ARABIC AND LATIN KEYBOARD CHOICE IN LEBANESE STUDENTS' DISCUSSIONS ON WHATSAPP

Ayman Halawi, Nasri Messarra and Raymond Bou Nader
Saint Joseph University: Université Saint-Joseph, Beirut, Lebanon
Corresponding Author: Ayman Halawi

(Received January 2021 – Accepted April 2021)

ABSTRACT

Halawi, A. Messarra, N. & Bou Nader, R. (2021). Arabic and Latin keyboard Choice in Lebanese Students' Discussions on WhatsApp. *Lebanese Science Journal,* 22(1), 127-147.

Written discourse in WhatsApp discussions has been addressed in several articles in Computer-Mediated Communication (CMC). The aim of this study is to determine Lebanese universities students' choice between Arabic and Latin keyboards while typing their online messages in WhatsApp groups and try to determine variables that affect their keyboards' choices. We joined 33 WhatsApp discussion groups from 7 major Lebanese universities and gathered 227,059 messages written by 1,112 multilingual students. The results showed that even though Arabic keyboard is not very popular amongst some Lebanese universities' students, it is still present in WhatsApp groups' discussions of students especially at some faculties of the official Lebanese University where courses are taught in Arabic language. The results showed also that Arabic is widely typed in Arabizi using a Latin keyboard. This seems to be the case in the majority of universities that took part in the study. In addition, students at universities with high tuition fees and those who are having their curriculum in foreign languages use Latin keyboard more than students at some faculties of the public university.

Keywords: Arabic and Latin script; *Arabizi*; *computer-mediated communication* ʻCMC ʼ; *keyboard-Switching*; *Latinized Arabic*; WhatsApp.

http://dx.doi.org/10.22453/LSJ-022.1.127-147

lsj.cnrs.edu.lb/vol-22-no-1-2021/

INTRODUCTION

Increased use of smartphones leads to new forms of communication, like voice notes, emoticons, gifs, abbreviations and, in the case of Arabic speaking countries, *Latinized Arabic* (Sullivan 2017; Alghamdi & Petraki 2018). Lebanese university students are no exception to that, especially when we look at their group discussions on WhatsApp. Lebanon is one of the few countries where foreign language has been taught since the first years of schooling making most Lebanese at least bilingual. French and/or English are understood and used by a large percentage of the Lebanese people (Shaaban, 1997). These are, after Arabic, the languages of education in all schools and universities alongside Arabic (Shaaban, 1997; Shaaban & Ghaith, 1999; Hoyek, 2004; Kanaan, 2011). Arabic is used in two major variants: *modern standard Arabic* and *dialectal Arabic* (Bassam, 2017; Wehbe, 2017). Therefore, *borrowing* words from English, French, and other languages is very common in the Lebanese dialect.

It is noteworthy that some of the Lebanese students use Latin characters to write in *dialectal Arabic*. They write *dialectal Arabic* words in Latin characters with Arabic numerals (2, 3, 5, 7 and 9) to designate some Arabic consonants not available in the English language. The '*Romanization*' of Arabic is also known as '*Arabizi*' (Yaghan 2008; Attwa 2012; El-Zahraa 2014; Tobaili 2016). The understanding of utilization of foreign languages and *Romanized Arabic* amongst Lebanese students is important for academics and professionals to understand the evolution of language from a linguistic perspective, as well as from a social and marketing perspective.

In the current work, the following research questions will be addressed:

1- **Is** Latin keyboard preferred amongst Lebanese students?

2- Does the level of tuition fees (Social class) (high, moderate, low) affect the students' keyboard choice?

3- Has the language of teaching (curriculum in English / French, mixed or in Arabic) an impact on Keyboard choice?

4- Are the students of LU learning in Arabic (Only for some faculties where we joined students' WhatsApp groups) using *Keyboard-Switching* more than other universities?

Theoretical background & related literature

CMC, especially at the level of writing, has been discussed widely in academic literature. The integration of Internet services in mobile phones has been a catalyzer in this field as well as a new trend setter. For example, Pérez-Sabater, (2015: p.5) discusses 'new trends in *CMC* Discourse Studies'. Other researchers, like Baron, Yus, Herring have studied the impact of mobile communication on oral and written language. A new terminology referring to the *'non-standard language'* of *'digital/online speech'*

http://dx.doi.org/10.22453/LSJ-022.1.127-147

and a '*new variety of language'* has been used in such articles. It includes the use of emoticons, lexicalization of vocal sounds, orthographic and punctuation mistakes, phonetic orthography and eye dialect, abbreviations, acronyms, clippings, ellipsis, use of contractions, one-word transmissions, short messages, short words, and short sentences.

The presence of nonverbal cues is a characteristic of *CMC* according to Bies, et al., (2014) and Eskander et al., (2014). It is distinguished from standard spelling by the repetition of letters to accentuate, informal vocabulary, intentional deviations from standard spelling, spelling errors, abbreviations, phonetic writing of sounds such as laughter and the emotions. See (Carter, 2003; Pérez-Sabater, 2015; Gulacti et al., 2016; Petitjean & Morel, 2017). Yus (2011) proposes the term *'oralised written'* when speaking of written online discourse.

Arabizi

In the 1990s, as a result of globalization, English has gained more importance in Lebanon and most Arab countries through domination of the new means of communication as most devices, operating systems and applications did not support Arabic. For instance, Internet Relay Chat (IRC), mobile phone messaging (SMS) and e-mails did not support the Arabic script at that time, which led to the *Romanization of Arabic* using the Latin alphabet (Allehaiby, 2013).

In this context, Bortzmeyer (2012) noted that the operating system of Android 9 did not manage the Arabic script or its right-to-left writing system.

Arabizi became the term to describe a system of writing Arabic using Latin characters with numbers replacing Arabic consonants that are not available in the Latin alphabet (Yaghan, 2008, Bianchi, 2012). According to the origin of this term, Yaghan (2008), Kenali et al., (2016: p.932) explain that it is composed of two Arabic words عربي (pronounced 'Arabi' for Arabic) and انكليزي (pronounced 'Englizi' for English).

In the literature, researchers defined *Arabizi* as *'Romanized Arabic'* or *'Latinized Arabic'* and described this phenomenon related to linguistics and communication as 'non-standard', 'spontaneous' and 'informal' (Darwish 2013; Masmoudi et al., 2015). Yaghan (2008: p.39) defines *Arabizi* as a «text messaging system used over the Net and cellular phone». Bianchi (2012) considers that *Arabizi* is mainly used in the context of *Computer-Mediated Communication (CMC).* However, in Lebanon at least, this language has moved beyond the scope of Internet-only, to reach graffiti's and even

http://dx.doi.org/10.22453/LSJ-022.1.127-147

restaurant names like Enab (grape), el Dunia Heik (this is life), Zaatar w Zeit (thyme and oil) and others.

Rodrigues (2012) classifies *Arabizi* as '*a non-standard chat alphabet*' in which the Arabic language is written using the Latin alphabet. He defines the chat alphabet as « a non-standard spelling system used for textual communication » (idem: ix). Tobaili (2016) talked about 'digital trend' to define *Arabizi* as a « texting Non-Standard Arabic using Latin script » (idem: 51).

Similarly, for Bies et al. (2014: p.94), *Arabizi* is an online communication system using the Latin alphabet: «it is a non-standard *Romanization* of Arabic script that is widely adopted for communication via Internet». Kenali et al. (2016) distinguish *Arabizi* at the oral level, spoken under the pretext of modernization, while written *Arabizi* was originally used by Arab expatriates around the world to communicate on the Internet and via SMS.

The non-standard aspect of *Arabizi* is underlined by other researchers, like van der Wees et al., (2016: 43): «*Arabizi* is mostly guided by pronunciation; it is also very sensitive to dialectal variations, which are more noticeable in spoken than in written Arabic».

The informal nature of *Arabizi*, dialectal variants and the impact of foreign languages may explain deviations from standard spelling (Bies et al., 2014). For example, English speakers write the phoneme / ش / in *Arabizi* with an 's' followed by an 'h' ('sh') while French speakers do so with a 'c' followed by an 'h' ('ch').

Thus Darwish (2013) emphasizes the importance of the word in context to distinguish between certain Arabic and English words that share a common spelling, such as the preposition من (from) which is written in *Arabizi* (men).

Al-Badrashiny et al., (2014) add the adjective 'spontaneous' to the definition of this system commonly used to write on social networks, in SMS and chat applications. For them « *Arabizi* is a spontaneous orthography used to write dialectal Arabic using the Latin script, the so-called Arabic numerals, and other symbols » (idem: 31).

Moreover, the phonological writing based on the correspondence between phonemes (Arabic dialects) and graphemes (in Latin alphabet) with the nonverbal indices of the *CMC* could explain the individual and non-standard nature of *Arabizi*. This irregular nature poses a real challenge for researchers working in the field of *Natural Language Processing (NLP).*

http://dx.doi.org/10.22453/LSJ-022.1.127-147

Many *Arabizi* researchers have been interested in the automatic identification of this Arabic writing system and have tackled the problems resulting from its non-standard nature to develop transliteration systems to *Modern Standard Arabic (MSA)* or English (Darwish, 2013; Al-Badrashiny et al. 2014; Naji & Allan 2016; Baly et al., 2017). In this context, Habash et al., (2012) tried to create the CODA system which is defined by the authors as «a conventional orthography for *Dialectal Arabic* […] designed primarily to develop computational models of Arabic dialects» (idem: 711).

Although researchers like Eskander et al. (2014: 3) limit the use of *Arabizi* « to write mainly in dialectal Arabic in social networks, SMS and chat applications ». Others notice the integration of this writing system into other media such as TV commercials, street banners, and posters (Attwa, 2012). She concludes that *Arabizi* is transformed into an essential trend especially in the CMC. Meanwhile, Bou Tanios (2016) sees that the reasons for this practice of writing vary from habit to the norms and requirements of the *CMC* and the unfamiliarity of users with Arabic keyboard.

The Arabic script

Arabic is the fourth language of the world with 315 million speakers according to the site www.ethnologue.com (Edition 21 consulted in July 2018).

The Arabic script is adopted in Indo-European languages like Urdu and Malay. It is the second phonemic system in the world (Titus, 2017).

The writing of Arabic is cursive. It is a graphical system that is written from left to right and does not include capital letters. The letters do not always connect in the same way and can have several graphic shapes depending on the position in the word initial, middle, end or isolated as shown in figure 1. This writing system includes about one hundred graphical forms of letters (Balius, 2013).

Single	Initial	Medial	Final
ا	ا		ـا
ب	بـ	ـبـ	ـب
ت	تـ	ـتـ	ـت

Single	Initial	Medial	Final
ك	كـ	ـكـ	ـك
ل	لـ	ـلـ	ـل
م	مـ	ـمـ	ـم

Figure 1. The different graphic forms of some Arabic letters according to the position in the word.

http://dx.doi.org/10.22453/LSJ-022.1.127-147

The Arabic script, also called '*abjad'* is a system rich in consonants. It is composed of 28 consonants and only 6 vowels, 3 long and 3 shorts. Short vowels are often omitted in writing although they can change the meaning of words. They are mostly used in the religious and literary context. The short vowels are written by marks above and below the letters (Titus, 2017). Without these signs, a word can have several meanings. The meaning of a word with missing short vowels is often guessed based on its position in the phrase or the context of the phrase, paragraph or document.

WhatsApp

WhatsApp is the world's most popular and fastest-growing platform with one million new users per day in 2014 (Pérez-Sabater, 2015). It is a very popular mobile phone messaging application that dominates mobile communication today. Conducting group chats is considered by many researchers as the secret of its success (Seufert et al., 2016).

Communication via WhatsApp is done through written messages, voice messages, internet calls, photos, and emojis. In 2014, WhatsApp exceeded the number of SMS exchanged in the world with more than 10 billion messages per day referring to 'arabesocialmediareport' [retrieved from Https://sites.wpp.com/govtpractice/.../arabsocialmediareport-2015.pdf (Consulted online on 5 July 2019). The report showed also that in Lebanon, since 2015, it has been the most preferred online messaging application. Statistics show also that Lebanon has the highest percentage of users (58%) that prefer WhatsApp among Arab countries.

WhatsApp is the most popular online messaging application in the world with 1.5 billion monthly active users in late 2017 against 1.3 billion for Facebook Messenger Https://datanews.levif.be (Consulted on 5 July 2019).

Previous works about the choice of writing languages

Keyboard preferences between Arabic and Latin in WhatsApp communication have not been addressed in scientific literature. A study about the percentage of *Arabizi* in a corpus collected on Twitter in Lebanon and Egypt by Tobaili (2016) shows that, in Lebanon, almost half of the tweets are in Arabic the rest is divided among the other languages of which English is dominant. The total percentage of *Arabizi* in tweets is 4.9 percent for Lebanon and 5.7 percent for Egypt. Tobaili finds that the percentage and usage schemes of *Arabizi* differs from one country to another: In Lebanon, Twitter users' alternate codes in the same tweet. In Egypt, users drop the vowels in order to type faster.

http://dx.doi.org/10.22453/LSJ-022.1.127-147

The author judges the rate of use of *Arabizi* on Twitter in Lebanon and Egypt as low. He assumes that users prefer not to write their tweets in *Arabizi* because this writing system is perceived as a mean of informal communication.

Another study was made by Bies et al. (2014) in Egypt but showed that the results of researches about *Arabizi* are not always corresponding. Authors analyzed a corpus of more than 100 thousand SMS and chat messages from 26 Egyptian Arab participants. The results showed that *Arabizi* dominates the means of CMC in the corpus. Only 15% of the conversations are written in the Arabic alphabet, 66% totally in *Arabizi*. The rest (19%) is a mixture of the two.

In Jordan, Bianchi (2012) analyzed a corpus of 460,220 web forum posts collected on the site Mahjoob.com. The results show that *Arabizi* is the most used code on this site. He concludes that the mixed code of « vernacular Arabic and English » (idem: 93) is the result of what seemingly is the conflictual contact between globalization and local culture.

Eskander et al. (2014) have developed the '3ARRIB' system of automatic processing of texts written in Roman characters (*Arabizi*). The group of researchers introduced this system which can transliterate *Arabizi* to Egyptian Arabic and classify Arabic text input (sounds, punctuation marks, names, foreign or Arabic words, emoticons). All this with an overall performance accuracy of 83.8% on randomly selected test messages. However, this system is still in its trial phase and is not available to the public.

MATERIAL AND METHODS

WhatsApp data collection

Because WhatsApp is so popular amongst Lebanese users, in this article, we use WhatsApp messages in university student groups to analyze keyboard switching between Arabic and Latin keyboards.

The corpus was collected from 33 WhatsApp study groups of major Lebanese universities' students.

In order to collect and use the messages, the authors were introduced to the groups and the purpose of the study was explained to the students. For privacy concerns, then we affirmed our commitment to use an encryption scheme to hide the identity of the participants so that the data remains anonymous and untraceable.

http://dx.doi.org/10.22453/LSJ-022.1.127-147

In our choice of universities, we tried to encompass the largest spectrum of education, culture and social groups. The university's students involved in this study, as shown in table 1, were students from the American University of Beirut AUB (Math department, full curriculum is being taught in English), Lebanese American University LAU (full curriculum where we collected WhatsApp conversations is being taught in English), Saint-Joseph University of Beirut USJ (curriculum in French), Lebanese International University LIU (English and Arabic curriculum), Al-Maaref University MU (English and Arabic curriculum), Phoenicia University PU (curriculum in English) and the Lebanese University LU (curriculum in Arabic with 1 foreign language course)(In other faculties at LU curriculum could be taught in foreign languages (French or English)).

The Lebanese University is the only public university in Lebanon and the largest one in number of students.

Table 1. List of universities and number of students in the sample and teaching language.

	University name	Language of teaching	Students in our sample	WhatsApp Groups	Total Messages
Private universities	American University of Beirut (AUB)	English	117	1	1,991
	Lebanese American University (LAU)	English	80	3	14,189
	University of Saint-Joseph (USJ)	French	27	2	872
	Lebanese International University (LIU)	English - Arabic	256	15	17,542
	Al-Maaref University (MU)	English - Arabic	121	4	26,091
	Phoenicia University (PU)	English	19	2	50,019
Public university	Lebanese University (LU)	Arabic*	492	6	116,355
Total			1,112	33	227,059

*Languages of teaching vary at LU from a faculty to another. Arabic is the teaching language at the Department of Psychology where students took part in our study.

As mentioned earlier, most citizens in Lebanon are at least bilingual, if not trilingual, at the school level. This means that, whatever the language of teaching,

http://dx.doi.org/10.22453/LSJ-022.1.127-147

students may decide to use a different language in their communication over WhatsApp.

Data processing

All the messages from the 33 WhatsApp groups were extracted using the 'export' method of the application and compiled into one database. The normalization of the extracted data went through different phases as we noticed that, depending on the device's operation system (iOS or Android) and its settings (mainly date settings), the extracted data differs significantly, as shown in table 2.

Table 2. Exported data structures.

10/24/18, 10:02 PM – John Doe: chill	The date is delivered in mm/dd/yy format and the time in 12 hours format. A hyphen follows with the name, a double dot and the message
4/22/18, 12:22 – John Doe: Thank you	The format is similar except for the time format, delivered in 24 hours style
[8/29/18, 11:33:23 PM] Jane Doe: Never	The date is delivered inside brackets and the seconds are added to the time in 12 hours format. The name follows with a double dot and the message
[27/10/2018, 12:26:39 am] Jane Doe: Shutup	A similar format except that the year is delivered with 4 digits and the 'am/pm' is served in lowercase

Moreover, depending on the device, media information is also delivered in different formats (see table 3).

Table 3. Structure of media message (iOS, Android).

12/5/17, 4:24 PM – John Doe: <Media omitted>	There is no specification of the type of media that was used. The message is '<media omitted>' without details
[8/31/18, 2:03:52 PM] John Doe : audio omitted **[26/10/2018, 10:47:58 pm] John Doe : image omitted**	The type of media is specified: audio, image, video, etc.

http://dx.doi.org/10.22453/LSJ-022.1.127-147

In each message, depending on the presence or absence of the user in the address book of the person collecting the information, the user information can be returned either as a name/nickname or as a phone number (international format).

It is also important to note that the character returned by different devices differs in terms of character code and fonts used. Some emoticons or special characters are returned differently and sometimes replaced by other graphic characters. Before processing the messages, we convert them to *UTF-8* format.

All system messages are removed from WhatsApp conversations. Those include group creation messages, new group users, users leaving the group, etc. Some examples follow:

- *10/16/18, 7:03 AM - Messages to this group are now secured with end-to-end encryption. Tap for more info.*
- *8/31/18, 12:30 PM – [name] created group "Math201-check description"*
- *10/16/18, 7:03 AM – [name] added you*
- *10/16/18, 7:03 AM – [name] left*
- *[name] changed their phone number to a new number. Tap to message or add the new number*

Messages with URLs were also removed (link sharing).

We then process and convert the text files into tab-delimited values files using the following regular expression:

\[?(\d{1,2})\/(\d{1,2})\/2?0?(\d{2}), (\d{1,2}):(\d{2})(.[APap][Mm]|:\d\d [APap][Mm])?\]? -? ?\??(.+?):\??(.)*

The expression translates as follows:

- Presence or absence of an opening bracket
- One or two digits representing the month followed by a slash
- One or two digits representing the day followed by a slash
- Presence or absence of the ‘2’ and ‘0’ representing the first two digits of the year
- Two digits representing the last two digits of the year followed by a comma and space
- One or two digits for the hour followed by double dot
- Two digits for the minutes
-

http://dx.doi.org/10.22453/LSJ-022.1.127-147

- Presence or absence of one of two options: a space followed by a lower or uppercases ‘am’ or ‘pm’, or a double dot with two digits representing the seconds followed by a lower or uppercases ‘am’ or ‘pm’
- An optional closing bracket
- An optional hyphen or an optional space
- An optional interrogation point: this is necessary as some files have a hidden character, probably due to phone numbers or user’s name written in Arabic
- Many characters exceeding one for the phone number or username followed by a double dot
- Another optional interrogation point inherited from the conversion to UTF-8 format for some text files
- Any number of characters representing the message

The conversion expression below converts the data to tab-delimited values based on the groups collected in the first expression and prepares the data for import and processing into Microsoft Excel® and SPSS®.

University Group 1 20$3-$1-$2 $4:$5$6 $7 $8

The messages are then processed in Excel® using the following regular expression to detect all words in Arabic script (Arabic keyboard usage):

[\u0600-\u06FF]+
The number of words in every message was counted using VBA:

```
Function WordCount(CellRef As Range)
Dim TextStrng As String
Dim Result() As String
Result = Split(WorksheetFunction.Trim(CellRef.Text), " ")
WordCount = UBound(Result()) + 1
End Function
```

Each word was checked for Arabic or Latin characters using regular expressions in order to count the words written with Arabic letters in a message.

We then compared the number of Arabic words to the total number of words in a message:

- Arabic word count = 0 : the message is in Latin characters with or without ‘emoticons’ – the Latin keyboard was used
- Arabic word count < total word count : the message is mixed – A keyboard switch has occurred

http://dx.doi.org/10.22453/LSJ-022.1.127-147

- Arabic word count = total word count: the message is in Arabic – the Arabic keyboard was used.

Also, we used a regular expression to detect some words in *Arabizi*. Basically, we detected all words that include numerals.

RESULTS AND DISCUSSION

The corpus is constituted, after media and system messages removal, of 204,994 messages. The total number of gathered messages was 227,059 messages. The total count of words in our corpus is 1,105,197.

The table 4 below shows the distribution of messages' typing keyboard (Arabic vs. Latin) per universities with percentages. Overall, 107,001 messages from the corpus were typed using Latin keyboard. The results showed also that the other 97,993 messages are composed with Arabic keyboard only or mixed with *Keyboard-Switching* in the same message.

Table 4. Distribution and percentages of keyboard choices among Lebanese universities.

Universities ROW Labels / message level	usage of Latin keyboard only	Percentage	Usage of Arabic keyboard (Arabic only or mixed)	Percentage	Grand total
AUB	1874	99.89%	2	0.11%	1,876
LAU	13,304	99.85%	19	0.14%	13,323
LIU	14,714	94.15%	913	5.85%	15,627
LU	41,706	39%	62,207	60.78%	103,913
MU	8639	36.29%	15,160	65.08%	23,799
PU	25,964	56.87%	19,683	44.27%	45,647
USJ	800	98.88%	9	1.12%	809
Grand total	**107,001**		**97,993**		**204,994**

On a word-level, table 5 showed that we detected 455,400 Arabic words in the corpus vs. 649,797 words typed with Latin characters among them 56,298 *Arabizi* words that include numerals. We note that all *Arabizi* words were detected based on the

http://dx.doi.org/10.22453/LSJ-022.1.127-147

occurrence of numbers (2,3,5,6,7 and 9) within a Latin word. While this is not an error proof method, given the fact that not all *Arabizi* words are written with numeral, it is somehow reliable because the Arabic letters replaced by these numbers are frequent in *Arabizi* (Mallas et al. 2008).

Table 5. *Arabizi* words detected based on the occurrence of numbers within a Latin word.

University	Detected Arabizi Words	Arabic Words	Sum of all Latin Words	Total Words	Total Messages
AUB	521	5	8208	8,213	1,876
LAU	1,015	42	72,065	72,107	13,323
LIU	7,254	4,626	74,815	79,441	15,627
LU	30,652	305,139	290,253	595,392	103,913
MU	3,158	70,401	109,105	119,506	23,799
PU	13,641	75,148	149,083	224,231	45,647
Grand Total	56,298	455,400	649,797	1,105,197	204,994

From a user perspective, we notice that the usage of the Arabic keyboard is not consistent with all students in the same university. While only 1 student over 117 in our AUB sample used an Arabic keyboard to type few messages, in LU, MU and PU, the numbers are higher (85 to 100%) with most users utilizing an Arabic keyboard at a time. It is also important to mention that 310 over 1,112 students from our sample have performed *Keyboard-Switching* in the same message and 657 students used Arabic keyboard at least one time as shown in table 6.

Table 6. Numbers and percentages of Arabic keyboard users and Keyboard-Switching.

University	Total Users	Arabic Keyboard users	Users with Keyboard-Switching	Percentage of Arabic Keyboard Users	Percentage of users with keyboard-Switching
AUB	117	1	0	1%	0%
LAU	80	9	5	11%	6.25%
LIU	256	94	25	37%	9.76%
LU	492	417	205	85%	41.66%
MU	121	115	59	95%	48.76%

http://dx.doi.org/10.22453/LSJ-022.1.127-147

University	Total Users	Arabic Keyboard users	Users with Keyboard-Switching	Percentage of Arabic Keyboard Users	Percentage of users with keyboard-Switching
PU	19	19	15	100%	78.94%
USJ	27	2	1	7%	3.70%
Total	1,112	**657**	**310**		
Average				**48.00%**	**27.01%**

To answer our research questions, we performed a statistical analysis to test the following hypothesis:

Hypothesis 1: Is Latin keyboard preferred amongst Lebanese students?

A *paired samples t-test* was run to determine if there were differences between the Keyboard preferences amongst Lebanese students. Scores were normally distributed, as assessed by *Shapiro-Wilk's* test ($p > .05$) and there were no outliers in the data, as assessed by inspection of a boxplot. The Mean usage score of the Latin keyboard (M=15286) and for the Arabic keyboard (M=13999) showed a non-statistically significant difference of 1287; $t(6) = 0.284$, $p > .05$. Consequently, our hypothesis has not been validated, and even if there is a difference between the use of two keyboards, this difference is still not significant enough to say that Lebanese students prefer the Latin keyboard to the Arabic keyboard.

Hypothesis 2: Do students at universities with high tuition fees (LAU + AUB + USJ = 224 students) vs moderate tuition fees (LIU + PU+ MU = 396 students) vs public university (LU = 492 students) prefer Latin keyboard?

In this part and in order to answer our hypothesis we have considered: AUB, USJ, LAU as a single variable representing the 'High tuition fees' universities, PU, LIU and MU as another single variable representing the 'Moderate tuition fees' universities and finally the Lebanese university as the only public university. Referring to figure 2, 99.81% of the students of the 'High tuition fees' universities use only the Latin keyboard, against 57.97% of the students of the 'Moderate tuition fees' universities and finally only 40.14% of the students of the Lebanese university use the Latin keyboard. This shows that our hypothesis has been partially validated since the difference between the usage percentages of 'High tuition fees' on the one hand and 'Moderate tuition fees' (41.84%) and public university (59.67%) on the other hand is remarkable. But the difference between 'Moderate tuition fees' and public university 'Low tuition fees'' is restrained (17.83%).

http://dx.doi.org/10.22453/LSJ-022.1.127-147

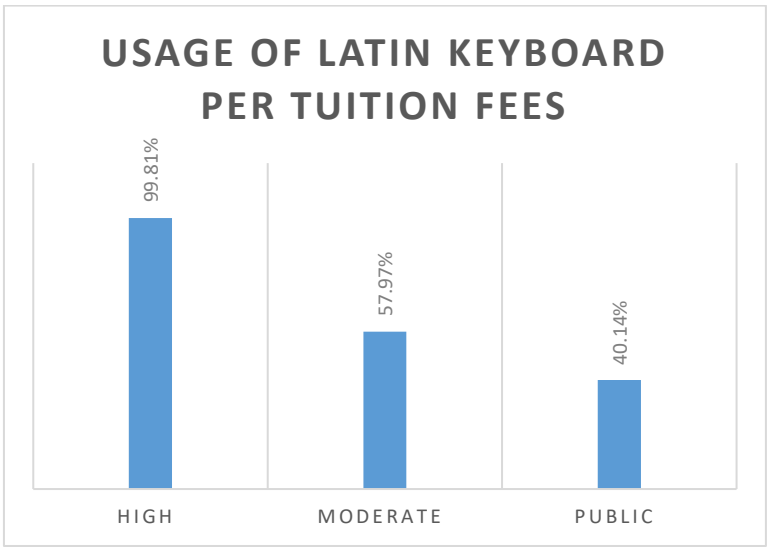

Figure 2. Usage of Latin keyboard per tuition fees.

In conclusion, students at universities with high tuition fees prefer and use Latin keyboard more than students at the public university.

Hypothesis 3: Do students learning their entire curriculum in English or French (AUB + LAU+ PU + USJ = 243 students) compared to mixed language curriculum (LIU + MU = 275) or in Arabic (LU = 492 students) use Latin keyboard more? (Descriptive)

In this part and in order to answer our hypothesis we have considered: AUB, USJ, LAU and PU as a single variable representing the 'English/French curriculum language' universities, LIU and MU as another single variable representing the 'Mixed curriculum language' universities and finally the Lebanese university as the only 'Arabic curriculum language' university. As shown in figure 3, 68.03% of the students of the 'English/French curriculum language' universities use only the Latin keyboard, against 59.23% of the students of the 'Mixed curriculum language' universities and finally only 40.14% of the students of the Lebanese university use the Latin keyboard. Which shows that our hypothesis has been partially validated since the difference between the usage percentages of 'English/French curriculum language' and 'Mixed curriculum language' universities (8.8%) is low and the difference between the usage percentages of 'English/French curriculum language' universities and public university (27.89%) is relatively high. On the other hand the difference between 'Mixed curriculum language' and public university is moderated (18.99%).

http://dx.doi.org/10.22453/LSJ-022.1.127-147

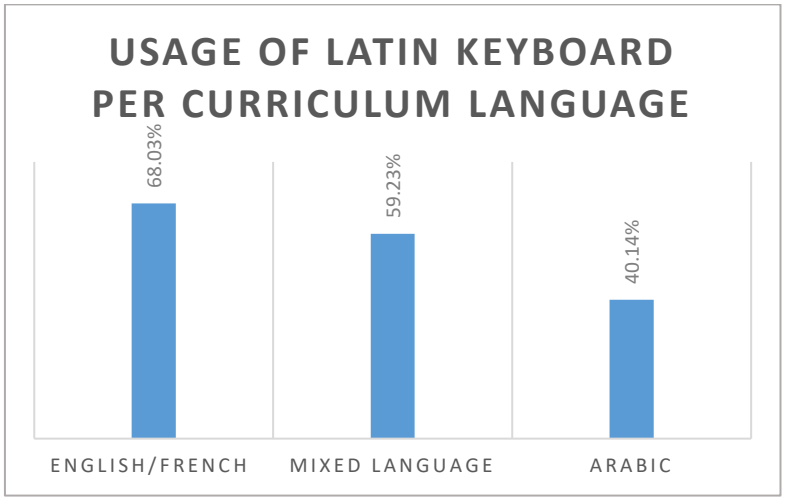

Figure 3. Usage of Latin keyboard per curriculum language.

In conclusion, students with 'English/French curriculum language' use Latin keyboard more than students at public university where curriculum is taught in Arabic.

Hypothesis 4: Students of LU learning in Arabic use keyboard-switching more than other universities?

In order to answer this hypothesis, a *one-sample t-test* was run to determine whether the percentage of *Keyboard-Switching* users at the private universities was different to the percentage of users at the Lebanese university (41.67%). The variable 'percentage of keyboard switching users' were normally distributed, as assessed by *Shapiro-Wilk's* test ($p > .05$) and there were no outliers in the data, as assessed by inspection of a boxplot. The Mean percentage score (24.57%) was lower than (41.67%), a non-statistically significant difference of 17.10%; $t(5) = -1.307$, $p > .05$ which shows that our hypothesis has not been validated, and that even if there is a difference between the Keyboard-Switching between private universities and the public university, this difference is still not significant to say that LU learning in Arabic use Keyboard-Switching more than other universities.

http://dx.doi.org/10.22453/LSJ-022.1.127-147

DISCUSSION

Based on our results, we believe that, while the use of the Arabic keyboard is not very popular amongst some Lebanese universities' students especially where tuition fees are expensive (AUB, LAU, USJ), Arabic is still present in WhatsApp groups discussions of students in Lebanon, however, it is often typed in *Arabizi* using a Latin keyboard. This seems to be the case in all universities, except for our AUB sample.

This article is one of the first studies made on a corpus of instant messages collected in Lebanon. While this research is limited in the number of students in some universities (19 from Phoenicia University and 27 from the Saint-Joseph University), it is rich enough to build an understanding of students' habits in terms of language of communication and keyboard usage and switching, and create leads for further research.

In the Lebanese context, there were no publications about the choices of typing keyboards in WhatsApp students' groups. However, the results of a research done by Tobaili (2016) on '*Arabizi* identification in Twitter data' meet the findings of the current work. He analyzed 60,364 Tweets collected in Lebanon and found that 47% of his corpus is typed in Arabic while the 'Non-Arabic' Tweets include all Tweets typed using Latin keyboard in *Arabizi*, English, French and other languages.

But the results of some studies made in Lebanon across *CMC* platforms are somehow discrepant. Bassam (2017) collected and analyzed a corpus of SMS from 58 subjects in 7 different Lebanese universities. On a message level, her results showed that 65% of the SMS were typed in *Arabizi*, 6% in Arabic script, 26% in English and 2% in French. The presence of the Arabic script in her SMS corpus is very weak (6% of the SMS) while in the current work is much higher (48%).

This could be explained by the context and terms of use of each *CMC* platform. While WhatsApp communication is usually casual, informal and fast, SMSs are short because it is limited in characters per message and the first mobile phones didn't support non-Latin scripts. This could explain the low percentage of Arabic script in the SMS study cited above.

One of the validated hypotheses of this study showed that students at universities with high tuition fees prefer and use Latin keyboard more than students at the public university. This could meet with many researchers' findings about the use of *CMC* means in the Arab world. In some of these countries, students type their messages on social media platforms using Latin keyboard and codeswitch to English because it is a social indicator of belonging to a high class (Salem & Atta, 2013; Akeel 2016; Elsayed & Abdulghaffar, 2016; Bassam, 2017).

http://dx.doi.org/10.22453/LSJ-022.1.127-147

The results of this work showed also that students with 'English/French curriculum language' use Latin keyboard more than students at public university where curriculum is taught in Arabic. Many researches in the *CMC* field explained that students tend to use keywords from their language of education as a « communicative strategy for instruction » (Ndlovu, 2015), and « to fill in the gap in their utterance or exchange » (Abdulbari et al. 2018).

Moreover, this work could be helpful in advancing the *Arabic Natural Language Processing* research by helping in the understanding of new Arabic writing practices in the *CMC* context. Finally, we find this information to be useful for language, marketing and communication scholars and professionals as it tackles an important aspect of language in electronic communication.

CONCLUSION

This study, which arose primary from *CMC* interests, is one of first attempts in the Lebanese context to collect and analyze a large corpus composed of students' WhatsApp groups' discussions. The aim is to determine Arabic and Latin keyboard choices by students of 7 major universities while typing messages in their daily communication. The corpus is composed of 204,994 messages collected by joining 33 WhatsApp groups with 1,112 multilingual students. 107,001 messages from the corpus were typed using Latin keyboard, and 97,993 messages typed with Arabic keyboard only or mixed with Keyboard Switching in the same message.

The total count of words in the corpus is 1,105,197 of which 649,797 words were typed with Latin characters, out of them 56,298 *Arabizi* words that include numerals and 455,400 Arabic words.

It is also important to mention that 310 over 1,112 students from our sample performed Keyboard-Switching in the same message and 657 students used Arabic keyboard at least one time.

The results showed also that students at universities with high tuition fees and those who are learning in foreign languages use Latin keyboard more than students at some faculties of the public university where curriculums are taught in Arabic.

http://dx.doi.org/10.22453/LSJ-022.1.127-147

REFERENCES

Abdulbar, A., Abdulmalik, A., Abdulraheem, M. & Alsabri, S. (2018). 'Types of code-switching between Yemeni dialect and English language among Yemeni undergraduate students at university of Sheba region'. *Language in India*, 18(49042), 1–10. *www.languageinindia.com.*

Akeel, E. S. (2016). 'Investigating code switching between Arabic/English bilingual speakers'. *English Linguistics Research*, 5 (2). https://doi.org/10.5430/elr.v5n2p57.

Al-Badrashiny, M., Eskander, R., Habash, N., Rambow, O. (2014). 'Automatic transliteration of romanized dialectal Arabic'. In Proceedings of the Eighteenth Conference on Computational Natural Language Learning.

Alghamdi, H. and Petraki, E. (2018). 'Arabizi in Saudi Arabia: A deviant form of language or simply a form of expression?' *Social Sciences*, 7(9). https://doi.org/10.3390/SOCSCI7090155

Allehaiby, Wid H. (2013). 'Arabizi: An analysis of the romanization of the Arabic script from a Sociolinguistic Perspective'. *Arab World English Journal AWEJ*.

Attwa, M. (2012). 'Arabizi: a writing variety worth learning?'. MA thesis, The American University in Cairo, School of Humanities and Social Sciences. Retrieved from http://dar.aucegypt.edu/handle/10526/3167

Balius, A. (2013). 'Arabic type from a multicultural perspective: Multi-script Latin-Arabic type design'. Thesis for the degree of Doctor of Philosophy, University of Southampton Research Repository ePrints Soton, Retrieved from https://eprints.soton.ac.uk/355433/1/Final%2520PhD%2520thesis_Andreu%2520Balius.pdf

Baly, R., Badaro, G., El-Khoury, G., Moukalled, R., Aoun, R. and Hajj, H. (2017). 'A characterization study of Arabic Twitter data with a benchmarking for state-of-the-art opinion mining models'. Proceedings of The Third Arabic Natural Language Processing Workshop (WANLP).

Bassam, L. (2017). 'Gender differences in SMS code-switching by Lebanese undergraduates'. Doctoral thesis, UNIVERSITAT ROVIRA I VIRGILI, Tarragona.

Bianchi, R. M. (2012). '3arabizi –When local Arabic meets global English on the Internet'. *Virginia Commonwealth University in Qatar*, 2(1), 89–100. https://doi.org/10.4312/ala.1.2.89-100

Bies, A., Song, Z., Maamouri, M., Grimes, S., Lee, H., Wright, J., Strassel, S., Habash, N., Eskander, R. and Rambow, O. (2014). 'Transliteration of Arabizi into Arabic orthography: Developing a parallel annotated Arabizi-Arabic script SMS/Chat corpus'. In Proceedings of the EMNLP 2014 Workshop on Arabic Natural Language Processing (ANLP). https://doi.org/10.3115/v1/W14-3612

http://dx.doi.org/10.22453/LSJ-022.1.127-147

Bortzmeyer, S. (2012). 'Multilingualism and Internet governance. Net.LaNg towards the Multilingual cyberspace'. *ReaserchGate*, C&F éditio. https://doi.org/CI-2005/WS/06 CLD 24821).

Bou Tanios, J. (2016). 'Language choice and romanization online by Lebanese Arabic speakers'. MA thesis Universitat Pompeu Fabra, Barcelona. Retrieved from http://repositori.upf.edu/handle/10230/27669

Carter, K. A. (2003). 'Type me how you feel: Quasi-nonverbal cues in computer-mediated communication'. *ETC: A Review of General Semantic*, *60*(1), 32.

Darwish, K. (2013). 'Arabizi detection and conversion to Arabic'. *ANLP*, page 217, Retrieved from http://arxiv.org/abs/1306.6755

Elsayed, A. S. A. (2016). 'Code switching in WhatsApp messages among Kuwaiti high school students'. Arab World English Journal. Master's Thesis. No. 185.

El-Zahraa, F. (2014). 'Al-Dzhawahir Al-'Arabiziyah : Muhawalah Li Ta'Alum Al-Lughah Infiradiyan Min Khilal Al-Internet'. *Arabiyat*. https://doi.org/10.15408/a.v1i2.1146

Eskander, R., Al-Badrashiny, M. Habash, N. and Rambow, O. (2014). 'Foreign words and the automatic processing of Arabic social media text written in Roman script'. In First Workshop on Computational Approaches to Code Switching.

Gulacti, U., Lok, U., Hatipoglu, S. and Polat, H. (2016). 'An analysis of WhatsApp usage for communication between consulting and emergency physicians'. *Journal of Medical Systems*, *40*(6). https://doi.org/10.1007/s10916-016-0483-8

Habash, N., Diab, M. and Rambow, O. (2012). 'Conventional orthography for dialectal Arabic'. Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC-2012), (January 2012), 711–718. Retrieved from http://www.lrec-onf.org/proceedings/lrec2012/pdf/579_Paper.pdf

Hoyek, S. (2004). 'Le français dans l'enseignement scolaire et universitaire du Liban', in Cahiers de l'Association internationale des études françaises, volume 56, pp. 49-56.

Kanaan, L. (2011). 'Reformulations, contacts de langues et compétence de communication: analyse linguistique et interactionnelle dans des discussions entre jeunes Libanais francophones'. Doctoral thesis. Université d'Orléans, 2011. Français. ffNNT : 2011ORLE1122ff. fftel-00747329f

Kenali, M., Yussof, N. M. R. N., Kenali, H., Kamarudin, M. S. and Yusri, M. (2016). 'Code-mixing consumptions among Arab students'. *Creative Education*, 7, 931-940. http://dx.doi.org/10.4236/ce.2016.77097

Mallas, T., Taifour, S. and Gheith A. (2008). 'Toward optimal Arabic keyboard layout using genetic algorithm'. Proc. 9th Int'l Middle Eastern Multiconf. on Simulation and Modeling (MESM 2008), 50-54.

Masmoudi, A., Habash, N., Ellouze, M., Estève, Y. and Hadrich Belguith, L. (2015). 'Arabic transliteration of romanized Tunisian dialect text: A preliminary

investigation'. In Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). https://doi.org/10.1007/978-3-319-18111-0_46.

Naji, N., & Allan, J. (2016). On Cross-Script Information Retrieval. In European Conference on Information Retrieval (pp. 796-802). Springer, Cham.

Ndlovu, A. (2015). 'Code-switching in WhatsApp chat messages in Botswana', 40-132.

Pérez-Sabater, C. (2015). 'Discovering language variation in WhatsApp text interactions'. *Onomazein*. https://doi.org/10.7764/onomazein.31.8

Petitjean, C. and Morel, E. (2017). 'Hahaha: Laughter as a resource to manage WhatsApp conversations'. *Journal of Pragmatics*, 110. https://doi.org/10.1016/j.pragma.2017.01.001

Rodrigues, P. (2012). 'Processing highly variant language using incremental model selection'. ProQuest Dissertations and Theses.

Salem, A. A. M. S. (2013). 'The impact of technology (BBM and WhatsApp applications) on English linguistics in Kuwait'. *International Journal of Applied Linguistics and English Literature*. https://doi.org/10.7575/aiac.ijalel.v.2n.4p.64.

Seufert, M., Hoβfeld, T., Schwind, A., Burger, V. and Tran-Gia, P. (2016). 'Group-based communication in WhatsApp'. In 2016 IFIP Networking Conference (IFIP Networking) and Workshops, IFIP Networking 2016. https://doi.org/10.1109/IFIPNetworking.2016.7497256

Shaaban, K. (1997). 'Bilingual Education In Lebanon'. In: Cummins J., Corson D. (eds) Bilingual Education. Encyclopedia of Language and Education, vol 5. Springer, Dordrecht

Shaaban, K., and Ghaith, G. (1999). 'Lebanon's language-in-education policies: From bilingualism to trilingualism'. Language Problems & Language Planning. https://doi.org/10.1075/lplp.23.1.01leb

Sullivan, N. (2017). 'Writing arabizi: Orthographic variation in romanized Lebanese Arabic on Twitter'. Master thesis. The University of Texas at Austin. Retrieved from: https://doi.org/10.1017/CBO9781107415324.004

Tobaili, T. (2016). 'Arabizi identification in Twitter data'. *Association for Computational Linguistics*. ISBN 9781510827608

Wehbe, O. (2017). 'Questions que pose une didactique plurilingue au Liban: pratiques et representations'. Doctoral thesis, Université Sorbonne Nouvelle - Paris 3.

Yaghan, M. A. (2008). 'Arabizi: A contemporary style of Arabic slang'. *Design Issues*, 24(2), 39–52. https://doi.org/10.1162/desi.2008.24.2.39

Yus, F. (2011). 'Cyberpragmatics. internet-mediated communication in context'. (E. A. F. U. of Würzburg, A. Editor, A. H. J. U. of Zurich, & Founding, Eds.). p.174, Amsterdam / Philadelphia: John Benjamins Publishing Company.

http://dx.doi.org/10.22453/LSJ-022.1.127-147