WRITING ARABIZI: ORTHOGRAPHIC VARIATION IN ROMANIZED LEBANESE ARABIC ON TWITTER

Natalie Sullivan

TC 660H
Plan II Honors Program
The University of Texas at Austin

May 4, 2017

_____

Barbara Bullock, Ph.D.
Department of French & Italian
Supervising Professor

_____

John Huehnergard, Ph.D.
Department of Middle Eastern Studies
Second Reader

ABSTRACT

Author: Natalie Sullivan

Title: Writing Arabizi: Orthographic Variation in Romanized Lebanese Arabic on Twitter

Supervising Professors: Dr. Barbara Bullock, Dr. John Huehnergard

How does technology influence the script in which a language is written? Over the past few decades, a new form of writing has emerged across the Arab world. Known as Arabizi, it is a type of Romanized Arabic that uses Latin characters instead of Arabic script. It is mainly used by youth in technology-related contexts such as social media and texting, and has made many older Arabic speakers fear that more standard forms of Arabic may be in danger because of its use.

Prior work on Arabizi suggests that although it is used frequently on social media, its orthography is not yet standardized (Palfreyman and Khalil, 2003; Abdel-Ghaffar et al., 2011). Therefore, this thesis aimed to examine orthographic variation in Romanized Lebanese Arabic, which has rarely been studied as a Romanized dialect. It was interested in how often Arabizi is used on Twitter in Lebanon and the extent of its orthographic variation. Using Twitter data collected from Beirut, tweets were analyzed to discover the most common orthographic variants in Arabizi for each Arabic letter, as well as the overall rate of Arabizi use. Results show that Arabizi was not used as frequently as hypothesized on Twitter, probably because of its low prestige and increased globalization. However, its consonants are relatively standardized, while its vowels show more variation. This thesis adds to the existing conversation about Romanized Arabic by presenting a detailed study of orthographic variation in Lebanese Arabic. The results could have useful implications for Arabic language ideology and technological endeavors, such as natural language processing or translation programs.

Acknowledgements

I would like to thank my advisers, Dr. Barabara Bullock, and Dr. John Huehnergard, for their advice, support, and feedback while completing this thesis, which has been invaluable throughout the whole project. I would also like to thank the UT Middle Eastern Studies Department, especially Dr. Mahmoud al-Batal, for helping check the accuracy of my methodology and translations and helping me acquire the necessary language skills to complete this project. Finally, thank you to Plan II for providing resources and a supportive community.

Table of Contents

Part 1: Introduction and Background

I. Introduction

Over the last decade, as global access to technology increased, Arabic-speaking youth across the Arab world began to write their spoken language on social media sites such as Facebook and Twitter, as well as in other forms of computer-mediated communication (CMC) such as text messaging. CMC, which is defined as "predominantly text-based human-human interaction mediated by networked computers or mobile telephony," (Herring, 2007:1) fundamentally changed how language was used in technology-based contexts. As Arabic speakers realized that early technology could not accommodate Arabic script, they adapted their language so it could be used in online spaces. Instead of writing in Arabic script, they used Latin characters and numerals (often using numbers to represent Arabic characters not found in English, such as the number 3 for the letter ع). However, this approach was not standardized: Since not every Arabic character has an exact English equivalent, the way people wrote the same sounds differed from country to country, depending on how spoken Arabic dialects differed in each location. This kind of writing, which came to be known as Romanized Arabic or Arabizi (see Figure 1), was a radical departure from traditional methods of writing Arabic, and made many older Arabic speakers uneasy that standard forms of Arabic, such as Modern Standard Arabic, were disappearing.

Although Arabizi originally arose from a need to facilitate the use of Arabic in technology-related contexts, it has now spread to other domains, such as advertisements and cartoons (See Figure 1) (Essawy, 2010:5).

Figure 1: Lebanese restaurant advertisement
[In Arabizi: "Na3na3" and below in Arabic script: "Mint"]

It has increased in use and popularity over the past few years throughout the Arab world, and is used especially by young people, who are typically more accustomed to dealing with CMC. Additionally, because of Arabizi's status as a technologically-linked language and the fact that English is one of the most popular languages used on the Internet, code-mixing (the use of two languages within a single utterance) between English and Arabizi frequently occurs in contexts where Arabizi is used (Warschauer et al., 2002:1). All of these developments suggest that Arabizi is a distinct form of writing that can signal certain characteristics about its users' identities and provide a lens through which to view language in the Arab world.

While the linguistic characteristics of Romanized Arabic have been well documented, not much research has been dedicated to examining its use on Twitter and how it is written in specific dialects, especially those in the Levant. Prior research on Arabizi suggests that although it is frequently used on social media, people do not think of it as a "real language" because its orthography is not yet standardized (Bahrainwala, 2011:16; Al- Khalil and Palfreyman, 2003). Additionally, forms of colloquial Arabic, such as those usually written in Arabizi, have traditionally been spoken, not written, so the extent to which Arabizi's writing is standardized is still a matter of debate— as Haggan (2007:442) notes, "to date, there is little research on how

widespread the Romanisation of Arabic is in electronic communication, whether there are inter-regional variations in the use of numerals and whether there may also be inter-media variations." Therefore, this thesis will analyze the rate and orthographic variation in Romanized Lebanese Arabic on Twitter. It will focus on three questions: 1) How often is Romanized Lebanese Arabic used on Twitter? 2) What variation occurs in Lebanese Arabizi orthography (to what extent is its writing standardized)?, and 3) How could this affect Arabic language ideology?

This thesis is organized as follows: First, it will provide background information relating to ideology and history of Arabizi, with a focus on defining key terms such as orthographic variation and diglossia. This section will also include information about the particular linguistic landscape of Lebanon, and how Lebanese Arabic phonology might relate to writing Arabizi. Then, it will present an exploratory study of Arabizi tweets collected from Lebanon, and analyze this data for the rate of Arabizi use and most common orthographic variants for each character. Finally, it will interpret these results in the context of the current technologic and linguistic environment of Lebanon, and evaluate how much Lebanese Arabizi orthography has become conventionalized and what the effect might be on Arabic language ideology in the future.

II. Background

2.1 Arabic Language Ideology

On the most basic level, language ideology involves "speakers' beliefs about social boundaries that shape their associations with language" (Irvine and Gal, 2000:1). Language ideology determines how speakers interpret and use language to signal their identity and affiliate themselves with or distance themselves from certain groups. Arabic has a particularly complex language ideology because it is diglossic, meaning that it has multiple varieties that are used

under different conditions within the same community. Ferguson (1959) cites Arabic as a textbook example of diglossia because it has a more formal, or "high" variety (Standard Arabic) as well as a "low" spoken variety (dialects). He defines diglossia as:

> A relatively stable language situation in which, in addition to the primary dialects of the language (which may include a standard or regional standards), there is a very divergent, highly codified (often grammatically more complex) superposed variety, the vehicle of a large and respected body of written literature, either of an earlier period or in another speech community, which is learned largely by formal education and is used for most written and formal spoken purposes but is not used by any sector of the community for ordinary conversation (338).

Although Ferguson first proposed the term *diglossia* to describe the relatively rigid divide between high and low language varieties (in this case, formal and informal Arabic), more recent research has shown that this divide is more flexible than originally believed. For example, in his book *Diglossia and Language Contact*, Lofti Sayahi (2014:202) argues that diglossia is stable in that there is a functional divide between high and low language varieties, but it is also variable because low varieties keep changing functionally and structurally depending on their ecologies. Similarly, as Brustad (2000) and Bassouiney (2009:198) show, the Arabic language can instead be viewed as a spectrum, spanning from Classical Arabic (*fusħa*) on one end as the highest, most formal language variety, to *ʕāmmiyya*, or colloquial Arabic (dialects), on the other (See Figure 2). An intermediate variety known as Modern Standard Arabic (MSA), ranks between these two forms, and is mostly used in news broadcasts and formal interactions, but rarely spoken elsewhere, and is not learned as a native tongue. Badawi (2012:3) points out that the boundaries between these categories are fluid, and native Arabic speakers will adapt their language forms based on context and communication needs.

Figure 2: Levels of Arabic Diglossia
Adapted from Bassouiney (2009)

Heritage Classical: Classical Arabic used in the Quran and other religious settings. Usually written, occasionally spoken. Uses vowel markings and case endings.
Contemporary Classical: Modern Standard Arabic used in news broadcasts, newspapers, and most literature.
Colloquial of the Cultured: Colloquial (dialect) Arabic influenced by MSA. Usually spoken by more educated people. Used in television, discussions, and in universities.
Colloquial of the Basically Educated: Colloquial Arabic used to talk to family and friends about everyday topics such as sports, movies, fashion, etc. Spoken in informal contexts.
Colloquial of the Illiterate: Used by less educated people in informal situations. Often involves slang or curse words. Used in soap operas, cartoon shows, and humorous situations.

However, most native Arabic speakers maintain that Classical and Modern Standard Arabic are superior to colloquial forms, and that the boundary between standard and colloquial Arabic is rigid. In fact, some speakers do not see colloquial forms of Arabic as "Arabic" at all (Hoigilt et al., 2013). This distinction helps explain why Arabizi is often involved with code-mixing; as Sayahi (2014:169) argues, the vernacular form is open to influence in language contact because it is seen as distinct from the standard form. Thus, language varieties such as Arabizi can be mixed with English or French. There are several reasons for this divide in ideology between formal and informal Arabic: Classical Arabic is the language of the Quran, the Islamic holy book, and is therefore afforded more prestige. Secondly, children learn to speak their own dialects of colloquial Arabic at home, and are only later taught to read and write MSA in school, giving it an association with education and higher social class. Because colloquial forms of Arabic are not usually written, no dialect has a standardized writing system, meaning that transcriptions are often ad hoc, whether they are in Latin or Arabic script. Due to this lack of

a formal writing system, many Arabs see dialects as having diminished value and as being a "dirty" or street form of Arabic. Lastly, globalization has contributed to a greater exposure to other languages and language mixing, so languages like French and English are usually associated with more job opportunities, but are also seen as a danger to standard Arabic.
As Eisele (2003:43) notes, the dominant Arabic language ideology seems to be that "Arabic unites all Arabs and should therefore be a single language for a single culture; it is in competition with foreign cultures and languages and needs to be protected from contamination by them and also by Arabic dialects, which represent corruptions of the norm." Consequently, many Arabic speakers tend to embrace the ideology that more formal Arabic should be preserved and act as a unifying factor in the Arab world, while colloquial forms should be disparaged.

Because of Arabizi's association with code-switching and Latin script, it is especially affected by Arabic language ideology. As Unseth (2005:23) points out, the Arabic script often acts as a unifying identity marker for Arabic speakers; even though they use different dialects, most Arabic speakers will say they all speak one language, in part because they all use the same script. In contrast, some language varieties such as Mandarin Chinese and Cantonese use the same script even though the two are not mutually intelligible. This partly explains the dismissive attitude many users hold toward Arabizi. Because it does not use Arabic script and is often mixed with other languages, most Arabizi users seem to think that it is "not a real language," and that it is used more out of convenience than anything else (Bahrainwala, 2011:16). Some Arabic researchers have also started a series of anti-Arabizi campaigns across the Arab world—for instance, the United Nations' World Language Day in 2014 focused on preserving the Arabic script, and openly discouraged the use of Arabizi (UNESCO, 2014). Surveys of Arabizi users

have shown that some feel guilty because they feel that Arabizi is "annihilating the script of the language" and reducing national identity (Bahrainwala, 2011:17). Interestingly, this contrasts with the idea of Arabizi as a unifying force among Arab youth that Bahrainwala also points out. This last factor, reduction of national identity, adds to the prevailing ideology that Arabic is in danger and needs to be preserved. On the surface, this statement seems absurd; after all, there are approximately 225 million Arabic speakers worldwide (Sayahi, 2014:20), and no indication that this number is decreasing. However, globalization and the increased code-mixing seen in language forms like Arabizi are viewed as threats by older, more traditional generations who fear that Classical Arabic and other more formal varieties of Arabic are dying out in favor of more prestigious languages like French and English, and that national identity and religion will fall apart with them.

However, younger generations seem to have more favorable attitudes toward Arabizi, but still support the preservation of Arabic. Some authors have even advocated for the use of Arabizi as a way to ensure the preservation of MSA; as Charbel El-Khaissi (2015:17) puts it, "the Latin script used in Romanised Arabic fills the 'script-gap' in spoken Arabic forms, allowing users to transcribe or transliterate their dialect in question, while simultaneously keeping the standard very distinct from the spoken varieties, thus keeping the prestige of the standard intact." In a study of Jordanian college students, 66% disagreed or strongly disagreed that "mixing Arabic and English will lead to the death of the Arabic language," but 82% agreed or strongly agreed that "the Arabic language must be promoted and protected in order for it to survive"(Al-Haq and Jaran, 2015:21). The authors take this as evidence that Arabic is in danger, but as Brustad (2015) points out, perhaps what is in danger is not the Arabic language itself, but the strict diglossic

ideology surrounding it; that is, the idea that formal and informal varieties of Arabic should remain separate and not mix. As she puts it:

> To what extent should Arabic exist and be used in public space only in its standard "correct" form? Can there be linguistic plurality? Increasingly, Arabic speakers are answering that in the affirmative through their usage, and this subtle ideological shift is where the real "danger" to Arabic lies (34).

This point is borne out by the increased use of both formal and informal varieties on social media, advertisements, and even in newspapers, suggesting that the divide between formal and informal Arabic is increasingly dissolving in a transition to a more relaxed language ideology. The data in the latter half of this thesis will be examined as possible evidence of this transition.

2.2 Written Arabic

Hoigilt et al. (2013) argue that attitudes about the Arabic language system have become tense in recent years, mainly caused by "the gap between official dominant language ideologies and actual language practices." That is, although people think that standard Arabic should be preserved and used, they rarely do so in real life. This tension is especially apparent in the increased variation and change seen in written Arabic, which has been accompanied by increased code-switching (Hoigilt et al., 2013), most prominently between English and Arabic. Doss (2006:63) and Badawi (2004) also note that colloquial Arabic is increasingly being written and mixed with MSA, in fiction, newspaper columns, and on social media. Similar to diglossia, this phenomenon of using two different script systems in the same community is called digraphia. Peter Unseth (2005: 36) acknowledges multiple forms of digraphia: 1) "situations where two scripts are used for (more or less) the same language by different communities," as in Hindi and Urdu, 2) "two scripts being used in the same multilingual community, but for different

languages," such as Arabic and French, or 3) "the same community using two (or more) different scripts at the same time period to write the same language, but using the script for different domains." This last definition applies most closely to Arabizi, because it is used mostly in CMC-related domains, while Arabic script is used in most other settings, so each script serves a particular social and communicative function. For instance, in this study, digraphia often occurred between colloquial Arabic written in Arabic script and Arabizi, as in the phrase (read right-to-left):

<w ma baref shou btaamle كل يوم بروح الصف >

[wa ma b-ʕrəf ʃu b-ətʕməli asˁaf b-əruħ jom kəl]

"*Every day I go to class and I don't know what you're doing.*"

Both parts of the phrase are written in colloquial Arabic (as evidenced by use of the b- prefix (ب) as an aspect marker for present tense before main verbs, which is not used in MSA), but the user switches from Arabic script to Arabizi between two independent clauses, which is a common boundary to switch between script systems (Unseth, 2005:37). Sometimes (as in the previous example) it is not entirely clear why someone chooses to write in two different scripts. However, script-switching is governed by a variety of social, political, and linguistic factors, which will be discussed further in the section on orthographic variation.

This script-mixing is done by a very specific subset of Arabic speakers—usually educated teens or young adults—to the point where "written language has never been presented to speakers of Arabic on such a grand scale and in such a variety as it is nowadays, and at the same time, it has never been less monopolized by traditional language authorities" (Kouloughli, 2010; cited in Hoigilt et al., 2013). Therefore, as Hoigilt et al. (2013) write, "The tension evident

in written Arabic can be interpreted as a tension between an official, formal notion of literacy, based on the dominant regime of authority (represented by Modern Standard Arabic), and a popular, unofficial notion (represented by colloquial Arabic) that is seemingly gaining a foothold in popular culture and beyond." Consequently, all of these factors—increased code-mixing and digraphia, increased acceptance of written colloquial Arabic, and youth control of certain language forms—have the potential to influence how Arabizi is written in online spaces.

III. Language in Lebanon

Lebanon is a polyglossic environment where many different languages interact with each other on a daily basis, including Arabic, French, and English, as well as some Armenian, Greek, and Kurdish. According to the Lebanese constitution, the country's official language is Arabic, although English and French are the main instructional languages in most colleges and universities (Bassam, 2014:115). More importantly, bilingualism is a way of life in Lebanon; as Chahine (2011: 1) notes, "it is not uncommon for people to greet each other on the streets by saying, "*Hi, Kayfik, Ca va*?*"* (In this case, the cédille is missing from ça va, as this is how it is popularly used). "Bilingualism is "mainly seen in the streets, on billboards, the way people address each other…Many people are bilingual, trilingual, if not multilingual" (Chahine, 2011:1). Bassam (2014:115) describes several reasons for this phenomenon: Lebanon's geographic location as a place where 'East meets West,' a long-standing history of immigration, French colonization, and globalization. Additionally, because of its strategic location on the Mediterranean coast and relative economic and political stability compared to other Middle Eastern countries, it has been a prime resettlement location for refugees, especially from Palestine, Iraq, and Syria. According to data from the European Commission for Humanitarian

Aid, Lebanon has the highest per-capita concentration of refugees in the world, and as of March 2017, was home to approximately 1.5 million Syrian refugees who emigrated to Lebanon due to the Syrian civil war and ongoing refugee crisis. It remains to be seen what effect, if any, this influx of refugees will have on Lebanon's linguistic environment, although it is likely that increased language mixing and acceptance of Syrian colloquial Arabic will occur. All of these factors make Lebanon an especially interesting place to study the interaction between multiple languages and their effects on orthographic variation as seen in Arabizi.

3.1 Arabic

According to Fischer and Jastrow (2000; cited in Elhij'a, 2014), Arabic dialects can be divided into five main geographic groups, each of which has its own distinguishing characteristics: North African, Egyptian-Sudanese, Levantine, Iraqi, and Peninsular (which includes Saudi Arabia, Yemen, and the eastern United Arab Emirates). Since it is located on the Mediterranean coast, with Syria to the east and Israel to the south, Lebanon is considered a part of the Levant, and its residents speak a dialect that falls under the umbrella of Levantine Arabic. Although MSA is used on TV and in literature, the main language of Lebanon is Lebanese Colloquial Arabic (LCA). LCA is grammatically structured differently than Classical Arabic or MSA and is acquired as a first language. It is used in TV shows, music, and everyday conversations. It is also not taught in schools, so until recently, it was not used in written literature (Bassiouney, 2009). Many Arab nations fear that the use of both formal and informal varieties of Arabic is declining; according to the 2015 Arab Youth Survey, 36% of Arab youth say they use English more than Arabic on a daily basis, especially in Gulf countries, where this is true of 56% of youth. This preference has been largely attributed to globalization, technology,

and more favorable attitudes toward French and English. LCA also has many variations in phonology, morphology, and lexicon that set it apart from MSA; these will be discussed further in the section on Arabic phonology.

3.2 French

When the Ottoman empire collapsed after World War I, France gained control over the provinces that would become present-day Lebanon, and retained control until the country declared its independence in 1943. Because of this, French has remained a linguistic bulwark in Lebanon. It is considered a prestige language linked to the upper middle class and higher educational opportunities, since most universities and higher-paying jobs require fluency in French. Almost 40% of Lebanese people are considered francophones, and another 15% "partial francophones." About 20% of the current population uses French on a daily basis (Barlow and Nadeau, 2008). However, even though about 70% of Lebanon's secondary schools use French as a second language of instruction, this does not guarantee French literacy; students' proficiency in French is largely governed by social class. Students from wealthier families tend to be fully bilingual in Arabic and French since they have spoken it from an early age, while lower class students often graduate with uneven language proficiency. Additionally, the use of French is also affected by religion. Lebanese Christians, which make up about 30% of the population, are more likely to use French or English due to influence from Christian missionaries, while Lebanese Muslims are more likely to use Arabic because it is the language of the Quran. Due to the prominence of French in Lebanon, it is expected that linguistic influence from French will have a significant impact on how Arabizi is written.

3.3 English

English is the third most popular language used in Lebanon, although it is slowly gaining traction due to globalization and the rise of social media, particularly among younger generations. This is evident by the relatively recent proliferation of English-language newspapers and TV channels, such as *The Daily Star* and *Al-Nahar English*. Additionally, about 21% of schools have English as a first foreign language, and this number is increasing (Shawish, 2010). Like French, English in Lebanon is viewed both positively and negatively: Many people seem to realize its importance on the global job market, while also being wary of its ties to Westernization and possible threat to Lebanese culture, identity, and language.

3.4 Arabizi

The word "Arabizi" is a blend of two words: Arabic and *ingleezi* (the Arabic word for English). Therefore, the term is used to refer to a form of Arabic written with Latin characters instead of Arabic script, although it is also sometimes used to refer to code-switching that occurs between Arabic and English. Throughout this thesis, the terms Arabizi or Romanized Arabic will be used interchangeably to refer to the first definition.

Figure 3: A tweet written in Arabizi

[Ghida @ghida_kh – Sep 11
X: Do you know how to swim?
Y: Yeah, normally.
X: What do you mean, "Normally?"
Y: I mean if you spray water on me I won't drown]

Arabizi uses numbers to represent certain sounds and letters that are available in Arabic, but not in English. Often, these numbers look like the letters they represent—for example, 7 for ح or 3 for ع (see Figure 4). Which numbers are used depends on the specific dialect of Arabic that is spoken in a region. Because different dialects vary in their pronunciation of Arabic, they also write Arabizi differently. A table with Arabic characters and their Arabizi counterparts is shown below:

Figure 4: Sound Symbol Correspondences in Romanized Lebanese Colloquial Arabic (RLCA)

RLCA	IPA	Arabic script
2	ʔ	ء
a	a	ا
b	b	ب
t	t	ت
Not usually written	θ	ث
j/g	ʒ	ج
7/h	ħ	ح
7’/5/kh	x	خ
d	d	د
d/z	ð	ذ
r	ɾ	ر
z	z	ز
s	s	س
sh/ch	ʃ	ش
s	$s^ʕ$	ص

d/D	dˤ	ض
t/T	tˤ	ط
th/z	ðˤ	ظ
3	ʕ	ع
3’/8/gh	ɣ	غ
f	f	ف
9/q	q	ق
k	k	ك
l	l	ل
m	m	م
n	n	ن
o/w/ou	w/u	و
h	h	ه
i/y/ey	y/i	ي

Although Arabizi orthography is not completely standardized, there are some general characteristics that define Arabizi as a writing code. In the representation of vowels and consonants, Yaghan (2008:42) notes that the use of vowels is optional in Arabizi, and they can even be omitted depending on the reader’s familiarity with the specific variety of Arabizi, the contextual clarity of the word, and sometimes the allowed number of characters per message. When vowels are used, the general trend is that “a” represents the *fatha* (ـَ), “i” or “e” represents the *kasra* (ـِ) and the “u”, “ou”, or “o” is used to represent the *damma* (ـُ). For consonants, Yaghan (2008:44) says that consonant sounds are represented by their English counterparts;

however, in a study of Egyptian Arabizi, Abdel-Ghaffar et al. (2011) note the orthographic conventions of a user's L2 affect the representation of consonants in Arabizi. For instance, if the user's L2 is English, the ج can be represented either with "g" or "j," but French L2 speakers must use <gu> to represent [g] because the <u> serves to indicate that <g> is a velar stop rather than a palatal fricative [ʒ]. Similarly, French L2 speakers use <ch> for ش and <ou> for *damma*, while English L2 speakers would be expected to use <sh> and <oo>. This difference is apparent in Arabic-speaking countries where French is a second or primary language, such as Morocco or Lebanon. Although some level of standardization has occurred (for example, the use of 7 for ح or 3 for ع), there is still no standardized way to write vowels, even for individual users. This is because Arabic uses a consonantal alphabet where vowels are written as diacritical markings on top of or below words, and are usually not printed. Additionally, vowel pronunciations often differ between dialects. Therefore, when trying to write these sounds using Latin script, vowels are sometimes omitted or written based on the user's preference. It is also important to note that the variety of Arabizi differs not only in each country in which it is used, but between different groups of speakers within a country as well.

IV. History

In order to understand the current context and use of Arabizi, it is useful to understand its history. No one seems to agree on when exactly modern-day Arabizi came into existence, although it has been closely linked with the development of the Internet. It first emerged during the mid-1900s when the first wave of technology was beginning to enter the Arab world, as a reaction to the Latin-script dominated world of electronic communication (Yaghan, 2008:41). Since the limited character set of ASCII code (American Standard Code for Information

Interchange) on which the Internet was based could only support English, it was a way to allow Arabic to be used on computers and other electronic devices, since they could not handle Arabic script, and Arabic keypads for mobile phones did not come into use until 2000 (Haggan, 2007:445). The dominance of English on the Internet can still be seen today in the use of only ASCII characters for domain names. Although Arabizi is not the first attempt at a Romanized Arabic transliteration system (Yaghan discusses such proposals that date back to the 1880s), it is unique in its use of numbers to represent characters not found in Arabic, which is a hallmark of the new technology and "text speak" such as *cu l8r* (*see you later*) that was widely used in the time around Arabizi's inception. Interestingly, this aspect of Arabizi seems to have evolved over time; early instances of Arabizi only contain the numbers 2, 3, and 7, instead of the range including 5, 6, 8, and 9 that is seen today (Haggan, 2007:442). Even though support for Arabic script is now available, many Arabic speakers continue to use Arabizi because they are used to typing on QWERTY keyboards and do not have the option to type in Arabic script. This is changing somewhat with the advent of international keyboards on cell phones, but many users still choose to use Latin script out of habit, which is one reason for Arabizi's permanence.

Arabizi has also shown its power symbolically as a protest language; during the Arab Spring and following years (from about 2011 onward), the use of Arabizi "practically exploded" as nationalistic messages on social media spread across the Arab world and Arabizi became the primary language used by protesters in digital spaces (Bahrainwala, 2011:3). The series of revolutions, which was powered by two factors, Arab youth and digital technology, appeared to create the perfect environment for Arabizi to take hold, proving that Arabizi "appears to have the rhetorical power to create and mobilize users in digital communities into social action"

(Bahrainwala, 2011:1). This marked a turning point for Arabizi in that it signaled youth autonomy in language choice and demonstrated the power of technology to unify the Arab world through language.

Lastly, Arabizi also indexes the social community and age of its users; it is typically used by younger, more technologically fluent generations, usually bilingual adolescents and young adults between the ages of 13-20 (Essawy, 2010:6). Surveys of Arabizi users have shown that their reasons for using Arabizi are varied, including: a) It is easier and faster to type in Latin characters than in Arabic characters, b) Users do not have access to a computer with Arabic script, c) Arabizi makes users look "cool" d) Users feel uncomfortable writing in Arabic script and want to avoid language policing (Essawy, 2010:6), e) Users can express personal content in their own language (Arabic) that they can't express well in English (Warschauer et al., 2002:13), f) Users want to create a "warm, friendly atmosphere" or identify as part of a community, g) It is more economical to write in Arabizi because of space or character limits, such as on Twitter, and h) Users are influenced by their peers and do it to "go with the flow" (Abdel-Ghaffar et al., 2011). Therefore, Arabizi use is affected by both social and linguistic factors. The process by which most youth learn Arabizi is also interesting; most Arabizi users say they pick it up from friends and are able to learn it without formal instruction, which suggests that writing Arabizi is somewhat intuitive (Gordon, 2011). According to Abdel-Ghaffar et al. (2011), most Arabizi users say it does not affect their sense of identity as Arabs, although it may contribute to a slight decline in their Arabic skills. Additionally, many users have said Arabizi helps them code-switch between Arabic and English more easily (Palfreyman. and al-Khalil, 2003; Yaghan, 2008:42; Essawy, 2010:7; Abdel-Ghaffar et al., 2011). Arabizi is not unique in this regard; other

languages with non-Latin writing systems, such as Greek and Hindi, have also developed Romanized scripts that promote code-switching. For example, Greek users of CMC often use Roman characters and Arabic numerals to represent the original orthography of their language in a hybrid known as "Greeklish" (Palfreyman, 2001). Interestingly, Arabizi's use across the Arab world and association with young people may give it an appeal that transcends traditional dialect boundaries and promotes a sort of pan-Arabism. Bahrainwala (2011:23) notes, "There are possibly more [Arabizi] users than speakers of any one dialect of Arabic…This is because [Arabizi] cuts across Arabic dialects, and is almost equally intelligible to the Arabic speaking digerati of any country." Thus, as Arabizi's history shows, it has positively impacted the Arab world in some ways by attempting to bridge dialect boundaries.

V. Arabic Phonology

Transcription vs. Transliteration

Phonology is the study of the system of sounds that make up a particular language and how they relate to each other. Since Arabizi requires writing down a mental representation of Arabic sounds based on Latin letters, there is an ongoing discussion about whether Romanized scripts such as Arabizi represent transcription (writing based on attempts to match pronunciation) or transliteration (writing based on replacing one character for another) of another language system. That is, do users of Arabizi see it as a written form of spoken LCA, or do they try to match Arabic characters to Latin letters? Gordon (2011:2) suggests that although the Romanization of Arabic involves both transcription and transliteration, it is primarily governed by attempts to match a character's pronunciation. Similarly, Aboelezz (2008:19) found that when users adapted native Arabic words into Latin script, they adapted primarily spoken forms of

Arabic, not written ones like MSA. Khalil (2012:18) also notes that users are "bringing the traditionally spoken form of the language into the written realm." This means that the most common form of Romanized Arabic is based on spoken colloquial dialects, not standard written ones like MSA. For example, the word "I was" (كنت), was more often written as <kent> in Arabizi, which mirrors its pronunciation in colloquial Arabic ([kənt]) rather than as <kuntu>, which is how it would be pronounced in MSA ([kuntu]). Since Romanization has been shown to be primarily a process of transcription, Arabizi in Lebanon will more closely mirror the spoken dialect of Lebanese Colloquial Arabic (LCA). Therefore, an understanding of Levantine Arabic phonology is useful.

5.1 General Arabic Phonology and Orthography

Unlike Latin script, Arabic is written right-to-left, and characters are written differently based on their position in a word (initial, medial, final, or isolated). In general, Arabic orthography is phonetic: One letter represents one sound, and silent letters (like the "k" in English "knight,") and digraphs (where two letters represent one sound, such as English "gh" or "sh,") do not occur. There are more consonants and fewer vowels in Arabic than in English, and Arabic also includes some sounds that do not occur in English, such as emphatic consonants, which are produced with secondary (usually velar) articulations, and gutturals (velar, uvular, and pharyngeal consonants). The Arabic alphabet consists of 28 characters, three of which are used to indicate long vowels (ا, و, and ي). Formal writing may also include diacritical marks to indicate short vowels (ـُ, ـِ, ـَ). The characters for the long vowels [aː] and [uː] can also occur as the consonants [y] and [w] if they take short vowel diacritics. Arabic uses a consonantal alphabet, which means that these short vowels are optional and can be placed on top of or below

letters as diacritical markings. Usually these markings are included only in more formal contexts, such as Quranic Arabic; without them, meaning is inferred based on context. For example, the word *ktb* (كتب) means "books" when written as *kutub* كُتُب and "he wrote" when written as *kataba* كَتَبَ.

5.2 Levantine Arabic Phonology

Levantine Colloquial Arabic (LCA) refers to the dialect of Arabic spoken in the Levant, which typically includes Syria, Lebanon, Palestine, and Jordan. As Gordon (2011:13) shows, Levantine Arabic tends to differ from Modern Standard Arabic (MSA) in two main ways: 1) new features, including distinct morphological and grammatical rules and lexical items unique to the dialect, and 2) phonological modifications to MSA. The first category includes differing verb conjugations, word order, and pluralizing processes for nouns that are distinct from MSA. The second involves pronunciation shifts, vowel deletion or changes, and changes in the pronunciation of certain grammatical features such as prepositions and subject prefixes. Since this thesis focuses on orthographic variation, the second category, phonological differences between MSA and LCA, will be most closely examined.

While Levantine has certain unique features, a lot of its phonemic variation is just MSA pronounced in a different way. For example, it often replaces interdental fricatives with dental stops (such as using [kitir] for the MSA [kiθir] (*a lot*)), pronounces *qaf* as a glottal stop (such as [məʔbul] instead of MSA [məqbul] (*accepted*)) and reduces the affricate to a fricative (such as [riʒaːl] (men) instead of MSA [ridʒaːl] (Barakat, 2009). However, these changes are not always regular; they may depend on the particular word in which they appear, or on the phonetic environment. Additionally, as Figure 5 demonstrates, not every Arabic letter in MSA is

represented only one way in dialect; phonemes in MSA may be pronounced as different sounds in different environments in the dialect.

Figure 5: **Potential Phonetic Changes from MSA to Levantine Dialect**
Adapted from Barakat (2009).

Arabic character	MSA Pronunciation	Levantine Pronunciation
ث	[θ]	[t], [s]
ذ	[ð]	[d], [z]
ظ	[ðˤ]	[dˤ], [z]
ق	[q]	[ʔ]
ء	[ʔ]	(omitted)
ج	[dʒ]	[ʒ]

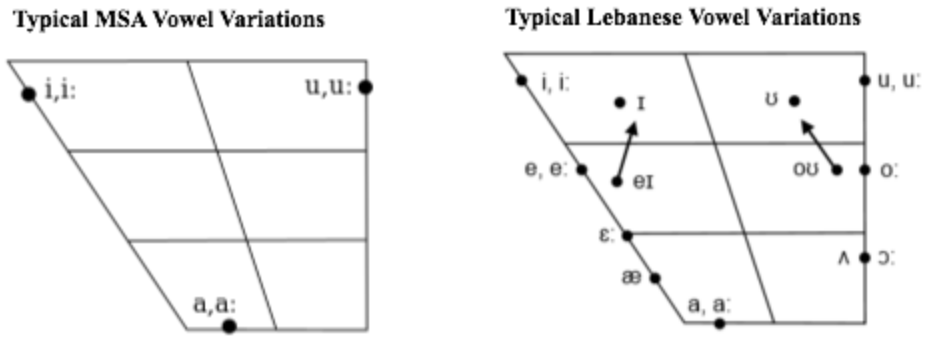
As Bassiouney (2009) notes, some non-phonological changes also occur in LCA, as follows:

- Use of the b- prefix (ب) as an aspect marker for present tense before main verbs
- Lack of case endings or mood markers in morphology
- Specific lexical items, especially for question words, such as [ʃu] for *what*, [leːʃ] for *why*, and [beddi] for *I want*.
- *Imaala*: An open vowel represented by *ta marbuta* (ة) is raised to [e] at the end of a word.
- The use of [ma] (ما) or [mu] (مو) for negation

Vowel variation

Significant variation occurs among vowels in Levantine Arabic, even between individual speakers. In Lebanon, variation also occurs depending on if the speaker is in a major city, like Beirut, or in the northern or southern part of the country. Vowel position within a word can also affect pronunciation; the emphatic consonants (ص, ض, ط, ظ) or ق will often deepen the surrounding vowel sounds. However, a general trend is apparent in that the majority of Lebanese Arabic speakers realize the diphthongs /aj/ and /aw/ as [eɪ] and [oʊ]. In urban dialects (such as in Beirut), [eː] has replaced /aj/ and sometimes medial /aː/, and [e] (similar to the French é) has replaced final /i/. Also, [oː] has replaced /aw/, and [o] has replaced some short /u/s. Vowel epenthesis, or insertion, is also common in Lebanese Arabic to break up consonant clusters—for example, the word *name* will be pronounced [ʔɪsm] in MSA but [ʔɪsɪm] in LCA (Lebanese Arabic Institute, 2017). MSA and LCA also differ in that Lebanese Arabic has vowel pronunciations like [e] and [oː] that do not appear in MSA (Lebanese Arabic Institute 2017), which increases the possible ways vowels can be written in colloquial Arabic orthographies like Arabizi. Since this thesis will analyze data from Beirut, typical central Lebanese vowel variants are shown, although they do not cover all possible pronunciations (see Figure 6):

Figure 6: MSA vs. Lebanese Vowel Variants
Adapted from the Lebanese Arabic Institute (2017)



The wide range of phonological variation evident in Levantine Arabic helps to explain why there is so much variation in Arabizi. Since there are more vowel sounds in Levantine Arabic than MSA and no clear one-to-one correspondence between sounds in MSA and LCA, there are many options to choose from when writing in LCA.

5.3 Orthographic Variation

One of the main concerns about Arabizi involves its orthographic variation; that is, how it is written differently depending on location, second language knowledge, or dialectical differences. Since not every Arabic character has a Latin script equivalent, trying to write it in Latin script is necessarily problematic: Does one pick the closest Latin script equivalent visually, or try to approximate its pronunciation? In other words, is Arabizi written based on what people say/hear, or what they read/see? For example, the word قمر (moon) [qamar] is sometimes written as <qamar>, which most closely matches its visual representation (writing <q>for ق), but it is also sometimes written as <amr> or <2mr>, which aligns more with its actual pronunciation in colloquial Arabic (the ق is often dropped or pronounced as a glottal stop, which is then omitted or represented as <2>, respectively). The problem with using Arabizi instead of Arabic script comes because this word (moon) could also possibly be written the same way as the word أمر [ʔamar] (order), which is also written as <amr> or <2mr> in Arabizi. One indicator of this confusion often appears when Western news outlets try to transcribe Arabic names — for example, the Libyan head of state Muammar al-Qaddafi's last name was variously written as <Qadafi>, <Gaddafi>, <Kadafi>, or <al-Gathafi>, among others when he appeared in news articles about the Arab Spring. As Aboelezz (2010:102) points out, there are three main reasons for this discrepancy: the lack of one-to-one correspondence between Arabic and Latin characters,

the necessity of writing short vowels that are absent in Arabic, and whether a text is intended for native or non-native Arabic speakers. Orthographic variation is also complicated by the fact that Latin-script languages have different pronunciations for the same character; for example, <j> could represent /ʒ/ in French, /x/ in Spanish, or /dʒ/ in English.

Many different factors can lead to orthographic variation, such as pronunciation differences, writer preference, gender, and age (Grenoble and Whaley, 2006:141). Orthographic variation can also provide a measure of a language's standardization (how consistently conventional forms are used throughout a certain community). Ideally, each character within a standard orthography would always represent the same sound, and every sound would be written the same way, but this never actually happens for many languages (for example, <ough> in English has many different pronunciations), and Arabic is no exception. However, having a generally standard orthography "can increase the functional domains of a language's use, which in turn increases its status within the community" (Grenoble and Whaley 2006:140). That is, languages with more standardized orthography tend to enjoy greater prestige than those that are not as standardized. Considering the strong ideological divides between formal Arabic (which has more prestige) and Arabizi (which has less prestige), looking at how much Arabizi is standardized could indicate if its prestige has changed within a community. Arabizi provides an especially useful study in orthographic variation because much of its writing is not yet standardized, which means that it can act as a sort of barometer of orthographic development over time.

Orthography choice is also linked with many social, psycholinguistic, and political factors that interact with a person's social and educational status and language ideology. From a

social and political perspective, the choice between Arabic and Latin script can be influenced by a number of factors, including formality, religion, age, and education. For instance, a study by Zoabi (2012) of Arabizi use on Facebook in Israel and the Arab world found that Arabic script was more commonly used for formal situations, such as writing status updates, while Latin script was used for more informal situations, such as replying to comments or posting on someone's wall. The use of Arabic script has also been closely associated with the spread of Islam, especially when it was used to represent languages unrelated to Arabic (such as Turkish and Persian) in Muslim societies. As Unger (2001; cited in Al-Khalil and Palfreyman, 2003) states, any decision about orthography constitutes a social and political statement, which is often related to technology access and power. This can be seen throughout history, especially when it comes to confrontations between Eastern and Western writing systems. For example, Turkey converted from Arabic to Latin script in the 1920s as part of its modernization efforts, and in the 1980s, the Syrian Ba'ath party destroyed shop signs written in Latin script to promote Arabization (Al-Khalil and Palfreyman, 2003). A different version of this struggle is seen with the current adoption of Arabizi. On one hand, it represents access to the West and globalization, but on the other, many fear that it will decrease Arab pride and participation in Arab language and culture. This point is demonstrated by Zoabi (2012), who found that Arabic script was used more often than Arabizi in countries where "Islam plays a significant role in defining national identity," such as Saudi Arabia, Iraq, Syria, and Oman, while Latin script was more common in Tunisia, Algeria, Morocco and Lebanon, which were all French colonies in the 20th century. In Lebanon, national identity is based on identifying with non-Arab Phoenician origins, so using Arabic script would work against this idea. Additionally, sociolinguistic factors such as age and education

have also been linked with Arabizi; younger, more educated social media users tend to write in Latin script, while people over 30 with less formal education tend to use Arabic script (Zoabi, 2012). Therefore, script choice is strongly associated with Arab nationalism, prestige, and colonization.

From a psycholinguistic perspective, there are additional benefits to using Latin script. Arabic orthography is visually complex, due to 1) the similar appearance of many Arabic letters, 2) the fact that they are written differently based on their position in a word, and 3) the limited representation of vowels, which means that the Arabic script has a higher perceptual load and is more difficult to mentally process, even when compared to other Semitic languages like Hebrew (Ibrahim, Eviatar, and Aharon-Peretz, 2002; Ibrahim, Khateb, and Taha, 2013; cited in Elhij'a, 2014:192). Thus, many Arabizi users may favor it over Arabic script because it is easier to read.

On a more practical level, the orthographic variation in Arabizi is also subject to linguistic constraints. As Ornan (2003:186) explains when talking about the Latinization of Hebrew, "the main difficulty in adapting an alphabet that is not the language's original one is that the target letters do not always supply or are not always suited to the store of sounds and phonemes of the converted language." Since Arabic includes some sounds that do not have regular correspondences in the Latin script, speakers of Arabic dialects writing their language in Latin script must adapt symbols that are not actual alphabetic letters, especially numbers, to represent them. As Palfreyman (2001) shows, the following factors are involved when choosing how to represent non-native sounds in Arabizi:

- Visual similarity (such as using 3 for ع or 9 for ق)

- Phonological Similarity (using @ for feminine plurals, which end in [æt], or 8 for the first person past tense of some verbs in colloquial Arabic, such as *hak8 (I spoke)* (although this is not seen as often)
- Ease of perception: some letters that look similar in Arabic, such as ح and خ , are either written with apostrophes or as digraphs to differentiate them. (as 7 and 7' or h and kh).
- Ease of production: Characters that would be hard to type (such as diacritical markings) are usually omitted, but sometimes doubled letters or upper case are used to signal the distinction between emphatic and non-emphatic consonants, for example.
- Orthography of other Roman-alphabet languages familiar to the writer: As shown previously, many French-Arabic bilinguals will use <ou> for /u/ or <ch> for /ʃ/.

Al-Khalil and Palfreyman (2003) point out that in general, Romanization tends to produce "competing alternate representations" of a given language, and therefore does not always solve the problem of making a language easier to write (or read) within a given context. In this way, Arabizi does not always offer a solution to the problems of the Arabic writing system, but instead tends to complicate them because it is not consistent among its users.

VI. Computer-Mediated Communication

According to the annual Arab Social Media report, the number of Internet users in Arabic-speaking regions is growing at a faster-than-average rate, with over 71 million active users on social media every day (2014). These users' communications fit into the category of computer-mediated communication, or CMC, which involves messages transmitted in technology-based contexts, especially on social media. This concept is beneficial in understanding how Arabizi is used on Twitter, since tweets are a mode of CMC—the site allows

users to interact with each other by exchanging a series of 140-character tweets. As prior research on CMC shows, it has several characteristics that are conducive to writing Arabizi, such as a more relaxed environment, space limitations, and synchronicity (users' ability to respond to each other in real time). For example, as Herring (2007:14) notes, Twitter is a mixed CMC mode because it can be used in both synchronous and asynchronous ways. Unlike text messaging, which is almost entirely synchronous, or email, which is asynchronous, Twitter users can choose to rapidly exchange tweets in a conversation, or merely tweet out their thoughts on a profile page without expecting a response. Other social media platforms such as Facebook also have similar divides— users can respond to each others' comments, or just post thoughts on a personal wall. This means that Twitter data can show both features of synchronous communication (such as abbreviations, informality, and less complexity) and asynchronous communication (such as more complexity and longer clauses). Twitter's 140-character limit may also contribute to orthographic informality, which is commonly seen in Arabizi, including abbreviations, non-standard spellings, and deletions. Additionally, according to Sayahi (2014:87), the more "lax and spontaneous" CMC environment provides a greater incentive for users to write in colloquial Arabic instead of MSA, While studies of Arabizi in CMC are becoming common (Aboelezz, 2012; Al-Khalil and Palfreyman, 2003; Mostari, 2009) few studies include CMC data from the Levant, or from Twitter. Those that do include Levantine Arabizi (Elhij'a, 2012; Zoabi, 2012) often focus on Facebook instead, or do not analyze orthographic variation. Therefore, this thesis attempts to reduce this gap by closely analyzing common orthographic variants in Romanized Lebanese Arabic on Twitter.

Although the emergence of "text-speak" over the past few decades has led to concerns that CMC is decreasing the standards of language used in technological contexts, several researchers have argued that CMC actually allows for more linguistic variety and creativity. Crystal (2006) suggests that "it seems likely that the Internet will speed up the process of language change," due to the rapid transmission of information online. Similarly, Mason and Allen (2003) argue that the Internet has introduced new growth dynamics to what have historically been oral languages, pointing out that Creole languages are undergoing a process of orthographic standardization because of their increased use online. This phenomenon can also be seen with Arabizi, where traditionally spoken language varieties are being written in similar ways because they are used in online spaces.

Consequently, CMC is often thought of as "a hybrid between speaking and writing" (Dorleijn and Nortier, 2008:128), because it uses written language but mirrors the way people talk, such as in text messaging. This is especially pertinent to studies of Arabizi, because Arabizi is an attempt to write down traditionally spoken varieties of Arabic. Considering the limitations imposed by CMC is also important when it comes to Arabizi because it has inverted the traditional boundary between formal and informal speech. Typically, written speech is considered more formal and spoken speech informal, but with the rise of text messaging and social media sites like Twitter, this paradigm is flipped. In CMC, users use the written form typically associated with formality to communicate messages that are typically informal. As El Essawi (2007:6) notes, CMC involves a "hybridized text that includes the relaxed informal style of talking combined with formal features of written texts." This can be seen in the way Arabizi

tends to mirror colloquial Arabic pronunciation, while also including features common to writing, such as condensed phrases.

However, despite its popularity, Arabizi still has several features that impede its standardization, such as a lack of consistency in writing the definite article and a wide range of vowel representations. As Yvon (2009:133) notes, forms of CMC such as Arabizi are "characterized by massive and systematic deviations from the orthographic norm." These deviations include phonetic transcription (in which *through* becomes *thru*), vowel deletion (in which *homework* becomes *hmwrk*), and the substitution of characters and numerals for their phonetic value (in which *great* becomes *gr8*). Some of these changes, such as transcription and vowel deletion, are commonly used in Arabizi to convey messages concisely, but are also a function of the technological constraints imposed by CMC—keyboard layouts and limits on message length make it easier and more cost-effective to use these strategies.

CMC also plays a role in language ideology by raising concerns about the dominance of English on the Internet. As mentioned in the section on Arabic language ideology, English is considered the lingua franca of the Internet, and is therefore seen as a threat to the Arabic language. Previous studies of CMC have found a preference for English over native languages in South Asia and Korea (Paolillo, 1996 and Yoon, 1996, respectively; cited in Herring, 1997), although Paolillo suggests that native languages "may fare better when computer networks are located entirely within the nation or region where the language is natively spoken, when fonts are readily available which include all of the characters of the language's writing system, and when there has been no colonial legacy of English within the home culture." The presence of Arabizi in Lebanon may be partially explained by these last two factors. Arabic script is now readily

available on computers, but still often cumbersome to use compared to English, and Lebanon's history of French colonialism promotes a trend toward French and English over Arabic in CMC. Consequently, Arabizi represents an attempt to bridge the divide between English and Arabic and mitigate the constraints posed by using Arabic script in CMC. Overall, exploring the characteristics of CMC is important because they provide an impetus and linguistic constraints for how Arabizi is written.

Part 2: Data Collection and Analysis

I. Research Questions

This thesis is designed as an exploratory study to examine the frequency and orthographic variation of Arabizi in Beirut on Twitter. It will explore three main questions:

1) How often is Arabizi used compared to other languages in Lebanon on Twitter?

Hypothesis: It is expected that Arabizi will be one of the most common languages used on Twitter, comprising at least 25% of the overall tweets. This is a conservative estimate drawn from previous studies of Arabizi rates in CMC (see Aboelezz, 2009; Haggan, 2007; Mostari, 2009; Palfreyman, 2001), where Arabizi use ranged from 20% to 98% of the time compared to other languages. The frequency of language use from greatest to least is expected to be: 1) Lebanese Colloquial Arabic (LCA), 2) Arabizi, 3) Any combination of LCA, English, or Arabizi, 4) English, 5) Modern Standard Arabic, and 6) French. This hypothesis is based on findings from Warschauer et al. (2002:16), which concluded that "Modern Standard Arabic was rarely used on the Internet," and from Aboelezz (2009) from a survey of Egyptian email groups, which found that the two most commonly used languages were Arabizi and English or just English,

followed by Arabic script (MSA) only. Additionally, in Mostari's review of Algerian texting, 59 out of 60 respondents chose to use Latin script in their texts, even though their phones allowed them to enter Arabic script (2009:380). Palfreyman and Al-Khalil note in their survey of Gulf Arabic in instant messaging that subjects who wrote in Arabic were split evenly between Arabic and Latin script (2007). As far as code-switching goes, in a corpus of Egyptian SMS messages (Bies et. al 2014:102), 66% were entirely in Arabizi, 19% Arabizi and English, and 15% Arabic script. Because of this, code-switching between English and Arabizi is also expected to occur frequently. The rate of French as compared to other languages is uncertain.

2) What variation occurs in Lebanese Arabizi orthography?

Hypothesis: It is predicted that orthographic variation in Arabizi will occur most often with vowels and emphatic consonants that do not appear in English, such as ظ، ص، ض، and ط, as well as consonants that have a wide range of available pronunciations, such as ق and ة. Conversely, consonants that have one-to-one correspondences to English letters (such as ب and ت) are expected to show the least variation. Emphatic consonants will probably be represented by capitalization or apostrophe use, and the use of numbers will probably be favored over digraphs (<5> instead of <kh>). For vowels, <a> is hypothesized as the most common variant for ا, <o> or <u> for و, and <i> for ي, while short vowels will not usually be written. A proposed list of likely variants is below:

Figure 7: Most Common Expected Orthographic Variants in Lebanese Arabizi

Arabizi	Arabic Script
a	ا
5	خ
d/z	ذ

d	د
sh/ch	ش
9	ص
9’	ض
6	ط
6’	ظ
8/3’	غ
2	ق
o/u	و (when used as a vowel)
i	ي (when used as a vowel)

This hypothesis is based on previous studies of Arabizi in SMS messages and on the Internet (Aboelezz, 2009; Haggan, 2007; Mostari, 2009; Palfreyman and Al-Khalil, 2007), as well as those that show a preference for numbers over digraphs in Levantine Facebook posts (Elhij’a, 2012) and significant variation in the representation of ق [q] in Arabizi (Bassam, 2014). Additionally, El-Khaissi (2015) has shown that Romanized Spoken Arabic follows a transcription process as opposed to a transliteration process, which suggests that the orthography for Arabizi will mirror spoken Lebanese Arabic and include traits such as the dropped ق and raising of an open vowel at the end of a word.

3) How could this orthographic variation affect Arabic language ideology?

Hypothesis: Since orthography is closely tied to language ideology, it stands to reason that the level of orthographic variation in Arabizi will affect Arabic language ideology. The degree of orthographic variation in a language is linked to its use and stabilization; as Grenoble and

Whaley (2006:140) show, languages with more stable orthographies tend to enjoy more prestige and use in a linguistic community. Currently, Arabizi (along with other varieties of CMC-language) has a relatively low prestige among users; it "[appears] to be perceived as modern, but also as somewhat sloppy and perhaps as a threat to the [Arabic] language" (Al-Khalil and Palfreyman, 2003). However, if Arabizi's orthography becomes more stabilized, it may in turn gain more prestige and acceptance in the Arab world. If, on the other hand, the data shows a lot of orthographic variation, it may indicate that Arabizi still has a low status in the Arab world.

II. Methodology

The tweets for this study were collected using a program called Twitter Archiver, which allows users to query the Twitter API using Twitter's advanced search parameters and returns data based on user input. The program was run for six weeks (during October and November 2016), and tweets were collected from users located in Beirut, Lebanon according to Twitter's location data by inputting the parameters "collect all data from geographic coordinates 33.8938° N, 35.5018° E, with a radius of 15 miles." Beirut was chosen because it is the capital city of Lebanon, and therefore was assumed to provide a good sample of Lebanese Arabic tweets. The minimum number of retweets was also set at five to minimize collection of spam (tweets with no analyzable linguistic content).

After deleting the remaining spam tweets, each tweet was coded according to language type (English, French, Modern Standard Arabic (MSA), Lebanese Colloquial Arabic (LCA) in Arabic script, Arabizi, or a mix of the above). Hashtags were not counted as part of the tweet text. Whether to code a tweet as LCA or MSA was not always clear, since the two share a lot of

vocabulary. However, if a tweet contained dialect markers associated with colloquial Arabic (such as شو instead of ماذا or the absence of case markings), it was coded as LCA, while if a tweet showed known features of MSA (such as the use of هل for questions, ليس, or case markings), it was coded as MSA. Tweets that showed no clear markers of either MSA or LCA and were therefore uncertain were shown to a native Arabic speaker for review. Data was then analyzed to find the overall rate of Arabizi use.

For orthographic variation, tweets were tokenized and examined for instances of suspected variation, such as those listed above (see Figure 5). Each Arabizi tweet was written out in Arabic script for comparison, and then analyzed to see which Arabizi characters matched up with which characters in Arabic script. The most common orthography for each variant was then determined and analyzed.

*As an outsider to this linguistic community (a non-native but proficient user of Arabic and Arabizi in the United States), I tried to be conscious of possible limitations or hidden biases I may have had while designing this methodology. I realize that outsiders to a linguistic community can sometimes make judgments based on their perceptions that may not always be correct. For the purpose of accuracy, I also had a native speaker of Arabic double check all the meanings of Arabic and Arabizi tweets, and provide input into the methodology design.

3. Results

Frequency of Arabizi Use

Overall, 16,473 tweets were collected and analyzed to discover how often each language was used. The results showed that Arabizi was actually the second-least used language type, with only 331 tweets. The order of frequency of each language type was: 1) LCA, 2) MSA, 3) English, 4) French, 5) Arabizi, and 6) A mix of multiple languages.

Tweet Analysis

Language Type	**Number**
Arabic script (LCA)	5987
Arabic script (MSA)	5323
English	3399
French	1106
Arabizi	331
Mixed	299
Total:	**16,473**

LCA made up 36% of the data, followed by MSA with 32%, and English at about 21%. Tweets written in just Arabizi made up only 2% of the data. This contrasts sharply with results from previous studies, such as Bies, et al. (2011) and Aboelezz (2009), both of which found that Arabizi and English were the majority languages in Egyptian SMS and email messages, respectively. While this study's data does provide support for English as one of the main

languages used in CMC contexts in Lebanon, it differs from these studies in its higher rate of MSA.

One major difference between these studies and this thesis' current data from Twitter is the medium; almost all of the MSA tweets on Twitter were from news sources, such as *Al-Mayadeen* or *Al-Manar News*, which would not be found in text messaging or email. Additionally, Egyptian Arabic only provides two languages to focus on (English and Arabic), whereas the Lebanese data also involved French, which provided a greater potential linguistic repertoire for participants. Therefore, if the tweets from news sources were removed, LCA and English would make up the majority of the corpus. However, this still does not explain the relative lack of tweets written in Arabizi. As previous studies show, Arabizi is commonly used on Twitter and other social media sites like Facebook, and in cases where multiple languages are used it makes up a majority of the text (Elhij'a, 2012; Aboelezz, 2009), so it was expected that it would make up more of this corpus. One could argue that some of the messages that could have been written in Arabizi were instead written in French, but given the low incidence of French in the dataset for this thesis (7%) this hypothesis does not seem likely. However, one other study of Arabizi in Lebanese chat rooms (El-Khaissi, 2015) also found a relatively low incidence of French (9.4%) in an Arabic, English, and French corpus, suggesting that the low rate of French in CMC may not be unusual. Some researchers have asserted that Arabizi use is decreasing or is not as popular as previously assumed (Al-Munziri, 2014:222; Kenali et al., 2016:933) due to the increase in multilingual keyboards on cell phones and computers, and a few 2016 Twitter users briefly popularized the hashtag #arabiziisdead, but the majority of studies on Arabizi conclude that it is still flourishing. Overall, the data in this thesis seems to indicate a preference for using

Arabic script (as opposed to Latin script) to write Arabic, since both LCA in Arabic script and MSA in Arabic script occurred more often than Arabizi.

Another interesting thing to note about the data for this thesis is that a significant portion of the tweets written with Arabizi also involved code-mixing, or the use of more than one language (see Figure 6). There were a total of 525 tweets with Arabizi; 331 (63%) were Arabizi-only, and 194 (37%) were a mix of Arabizi and another language. The most common languages for code-mixing were Arabizi and English (which represented 184/525 = 35% of Arabizi tweets, 184/194 = 95% of the mixed-Arabizi tweets, and 184/299 = 62% of the total of mixed-language tweets), followed by English and LCA (which represented 93/525 = 18% of Arabizi tweets, 93/194 = 48% of the mixed-Arabizi tweets, and 93/299 = 31% of the total of mixed-language tweets). This aligns with other studies (Aboelezz, 2009; Attwa, 2012; Warschauer et al., 2002) which show that Arabizi is associated with a high percentage of code-mixing, especially with English. Arabizi's common association with English is demonstrated by the fact that it code-mixed almost exclusively with English, and the lack of any tweets with code-mixing between MSA and English. It is interesting to note that many of the tweets that used both English and Arabizi also employed abbreviations commonly seen in English SMS messaging, such as *lol* and *idk*, which provides further evidence of English's close relationship with Arabizi and Arabizi users' familiarity with English CMC conventions. The incidence of tweets that mixed Arabizi and LCA provides evidence that choice of orthography matters; although all of these tweets were written in Arabic, some words, such as *habibi* (a term of endearment used for close friends or significant others) were usually written in Arabizi, not Arabic script, which attests to Arabizi's use as a marker of social identity.

Figure 8: Breakdown of Tweets Involving Multiple Languages

Language Type	Number of Tweets
Arabizi and English	184
English and LCA	93
English and French	9
Arabizi and French	4
Arabizi and LCA	3
French and LCA	3
English, Arabizi, and French	1
English, Arabizi, and LCA	2
Total	299

Orthographic Variation

In total, the 525 tweets that involved Arabizi were analyzed to determine the extent of orthographic variation in Lebanese Arabizi and which spellings were most common for each Arabic character. The most standardized characters (with little variation) were ع, ح, and characters that had direct correspondences with English sounds (م, ل ,ك, ف, س, ز, ر, ت, ب and ن) while the most variable were vowels, emphatic consonants, and ق. Overall, users writing in Arabizi seemed to favor a transcription process (writing based on attempts to mirror pronunciation) rather than transliteration (graphic transposition) for most characters. A table of the results is shown below:

Figure 9: Orthographic Variants For Each Character in Lebanese Arabizi

Arabic Character	Most Common Lebanese Arabizi Variant	Alternatives (Percentage of time occurred)
أ	a	e (11%) , i, ay, ei (each <1%)

ب	b	
ت	t	
ث	t/s (depending on the word)	
ج	j	
ح	7	
خ	kh	5 (21%)
د	d	
ذ	d/z (depending on the word)	
ر	r	
ز	z	
س	s	
ش	sh	ch (14%)
ص	s	S (7%)
ض	d	
ط	t	
ظ	z	
ع	3	
غ	gh	8 (5%)
ف	f	
ق	2 (usually pronounced as a glottal stop)	dropped (9%)
ك	k	
ل	l	
م	m	
ن	n	

و	w (as a consonant), ou (as a vowel)	u (16%), o (15%), aw (8%), oo (2%), ow (<1%) (as a vowel)
ه	h	
ي	y (as a consonant), i (as a vowel)	e (31%), ay (11%), ee (2%), ei (2%), ea (<1%), iy (<1%) (as a vowel)
ء	2	
(ـَ)	a	e (31%)
(ـُ)	o	e (28%), u (6%), i(3%)
(ـِ)	e	i (40%), a (6%)

Additionally, the following patterns were seen in the data:

Conventionalized Consonants

The following 13 consonants were represented the same way 100% of the time, which seems to indicate that they have conventionalized in Lebanese Arabizi: د ,م ,ل ,ك ,ف ,س ,ز ,ر ,ت ,ب ,ع ,ح and ن. These results matched the expected hypothesis, since all except ع and ح have direct equivalents in English. However, ع was always represented by <3>, and ح was always represented by <7>.

<Sh> vs. <ch> (ش)

The use of <sh> to write ش occurred much more frequently than <ch>; <sh> was used 138 times, or 87% of the time, while <ch> was used only 22 times, or 14% of the time. This contrasts with results from a previous study of ش in Facebook posts—Elhij'a (2012) found that <ch> was used slightly more than <sh> in Lebanon, with ratios of 57 and 43%, respectively. However, Elhij'a's

study had a smaller sample size (only 107 total instances of ش were analyzed), and considering the relatively low overall incidence of French in my data set (7%), it makes sense that because not as many tweets were written in French, there is also a lower incidence of French-influenced Arabizi, as represented by <ch>. As previously mentioned, Abdel-Ghaffar et al. (2011) explain that this difference between <sh> and <ch> may be due to L2 influence, because speakers whose second language is French are more likely to use <ch>, while English L2 speakers are more likely to use <sh>. Interestingly, Elhij'a (2012:94) also found that this distinction could also be a marker of religion; Christians, who are more likely to attend French schools, use <ch> more often, while Muslims, who usually attend English-speaking schools, are more likely to use <sh>. The difference between <sh> and <ch> also seems to indicate that Arabizi was written based on attempts to match pronunciation, since <ch> would be the more likely French pronunciation and <sh> the English one. Elhij'a also found that some Jordanian students were starting to use <$> to avoid the digraph <sh>; however, this phenomenon did not occur in my data, and there is no evidence that it has spread to Lebanon. Therefore, this thesis' data likely captured a more Muslim, English-Arabic speaking demographic of Twitter users in Lebanon.

***Jeem*: (ج)**

Previous studies of Romanized Arabic have noted that ج is still not fully stabilized; it can be represented as <g> or <j>, although <g> is generally more popular in Egypt and sometimes Israel/Palestine due to dialectal differences. It has two pronunciations in the Levant: [ʒ] and [dʒ], but is pronounced as [ʒ] in Lebanon. In my data, ج appears to have completely conventionalized as <j>, because <g> did not appear at all, and <j> appeared 100% of the time. This concurs with Elhij'a's (2012) and El-Khaissi's (2015) data, which also found a 100% occurrence of <j> for [ʒ]

in Lebanese Arabizi. The use of <j> instead of <g> to represent ج also presumably favors a transcription process, because [ʒ] would be the expected French pronunciation of <j>.

Emphatic Consonants (ظ, ص, ض, ط)

One of the most interesting results was in how the emphatic consonants ظ [ðˤ], ص [sˤ], ض [dˤ], and ط [tˤ] were realized in Lebanese Arabizi. Each of these consonants also has a non-emphatic counterpart (ذ [ð], س [s], د [d], and ت [t]), but in the data, both emphatic and non-emphatic consonants were generally written the same way (for instance, [tˤ] and [t] were both usually written as <t>). This lack of differentiation at first seems confusing (how is one supposed to tell the difference between words that differ only in their use of emphatic/non-emphatic consonants?) but most differences are readily apparent through context, and usually vowels following an emphatic consonant would be doubled to show the different pronunciation, such as writing <darab> for درب [darab] (*path*) but <daarab> [dˤarab] for ضرب (*hitting*). This representation of emphasis on the vowels [aa] could reflect cue-trading, or speakers' perceptions that vowels, not consonants, carry pharyngealization. Occasionally capitalization was used to mark emphatic consonants, but this occurred rarely: only 8% of the time for [sˤ] and never for [ðˤ], [dˤ], and [tˤ]. Both Elhij'a (2012:99) and El-Khaissi (2015) note that this lack of capitalization may be due to the Lebanese tendency not to clearly differentiate between emphatic and non-emphatic consonants when speaking; El-Khaissi says:

> While the pronunciation of pharyngealised phonemes may be maintained in many other dialects of Arabic, this is certainly not the case for the Levant dialect...Romanised Arabic in the chat room communication does not distinguish Arabic pharyngealised consonants (also known as emphatic consonants in the traditional literature) from their non-pharyngealised counterparts (69).

Similarly, Elhij'a notes that "Lebanese youngsters as well as Egyptians, rarely express pharyngeals." Thus, the lack of differentiation may be reflective of pronunciation differences between dialects. It also suggests that in this case, users preferred to write Arabizi based on a character's pronunciation, rather than differentiating between characters visually by capitalizing them. Additionally, some countries, such as Jordan, use numbers to indicate emphatic consonants (such as <6> for [tˤ]), but this did not occur in the Lebanese data.

***Qaf* (ق)**

The *qaf* (ق) is one of the most complex characters in Arabizi because it can be represented so many different ways. When pronounced as [q] (as in MSA and Jordanian dialects), it is often written as <q> or <9> (due to visual similarity), while in other dialects it is pronounced as a glottal stop (ء) and represented as <2>. In regions where it is pronounced as <g>, it is written as <8>. Often it is not pronounced at all, and so is not written. Whether it is pronounced or not can also depend on the specific word in which it is used. For example, high-register "educated" words borrowed from MSA into colloquial Arabic such as مقاومة [muqawəmə] (*resistance*) retain pronunciation of the *qaf* as a voiceless uvular stop. In Lebanon, it has various pronunciations, but is usually pronounced as a pharyngealized voiceless velar stop [kˤ] or [ʔ]. In the data set for this thesis, <2> and no pronunciation were the most common choices for representing *qaf*, with 121 and 12 instances, respectively. <9> and <q> were not used at all. For instance, the word قلبي (my heart/sweetheart) was variously written as <albi> or <2albi>, but not as <qalbi> or <9albi>. A study by Gordon (2011:30) also found similar results, noting, "the letter *qaf*, represented as *9* in Romanized orthography, hardly ever appears in Levantine [CMC]." This seems to indicate that

the Arabizi users in my data preferred phonetic transcription over graphic transposition for *qaf*, since they wrote it how it would be pronounced in Lebanese dialect, not how it looks on screen.

***Hamza* (Glottal Stop) (ء)**

As noted above, Lebanese pronunciation favors using the glottal stop, which is written as <2>, in place of *qaf*. Generally, <2> was almost always written for *hamza* in word-medial position and word-final position, but was rarely written in word-initial position. For example, as shown previously, قلبي (my heart/sweetheart) was often written without the initial [q] as <albi>, while words pronounced with a word-medial glottal stop, such as رئيس (*president*) [raʔiːs] were always written with a 2, as <ra2is>. This could possibly be because English orthography does not allow for a glottal stop in the middle of a word, so the use of a number was necessary to comply with orthographic rules. In this case, the use of <2> for *hamza* both mirrored typical Lebanese pronunciation and took into consideration the orthographic conventions of the script's language.

***Ghayn* (غ)**

Ghayn (غ) is similar to *jeem* in that it appears to be highly conventionalized in Lebanon but not in other countries. It is usually represented as <gh> or <8>. <3'> or <g'> have also been proposed as an alternative in some countries, but as noted in the section on apostrophes, this practice no longer seems to be popular. In the data for this thesis, the digraph <gh> showed a clear preference, because it was used 95% of the time compared to <8>, which was only used 5% of the time. For example, words like <ghayer> (*change*) or <aghniyye> (*song*) were seen more often than words like <a8la> (*most expensive)*. Elhij'a (2012:81) also found a similar pattern, with <gh> being the most popular form in Lebanon, followed by <8> and <g>. Therefore, *ghayn* seems to be one of the more stable consonants in Lebanese Arabizi.

***Khaa'* (خ)**

Khaa' (خ) can also be represented in various ways in the Levant: <kh>, <5>, or <7'>. <Kh> proved to be the most common form in the data for this thesis, occurring 79% of the time, while <5> was the next most common at 21%. For example, <khalas> (*finished*) or <akhi> (*my brother*) occurred frequently, while <5alas> was only seen twice. <7'> did not appear at all. This contrasts with Elhij'a's (2012:80) data, which found that <5> was the most common representation for *khaa'* across the Levant, followed by <kh>. She found that age was a significant factor in the writing of *khaa'* and that older writers (25 and older) tended to use <kh> more than <5>, while the reverse was true for those under 25. In the data set for this thesis, it is interesting that <kh> was the most popular variant, since it could possibly be confused as representing two different phonemes, /k/ and /h/, instead of the clearer <5>. However, this combination does not usually appear in Arabic, and background knowledge and context would presumably help users differentiate between the two. Elhij'a proposes that the prevalence of <5> over <kh> means that <5> will replace <kh> in the future, but judging from the current data, this does not seem to have happened. The use of digraphs over the more visually-based <5>, similar to *ghayn*, again seems to indicate a preference for transcription over transliteration, since <kh> more closely represents how خ is pronounced.

***Thaa'* (ث) and *dhal* (ذ)**

Thaa' [θ] is pronounced in two ways in Lebanon: [t] or [s], depending on the word. (For example, *kteer (كثير) (a lot)* is pronounced with a *t*, while *t2ssir (تأثير) (effect)* is pronounced with an *s*). Similarly, *dhal* [ð] is pronounced as either [d] or [z]. However, there are no formal rules for which words are pronounced which way, so native speakers make judgements based on their

intuition. The reason for this lexical variation is not entirely clear, but the typical explanation seems to be that at some point historically, the interdental sounds [θ] and [ð] merged with dental ones ([d] and [t]) in most urban Levantine dialects. This meant that words like [θalaθa] (ثلاثة)(*three*) were pronounced like [talata], or [kaðab] (كذاب) (*liar*) were pronounced like [kadab]. Then when these Lebanese speakers heard people using MSA (at mosques, or later, on TV), some of them started to imitate the more prestigious interdental pronunciation, which to them sounded like [s] and [z]. So some more formal or academic words use the perceived interdental pronunciations ([s] and [z]), while others kept the original dental ones ([t] and [d]), although the boundary became more blurred over time as words that were once exclusive to MSA made their way into the colloquial sphere (Lebanese Arabic Institute). Therefore, in order to determine how closely the written Lebanese Arabizi data matched up with actual pronunciation, a native Lebanese Arabic speaker was asked which words in the data set were typically pronounced with <t> and <s> or <d> and <z>, and the data was coded accordingly. Overall, the results indicated that writing generally matched pronunciation: Words usually written with <s> or <t> were pronounced as [s] or [t], 98% of the time, respectively, as were [d] and [z]. El-Khaissi (2015) and Elhij'a (2012) also mirror this preference for phonetic transcription of ث and ذ instead of transliteration, noting that their representations in Arabizi depend on pronunciation.

Number Use

Another interesting aspect to note is the variation in number use with Arabizi. In this aspect, Arabizi users seem to take a transliteration-based approach, since certain numbers, such as 3 and 7, have conventionalized as characters that have one-to-one relationships: 3 unequivocally and exclusively represents ع, while 7 represents ح. The same appears to be true of 2, which

represents the glottal stop (ء). If users had taken a transcription-based approach, presumably they would have tried to represent these characters' pronunciations, such as by writing <h> for ح. Other numbers were also used occasionally in this thesis, such as 8 for غ, 9 for ق, or 6 for ط, although they were not common. Therefore, contrary to the result expected from the hypothesis, Lebanese Arabizi seems to favor the use of digraphs over numbers (for example, <gh> instead of <8>), and tends to use mostly 7, 2, and 3.

Apostrophe Use

In previous studies on Arabizi, various authors have noted the use of apostrophes to represent the dots on top of certain characters such as ظ, ض, خ and غ (Aboelezz, 2009; Haggan, 2007; Mostari, 2009; Palfreyman and Al-Khalil, 2007). This helps differentiate between letters that look similar to each other (for example, using 7 for ح and 7' for خ). However, this practice seems to have fallen out of favor, because very few apostrophes have been used in recent Arabizi data (Gordon, 2011; Elhij'a, 2012), and no apostrophes were found in the data for this thesis either. This seems to indicate that the Arabizi users in this thesis have moved away from a graphic transposition approach based on visual similarity.

Condensed Phrases

Another hallmark that appeared fairly often in the data was the use of condensed phrases, where two or more words or phrases that would usually be separated were written as one word. Often this also involves the deletion or simplification of certain sounds to facilitate pronunciation, as well as the use of abbreviations, as in the English SMS phrase *cya* (*See you*). As Kul (2007:43) suggests, this occurs often in English language text messaging with phrases like *inorite (I know, right?)* and *idk* (*I don't know*), which implies that Arabizi users do not solely write based on

pronunciation, but also take into account written conventions of the script language. Condensed phrases and abbreviations also occurred frequently in the Arabizi data, as in the following examples:

- *missaalkheir* [mɪsaɛlxejr] (مساء الخير) (*Good evening*)
- *Yaret* [jaːreɪt] (يا ريت) (*I wonder*)
- *shitane* [ʃiːtani] (شي تاني) *(Something else)*
- *mabefham* [mabɛfhɛm] (ما بفهم) *(I don't understand)*
- *3ambye5las el nhar* (ʔambɪjɛxlasʕl nɛhar) (عم بيخلص النهار) (*I'm done for the day*)

Especially in the last two examples, it was interesting to note that grammatical forms such as negation (the use of ما in *I don't*) and present progressive (the use of عم to indicate *I am)* were condensed and joined to other words, since this does not typically occur when writing either formal or informal Arabic.

Vowels

Vowels in Arabizi had a much greater range of variation than the consonants. Overall, the three long vowels were written fairly consistently (<a> for /aː/, <ou> for /uː/ and <i> for /iː/), but the short vowels, if written, showed much more variation. There was also a general trend toward the use of <e> to represent almost all vowels (it was usually the first or second choice for each vowel), which perhaps reflects the typical collapse and deletion of short vowels into [ə] and the raising of /a/ to [e] (*imāla*) often seen in Lebanese Arabic. For instance, *kasra, alif,* and *yaa'* were all commonly represented by <e>, as seen in the word <lebnene> *(Lebanese)* [ləbnanə], where the first <e> represents a *kasra* in Lebanese pronunciation, the second <e> represents an *alif*, and the third represents *yaa*'. While consonants in Arabizi were generally based on

transcription, the vowel data was more variable, suggesting that sometimes vowel orthography is based on transcription, while at other times it is based on transliteration. Some general vowel orthography patterns are discussed below.

***Yaa'* (ي)**

The most common way to write *yaa'* as a vowel was <i> (which occurred 43% of the time), followed by <e> (which happened 34% of the time). <i> was more often used to represent [i], as in *kitir* [kɪtir] (*a lot),* while <e> appeared more often to represent the diphthong [aj] in words like *lesh* [leːʃ] *(why), hek* [heːk] *(like this),* and *le* [leː] *(why*), which fits with the pattern of monophthongization of /aj/ to /eː/ characteristic of Lebanese Arabic pronunciation. Interestingly, <i> was also sometimes used to represent [j] (that is, *yaa'* being used as a consonant and not a vowel) instead of being written as <y> as in words like *liom (today).* This suggests that while most of the time *yaa'* was written based on how it would be pronounced, in some cases, Arabizi users favored a one-to-one representation of ي regardless of if it was used as a vowel or consonant, in a transliteration-based approach. <Ay> and <ey> were also the two most common ways to write the diphthong of *yaa'* + *fatha*, as in *layl* [leːl] *(night)* or *heyet* [ħeːjeːt] *(life),* although <ee>, <ea>, <ai>, and <iy> also appeared.

***Waw* (و)**

<Ou> was the most popular way to write و as a vowel, and was used 48% of the time. The next most popular variants were <u> and <o>, which were used 16% and 15% of the time, respectively. The French representation <ou> was more popular than the English representation <oo> (which occurred only 2% of the time). This was somewhat surprising given the higher incidence of English (21%) than French (7%) in the data, since presumably a higher incidence of

English would also indicate more influence from English pronunciation on Arabizi. However, this finding concurs with the results from another study of Lebanese Arabizi text messaging, which found that <ou> was also the most common variant to represent /uː/ (Bou Tanios, 2016). This seems to indicate that the representation of و tended to follow pronunciation conventions, since <ou> is the French representation of /uː/. <W> was used most often to represent و as a consonant, and <aw> was most often written to represent the diphthong of و + *fatha.*

***Alif* (أ)**

Even though *alif* has perhaps the most possible pronunciations out of any vowel in Lebanese Arabic (it can range from [æ] to [ɔ] depending on the surrounding letters), its representation in Arabizi was surprisingly consistent, possibly because it is not used in diphthongs. The most popular way to write *alif* was <a> (which occurred 445/507 = 88% of the time), although 11% of the time <e> was used, such as in the words <kamen> (كمان) [kamɛːn] *(also)* and <ensen> (إنسان) *(person)* [insɛːn]. The use of <e> probably represents the user's pronunciation, since *alif* is usually pronounced as [aː] but in Lebanese dialect can be pronounced as [ɛː] when it is preceded by a labial or alveolar consonant (such as ب [b], ت [t], د [d], ز [z], س [s], ف [f], م [m], or ن [n]), and not followed by an emphatic consonant, as is true in these examples.

***Fatha* (ـَ)**

When it was written, *fatha* most commonly appeared as <a> (451/651 = 69% of the time), although there was some trend toward the use of <e> similar to that noted above in the section on *alif* (31% of the time). This aligns with El-Khaissi's data, which also showed <a> and <e> as the most common variants for fatha, respectively. Some of this variation is probably due to different pronunciations of /a/ in different environments; for instance, it can be pronounced as as [a], [ɛ],

[e] depending on the surrounding consonants. *Fatha* was written the most consistently out of all the vowels, with only two variants (<a> or <e>), while all other vowels had at least three representations. However, sometimes both <a> and <e> would be used to write *fatha* within the same word (as in <nefham> [nɛfhɛm] (نَفهَم) (*we understand*)), which was rather confusing. However, this probably reflects a tendency for users to write Arabizi how they would pronounce it, since the phoneme /a/ (or the letter *fatha*) is pronounced [ɛ] or [e] in most environments, but [a] or [ɑ] before and after gutturals, including /h/.

***Damma* (ُ)**

Damma was usually written as <o> (94/149 = 63% of the time), followed by <e> (41/149 = 26% of the time). This differs from El-Khaissi (2015), who found <ou> as the most common variant, followed by <o>, similar to the representation of *waw*. The occurrence of <e> as a representation of *damma* is probably due to the tendency in Lebanese Arabic to collapse short vowels such as /i/ and /u/ into [ə], such as in the word <ente> [əntə] (*you*), which again mirrors user pronunciation.

***Kasra* (ِ)**

Kasra was usually written as <e> (54% of the time), although <i> was also popular (40% of the time). This aligns with Bou Tanios's (2016) study of Lebanese Arabizi in text messaging, which also found that the most common variant for *kasra* was <e>. Likewise, El-Khaissi's results also found <e> and <i> to be the most common representations for *kasra* in Lebanese Arabizi (2015). The use of <e> for *kasra* also probably reflects the Lebanese Arabic tendency to shorten vowels and pronounce /i/ as [ə]. <e> was also commonly used to represent *kasra* when paired with *alif* at the beginning of a word, such as in *ente* [əntə] (*you feminine*). Since <e> was commonly used to represent both *kasra* and *fatha*, sometimes it could be confusing if they were both written as <e>

within the same word. However, most users found a way around this problem by using one of two variations consistently: either <e> for *fatha* and <i> for *kasra*, or <a> for *fatha* and <e> for *kasra*.

***Ta marbuta* (ة)**

Ta marbuta (ة) is used in Arabic as a variant of the letter *taa* (ت) at the end of words to indicate feminine words. It always follows a fatha (ـَ), and is usually pronounced as [e] in Lebanese Arabic. In genitive constructions (*iDaafas*) it is pronounced as [t], and is written as (ت) when a suffix is added. (For example, the word حياة [ħeːjeː] *(life)* ends in a ta marbuta, but when it is made possessive by adding ي (e.g. "my life"), the ta marbuta is written as ت and pronounced حياتي [ħeːjeːti]). In this thesis, ta marbuta was most often written as <e> (92/130 = 71% of the time), followed by <a> (29/130 = 22% of the time). This reflects the overall trend in Lebanese Arabic of (*imāla*), or the raising of /a/ to [e] at the end of a word. Thus, the use of <e> mirrors user pronunciation, since *ta marbuta* is pronounced as [a] after a guttural or emphatic consonant but as [e] after other consonants. For example, the word <madrase> (مدرسة) [madrasa] (*school*) was written in the data with an <e> because [s] is not an emphatic or guttural consonant, while words like <mnii7a> (منيحة) [mɛniːħa] *(good)* were written with an <a> because [ħ] is pharyngeal.

Overall, two trends appeared consistently in the data related to vowel orthography: repeated vowels and vowel deletion, which suggest that although Arabizi may be a mostly transcription-based process, it also takes into account written conventions of CMC.

Repeated Vowels

As in English CMC, the features of spoken language are approximated in Arabic CMC, departing from the conventions of written Arabic to convey more informal types of speaker meaning. One example of this was vowel repetition within words to convey stressed and elongated pronunciation. Some examples of repeated vowels found in the data for this thesis are shown below:

- *Bhebak kitiiiiiir* [bħɛbək kɪtiːr](بحبك كتير) *[I love you a lot]*
- *Ya jamelaaaa* [ja dʒɛmiːla] (يا جميلة) *[Hey beautiful]*

Vowel Deletion

Another main feature of vowel use in the data was vowel deletion, or the removal of vowels that would normally have been spoken. As Crystal (2006) notes, this is also common in English CMC messages as a more economical way to write longer words, although Arabizi complicates it a little because short vowels are not usually written in colloquial Arabic anyway. Consequently, vowel deletion occurred most often with short vowels that would have been written as diacritics (*fatha, damma*, and *kasra*) and less frequently with long vowels (*alif*, *waw*, or *ya*). These representations also possibly indicate a more transliteration-based approach to writing Arabizi, since the short vowels would be pronounced, but not normally be written in Arabic anyway. Some examples of vowel deletion are shown below:

- *Ya 3ame ente ma **btkbare**?* [direct address: You're not getting older?]

  (يا عمي انتي ما **بِتكبري**؟) *(Removal of 2 kesras)*
- ***Msh** ma32ol jamelaaaaaaaaa* [That's nonsense, beautiful]

  (**مِش** معقول جميلة) *(Removal of kasra)*

Overall, despite these orthographic variations, the Arabizi users in this sample seemed to be particularly adept at spelling lexical items unique to a dialect in a consistent manner even though they were part of an unwritten register of speech, as there were many instances of the same words spelled consistently, such as <hayat> *(life)* and <lesh> *(why).* On the other hand, different variants were often used to represent the same sound(s) within the same word, and the same words were often spelled multiple ways within the data, such as <mout>/<moot> (*death*) and <layl>/<leil> (*night*), suggesting that Lebanese Arabizi orthography has not fully conventionalized. Additionally, it is important to note that the orthographic variation seen in the data for this thesis could also be due to a number of outside factors besides linguistic constraints, such as personal preference, L2, the presence of non-Lebanese immigrants in the data, or gender.

4. Discussion and Conclusion

I. Discussion

Overall, the results from this study illuminated three main findings: 1) Romanized Arabic was not used as often as was previously thought on Twitter in Lebanon, and was instead surpassed by colloquial Arabic, 2) Romanized Arabic is still associated with a high percentage of code-mixing with other languages, especially English, and 3) Romanized Arabic orthography in Lebanon is a complex writing system that seems to adopt a mostly transcription-based approach, in which its consonants have become fairly conventionalized. All three of these findings could also indicate a shift to a more mixed language ideology in which the traditional divide between colloquial and formal Arabic is relaxed and colloquial Arabic is becoming more accepted.

First, the results showed that Romanized Arabic was used less frequently than hypothesized on Twitter in Lebanon. Instead, colloquial Arabic, English, and even French were favored over Romanized Arabic. This could be caused by several factors: the general trend toward English as a global language online, the greater availability of Arabic language keyboards compared to previous decades, a change in attitude about using Arabic script online (either an increased desire to use Arabic script, or decreased motivation to use Latin script), or other sociolinguistic factors. Most likely all of these factors contributed to this phenomenon, although how much each one may have contributed is uncertain due to the difficulty of determining Twitter users' motivations for using certain scripts. This is interesting because as previously mentioned, traditional Arabic language ideology maintains that colloquial Arabic is not written and therefore has less prestige, so the fact that it was the most popular way to write Arabic on Twitter could indicate a shift in this ideology towards a more accepting view of colloquial Arabic. More research could be done to see if this lower occurrence of Arabizi also appears on other social media sites such as Facebook or Instagram, and in other Arabic dialects as well. It would also be interesting to see if Twitter's 140-character limit or use of hashtags affects whether users write in Latin or Arabic script.

Second, the data also showed that Romanized Arabic is still associated with a high degree of code-mixing, especially with English. Arabizi was most commonly code-mixed with English, followed by LCA. Although this thesis did not closely examine code-mixing, from a cursory review of the data, it appears that most of the code-mixing appeared at sentence or clause boundaries and involved primarily noun phrases or proper nouns. Additionally, influence from French and English was clearly visible in both Arabizi vowel and consonant representations, as

shown by the French <ou> vs. English <oo> and French <ch> versus English <sh>. This is consistent with other studies of Arabizi-English and Arabizi-French code-mixing, which have shown that English is commonly mixed with Arabic, and that words and grammatical features from English and French are increasingly being integrated into the Arabic lexicon (Doss, 2006; Post, 2015). The fact that language mixing appeared both ways—features from French and English are newly appearing in primarily Arabic text, and Arabic words are being mixed with mainly French or English text—could show that the previously strict boundaries between languages in Arabic language ideology may be becoming more relaxed compared to previous years. Additionally, Arabizi and LCA were the second most-commonly code-mixed languages, although they only occurred three times in the data. Although these instances did not occur frequently, it would be interesting to see if other studies also find any code-mixing between Arabizi and LCA, since it could provide evidence for whether Arabizi users see Arabizi as just a way to write LCA, or as a separate language form. Much research has been done on code-mixing between Arabic and other languages, but more research is needed to further analyze the ideological, sociolinguistic, and practical implications of code-mixing between Arabizi and other Arabic dialects, which could provide more evidence for whether users think of Arabizi as a distinct language variety and how it interacts with other dialects.

Lastly, Lebanese Arabizi orthography showed a surprisingly high level of conventionalization, at least for consonants—18 of the 25 consonants were represented the same way 100% of the time, and the remaining consonants were represented the same way (by one variant) more than 75% of the time. The consonants also seemed to follow primarily a transcription, rather than transliteration process when written in Arabizi, since characters like ق, ذ

ث, and خ were written based on how they would be pronounced, not as a one-to-one correspondence based on graphic transposition. The exception to this was ع and ح, which were always represented the same way with <3> and <7>. The fact that the consonants were relatively conventionalized could indicate that Arabizi is mostly seen as a way to transcribe spoken colloquial Arabic.

However, the vowels in Lebanese Arabizi were more variable—*alif*, *fatha*, and *damma* were written the same way more than 50% of the time, but the other vowels often had competing representations that had not yet conventionalized. While many of the vowels (such as *ta marbuta* and *fatha*) showed evidence of a transcription mindset, as they varied based on pronunciation, other factors pointed to a transliteration-based approach, such as the consistent representation of *yaa'* as <i> and the tendency to delete short vowels which would be pronounced, but not written. Despite this, there seemed to be a general trend towards <e> for *ta marbuta* and *kasra* and a fairly stable choice of <ou> and <o> to represent *waw* and *damma*, respectively. Since the vowel patterns of Lebanese Arabic are fairly distinctive, these results could be used to distinguish Lebanese Arabizi from other Arabic dialects. The hallmarks of Lebanese Arabizi could be generalized as the use of <ou> for *waw* (which is more commonly represented as <o> in other dialects) and the use of <e> for *kasra* and *ta marbuta* (which are usually written as <i> and <a> in other dialects, respectively). As previously mentioned, orthographic variation can serve as an indicator of the amount of linguistic prestige a certain language variety has, but since Arabizi made up only 2% of this thesis' data and it shows a lot of orthographic variation, it seems unlikely to significantly gain prestige in the future.

II. Limitations

Although this thesis attempted to control for outside factors that may have influenced the data (such as location and spam tweets), it still had some limitations due to the relatively small sample size and possible error in the Twitter Archiver collection program. First, this study examined only a limited sample of tweets in Lebanon, and therefore provides only a snapshot of possible trends in Arabizi orthography. Although it focused on Beirut, Twitter Archiver's location query occasionally may have collected tweets from outside Lebanon that were not representative of Lebanese Arabizi. Twitter Archiver's filters may have also let some spam tweets into the data or tended to collect at certain times of day, which would have captured the Twitter environment only at particular times, not Twitter in Lebanon in general. Second, it is hard to determine a user's personal and sociolinguistic information from a Twitter profile, so even though this thesis explored reasons for orthographic variation in Arabizi, some of this variation may have been caused by extralinguistic factors such as age or gender. This study also tended to collect a large amount of tweets that were from news organizations, which may have skewed the data towards tweets written in MSA. It would be useful to repeat the collection process and exclude all news tweets to see if the data showed different trends in language frequency. Lastly, tweets that did not show obvious signs of either Standard or colloquial Arabic were shown to a native speaker for review, so the decision to categorize some tweets as MSA or LCA could have turned out differently if more native speakers were asked about a tweet's language type.

III. Implications and Further Research

Despite its low rate of use on Twitter in Lebanon and lack of prestige in the Arab world, Arabizi (and colloquial Arabic) is still used in various online spaces, and understanding how to use, read, and translate it still poses a challenge for Internet users. Understanding how Arabizi is written, how it can indicate a user's gender or second language, and how it interacts with other languages and dialects are all issues at the forefront of linguistic research, and will have significant impacts on language processing, translation, and national security. Many word-processing programs (such as Microsoft Word) are still unable to properly render Arabic script, so it is likely that Arabizi will continue to be used in the future. Moreover, as Arabizi continues to appear in various contexts across the Arab world, the demand for translating it will presumably increase, but as of now, no accurate widespread public program for detecting or translating Arabizi is available. Some attempts have been made to develop a system to detect Arabizi and convert it to Arabic or English (see Darwish, 2013; Voss et al., 2014 and Tobaili, 2016), but the amount of orthographic variation between different Arabizi dialects and its tendency to code-mix with other languages makes this difficult. However, some online sites are becoming more aware of the need to accommodate users' preferences for searching the web and finding content written in Arabizi. This can be seen through the increased availability of sites like Yamli, Onkosh, and Google Translate, which allow users to enter text or search the Internet in Arabizi (Aboelezz, 2009:20).

Arabizi use also has implications for second language learning. In a study of L2 Arabic learners, Attwa (2009:22) found that 95% believed that they needed to learn Arabizi to effectively communicate with Egyptians on CMC, although 65% said it hindered their

communication in Arabic via CMC. Most of the learners said that the complexity of communicating in Arabizi was caused by the fact that it is not standardized, and that they take longer to read Arabizi and to be understood when writing in Arabizi. Therefore, learning Arabizi appears to pose a significant but necessary challenge for L2 Arabic learners. Further research in understanding how Arabizi is written and the extent to which it is standardized, as attempted in this thesis, would help these learners become more accustomed to using and reading Arabizi. It could also help teachers develop effective methods to teach students how to read and write Arabizi, which is not commonly addressed in most L2 Arabic classrooms.

Taking all of these implications into account, the question arises: How will Arabic speakers deal with Arabizi in the future? Will they encourage the continued use of Arabizi and its mixing with other languages, or will Arabizi eventually fade away as technology gets better at dealing with Arabic script? And more importantly, how will this affect Arabic language ideology? On one hand, this study showed a lower than expected rate of Arabizi on Twitter, and developments are being made to encourage the use of Arabic script and adapt it to Western technology. For instance, typographer Abou Rjeily has developed the Mirsaal typeface, which is designed to bridge the gap between Arabic script and Arabizi by using Arabic letters that are detached from each other (Saghbini and Zaidi, 2011). On the other hand, many Arabizi users have grown accustomed to using Arabizi and like the convenience it provides to use English and French devices and programs. Most users do not seem to care too much about the overall standardization of Arabizi, as long as it is still convenient for them to use. As Attwa (2009:19) points out, “Some of the users even do not see a need to stop using [Arabizi] as long as it is economical and enjoys some level of standardization.” Similarly, Abdel-Ghaffar, N., et al.

(2011) acknowledge that "even natives do not find a reason to stop using Arabizi as long as it is efficient, economical, and, above all, mutually comprehensive." It is impossible to know the future direction of Arabizi for sure, but at least from an orthographic perspective, the limited evidence from this study suggests that Lebanese Arabizi is here to stay for the foreseeable future, and will probably become more conventionalized over time as it appears in more contexts and its users develop a consensus on how it is written.

Regardless of the direction Arabizi takes in the future, understanding its potential effects on the linguistic environment of the Arab world is important, because it affects how Arabic language ideology will be shaped and thus how the Arab world views the West. Of course, Arabizi use (or not) will not significantly affect relations between the East and West, but the consequences are interesting to consider. Arabic and Arabic script will probably always be associated with Arab identity and Islam, and the Latin script with Westernization. Arabizi represents a unique attempt to bridge this gap and join both worlds. In the increasingly turbulent political and social climate of the Middle East, the increased influx of refugees to the West, and often discordant relationships between the West and the Arab world, it remains to be seen how Arabizi will influence or be influenced by these situations. If Arabizi continues to be seen as a bad influence and a threat to the Arabic language, there may be further backlash against globalization and Westernization, while if the transition to a more relaxed language ideology continues and Arabizi assumes greater prestige, it may help promote more overlap between the two cultures.

References

Abdel-Ghaffar, N., et al. (2011). Arabizi or Romanization: The Dilemma of Writing Arabic Texts. *Jil Jadid Conference.* University of Texas, Austin.

Abolezz, M. (2010). A Latinised Arabic for All? Issues of Representation, Purpose and Audience. *The International Symposium on Arabic Transliteration Standard: Challenges and Solutions,* 100-110.

Abolezz, M. (2009). Latinised Arabic and Connections to Bilingual Ability. Lancaster University Postgraduate Conference in Linguistics & Language Teaching.

Albirini, A. (2016). *Modern Arabic Sociolinguistics: Diglossia, Variation, Codeswitching, Attitudes and Identity*. Routledge, 105-120.

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

Al-Munziri, R. (2014). Mustawa Istikhdam Al-'Arabizi lada Ash-Shabab Al-Ammani Fi Maeqi' At-Tawasol Al-Ejtima'i. *Lughatul Syabab Al-;Arabi Fi Wasael Al-Tawasul Al-Hadithah*, 205-234.

Al- Khalil, M., and Palfreyman, D. (2003). "A Funky Language for Teenzz to Use:" Representing Gulf Arabic in instant Messaging. *Journal of Computer Mediated Communication*, 9 (1).

Al-Khatib, M. A., & Sabbah, E. H. (2008). Language Choice in Mobile Text Messages among Jordanian University Students. *SKY Journal of Linguistics*, (21): 37–65.

Allen, J. and Mason, M. (2003). Computing in Creole Languages. *Multilingual,* 14 (1).

Attwa, M. (2009). *Arabizi: A Writing Variety Worth Learning?* The American University in Cairo.

Badawi, E., Carter, M. G., & Gully, A. (2004). *Modern Written Arabic: a Comprehensive Grammar*. London ; New York: Routledge.

Bahrainwala, L. (2011). You say Hello, I say Mar7aba: Exploring the Digi-speak that Powered the Arab Revolution. Michigan State University.

Barakat, Dima, (2009). *Yasmeen Al-Sham.* Damascus: Counsel on Arabic as a Foreign Language.

Barlow, J., & Nadeau, J.B. (2008). *The Story of French*. New York, NY: St. Martin's Griffin.

Bassam, L. (2014). Gender and linguistic background in SMS code-switching by Lebanese students. Tarragona: *Intercultural Studies Group,* 113-126.

Bassiouney, R. (2009). *Arabic Sociolinguistics: Topics in Diglossia, Gender. Identity, and Politics.* Washington, DC: Georgetown UP.

Bies, A., Song, Z., Maamouri, M., Grimes, S., Lee, H., Wright, J., … Rambow, O. (2014). Transliteration of Arabizi into Arabic Orthography: Developing a Parallel Annotated Arabizi-Arabic Script SMS/Chat Corpus. *Association for Computational Linguistics*. 93-103.

Bou Tanios, J. (2016). Language Choice and Romanization Online by Lebanese Arabic Speakers. Retrieved from http://repositori.upf.edu/handle/10230/27669

Brustad, K. E. (2015). The Question of Language. *The Cambridge Companion to Modern Arab Culture*. Cambridge: Cambridge University Press, 19-35.

Brustad, K. E. (2000). *The Syntax of Spoken Arabic: A Comparative Study of Moroccan, Egyptian, Syrian and Kuwaiti Dialects.* Washington, D.C.: Georgetown University Press.

Chahine, D. (2011). Children in the Early Years Classrooms Code-Switching. Lebanese American University.

Crystal, D. (2006). *Language and the internet* (2nd ed.). Cambridge, UK; New York: Cambridge University Press.

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

Dorleijn, M., & Nortier, J. (2009). Code-switching and the internet. In B. Bullock & A. J. Toribio (Eds.), *The Cambridge Handbook of Linguistic Code-Switching*. Leiden: Cambridge University Press. 127-141.

Doss, M. (2006). Cultural Dynamics and Linguistic Practice in Contemporary Egypt. *Cairo Papers in Social Science*, 27, 51-68.

Eisele, J. (2003). Myth, values, and practice in the representation of Arabic. *International Journal of the Sociology of Language*, (163), 43-59.

Elhij'a, D. (2014). A new writing system? Developing orthographies for writing Arabic dialects in electronic media. *Writing Systems Research*, 6:2, 190-214.

Elhij'a, D. (2012). Facebook Written Levantine Vernacular Languages. *The Levantine Review*, 1(1). 68-105.

El-Khaissi, C. (2015). The Romanisation of Arabic: A Comparative Analysis of Romanized Spoken Arabic and Romanized Modern Standard Arabic. La Trobe University.

Essawy, R. (2010). Arabic in Latin Script: Who is using it and why in the Egyptian Society. *Global English: Issues of Language, Culture, and Identity in the Arab World*. Peter Lang Publishers, 5-10.

Gordon, C. (2011). From Speech to Screen: The Orthography of Colloquial Arabic in Electronically-Mediated Communication. Swarthmore College.

Grenoble, L. and Whaley, L. (2006). Orthography. *An introduction to language revitalization*. Cambridge: Cambridge University Press, 137-159.

Haggan, M. (2007). Text messaging in Kuwait. Is the medium the message? *Multilingua, 26*(4), 427–449.

Herring, S. C. (2007). A Faceted Classification Scheme for Computer - Mediated Discourse. *Language@Internet*, 1, 1–37. Retrieved from http://www.languageatinternet.de/articles/2007/761/index_html

Herring, S. C., Ed. (1997). *Computer-Mediated Discourse Analysis*. Special issue of the *Electronic Journal of Communication, 6* (3).

Hoigilt, J. et al.. (2013). The Ideology and Sociology of Language Change in the Arab World. Georgetown University, The University of Oslo, The University of Texas at Austin, City University New York, Cairo University.

Irvine, J., and Gal, S. (2000). Language ideology and linguistic differentiation. *Regimes of language: Ideologies, polities, and identities.* Santa Fe: School of American Research Press, 35-84.

Kenali, A. et al. (2016). Code-Mixing Consumptions among Arab Students. *Creative Education*, 7, 931-940.

Khalil, S. (2012). The evolution of the Arabic language through online writing: the explosion of 2011. British Society for Middle Eastern Studies.

Kul, M. (2007). Phonology in Text Messages. *Poznan Studies in Contemporary Linguistics*, *43*(2), 43.

Lebanese Arabic Institute. (2017). *The Arabic Alphabet: A Guide to the Phonology and Orthography of MSA and Lebanese Arabic.*

Mostari, H. (2009). What do mobiles speak in Algeria? Evidence from language. *Current Issues in Language Planning, 10*(4), 377-386.

Ornan, U. (2003). Latin Conversion of Hebrew: Grammatical, Full, and Deficient. *Hebrew Studies,* 44(1), 185-202.

Palfreyman, D. (2001a). *LINGUIST List 12.2760: Informal Romanized Orthographies*. http://linguistlist.org/issues/12/12-2760.html#1

Saghbini, S., & Zaidi, R. (n.d.). Changing the Face of Arabic. Retrieved from http://languagemagazine.com/LangPages/Arabic_Script_Aug11.pdf

Sayahi, L. (2014). *Diglossia and language contact: Language variation and change in North Africa*. Cambridge: Cambridge University Press.

Shawish, H. (2010). Campaign to save the Arabic language in Lebanon. *BBC News.* Retrieved from http://www.bbc.com/news/10316914

Tobaili, T. (2016). Arabizi Identification in Twitter Data. *Association for Computational Linguistics,* 51.

UNESCO and Qatar Foundation. (2007). Literacy Challenges in the Arab Region: Building Partnerships and Promoting Innovative Approaches. Doha: UNESCO and Qatar Foundation.

Unseth, P. (2005). Sociolinguistic parallels between choosing scripts and languages. *Written Language & Literacy*, 8(1), 19-42.

Voss, C. R., Tratz, S., Laoudi, J., & Briesch, D. M. (2014). Finding Romanized Arabic Dialect in Code-Mixed Tweets. In *LREC*, 2249–2253.

Warschauer, M., El Said, G. R., & Zohry, A. G. (2006). Language Choice Online: Globalization and Identity in Egypt. *Journal of Computer-Mediated Communication*, 7(2002), 1–18.

Yaghan, M. A. (2008). "Arabizi": A Contemporary Style of Arabic Slang. *Design Issues*, *24*(2), 39–52.

Yvon, F. (2010). Rewriting the Orthography of Text Messages. *Natural Language Engineering, 16*(2), 133 – 159.

Zoabi, Z. (2012). *A'amiya : kef mnektibha ? Alphabet Choice in Electronic A'amiya in Israel and The Arab World*. University of Haifa.

List of Figures

Biography

Natalie Sullivan was born in Maryland in 1994 and moved to The Woodlands, Texas in 2005. She entered the University of Texas at Austin in 2013, majoring in Plan II Honors, English, and Middle Eastern Languages & Cultures with a minor in Spanish. She studied abroad after her sophomore year in Rabat, Morocco and Seville, Spain. In college, she worked as a reporter and editor for *The Daily Texan*, a consultant in the University Writing Center, and a research assistant in the LLAMA Linguistics lab. She graduated in May 2017. After graduation, she plans to teach English in Morocco on a Fulbright U.S. Student grant and then go to graduate school.