Arabizi Identification in Twitter Data

Taha Tobaili
Knowledge Media Institute
The Open University
taha.tobaili@open.ac.uk

Abstract

In this work we explore some challenges related to analysing one form of the Arabic language called *Arabizi*. Arabizi, a portmanteau of *Araby-Englizi*, meaning Arabic-English, is a digital trend in texting Non-Standard Arabic using Latin script. Arabizi users express their natural dialectal Arabic in text without following a unified orthography. We address the challenge of identifying Arabizi from multi-lingual data in Twitter, a preliminary step for analysing sentiment from Arabizi data. We annotated a corpus of Twitter data streamed across two Arab countries, extracted linguistic features and trained a classifier achieving an average Arabizi identification accuracy of 94.5%. We also present the percentage of Arabizi usage on Twitter across both countries providing important insights for researchers in NLP and sociolinguistics.

1 Introduction

Arabizi comprises a portion of the Arabic social media, thus any large dataset crawled from an Arabic public source may contain Modern-Standard Arabic (MSA), Non-Standard Arabic (NSA), Arabizi and other languages such as English and French. MSA is the formal Arabic that is mostly used in news broadcasting channels and magazines to address the entire Arab region. NSA is informal, dialectal and esoteric to each region. It varies among North Africa, Egypt, Levant, and the Arabian Gulf. Arabizi is a descendant of NSA, where dialectal words are expressed in Latin script such as حبيبي which means *darling* written as *7abibi*. Apart from being dialectal, people express their natural voice in text without following a unified orthographical and grammatical regulations.

A. Bies et al. mention that the use of Arabizi is prevalent enough to pose a challenge for Arabic NLP research (2014). Basis Technology, a company that specializes in computational linguistics for digital forensics stated that Arabizi poses a problem for government analytics since it has no structure (2012). The way Arabs use Arabizi is a significant challenge for data scientists for the following reasons: it is written in Latin script, it varies among regions, it is not written in a unified orthographical, syntactical or grammatical structure, it could be mixed with other languages in a single sentence, and it often exist within multilingual datasets.

Identification of Arabizi advances the Arabic NLP, specifically in sentiment analysis for Arabic. Being able to identify and analyse Arabizi will fill an important gap in processing Arabic from social media data. Several researchers working on sentiment analysis for Arabic filter out Arabizi from their datasets, mainly due to the non-availablitiy of public resources such as word lexicons, stemmers, and POS taggers to process this type of text. In a recent survey about sentiment analysis for Arabic, S. Ahmed et al. mention that the use of dialect and Arabizi have not been addressed yet in existing literature (2013). In this paper we address the following questions: How frequent is the usage of Arabizi in Egypt and Lebanon on Twitter, and which methods could be used to automatically identify Arabizi within multi-lingual Twitter streams.

Public social media data, particularly Twitter, is important for sentiment analysis as it contains and reflects the public's opinion. The type of data in Twitter is large-scale and diverse, not biased to certain groups of people or topics. We collect Twitter data from Egypt and Lebanon, pre-

Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics – Student Research Workshop, pages 51–57,
Berlin, Germany, August 7-12, 2016.

process it, and annotate sample datasets. First, we present the amount of Arabizi usage in each of Egypt and Lebanon by reporting the percentage of Arabic, English, and Arabizi tweets. Second, we extract some linguistic features using Langdetect[1] (Nakatani, 2010), a language detection library ported from Google's language-detection, and train an SVM classifier to identify Arabizi from multi-lingual Twitter data. We believe that being able to identify Arabizi from multi-lingual data on social media brings us closer to addressing sentiment analysis for Arabic, inclusive of NSA and Arabizi.

The rest of the paper is structured as follows: In Section II, we investigate what other researchers have done for analysing Arabizi. In Section III, we collect, pre-process and annotate Twitter data and present our approach of extracting linguistic features and training the classifier. In Section IV, we present the results and a discussion. In Section V, we conclude and add a future work plan.

2 Related Work

In this section we survey papers and present the efforts of other researchers on the percentage of Arabizi usage, the motive for analysing sentiment from Arabizi data, and related work in Arabic dialect and Arabizi detection.

2.1 Percentage of Arabizi Usage

Several sociolinguistic studies focus on the Arabizi phenomena, where the researchers tend to explore how this texting style developed and became a trend in the Arab region. S. Jaran and F. Al-Haq (Jaran and Al-Haq, 2015) presented how natives coin and trend Arabizi words by adopting an English word and conjugating it in their NSA such as: I miss you → *missak*, *ak* is a suffix added when referring to the pronoun *you* in the masculine form in several NSA dialects. In (Muhammed et al., 2011; Yaghan, 2008; Aboelezz, 2009; Alabdulqader et al., 2014; Gibson, 2015; Jaran and Al-Haq, 2015; Keong et al., 2015) the authors collected information about people who use Arabizi such as age, gender, and level of education and reported the frequency and context of Arabizi usage within certain groups of people. In Table 1 we present these studies that were conducted by monitoring mobile messaging, distributing surveys among university students, or analysing online forum comments. The percentage of Arabizi usage varies in each of these studies depending on the year the study was conducted, the region, the medium, and the users. However, most of these studies are based on private mobile messages, we on the other hand report the percentage of Arabizi usage on a public social medium. We investigated Arabizi from Twitter data across 2 Arab countries, Egypt and Lebanon; our method can be applied to any other Arab country.

2.2 Arabizi in Sentiment Analysis

We point to few studies where researchers collected Arabic data for sentiment analysis but filtered out Arabizi, saying that their tools are incapable of handling Arabizi. M. Al-Kabi et al. carried out a research in sentiment analysis on a dataset of 1,080 NSA reviews collected from social and news websites filtering out Arabizi (2013; 2014). R. Duwairi and I. Qarqaz collected 500 Facebook comments in Arabic for sentiment analysis filtering out Arabizi as well (2014). R. Duwairi et al. also mention that Arabizi is common in Arab blogs, highlighting the fact that there are no Arabizi benchmark datasets nor sentiment lexicons available (2015).

2.3 Arabic Dialect Detection

Recent studies are focused on detecting dialectal Arabic from a given text. Most of these studies rely on annotated dialectal Arabic corpora and training a classifier with character and word n-gram features. O. Zaidan and C. Callison-Burch annotated over 100,000 sentences and trained an n-gram probabilistic classifier that detects the dialect of input sentences (2014). S. Malmasi et al. used 2 classifiers to predict whether a given sentence is dialectal or MSA (2015). First, a Conditional Random Field (CRF) classifier trained with word-level features using MADAMIRA (2014) and other tools. Second, a trained sentence level classifier covering dialectal statistics, writing style, and word relatedness features. Other recent efforts on dialect detection (Cotterell and Callison-Burch, 2014; Elfardy and Diab, 2013; Sadat et al., 2014; Al-Badrashiny et al., 2015) include creating dialectal corpora.

To the best of our knowledge, K. Darwish presented the only work on Arabizi detection in the literature (2014). However, his work focuses on word-level detection. He collected Arabizi words from tweets and trained a character-level language

[1] https://goo.gl/xn1jJr

Reference	Year	Location	Participants	Data	Size of Data	Arabizi	English	Arabic
(Keong et al., 2015)	2015	Malysia	20 Arab Post Graduates	SMS	200 Messages	35%	50%	10%
(Bies et al., 2014)	2014	Egypt	26 Native Arabic Speakers	SMS	101,292 Messages	77%	-	23%
(Alabdulqader et al., 2014)	2014	Saudi Arabia	61 Students and Non-students	SMS, BBM, and Whatsapp	3236 Messages	15%	8%	74%
(Bianchi, 2012)	2012	Jordan	-	Online Forum	460,220 Posts	35.5%	17.5%	32%
(Al-Khatib and Sabbah, 2008)	2008	Jordan	46 Students	SMS	181 Messages	37%	54%	9%

Table 1: Percentage of Arabizi Usage in Related Work

model and a statistical sequence labelling algorithm. In our work, we extract sentence-level features using Langdetect and train a classifier to identify Arabizi tweets.

3 The Approach

3.1 Data Collection and Annotation

We use geographic information[2] to stream tweets coming from within Lebanon and Egypt, where we specified the coordinates of each region separately. We collect two datasets, one from each country, and split each into Arabic and Non-Arabic, shown in Table 2. The Non-Arabic data includes any tweet written in Latin script. We take the Non-Arabic data segment, pre-process it, and annotate a sample of 5,000 tweets to be used for reporting the percentage of Arabizi usage and as a training dataset for the Arabizi identification classifier.

Country	Tweets	Arabic Tweets	Non-Arabic Tweets
Lebanon	60,364	28,340	32,024
Egypt	249,149	174,821	74,328

Table 2: Distribution of Tweets

Twitter data contains a lot of noisy tweets such as tweets that contain only URLs, hashtags, user mentions, laughter, and spam. We pre-process the data to maximize the number of textual tweets in each of the Non-Arabic dataset. We filter out URLs, hashtags, user mentions, laughter, exaggeration, and emojis from tweets. Though some of these features can be used in an extended work for sentiment analysis, for this task we only aim for language identification. We filter out hashtags because most of them are in English. We filter out laughter expressions and exaggerated words because Langdetect misdetects sentences containing words with repetitive letters. From the resulting data, we deleted tweets that contain no text and duplicated tweets. We observed from our datasets that many tweets aim to gather followers with expressions such as: *follow me*, *follow insta*, and *follow plz*. We consider such tweets as spam and filter out any tweet containing the word *follow*. Our pre-processed Non-Arabic datasets lessened from 32,024 to 21,878 for Lebanon and from 74,328 to 36,970 for Egypt.

We extracted and annotated 5,000 tweets which was done manually by one Arab native. Since Arabizi users might switch between Arabizi and English within a single sentence, we tag Arabizi tweets if the number of Arabizi words are sufficient to imply that the dominant language used is Arabizi. To tag an Arabizi tweet it should have more Arabizi words than English words and the Arabizi words should consist of nouns and verbs not just connectors and stop words. For example:

Tweet: *honestly allah y3afeke (recovery wish) that you still cant get over a story thats a year not my fault ur ex boyfriend was a *** sara7a (honestly)*
Arabizi Words $<$ English Words
Tag: Not Arabizi

Tweet: *kel marra b2oul monday bade ballesh diet bas emta ha yeje hayda lnhar/Everytime I plan to start a diet on monday but when will this day come*
Arabizi Words $>$ English Words
Tag: Arabizi

Tweet: *eh (yes) God bless your mom w (and) your family*
Tag: Not Arabizi

Out of each sample dataset, we tagged 465 Arabzi tweets from Lebanon and 955 Arabizi tweets from Egypt. However, each sample dataset is multi-lingual containing tweets in languages other than English and Arabizi. The annotated

[2] https://goo.gl/PFJj3H

datasets can be found on project-rbz.com[3]

3.2 Arabizi Identification

We utilize Langdetect, a language detection library that detects one or more languages for each input sentence, and returns the language with the highest confidence score. Though it does not detect Arabizi sentences, it detects irrelevant languages when tested against Arabizi. It may detect 3 or more languages, or one irrelevant language with high confidence. We use those detections as input features to train a classifier to identify Arabizi from Twitter data. For example:

Tweet: *never been so scared for an exam*
Languages Detected: {en:0.99}

Tweet: *kan yom eswed yom ma 3reftk /It was a bad day I didn't recognize you*
Languages Detected: {so: 0.42, cy: 0.42, sv: 0.14}

We use the irrelevant languages detected, the number of irrelevant languages, and their confidence scores as input features for the Arabizi identification classifier.

3.2.1 Feature Selection

We extracted the following features during the streaming and pre-processing of tweets: Language detected by Twitter API, languages detected by Langdetect, location of the tweet, country of the user, language of the user, number of words per tweet, and count of word occurrences per tweet.

We extracted (language detected by Twitter API, tweet location, country and language of the user) from each tweet stream, for example:

id	twt	lang	twt_country	usr_id	usr_lang	usr_country
001468231	Hello World	EN	EGY	48933812	EN	EGY

We tested all the features on several classifiers and found that the best results are obtained from an SVM classifier using (languages detected by Langdetect, the language detected by Twitter API, and the count of word occurrences per tweet). The languages detected by Langdetect include: languages predicted, number of predicted languages, and the confidence score of each. Although, Langdetect is more accurate than Twitter API when tested against our data, adding the language detected by Twitter API to the set of features improved the overall accuracy of Arabizi identification. The count of word occurrences per tweet helps the classifier identify words that are frequently used in Arabizi. We disregarded the other features (location of the tweet, country and language of the user, and the number of words per tweet) because they did not have any effect on the classification results.

3.2.2 Classification

We run Langdetect against our annotated sample datasets; in Table 3 we present the distribution of languages detected with high confidence scores apart from the manual Arabizi annotation. We note that the other languages detected are mainly Far-Eastern languages written in Latin script. Though there are very few tweets in Spanish and Dutch, they are negligible. Far-Eastern expatriates living and working in the Arab region constitute a large part of the population. Our findings show that most of the other languages detected in our Twitter datasets in Lebanon are Filipino, and Indian in Egypt. For this experiment we filter out all languages other than English and Arabizi that have confidence score of 0.7 or higher from our sample datasets. The annotated datasets are lessened to 3,707 tweets for Lebanon and 3,823 for Egypt. We note that the remaining datasets contain multi-lingual tweets however those tweets were not given high confidence scores by Langdetect.

Country	Dataset Size	English	Arabizi	French	Other
Lebanon	5,000	3,242	465	158	1,135
Egypt	5,000	2,868	955	0	1,177

Table 3: Distribution of Languages in Sample Data

We carry out two experiments, one with our annotated datasets that are filtered from other languages and another with balanced datasets. Since the ratio of English to Arabizi tweets is very high, we under-sample the annotated-filtered datasets to have an almost equal number of English to Arabizi tweets. We applied a 10-fold cross validation technique in which we split the data into 80% and 20% for training and testing respectively, and average the validation results for all folds.

[3] http://www.project-rbz.com/

4 Results and Discussion

4.1 Arabizi Usage

In Table 4 we present the percentage of Arabic vs Non-Arabic tweets in each country. In Table 5 we present the distribution of languages in each of the Non-Arabic sample dataset.

Country	Tweets	Arabic	Non-Arabic
Lebanon	60,364	47%	53%
Egypt	249,149	70%	30%

Table 4: Arabic vs Non-Arabic Tweets

Country	Tweets	English	Arabizi	French	Other
Lebanon	5,000	65%	9.3%	3%	22.7%
Egypt	5,000	57%	19%	-	23%

Table 5: Distribution of Languages in Non-Arabic Tweets

As it can be seen from the results, the percentage of Arabizi usage differs in each Arab country. In Lebanon, Arabic and Non-Arabic tweets are almost equal, however English is dominant in Non-Arabic tweets. On the other hand, Arabic is dominant in tweets from Egypt. The total Arabizi usage is 4.9% for Lebanon and 5.7% for Egypt. We also observed that not only do the percentage of Arabizi usage differs between countries but also they way it is used in text. In Egypt, most of the Non-English tweets are written either in English or in Arabizi rather than mixing both in single tweets as compared to Lebanon. Also, in Egypt people tend to abbreviate Arabizi words in many cases by avoiding to write the vowels. For example:

Tweet from Egypt:
na w nta hn3ml duet m3 b3d /me and you will perform a duet together
Abbreviations: ana → na, enta → nta, hane3mal → hn3ml, ma3 → m3, ba3d → b3d

Tweet from Lebanon:
bonsoir 7ewalit a3melik add 3ala fb bass i didnt find you can you give me your account /Morning I tried adding you on fb but...
Languages: English, French, and Arabizi.

4.2 Arabizi Identification

We filtered our sample datasets from languages other than English and Arabizi that were detected by Langdetect with high confidence scores. We selected an SVM classifier with the following features: Languages detected by Langdetect, the language detected by Twitter API, and the count of word occurrences per tweet. We present the averaged 10-fold cross validation results in Table 6.

Country	Tweets	Recall	Precision	F-Measure	Accuracy
Lebanon	3,707	91	88	88	93
Egypt	3,823	96	78	85	96

Table 6: Averaged K-Fold Validation Results for Sample Datasets

Since Arabizi is only 12% for Lebanon and 25% for Egypt from the Non-Arabic datasets that are filtered from other languages, shown in Table 3, these datasets are considered imbalanced. We balanced the datasets by undersampling the English tweets and repeated the experiment. We present the averaged validation results for the balanced datasets in Table 7.

Country	Tweets	Recall	Precision	F-Measure	Accuracy
Lebanon	1,150	97	97	97	97
Egypt	2,200	97	97	97	97

Table 7: Averaged Validation Results for Balanced Sample Datasets

4.3 Discussion

Our results show that the percentage of Arabizi usage in Twitter data across both Lebanon and Egypt is lower than the findings by other researchers in mobile messaging, as shown in Table 1. We hypothesize that people prefer to text in Arabizi on private mediums since Arabizi is generally perceived as an informal way of communication. However, 4.9% or 5.7% of a country's Twitter data is Arabizi, which is a large amount of data that might contain valuable information. Therefore it is important to generate NLP resources to identify, analyse, and process Arabizi data.

We found that most of the unidentified Arabizi tweets are tweets from Lebanon written in both English and Arabizi. Langdetect identifies most of such mixed tweets as English. As for the false identification, it was due to misidentification of Far-Eastern tweets. Neither Langdetect nor Twitter API was able to correctly identify all Far-Eastern tweets. The classifier could be enhanced to overcome those errors by extracting word-level features from tweets, such as TF-IDF, n-grams, word lengths, and vowel-consonant ratio, and by training it to classify mixed and Far-Eastern tweets.

The analysis of Arabizi usage on Twitter for different Arab countries provides an insight for researchers who tempt to analyse sentiment from Arabic data and for sociolinguistic researchers who study a language in relation to social factors such as region and dialect. We believe that creating tools to automatically identify Arabizi is a necessary step towards sentiment analysis over this type of text. Arabizi identification could be applied in automatic creation of an Arabizi corpus that could be used for classification tasks and in automatic language detection for machine translation tools.

Another aspect of research in Arabizi includes transliteration to Arabic. There are some tools available such as Yamli[4], Microsoft Maren[5], and Google Input Tools[6], however those tools are designed to help Arab speakers get MSA text by typing Arabizi. Using transliterators to convert the natural Arabizi text, such as tweets, may result in broken Arabic words. Some researchers are working on Arabizi transliteration as in (Bies et al., 2014; May et al., 2014; Chalabi and Gerges, 2012; Darwish, 2014).

5 Conclusion and Future Work

In this work we have studied the usage of Arabizi on Twitter and the creation of tools to automatically identify Arabizi from multi-lingual streams of data. We collected Twitter data from Lebanon and Egypt and presented the percentage of each language, particularly Arabizi, providing an important insight for researchers working on the analysis of natural text for the Arab region. We trained an Arabizi identification classifier by annotating sample datasets and extracting features using Langdetect, an existing language detection library. We achieved an average classification accuracy of 93% and 96% for Lebanon and Egypt datasets respectively. Our Arabizi identification classifier relies on sentence-level features; it could be improved by extracting word-level features from text. Our aim is to advance the Arabic NLP research by facilitating analysis on social media data without the need to filter out complex or minority languages.

Several researchers have contributed to the analysis of MSA in the literature, which is useful for analysing formal blogs and news pages. However analysing the public's sentiment requires effort for NSA and Arabizi as most people from the Arab region express their opinion on social media using their mother tongue in text. Though NSA and Arabizi are written in different scripts, they can be addressed simultaneously since both share similar challenges. We plan to extend this work by exploring the usage of NSA and Arabizi across several regions, and by identifying dialect on social media. We follow by trying to extract sentiment from NSA and Arabizi which might require the creation of dialect sentiment lexicons and parsers to process the heterogeneous Arabic social data.

[4]`http://www.yamli.com/`

[5]`https://goo.gl/3zLLOn`

[6]`http://www.google.com/inputtools/`

References

M Aboelezz. 2009. Latinised Arabic and connections to bilingual ability. In *Papers from the Lancaster University Postgraduate Conference in Linguistics and Language Teaching*.

Shehab Ahmed, Michel Pasquier, and Ghassan Qadah. 2013. Key issues in conducting sentiment analysis on Arabic social media text. In *9th International Conference on Innovations in Information Technology (IIT)*, pages 72–77. IEEE.

Mohamed Al-Badrashiny, Heba Elfardy, and Mona Diab. 2015. Aida2: A hybrid approach for token and sentence level dialect identification in Arabic. *CoNLL 2015*, page 42.

Mohammed Al-Kabi, Amal Gigieh, Izzat Alsmadi, Heider Wahsheh, and Mohamad Haidar. 2013. An opinion analysis tool for colloquial and standard Arabic. In *The 4th International Conference on Information and Communication Systems (ICICS)*.

Mohammed N Al-Kabi, Izzat M Alsmadi, Amal H Gigieh, Heider A Wahsheh, and Mohamad M Haidar. 2014. Opinion mining and analysis for Arabic language. *International Journal of Advanced Computer Science and Applications (IJACSA)*, 5(5):181–195.

M. A. Al-Khatib and E. H. Sabbah. 2008. Language choice in mobile text messages among Jordanian university students. *SKY Journal of Linguistics*, 21:37–65.

Ebtisam Alabdulqader, Majdah Alshehri, Rana Almurshad, Alaa Alothman, and Noura Alhakbani. 2014. Computer mediated communication: Patterns & language transformations of youth in Arabic-speaking populations. *Information Technology & Computer Science (IJITCS)*, 17(1):85.

Basis-Technology. 2012. The burgeoning challenge of deciphering Arabic chat.

R. M. Bianchi. 2012. 3arabizi-when local Arabic meets global nglish. *Acta Linguistica Asiatica*, 2(1):89–100.

Ann Bies, Zhiyi Song, Mohamed Maamouri, Stephen Grimes, Haejoong Lee, Jonathan Wright, Stephanie Strassel, Nizar Habash, Ramy Eskander, and Owen Rambow. 2014. Transliteration of Arabizi into Arabic orthography: Developing a parallel annotated Arabizi-Arabic script sms/chat corpus. *ANLP 2014*, page 93.

A. Chalabi and H. Gerges. 2012. Romanized Arabic Transliteration. In *24th International Conference on Computational Linguistics*, page 89.

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

K. Darwish. 2014. Arabizi Detection and Conversion to Arabic. *ANLP 2014*, page 217.

R. M. Duwairi and I. Qarqaz. 2014. Arabic sentiment analysis using supervised classification. In *International Conference on Future Internet of Things and Cloud (FiCloud)*, pages 579–583. IEEE.

RM Duwairi, Nizar A Ahmed, and Saleh Y Al-Rifai. 2015. Detecting sentiment embedded in Arabic social media–a lexicon-based approach. *Journal of Intelligent and Fuzzy Systems*.

H. Elfardy and M. T. Diab. 2013. Sentence Level Dialect Identification in Arabic. In *ACL (2)*, pages 456–461.

M. Gibson. 2015. A framework for measuring the presence of minority languages in cyberspace. *Linguistic and Cultural Diversity in Cyberspace*, page 61.

S. A. Jaran and F. A. Al-Haq. 2015. The use of hybrid terms and expressions in colloquial Arabic among Jordanian college students: A sociolinguistic study. *English Language Teaching*, 8(12):86.

Yuen Chee Keong, Othman Rahsid Hameed, and Imad Amer Abdulbaqi. 2015. The use of Arabizi in English texting by Arab postgraduate students at UKM. *The English Literature Journal*, 2(2):281–288.

Shervin Malmasi, Eshrag Refaee, and Mark Dras. 2015. Arabic dialect identification using a parallel multidialectal corpus. In *Proceedings of the 14th Conference of the Pacific Association for Computational Linguistics (PACLING 2015), Bali, Indonesia*, pages 209–217.

Jonathan May, Yassine Benjira, and Abdessamad Echihabi. 2014. An Arabizi-English Social Media Statistical Machine Translation System.

Randa Muhammed, Mona Farrag, Nariman Elshamly, and Nady Abdel-Ghaffar. 2011. Summary of Arabizi or Romanization: The dilemma of writing Arabic texts. In *Jil Jadid Conference*, pages 18–19. University of Texas at Austin.

S. Nakatani. 2010. Language Detection Library for Java.

Arfath Pasha, Mohamed Al-Badrashiny, Mona T Diab, Ahmed El Kholy, Ramy Eskander, Nizar Habash, Manoj Pooleery, Owen Rambow, and Ryan Roth. 2014. Madamira: A fast, comprehensive tool for morphological analysis and disambiguation of Arabic. In *LREC*, pages 1094–1101.

Fatiha Sadat, Farnazeh Kazemi, and Atefeh Farzindar. 2014. Automatic identification of Arabic language varieties and dialects in social media. *Proceedings of SocialNLP*.

M A Yaghan. 2008. ”arabizi”: A contemporary style of Arabic slang. *Design Issues*, 24(2):39–52.

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