You Tweet What You Speak: A City-Level Dataset of Arabic Dialects

Muhammad Abdul-Mageed, Hassan Alhuzali, Mohamed Elaraby
Natural Language Processing Lab
University of British Columbia
muhammad.mageeed@ubc.ca, {halhuzali,mohamed.elaraby}@alumni.ubc.ca

Abstract

Arabic has a wide range of *varieties* or *dialects*. Although a number of pioneering works have targeted some Arabic dialects, other dialects remain largely without investigation. A serious bottleneck for studying these dialects is lack of any data that can be exploited in computational models. In this work, we aim to bridge this gap: We present a considerably large dataset of $> 1/4$ billion tweets representing a wide range of dialects. Our dataset is more nuanced than previously reported work in that it is labeled at the fine-grained level of city. More specifically, the data represent 29 major Arab cities from 10 Arab countries with varying dialects (e.g., Egyptian, Gulf, KSA, Levantine, Yemeni).

1. Introduction

The Arab world covers a vast region across the two contents, Africa and Asia. The term *Arabic* itself refers to a collection of varieties, possibly comprised by three major categories: (1) Modern Standard Arabic (MSA), (2) Classical Arabic (CA), and (3) Dialectal Arabic (DA). MSA (Badawi, 1973) is the modern variety of the language used in educational settings and some pan-Arab networks like AlJazeera (Abdul-Mageed, 2008; Abdul-Mageed and Herring, 2008). CA is the language of the Qura'an (the Holy Book of Islam) that is employed in religious and elite literary works. MSA and CA differ mainly lexically and morphologically, with fewer structural and syntactic differences (Bateson, 1967; Ryding, 2005). DA is a collection of arbitrarily defined (Versteegh, 2001; Habash, 2010) variations, although geography does play a role in the classification of DA.

Most Arabic varieties remained primarily spoken for a long time. With the advent of the internet and the proliferation of social media, Arabic dialects found their way to online written form (Abdul-Mageed, 2015). Early computational studies of Arabic dialects have depended on data collected from blogs and comments on online news sites, e.g., (Diab et al., 2010; Elfardy and Diab, 2012). Due to the costly efforts associated with labeling the data with dialect tags, these pioneering works have focused on a few varieties like Egyptian or Levantine (Elfardy and Diab, 2013; Zaidan and Callison-Burch, 2011; Zaidan and Callison-Burch, 2014). User-provided location information in the Twitter microblogging platform have made it possible to collect data with nuanced geographical labels in ways not previously possible, see e.g., (Jurgens et al., 2017). We depend on these cues to label our dataset with location tags as a proxy for the relevant dialects.

Although no agreement exists as to where dialectal boundaries should be drawn, some proposals have been made. Figure 1 shows only one such classifications of Arabic dialects. The vast geographic extension the Arab world constitutes naturally translates into rich and varied linguistic tradition, thus making nuanced study of Arabic dialects an attractive object of scientific investigation. In this work, we take a step in this direction by collecting a large dataset of Arabic dialects. We focus on the Eastern part of the region,

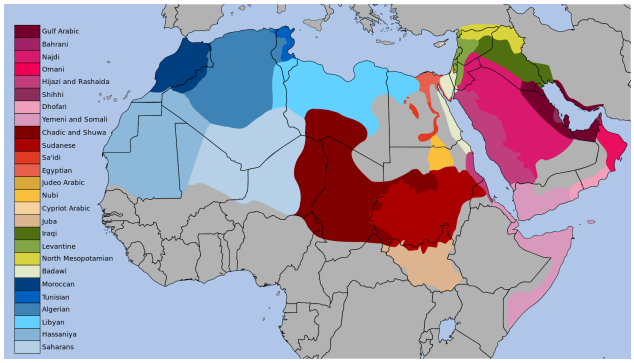

Figure 1: A classification of Arabic varieties. [1]

covering 10 different countries. These includes countries for which no (large) datasets are available (e.g., KSA, Jordan, Qatar, Yemen, and UAE).

Overall, we offer the following contributions: **(1)** We build a large-scale dataset for a variety of Arabic dialects with automatically labeled city- and country-level tags, as a proxy for respective dialects, **(2)** we manually verify the dialect labels on a pilot section of our data, **(3)** to show the utility of our data, we build a word-level embeddings model exploiting the data. We illustrate the capacity of this distributed representation model of words by both comparing its coverage to a standard publicly available model built on MSA and providing examples of word relationships it is able to capture.

The rest of the paper is organized as follows: Section 2. reviews related work. In Section 3., we describe our dataset. Section 4. is a description of our pilot dialect annotation study. In Section 5., we describe a distributed representation of words model we built exploiting our data. Section 7. is where we conclude and discuss future directions.

2. Related Works

2.1. Arabic Dialects.

Arabic dialects differ in various ways from MSA. These include phonological, morphological, lexical, and syntactic differences (Bassiouney, 2009; Holes, 2004; Palva, 2006). Although, in theory, Arabic dialects can be classified in various ways (Palva, 2006), categorizations of Arabic dialects

remain arbitrary and primarily based on geographical divisions, e.g., (Habash, 2010; Versteegh, 2014). Habash (2010) provides the following classification of Arabic dialects, indicating it is only one of many categorizations:

- **Egyptian (EGY):** Includes dialects of Egypt and the Sudan (Nile valley);
- **Gulf (GLF):** Covers dialects of Bahrain, Kuwait, Oman, Qatar, Kingdom of Saudi Arabia, and United Arab Emirates;
- **Iraqi (IRQ):** Includes elements from both Levantine and Gulf;
- **Levantine (LEV):** Includes dialects of Lebanon, Syria, Jordan, and Palestine;
- **Maltese (MLT):** Is a variant that is not always considered an Arabic dialect, but rather a separate language, and is written in the Roman script;
- **North African (MAG):** Encompasses dialects of Algeria, Libya, Mauritania, Morocco, and Tunisia;
- **Yemenite (YEM):** often considered its own class (i.e., does not include other varieties and hence stands as a category by itself).

The classification above is perhaps the most common in the literature[2]. Differences between dialects within the categories above are most pronounced. The following tweet illustrates a morphological characteristic of the the Iraqi dialect, for example. In specific, it is focused at the practice of using the final "ج" for "second person singular" in the word "طبيعتج" in Iraqi Arabic:

- (1) سخافة اتصرفي ع طبيعتج دادة، حتى لو تصرفات عبوسي وبعفويتج احلى مـن تصرفات مصطنعة وكلها فوطوشوب، نصيحتي كيوت لاتكلمني البنت لماتدلع

  Eng. "Enough with fake behaviour, girl! Be yourself, even if you think it's not attractive, being natural is way better than the "I'm cute, artificial behaviour. My advice is to ignore girls that are artificial."

This contrasts with use of "ك" for "second person singular" in the word "طبيعتك" in Gulf (example from Palestinian) below:

- (2) بصي ياحبي العيب منك و منهم بمعني انك خليكي ع طبيعتك و طيبة قلبك بس بلاش الثقه ف اي حد او الحب ال بنوزعه عمال ع بطال ع الناس لان اقسم بالله بنتشنير شنيره بنت لذين بسبب طيبتنا دي..خليكي فاللون الرمادي حبي واكتمي ولو حسيتي الحب متبادل يبقا اتكلي ع الله

  Eng. "Listen, [female] dear, the problem lies both in you and them. In other words, be yourself, as kind as you are, but don't put extra trust in anyone. So forget about the overflowing love we distribute right an left for this naive kindness ends up causing us to suffer tons... So just stick to the grey area, dear, and keep your emotions to yourself. If you start feeling it's mutual love, then go ahead [and express your feelings]!"

The following two examples illustrate lexical differences between the Gulf (example #3, from the Qatari variety) and Egyptian (example # 4). In example (3), the equivalent of the English word "I want" is "ابي", whereas in example (4), the equivalent is "عايز":

- (3) وحده داقه علي موظف بنك : ممكن استفسر : تفضلي عن شنو اختي :ابي اشوف كم برصيدي : اختي شنو اخر عملية سويتيها :ولاده. يقولك الموظف تقاعد

  Eng. "One time, a lady called a bank employee: "Hello, can I inquire about my balance?" "Of course sister, what was the last 'operation' you made?" "A delivery!" People say that he [the employee] quit his job after this call."

- (4) انا بس عايز فرصة كمان وانا هاثبتلك انى مش هتغير وانك تروح تضرب دماغك فى الحيط اسهل

  Eng. "I only need another chance to prove to you I won't change, and that you'd better go hit your head on the wall."

Classifications of dialects in general can easily gloss over distinctions between language variants. Many classifications of Arabic dialects, the one classification used by Habash (2010) being no exception, does. The limitations of some of these classifications include that differences between variants, especially across countries or regions of countries, can be significant. For example, varieties of Arabic in Egypt and the Sudan can be very different at various linguistic level (e.g., lexical, morphological, syntactic). These types of regional variations (Gonçalves and Sánchez, 2014) are pervasive, and even within each of these countries there exists further, more nuanced variations. The Arabic of Egypt's Sinai (north east of the country) is different from that of Cairo (the capital, which is situated toward the North), which is still different from that of Alexandria (north west, on the Mediterranean). Indeed, the linguistic literature shows how language can vary even within different parts of the the same city (Labov, 1964; Orton et al., 1998; Trudgill, 1974), thus creating micro-dialects within the same dialect. To the best of our knowledge, these fine-grained variations within countries has not been studied in Arabic NLP. Our work seeks to take a first step toward enabling the bridging of this important gap.

2.2. Computational Treatment of Arabic Dialects.

Early NLP work on Arabic dialects focused on collecting datasets that would enable the investigation of these

[2]Nizar Habash (personal communication, December, 2011) also points out he found this classification to be the 'most common' in the literature and hence he opted for it in his book.

dialects. A number of these pioneering studies focused on collecting data from blogs (Diab et al., 2010; Elfardy and Diab, 2012; Al-Sabbagh and Girju, 2012; Sadat et al., 2014), the general Web (Al-Sabbagh and Girju, 2012), comments on online news sites (Zaidan and Callison-Burch, 2011), or building dialectal lexica (Diab et al., 2014). Other works have dealt with detecting one or more of the Levantine, Gulf, and Egyptian dialects (Elfardy and Diab, 2013; Zaidan and Callison-Burch, 2011; Zaidan and Callison-Burch, 2014; Cotterell and Callison-Burch, 2014). Works covering other dialects (e.g., from countries like Tunisia, the Sudan, Qatar, Bahrain) include (Sadat et al., 2014), although they exploited small datasets (mostly $< 5K$ sentences from each country). Closer to our work is (Mubarak and Darwish, 2014) who report collecting a dataset of 123 million tweets covering Egyptian, Levantine, Iraqi, Maghrebi dialects. Our work follows (Mubarak and Darwish, 2014)'s lead, while developing a dataset almost twice the size (and from as twice countries). In addition, our work compares preferably to (Mubarak and Darwish, 2014) in that our data has more nuanced labels (i.e., at the city level).

Also related to our research is recent work on discriminating similar languages, e.g., via the VarDial workshop (Malmasi et al., 2016; Zampieri et al., 2017) where some works focused on Arabic (Malmasi and Zampieri, 2017; Ionescu and Butnaru, 2017). Some works also focus on Arabic dialect identification in speech transcripts, e.g., (Malmasi and Zampieri, 2016). Again, our work has wider scope and coverage. We now turn to describing our dataset.

3. Dataset

In order to develop our dataset, we exploit several in-house corpora (i.e., a total of > 1 billion tweets) covering the 10 Arab countries from the set {*Oman, Egypt, Iraq, Jordan, Kuwait, Palestine, Qatar, KSA, UAE, and Yemen*}. Our in-house data were collected using Twitter API using several bounding boxes over multiple Arab countries. As such, the data are diverse as we do not use any specific seeds to crawl. In addition, the data cover $\sim$ the last 5 years (i.e., $2013 - 2018$). To acquire location labels on the data, we use the Python geocoding library geopy [3], which helps locate the coordinates of addresses (e.g., 2103 Charleston Rd, Mountain View, CA 55321, USA), cities (e.g., Seattle), countries (e.g., Yemen), and landmarks (in the form of a co-ordinates, e.g., $49.264031, -123.246179$) based on third-party geocoders and a number of other data sources [4]. More specifically, we use "OpenStreetMap Nominatim" [5] as a third party tool. We acquire a total of $234,801,907$ tweets from 29 Arab cities, representing 10 Arab countries. The 29 cities in our data are shown in Figure 2. In addition, statistics of the dataset are provided in Table 1.

[3] `https://github.com/geopy/geopy`.

[4] A list of these third-party geocoders and other sources can be found at: `https://github.com/geopy/geopy/tree/master/geopy/geocoders`.

[5] `https://nominatim.openstreetmap.org/`

Country	City	# Users	# Tweets
Egypt	Alexandria	6,613	9,839,453
	Cairo	20,544	29,597,031
	Giza	2,499	3,252,507
Iraq	Baghdad	2,447	2,617,790
	Karbala	235	223,885
	Zubair	238	266,777
Jordan	Amman	2,943	4,132,434
	Aqaba	53	57,066
	Irbid	332	431,016
Kuwait	Ahmadi	396	678,050
	Hawally	142	200,757
	Kuwait City	2,827	5,071,420
Oman	Muscat	2,247	2,883,711
	Salalah	298	339,296
	Sohar	256	340,806
Palestine	Gaza	1,931	2,754,851
	Nablus	113	146,967
	Ramallah	167	216,245
Qatar	Al-Rayyan	466	694,715
	Doha	4,025	6,394,218
KSA	Dammam	5,560	8,483,462
	Jeddah	29,045	42,840,379
	Riyadh	61,697	90,410,407
UAE	Abu Dhabi	5,074	8,093,556
	Al Ain	497	822,870
	Dubai	7,050	11,436,814
Yemen	Aden	481	674,345
	Sana	1,200	1,610,875
	Taiz	229	290,204
All	–	**159,605**	**234,801,907**

Table 1: Data statistics: Number of users and tweets per city, for 29 cities covering 10 different countries representing the Eastern part of the Arab world.

4. Dialect Annotation

We perform a pilot dialect annotation task with varieties from the 10 Arab countries in our data. We provide each annotator with data representing a single country at a time, after explaining how the data were collected and the goal of our work. We then cast the dialect annotation task as a 3-way decision where judges choose whether a tweet (1) represents the dialect of the given country **(DA)**, (2) **(MSA)**, or (3) any other dialect **(OTHER)**. In the case of **(OTHER)**, we do not ask annotators to specify what other dialect the tweet exactly belongs to, thus reducing cognitive overload and keeping the task simple. Judges labeling the data are college-educated native Arabic speakers. A total of 5 annotators performed the task and we ensured each annotator is fluent with the variety they worked on. In almost all cases, the annotator either comes from the country from which the data are derived or from a directly neighboring country. We asked annotators to follow a number of steps for labeling tweets with language they cannot understand, including consulting with one another and online. To ensure quality, any tweet whose language

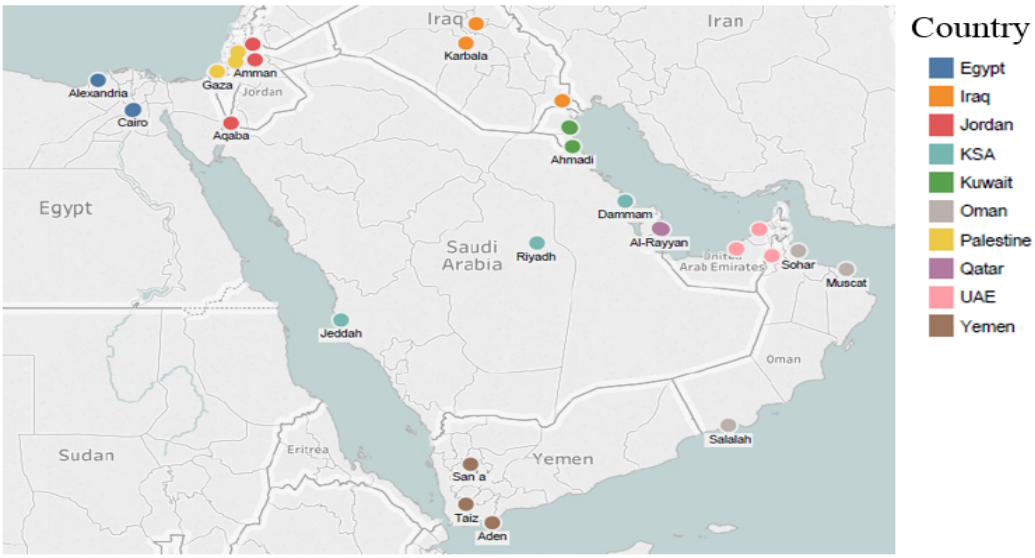

Figure 2: Cities represented in our data. Each city is shown as a dot; cities belonging to the same country are shown in similar color.

was still judged unintelligible after following these steps was excluded from the data. All non-Arabic tweets were removed from the data automatically before annotation using a simple character count method. In addition, we asked annotators to manually remove any non-Arabic tweets that may have remained after automatic filtering.

For this pilot annotation, we select a sample of 250 tweets per country (a total of $2,500$ tweets). Each tweet was labeled by two judges. Table 2 shows inter-annotator agreement on the task. As table 2 shows, annotators agree with a Cohen's Kappa $(K) = 67\%$, on average. This reflects 'substantial' agreement (Landis and Koch, 1977). Table 2 also shows that annotators agreed less on the cases of Yemen ($(K) = 40\%$), Oman ($(K) = 55\%$), and Qatar ($(K) = 58\%$), and Jordan ($(K) = 60\%$). This may be due to one or more of several factors. For example, annotators reported not being able to distinguish the dialects coming from some cities that closely neighbor other countries. For example, annotators had difficulty distinguishing tweets from Al Ain (UAE) and Sohar (Oman). In addition, annotators reported less acute difficulty working on data from countries in which there seems to be users originally from other countries. For example, users with Egyptian dialect tweeted from Qatar. For these reasons, we believe the political situation and immigration waves in the Arab world are important factors for dialect data collection. Conceivably, there would be cases where there are cities near borders where more than one dialect are used. Our research did not investigate these cases. However, we note this as an important direction for future research.

Country	Cohen's Kappa *(K)*
Egypt	0.73
Iraq	0.71
Jordan	0.60
KSA	0.89
Kuwait	0.71
Oman	0.55
Plastine	0.88
Qatar	0.58
UAE	0.60
Yemen	0.40
Avg.	**0.67**

Table 2: Inter-annotator agreement for the human dialect identification task.

5. Distributed Representations of Dialects

5.1. Building Word Vectors Model

Distributed representations of language at various levels of granularity, e.g., words and phrases (Mikolov et al., 2013; Pennington et al., 2014) or sentences (Kiros et al., 2015) boost performance on various NLP tasks. Zahran et al. (2015) pioneered efforts to build word embedding models for Arabic. In spite of the usefulness of a model built on MSA, it is expected to suffer coverage issues (i.e., sparsity) when applied to dialectal data. To alleviate this problem, we build a word vectors model exploiting our data. We adopt the pre-processing pipeline of (Zahran et al., 2015). Namely, we remove any non-unicode characters, normalize *Alif maksura* to *Ya*, reduce all *hamzated Alif* to plain *Alif*, and remove all non-Arabic characters. Additionally, to clean noise associated with social media non-standard typography, we reduce all letter repetition of > 2 characters to only 2. We build a skip-gram model with 300 dimensions, with a minimal word count $= 100$ words, and a window size of 5 words on each side of a target word. We use the gensim [6] implementation for the word2vec tool [7].

5.2. Hand-Picked Examples

In order to demonstrate the capacity and richness of our word vectors model, particularly in terms of dialectal word coverage, we ask our annotators to identify a list of dialectal words from the data

[6] https://radimrehurek.com/gensim/models/word2vec.html.

[7] https://code.google.com/archive/p/word2vec.

Country	Arabic	English	Most similar words
Egypt	ازاي	how	كده، بردو، دلوقتى، علشان، دلوقت
	معرفش	I don't know	مشعارف، دلوقتى، معرفوش، بردو، بجد
	جدعان	men	جماعه، خونا، شوباب، مسريين، كماعه
	بتعدي	pass by	هتعدى، بتمر، بتيجى، بتمشى، بتجرى، بتبقى
	مفيش	nothing	ومافيش، مفهاش، معندناش، مقيش، معنديش
KSA	ماننام	don't we sleep	ننام، ننوم، نرقد، مانموت، ننخمد
	هالكلام	this talk	هالحكى، هالخبر، هالحجه، الشئ، هالتصريح
	سم	name it	تسم، سمى، ذبو، آسلوب، مرب
	ورع	boy	بزر، هطف، مهابطى، طعس، شايب، زلابه
	ايوه	yes	ايوا، ايون، ايبه، اهوه، ليه
UAE	يالس	sitting	قاعد، جالس، داش، مستعيل، مجابل
	شحالك	how are you?	شلونك، شخبارك، شحالج، شلونج، شالاخبار
	سير	go	معرقل، وعرقله، وتسيير، سيرها، حادثى
	غرشه	sip	قلاص، قوطى، قاروره، علبه، قنينه
	يرمسون	they talk	يتكلمون، يرمس، يحوفون، شيقولون، يتحرطمون

Table 3: A set of hand-picked words from the top 3 countries (Egypt, KSA, and UAE) and their most similar words in our word embeddings model.

(10 words from each of the 10 countries for a total of 100 words). We then pick a random sample of 5 words from each of the top 3 countries, i.e., {*Egypt, KSA, UAE*}, in our data. Next, we use each item in this list to query the model for the 5 most similar words. Table 3 shows some examples. As Table 3 illustrates, for each dialectal word, the model captures not only morphological variants of the word: e.g., "ومافيش" and "مفهاش" (Eng. "there is nothing") for EGY, but also its orthographic variants: e.g., "شحالك" and "شحالج" (Eng. "how are you?"), for UAE. The model also captures similarity inter-dialectaly. For example, given the query "هالكلام" (Eng. "this argument) in KSA, the model returns "هالحكى" which belongs to the Levantine dialect. Moreover, given a query word, the model identifies syntactically and semantically related words. For example, the words "جدعان" (Eng. "men") and "مسريين" (Eng. sarcastic in a political context for "Egyptians") frequently occur after the vocative particle "يا" (Eng. "oh, you") and hence share syntactic context. These retrieved words also often times occur in humorous/sarcastic contexts, which implies the model may be capturing some pragmatic relationships in the data.

5.3. Lexical Coverage

Under the training parameters listed earlier, our word embeddings model ended with about a total of $500K$ words. Clearly, there is lexical overlap between different Arabic varieties and each dialect would employ words that are functional in MSA, with varying degrees of semantic relatedness. Our interest here is investigating this space, simplistically as a starting point: We query an embeddings model built with corpora that are overridingly MSA, to check how much coverage it affords for entries in our model. Our intuition is that the more frequent a word is in our data, the higher the likelihood it will be covered in an MSA-based model, and *vice versa*. This also implies that a list of words randomly sampled from our data should have a coverage in an MSA model that is neither as high as these most frequent in our data nor as low as those least frequent in our data. We test this intuition, finding it holding true as shown in Figure 3. In other words, our data have wider lexical coverage than is captured in the (Zahran et al., 2015) model.

Figure 3 shows distribution of coverage in the MSA model (i.e., (Zahran et al., 2015)) with highest, lowest, and random word frequencies as extracted from our data. In each case, we limit to frequencies from the set {*1k,2k,3k,4k,5k*}. Interestingly, this simple test also shows how certain dialects overlap with MSA more than others. For example, while coverage is $< 50\%$ as is clearly shown by the least common line in the Yemen data (Figure 4), coverage is $> 50\%$ for both EGY and KSA. Since the data representing each dialect are not of equal size, we cannot make claims as to semantic distance between dialects at this point, although this is one question we would like to eventually be able to answer.

6. Conclusion

In this paper, we reported the development of a new large-scale dataset for a number of Arabic dialects. The data are tagged at the city level. We also reported a pilot annotation study, identifying some of the challenges associated with fine-grained 'country-based' dialect annotation. Finally, we investigated the dialectal coverage of our data using a word vectors model. The distributed representations enabled by the model, as we show, have richer coverage than available models. Together with similar works, we believe this line of research opens up interesting frontiers for dialectal Arabic NLP. In the future, we will perform a wider scale annotation of the data and evaluate the distributed representations models in a number of downstream tasks.

7. Acknowledgement

This work was partially funded by UBC Hampton Grant #12R74395 and UBC Work Learn Award to Muhammad Abdul-Mageed. The research was also enabled in part by support provided by WestGrid (`https://www.westgrid.ca/`) and Compute Canada (`www.computecanada.ca`).

8. Bibliographical References

Abdul-Mageed, M. M. and Herring, S. C. (2008). Arabic and english news coverage on aljazeera. net.

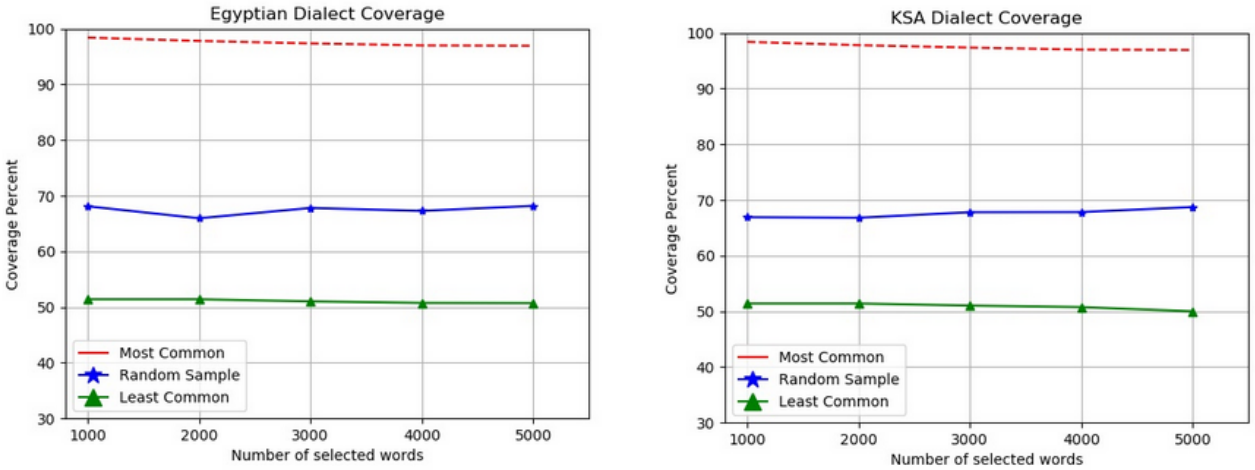

Figure 3: Distribution of dialectal lexical coverage in an MSA word vectors model (Zahran et al., 2015) on words from 2 countries (Egypt and KSA) in our data. The more frequent a word is in our data, the higher the likelihood it is covered in the MSA model: **`Red, dotted line`**: most common words in our data are covered best (i.e., close to 100%) in MSA model. **`Blue, starred line`**: Random frequency words are covered at values between 60% and 70% in MSA model. **`Green, triangled line`**: Least frequent words are sparse in MSA model (i.e., around, *or* $< 50\%$).

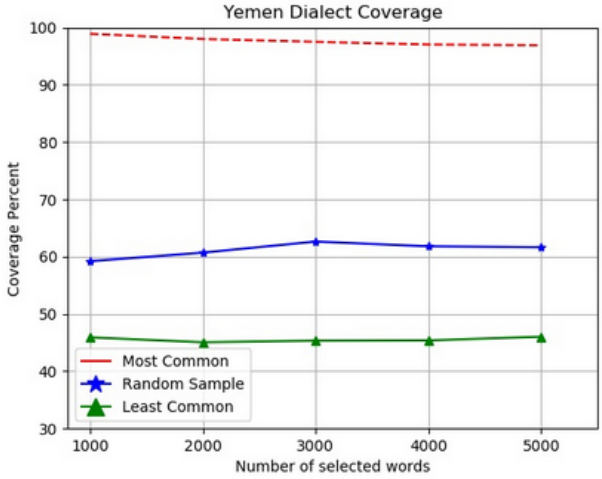

Figure 4: Distribution of dialectal lexical coverage in an MSA word vectors model (Zahran et al., 2015) on words from Yemen in our data.

Abdul-Mageed, M. M. (2008). Online news sites and journalism 2.0: Reader comments on al jazeera arabic. *tripleC: Communication, Capitalism & Critique. Open Access Journal for a Global Sustainable Information Society*, 6(2):59–76.

Abdul-Mageed, M. (2015). *Subjectivity and sentiment analysis of Arabic as a morophologically-rich language*. Ph.D. thesis, Indiana University.

Al-Sabbagh, R. and Girju, R. (2012). Yadac: Yet another dialectal arabic corpus. In *LREC*, pages 2882–2889.

Badawi, M. (1973). Levels of contemporary arabic in egypt. *Cairo: Dâr al Maârif*.

Bassiouney, R. (2009). *Arabic sociolinguistics*. Edinburgh University Press.

Bateson, M. C. (1967). *Arabic language handbook*, volume 3. Georgetown University Press.

Cotterell, R. and Callison-Burch, C. (2014). A multi-dialect, multi-genre corpus of informal written arabic. In *LREC*, pages 241–245.

Diab, M., Habash, N., Rambow, O., Altantawy, M., and Benajiba, Y. (2010). Colaba: Arabic dialect annotation and processing. In *Lrec workshop on semitic language processing*, pages 66–74.

Diab, M. T., Al-Badrashiny, M., Aminian, M., Attia, M., Elfardy, H., Habash, N., Hawwari, A., Salloum, W., Dasigi, P., and Eskander, R. (2014). Tharwa: A large scale dialectal arabic-standard arabic-english lexicon. In *LREC*, pages 3782–3789.

Elfardy, H. and Diab, M. T. (2012). Simplified guidelines for the creation of large scale dialectal arabic annotations. In *LREC*, pages 371–378.

Elfardy, H. and Diab, M. T. (2013). Sentence level dialect identification in arabic. In *ACL (2)*, pages 456–461.

Gonçalves, B. and Sánchez, D. (2014). Crowdsourcing dialect characterization through twitter. *PloS one*, 9(11):e112074.

Habash, N. Y. (2010). Introduction to arabic natural language processing. *Synthesis Lectures on Human Language Technologies*, 3(1):1–187.

Holes, C. (2004). *Modern Arabic: Structures, functions, and varieties*. Georgetown University Press.

Ionescu, R. T. and Butnaru, A. M. (2017). Learning to identify arabic and german dialects using multiple kernels. *VarDial 2017*, page 200.

Jurgens, D., Tsvetkov, Y., and Jurafsky, D. (2017). Incorporating dialectal variability for socially equitable language identification. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, volume 2, pages 51–57.

Kiros, R., Zhu, Y., Salakhutdinov, R. R., Zemel, R., Urtasun, R., Torralba, A., and Fidler, S. (2015). Skip-thought vectors. In *Advances in neural information processing systems*, pages 3294–3302.

Labov, W. (1964). *he social stratification of En- glish in New York City*. Ph.D. thesis, Columbia uni- versity.

Landis, J. R. and Koch, G. G. (1977). The measurement of observer agreement for categorical data. *biometrics*, pages 159–174.

Malmasi, S. and Zampieri, M. (2016). Arabic dialect identification in speech transcripts. *VarDial 3*, page 106.

Malmasi, S. and Zampieri, M. (2017). Arabic dialect identification using ivectors and asr transcripts. *VarDial 2017*, page 178.

Malmasi, S., Zampieri, M., Ljubešic, N., Nakov, P., Ali, A., and Tiedemann, J. (2016). Discriminating between similar languages and arabic dialect identification: A report on the third dsl shared task. *VarDial 3*, page 1.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

Mubarak, H. and Darwish, K. (2014). Using twitter to collect a multi-dialectal corpus of arabic. In *Proceedings of the EMNLP 2014 Workshop on Arabic Natural Language Processing (ANLP)*, pages 1–7.

Orton, H., Sanderson, S., and Widdowson, J. (1998). *The linguistic atlas of England*. Psychology Press.

Palva, H. (2006). Dialects: classification. *Encyclopedia of Arabic Language and Linguistics*, 1:604–613.

Pennington, J., Socher, R., and Manning, C. D. (2014). Glove: Global vectors for word representation. In *EMNLP*, volume 14, pages 1532–1543.

Ryding, K. C. (2005). *A reference grammar of modern standard Arabic*. Cambridge university press.

Sadat, F., Kazemi, F., and Farzindar, A. (2014). Automatic identification of arabic language varieties and dialects in social media. *Proceedings of SocialNLP*, page 22.

Trudgill, P. (1974). Linguistic change and diffusion: Description and explanation in sociolinguistic dialect geography. *Language in society*, 3(2):215–246.

Versteegh, K. (2001). Linguistic contacts between arabic and other languages. *Arabica*, 48(4):470–508.

Versteegh, K. (2014). *The arabic language*. Edinburgh University Press.

Zahran, M. A., Magooda, A., Mahgoub, A. Y., Raafat, H. M., Rashwan, M., and Atyia, A. (2015). Word representations in vector space and their applications for arabic. In *CICLing (1)*, pages 430–443.

Zaidan, O. F. and Callison-Burch, C. (2011). The arabic online commentary dataset: an annotated dataset of informal arabic with high dialectal content. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers-Volume 2*, pages 37–41. Association for Computational Linguistics.

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

Zampieri, M., Malmasi, S., Ljubešić, N., Nakov, P., Ali, A., Tiedemann, J., Scherrer, Y., and Aepli, N. (2017). Findings of the vardial evaluation campaign 2017.