused with Levantine, making them the most confusable pair. The researchers believe that this is related to the fact that Magharebi speakers, in most of the cases, are talking from eastern channels and addressing people from different Arabic varieties. Magharebi dialect is considered to be asymmetrically intelligible with other varieties. This means that people from the East have more difficulty understanding Magharebi speakers than Magharebi speakers have understanding speakers from the East. This induces Magharebi speakers to detach themselves as much as possible from their own dialect in order to be comprehended by others. This new dialect (white dialect) lacks most of the distinctive features of Magharebi dialect borrowing many linguistic features from other varieties, especially Levantine.

Results also suggest that Levantine dialect forms a strong confusion pair with all other dialects. This is likely due to the great lexical similarity between MSA and Levantine. As a result, when speakers talk in MSA, they are more likely to be identified as Levantine, unless they show some of their dialectal characteristics. This makes Levantine dialect exhibit the highest percentage of competition in the provided judgments.

8 Applications

The speakers' regional dialect has a great effect on all aspects of speech such as phones, vocabulary, and syntax. This fact makes it crucial to use dialect-based corpora when dealing with Arabic.

Assigning a dialectal label for each speaker in a large dataset such as GALE (phase 3) would make it possible to create a collection of reasonably large monolingual datasets. Such multi dialect corpora can be used in the creation of many NLP systems

that deal with dialectal Arabic content. For instance, the creation of dialect identification tool, improving the performance of ASR and Text to Speech systems, Arabic information retrieval, and developing speaker identification tool. Identifying speakers' dialect can also help machine translation systems to deal with dialectal Arabic. Generally speaking, the presence of informative speech resources is essential for any Arabic NLP application, and any shortage in this data might lead to deficiencies in the final accuracy.

9 Conclusion

Despite sharing a considerable number of linguistic features with one another and with MSA, Arabic dialects do substantially differ from NLP point of view in almost all language aspects. This difference leads researchers to treat Arabic varieties as different languages. With the increased interest in dialectal Arabic in the field of NLP, the availability of related data sources is becoming essential. This paper presents the first phase of an effort towards building an efficient Arabic ASR system. This phase aims at providing information about the speakers' regional origin, which is believed to be a great source of variation in speech. The data in this study are obtained from Linguistic Data Consortium (LDC) and the annotation process is applied to GALE (phase 3) broadcast news and broadcast conversational speech. The annotation process resulted in assigning a dialect label for about 2900 speakers, who contributed in this substantial Arabic resource.

The paper introduced the methodology followed in collecting annotations, information about the annotators who participated in the dataset development, and information about the annotated speakers. Besides providing a comprehensive evaluation for the annotators and the collected annotations, the paper gives a sufficient study of Arabic varieties and linguistics differences between MSA and dialects and between dialects themselves. The paper also presented a thorough analysis of the collected annotations. The analysis covers the following topics: distribution of dialectal label in the GALE data, sources of low agreement as founded in some cases, and interdialectal confusability measurement.

The final evaluation of the annotations shows that it achieved a substantial level of agreement. The annotations are fully available online. A set of access tools are also shared to help extract specific information from the database. However, due to membership agreement by LDC, the associated wave and text files cannot be shared. Having access to such dialect specified corpora will definitely help the research community to improve the current Arabic NLP technologies.

Nots:

(1) https://github.com/AllanRamsay/ACCENTS

(2) The letter A is silent

(3) Information about the origins of broadcasters working at Al jazeera and Al Arabiya channels were extracted from their websites.

Works Cited

Al-Sabbagh, Rania, and Roxana Girju. "YADAC: Yet anotherr Dialectal Arabic Corpus." LREC. 2012.

Alsharhan, Eiman, and Allan Ramsay. "Improved Arabic Speech Recognition System Through the Automatic Generation of Fine-grained Phonetic Transcriptions." Information Processing & Management 56.2 (2019): 343-353.

Alghamdi, Mansour. "Arabic Phonetics." Al-Toubah Bookshop, Riyadh (2001).

Almeman, Khalid, Mark Lee, and Ali Abdulrahman Almiman. "Multi Dialect Arabic Speech Parallel Corpora." 2013 1st International Conference on Communications, Signal Processing, and their Applications (ICCSPA). IEEE, 2013.

Aoun, Joseph, Elabbas Benmamoun, and Dominique Sportiche. "Agreement, Word Order, and Conjunction in Some Varieties of Arabic." Linguistic inquiry (1994): 195-220.

Biadsy, Fadi, Pedro J. Moreno, and Martin Jansche. "Google's Cross-dialect Arabic Voice Search." 2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2012.

Cotterell, Ryan, and Chris Callison-Burch. "A Multi-Dialect, Multi-Genre Corpus of Informal Written Arabic." LREC. 2014.

Elfardy, Heba, and Mona Diab. "Sentence Level Dialect Identification in Arabic." Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). Vol. 2. 2013.

Fleiss, Joseph L., and Jacob Cohen. "The Equivalence of Weighted Kappa and the Intraclass Correlation Coefficient as Measures of Reliability." Educational and Psychological Measurement 33.3 (1973): 613-619.

Kirchhoff, Katrin, and Dimitra Vergyri. "Cross-dialectal Data Sharing for Acoustic Modeling in Arabic Speech Recognition." Speech Communication 46.1 (2005): 37-51.

Landis, J. Richard, and Gary G. Koch. "An Application of Hierarchical Kappa-type Statistics in the Assessment of Majority Agreement among Multiple Observers." Biometrics (1977): 363-374.

Masmoudi, Abir, et al. "A Corpus and Phonetic Dictionary for Tunisian Arabic Speech Recognition." LREC. 2014.

McCarthy, John J. "The Length of Stem-final Vowels in Colloquial Arabic."

Perspectives on Arabic Linguistics XVII–XVIII (2005): 1-26.

Rosenhouse, Judith, Noam Amir, and Ofer Amir. "Vowel Systems of Quantity Languages Compared: Arabic Dialects and other Languages." Proceedings of Meetings on Acoustics 167ASA. Vol. 21. No. 1. ASA, 2014.

Selouani, Sid Ahmed, and Malika Boudraa. "Algerian Arabic Speech Database (ALGASD): Corpus Design and Automatic Speech Recognition Application." Arabian Journal for Science and Engineering 35.2 (2010): 157-166.

Shoufan, Abdulhadi, and Sumaya Alameri. "Natural Language Processing for Dialectical Arabic: A Survey." Proceedings of the Second Workshop on Arabic Natural Language Processing. 2015.

Simons, Gary F., and Charles D. Fennig. "Ethnologue: Languages of the World, Dallas, Texas: SIL International." Online version: http://www.ethnologue.com (2017).

Versteegh, Kees. Arabic Language. Edinburgh University Press, 2014.

Zaidan, Omar F., and Chris Callison-Burch. "Arabic Dialect Identification." Computational Linguistics 40.1 (2014): 171-202.

المجلة التربوية

مجلة فصلية، تخصصية، محكّمة

تصدر عن مجلس النشر العلمي - جامعة الكويت

رئيس التحرير: أ. د. عبدالله محمد الشيخ

تنشر:

- البحوث التربوية المحكمة
- مراجعات الكتب التربوية الحديثة
- محاضر الحوار التربوي
- التقارير عن المؤتمرات التربوية وملخصات الرسائل الجامعية

- تقبل البحوث باللغتين العربية والإنجليزية.
- تنشر لأساتذة التربية والمختصين بها من مختلف الأقطار العربية والدول الأجنبية.

الاشتراكات:

في الكويت: ثلاثة دنانير للأفراد، وخمسة عشر ديناراً للمؤسسات.
في الدول العربية: أربعة دنانير للأفراد، وخمسة عشر ديناراً للمؤسسات.
في الدول الأجنبية: خمسة عشر دولاراً للأفراد، وستون دولاراً للمؤسسات.

توجه جميع المراسلات إلى:

رئيس تحرير المجلة التربوية - مجلس النشر العلمي ص.ب. ١٣٤١١ كيفان - الرمز البريدي 71955
الكويت هاتف: ٢٤٨٤٦٨٤٣ (داخلي ٤٤٠٣ - ٤٤٠٩) - مباشر: ٢٤٨٤٧٩٦١ - فاكس: ٢٤٨٣٧٧٩٤

E-mail: joe@ku.edu.kw