A Study of Arabic Keyboard

Noor Sawalha[1], Majdi Sawalha[2]

[1] Ministry of Education, Jordan
[2] King Abudllah II School for Information Technology, The University of Jordan, Amman, Jordan
E-mail: n_sawalha@yahoo.com, sawalha.majdi@gmail.com

Abstract

Computer keyboard is considered to be the main input device. This study investigates the layout of Arabic keyboard and its letter position relationship with letter frequencies. The main idea of this research depends on computing probabilities of bi-gram letters. It had been found that there is a strong relationship between the letters in the center of the keyboard and their frequencies in the language. The question, "why are letters arranged in the recently used Arabic keyboard?" was answered. The results indicate that the letters of home row have highest frequency scores compared with other letters. Also, the results show that letters at the positions of the index fingers (*i.e.* letters at the center of the keyboard) have highest bi-gram probabilities. This proves that there is a strong relationship between the groups of letters that are assigned to each finger in the typing process. The second question investigated in this research was "what was the base of arranging letters on Arabic keyboard?" This question was answered by counting letter frequency scores and bi-gram frequency scores of home row letters and the letters positioned in the center of first and third rows. The results show that these letters scored highest letter and bi-gram frequency scores. This may indicate that letter-frequency is the base in designing the Arabic keyboard.

Keywords: Arabic Keyboard, Letter frequency, Typing Process.

1. INTRODUCTION

Every day each person who has a computer uses the keyboard as his/her main input device." Typing is the process of inputting text into the typewriter, computer, or calculator, by pressing keys on the keyboard"[2]. So, there is a close relation between the keyboard and typing, but we ask ourselves: is there any relation between the arrangement of letters on keyboard and the way of typing? Figure 1 shows Arabic keyboard with right and left hand layouts. Right hand is labelled with R and left hand is labelled with L. Each finger is labelled with a number. For example, R4 denotes the little finger in the right hand. As we see in the figure each group of letters that corresponds to a certain finger in the typing process is grouped together and assigned a unique colour. For example, the finger labelled with R4 is responsible for typing the group of letters (د، ج، ح، ط، ك، ظ) which is designated by red colour. As we have seen in the figure each finger in each hand is responsible for typing a certain group of letters. The question that arises is there any relationship between the letters in each group? For example, is there a relationship between the letters (د، ج، ح، ط، ك، ظ) that R4 finger responsible for typing?

1.2 Objectives

The way of typing and the arrangement of letters on the Arabic keyboard is the main motivation toward doing this study.

• Why are the letters arranged as in the recently used Arabic keyboards?

• What is the base of arranging letters on the Arabic keyboard?

• Was the letter frequency adopted to be the base of the Arabic keyboard arrangement?

• Is there a relationship between the groups of letters that are assigned to each finger in the typing process?

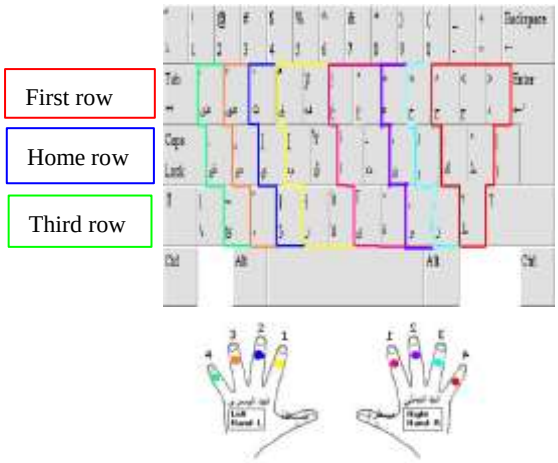


Figure 1: The relationship between each finger and group of letters.

2. LITERATURE REVIEW

The layout of the computer keyboard which we use nowadays was developed hundred years ago. The layout of the computer keyboard came from the typewriter keyboard. The first person who was very interested in designing an efficient layout for typewriter keyboard was Christopher Sholes in 1874 [1, 2]. One problem with the first typewriter machine was that the keys jammed when the operator types at any speed. So Sholes invented what was to become known as the Sholes keyboard or QWERTY. Shole depended on a study of letter-pair-frequency was Amos Densmore did it [1] .

The QWERTY keyboard greatly reduced the problem of jammed keys . After that in 1878 Remington's [1] added shift keys which offer typing upper and lower case letters. Alternative keyboards were designed and developed decades after that . Blickensderfer arranging letters according to frequency which was the base of Devorak's layout. Dvorak's home row uses all five vowels and the five most common consonants: AOEUIDHTNS .Devorak put the vowels on one side and consonants on the other,the problem of Devorak's layout is when each hand would alternate between sides the typing process will be rough and not easy.

Hartmut Goebel [3] presented a layout called NEO developed in 2004. It is an ergonomic layout, established for the German area (and western/European area/Unicode) linguistic and usable for English and with small adjustments of special characters for other languages. Goebel designed certain paradigms for his layout. He depended on the cryptographic statistics of the frequency distribution of the letters in the German language, the resulting arrangement of the keys are more or less.

Goebels [3] developed a small statistic program, depends on word statistics and ergonomic scientific investigations to obtain the results .Goebels matched German words on the QWERTY keyboard and scores 75 words, then on Dvorak 1400, and on the NEO design which scores over 3600.

Shumin Zhai and Per-Ola Kristensson [5] proposed a method for enhancing the speed of computer-based writing, called SHARK (shorthand aided rapid keyboarding), which enhances using stylus keyboarding with shorthand gesturing. Barton A.Smith and Shumin Zhai[6]designed anew keyboard layout based on alphabetical ordering tendency and with a little movement efficiency cost. Their experiment improves novice users' performance and was used by most participants. Issues of keyboard layout, and issues of large keyboards were considered by Martin Hosken [7] large keyboards are those with more characters to be typed than keys to type them, Hosken looked at sequence checking to express different types of keyboard as rule systems.

There are no studies in literature denoted to using letter frequencies in Arabic keyboard design, but there are some studies about letter frequencies in another areas. Essam El-Dessoki and Darwish Abo-Ghuneem [4] studied arabization issues in communication systems , they suggested algorithms for encoding the Arabic characters to be used communication in order to reduce the total number of bits that must be transmitted. In order to do that they calculated the frequency of Arabic letters in two different corpuses from Qura'n Suras and military sub.

3. THE METHODOLOGY STEPS

Typing is not a random process. There are sophisticated rules that the typists must use. As we know, the alphabetical letters are distributed on the keyboard on three rows: the middle row is called the home row which performs the reference point in typing because the typist puts the fingers of his/her hands on each key of this row. For example, the home row of the QWERTY keyboard consists of the letters and punctuation marks (A,S,D,F,G,H,J,K,L,;,') and for the Arabic keyboard the home row comprises the letters (ش ، س ، ي،ب ،ل ، ا ، ت ، ن ، م ، ك ،ط). The basic idea underlying typing is to put fingers on the keys of the home row and the thumbs on the space key and then move fingers up and down to press the key you want to type. As you can see in Figure 5 each finger is responsible for typing a certain group of letters. Table 1 shows each finger label and the group assigned to it. An important issue that has to be addressed is the arrangement of letters on each row. For example, the finger R4 is responsible for typing the letters (ح ، ج ، د) in the first row and (ك ، ط) in the home row and (ظ) in the third row. So, what is the relationship between these letters? Is there a relationship between letters in the same row?

Table 1: Groups of letters that are assigned to each finger in both hands.

Finger	Group of letters
R4	ظ، ك، ط ، ح ،ج ،د
R3	ز،م ، خ
R2	و ، ن ، ه
R1	ة ، ا ، ت ، غ ، ع
L4	ئ،ش ، ذ ، ض
L3	ء ، س ، ص
L2	ؤ ، ى ، ث
L1	ر،لا،ب ، ل ، ق ، ف

The goal of this paper is to study the relationship between the Arabic keyboard layout and the typing process according to letter frequencies. To achieve this objective and others a methodology was designed and implemented carefully in order to achieve good results Figure 2 shows the steps of the methodology.

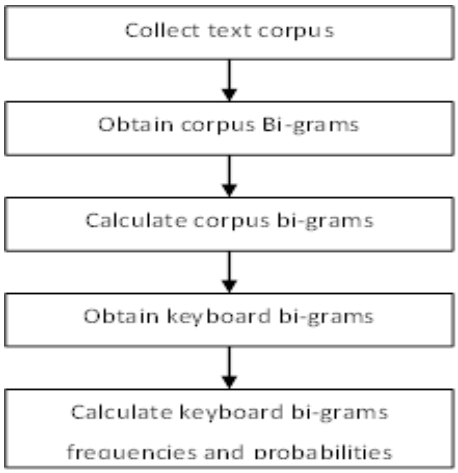


Figure 2: Methodology steps

3.1 Text Corpus

Studying Arabic letter frequencies is the base of this methodology. In order to do that, I collected a corpus of texts which were sampled from an online news articles related to Alrai newspaper (www.Alrai.com) which is a Jordanian daily newspaper. we collected 425 articles covering a range of topics from sports, politics, economics, arts, and social subjects. The articles were drawn from editions of the newspaper. The size of the corpus is about 201532 words and about 806129 characters.

3.2 N-gram

Language is made of words, each with its own separate piece of information; it consists of sequences of words, (*i.e* individual words and phrases of two, three or more words together). N-grams are any substring of length n.

3.3 Corpus Bi-Gram Frequencies

The most important step in this approach is the process of obtaining bi-grams frequencies of Arabic texts. Bi-gram refers to words consisting of two letters. We need to study bi-grams because the structure of keys on keyboard is a sequence of adjacent keys. In the typing process the fingers move left or right up or down to reach a specific key, so it would be assumed that the best arrangement of the keys is to put the bi-grams with high probability adjacent to each other. The bi-grams obtained from the corpus by implementing a program that take the text corpus as input, and then tokenizing each word in it to its bi-grams. After obtaining all corpus words bi-grams the frequency of all distinct bi-grams is calculated and the probability of each bi-gram is also computed. The probability *P*(A and B) is numerically calculated by the ratio between the frequency of *A-B* big-gram and the total frequencies of bi-grams in the text corpus.

$$P(A\,and\,B) = \frac{Frequency\ of\ AB\,bi\_gram}{Total\ \ frequencies\,of\ bi\_grams} \quad \ldots (1)$$

The conditional probability P(B/A) which means the probability of B given that A has already occurred is given by:

$$P(B/A) = \frac{P(A\,and\,B)}{P(A)} \quad \ldots\ldots\ldots\ldots\ldots (2)$$

Where:

- P(A and B): is the probability of A and B occurring together.
- P(A): is the probability of A.

Conditional probability helps in estimating the probability of the occurrence of which letter coming after a letter already occurred.

3.4 Keyboard Bi-Grams

The last important step in the methodology is obtaining the keyboard bi-grams. First, the bi-grams of each row on the keyboard was obtained, then for each cluster of letters of each finger as given in Figure 1 and Table 1 .Finally, the frequencies and the conditional probability of each bi-gram was calculated. This step is considered to be the most important one to get the results. All possible bi-grams on the keyboard along with their probabilities have been listed in Table 3. These bi-grams have been derived from the keyboard in two directions from right to left (RL) and from left to right (LR), also with horizontal (H) and vertical (V) orientation. Figure 8 shows the directions of obtaining bi-grams. Deriving bi-grams horizontally means taking two adjacent keys in the same row. For example, the bi-grams derived horizontally from the first row are (دج، جح، حخ، خه، هع،عغ ،غف ، فق ، قث ، ثص ، صض) while deriving bi-grams vertically means taking two adjacent] keys in adjacent rows. For example, the bi-grams derived vertically from the first and second rows are(دط، جط، كج، حك، حم، خن،خم، هن، هت، عت، غا، عا، غل، فل، فب، قب، قي، ثي، ثس، صش، ضش). In Table 3 row refers to the position of the first letter of bi-gram.

Table 3: Some of Keyboard Bi-grams

Bigr	Row	prob_	Prob_B_aft	Directi	Orientat
ال	2	0.087	0.5067	LR	H
عا	1	0.005	0.1686	LR	V
من	2	0.010	0.1612	LR	H
بي	2	0.005	0.1592	LR	H
سي	2	0.003	0.1523	RL	H
لا	2	0.017	0.1479	RL	H
غا	1	0.000	0.1429	LR	V
قي	1	0.002	0.1176	LR	V
رل	3	0.000	0.0021	RL	H
يء	2	0.000	0.0013	LR	V
ؤء	3	0	0	LR	H

4. RESULTS

The results indicate that there was about 64255 bi-grams. Table 4 shows the first ten bi-grams which have the highest frequencies in the text corpus, as we notice that the bi-gram with the highest frequency is (ال) (AL) with approximately 0.0873.

Table 4: highest 10 Bi-gram frequencies

Bi-gram	Frequency	Probability
ال	58310	0.0873
لا	11844	0.0177
لم	11390	0.0171
ية	10653	0.0159
وا	9274	0.0139
ان	8511	0.0127
ات	7906	0.0118
لة	7763	0.0116
في	7678	0.0115
ار	7283	0.0109

4.1 Conditional probability of bi-grams P(B/A)

The conditional probability was calculated by applying equation No.2. Table 5 shows a sample of the conditional probability for two letters (ل ،ا), i.e. (A is (ال) which is the first letter of bi-gram). Bi-grams are ordered according highest conditional probability.
For Table 5:

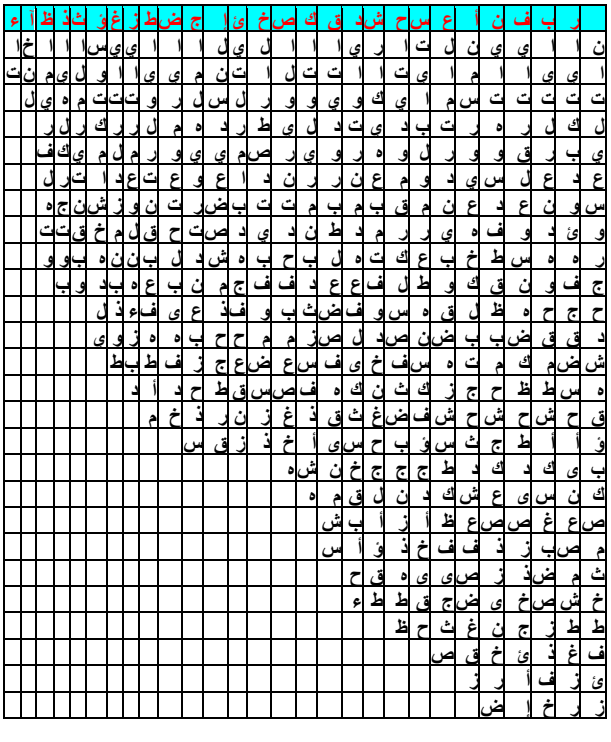
A: represents first letter in bi-gram.
B: represents second letter.
AB: represents bi-gram.
P_A: the probability of the first letter.
P_AB: the probability of bi-gram.
P_B_after_A: the probability of B coming after A.

Table 5: Conditional probability

A	B	AB	P_A	P_AB	P_B_after_A
ا	ل	ال	0.1723	0.0873	0.5067
ا	ن	ان	0.1723	0.0127	0.0737
ا	ت	ات	0.1723	0.0118	0.0685
ا	ر	ار	0.1723	0.0109	0.0633
ا	م	ام	0.1723	0.0084	0.0488
ا	س	اس	0.1723	0.006	0.0348
ا	د	اد	0.1723	0.0054	0.0313
ا	ع	اع	0.1723	0.0048	0.0279
ا	ب	اب	0.1723	0.004	0.0232
ا	ق	اق	0.1723	0.0039	0.0226
ا	ء	اء	0.1723	0.0036	0.0209
ا	ف	اف	0.1723	0.0034	0.0197

Table 6 summarizes the results in Table 4; it shows if a letter occurred what the probable letters will come after it. For example letter (ء) the probable letters to come after it are (ا ، ت) .

Table 6: bi-grams arranged according conditional probability



4.2 Probability of single symbol P(A)

In equation 2, there was a need to find P(A) which means the relative number of occurrences of a single symbol A in text corpus it is given by the equation:

$$P(A) = \frac{\text{No. of occurences of } A}{\text{Total no. of letters in the sample text}}$$

Table 7 represents the probability for each symbol on keyboard.

Table 7: symbol probability

Letter	Freq.	Probability
ا	138909	0.1723
ل	96526	0.1197
ي	60654	0.0752
م	51519	0.0639
و	42864	0.0532
ن	40253	0.0499
ر	38140	0.0473
ت	37298	0.0463
ة	28620	0.0355
ع	28185	0.035
ب	25282	0.0314
د	25170	0.0312
ف	19842	0.0246
س	19567	0.0243
ق	17779	0.0221
ه	17427	0.0216
ح	14160	0.0176
ك	13703	0.017
ج	11161	0.0138
ش	8169	0.0101
أ	8026	0.01
ط	7955	0.0099
ص	7303	0.0091
ى	7073	0.0088
خ	6272	0.0078
ز	5261	0.0065
ض	5225	0.0065
ذ	4685	0.0058
ث	4279	0.0053
ئ	3432	0.0043
غ	2811	0.0035
ء	2663	0.0033
إ	2531	0.0031
ظ	1574	0.002
ؤ	1375	0.0017
آ	436	0.0005

Hassan E. and Al-Soualhi A. [8] EL-Dessoki O. and Abo Ghoneem D. [4] studied Arabic letters frequencies taking different samples of texts and the results were compared with the results in table 7. It appears that the results of this study are approximately similar to the results reported in these two studies.

4.3 Evaluation

The purpose of this step was to find the relationship between letters frequency and their arrangement on keyboard. In the typing process, the most important row is the second one which lies on the center of the keyboard on which fingers deal with as the reference point in the typing process.

we conclude that the bi-grams with highest probability are those which lie at the center of keyboard. For example the first tenth highest bi-grams are (ال، عا، من ، سي، بي ، لا، غا ، قي ، ثي ، غل) which lie exactly at the center and have high probability to occur in the Arabic language. This answers the question is there any relationship between letters frequencies and their arrangement on keyboard.

In order to show if there is a relationship between the typing process and the letters on keyboard, all possible bi-grams were derived from the letters in each cluster as shown in Table 8.

Table 8 : Clusters of bi-grams for the Left hand and their probabilities

Finger	Bigram	prob(B after A)	Orient	Finger	Bigram	prob(B after A)	Orient
L4	ضش	0	V	L1	بل	0.0764	V
	شئ	0	V		بر	0.0732	V
	شض	0	V		قب	0.0588	V
L3	صس	0	V		فق	0.0407	H
	ءس	0	V		فل	0.0407	V
L2	ثي	0.0943	V		قل	0.0407	V
	ؤي	0.0588	V		لب	0.0309	H
	يث	0.0093	V		لف	0.0284	V
	يؤ	0.0027	V		بق	0.0252	V

It seems from the table that the bi-grams which have the highest probabilities lie in the center of the keyboard for which the fingers L1 and R1 responsible in typing. As we move from the center to the edges we notice that the probability of bi-grams is decreases and for some bi-grams it is zero, which means that they do not occur. For example, all the probabilities of the bigrams cluster which L4 finger is responsible for typing is zero. So, it is clear that there is a relationship between letters frequency, typing process, and Arabic keyboard layout.

5. CONCLUSION

The organization of letters is very important to facilitate the typing process; the ideal organization is to put the most frequent letters at the home row. The question why are letters arranged in the recently used Arabic keyboard, was answered, the results indicates

that the letters of home row have high frequencies compared with another letters. For example, the letters (ا، ل، ب، ت، ي، ن، م) have highest frequency in Arabic language according to the text corpus which was used also the bi-grams of the home row were frequently appeared in the text corpus.

Other questions what was the base of arranging letters on Arabic keyboard, and was the letter frequency adopted to be the base of the Arabic keyboard arrangement, these two questions were the main key questions in the study, when the keyboard bi-grams was obtained and arranged according to highest frequency, the bi-grams of the home row and the center of first and third rows scored highest frequencies. For example the highest ten bi-grams on keyboard are (ال، لا، عا، من، تا، سي، بي، غا، قي، غل، ثي) which are the center of keyboard bi-grams, from this results it may be that letters frequency was the base of arranging letters on the Arabic keyboard.

An important issue was discussed which is the relationship between the groups of letters that are assigned to each finger in the typing process, the results shows that the index fingers L1 and R1 which responsible for typing the letters at the center of the keyboard have bi-grams with highest probabilities to occur.

Another main research question raised in this study was: Is there any relationship between letters frequency, the typing process, and Arabic keyboard layout?.

Bi-gram frequency was taken as a base of this study. The frequencies of bi-grams were extracted from the text corpus and the bi-grams on the keyboard were computed. The results indicated that the most frequent bigrams in the Arabic language (within the limits of the used corpus) lies on the center of keyboard, on which the fingers are placed during typing process.

Computer keyboard is still a rich research area. This study may represent a base for other studies. It may lead to design a new layout for the Arabic keyboard more efficient than the existing one. By rearranging the frequent bi-grams and tri-grams also using Fitt's Law [9] as a tool for improving it, Fitt's law is a robust model of human psychomotor behavior developed in 1954.

Fitts law measures how many words we can type in time unit this will help in testing the new layout and improve the speed of typing. Through this study the main concerns was about the clusters of letters that each finger responsible for typing but what about the letters between each clusters for example the letters between the clusters that R4 and R3 responsible for typing are (حخ، كم، ظز) some studies indicates that the letters between clusters must have heights frequency, this issue can be tested on Arabic keyboard and improved. Designing ergonomic keyboard is another important issue that can be developed in the future.

6. REFERENCES:

[1] Yamada.H (1980),historical study of typewriters and typing methods, journal of processing.

[2](2016,April)wikipedia,[online] https://en.wikipedia.org/wiki/Christopher_Latham_Sholes.

[3] Goebel H.(2005) Ergonomic layout of a standard keyboard"NEO".

[4] Dessouki O. and Abo Ghoneem D. Arabization "Issues in communication systems", Electronics and communication eng. Dept., Cairo university,1997.

[5] Zhai S. and Per-Ola Kristensson ,"Short hand Writing on Stylus Keyboard", Human-Computer Interaction,2002.

[6] Smith A. B. and Zhai Sh., "Optimized Virtual Keyboards with and without Alphabetical Ordering – A Novice User Study", In proceedings of *INTERACT'2001 – IFIP TC13* International Conference on Human-Computer Interaction, Tokyo, Japan, p92-99,2002.

[7] Hosken M., "An introduction to keyboarddesign theory: What goes where, NRSI: computer and writing systems web site,2002.

[8] Hassan E. E. and Al-Soualhi A,"On the synchronization of Arabic Codes", *the 11th National computer conference and exhibition*: computers and productivity, Dhahran: King Fahd university of petroleum and minerals,1989.

[9] Mackenzie, I.S,"Fitt's law as a research and design tool in human computer interaction". *Human computer interaction*,1992.