AN IMPROVED ARABIC KEYBOARD LAYOUT

[1]Amjad Qtaish, [2]Jalawi Alshudukhi, [3]Badiea Alshaibani, [4]Yosef Saleh, [5]Salam Bazrawi

College of Computer Science and Engineering, University of Ha'il, Ha'il, Saudi Arabia.

[1]am.qtaish@uoh.edu.sa, [2]j.alshudukhi@uoh.edu.sa, [3]b.alshaibani@uoh.edu.sa, [4]y.saleh@uoh.edu.sa, [5]s.albazrawi@uoh.edu.sa

ABSTRACT: *One of the most important human–machine interaction (HMI) systems is the computer keyboard. The keyboard layout (KL) dictates how a person interacts with a physical keyboard through the way in which the letters, numbers, punctuation marks, and symbols are mapped and arranged on the keyboard. Mapping letters onto the keys of a keyboard is complex because many issues need to be taken into considerations, such as the nature of the language, finger fatigue, hand balance, typing speed, and distance traveled by fingers during typing and finger movements. There are two main kinds of KL: English and Arabic. Although numerous research studies have proposed different layouts for the English keyboard, there is a lack of research studies that focus on the Arabic KL. To address this lack, this study analyzed and clarified the limitations of the standard legacy Arabic KL. Then an efficient Arabic KL was proposed to overcome the limitations of the current KL. The frequency of Arabic letters and bi-gram probabilities were measured on a large Arabic corpus in order to assess the current KL and to design the improved Arabic KL. The improved Arabic KL was then evaluated and compared against the current KL in terms of letter frequency, finger-travel distance, hand and finger balance, bi-gram frequency, row distribution, and most frequent words. The comparisons proved that the improved Arabic KL was able to outperform the current KL. Based on these results, some conclusions are made and a number of recommendations for future work are suggested.*

Keywords: Human–machine interaction (HMI), Arabic keyboard layout, letter frequency, bi-gram frequency, hand and finger balance

1. INTRODUCTION

Nowadays, most people are familiar with the use of computers. Due to the wide range of everyday tasks in which computers are being used, the issue of human–machine interaction (HMI) is becoming more significant. One of the most important HMI systems is the keyboard. The keyboard has proved to be the most valued computer input device [1, 2]. Typing is the process of inputting text into a typewriter, computer, or calculator, by pressing keys on a keyboard [3]. The way in which the arrangement of the letters on a keyboard, or keyboard layout (KL), affects typing ease and finding a way to improve this for users of Arabic-language keyboards is the main motivation for this study. Efforts to improve the KL date back to the 1930s. Numerous research studies have used several techniques to improve the KL based on the statistical analysis of a given language such as English or Arabic, genetic algorithms, physical keyboard shape or the anatomy of the human hand [4-8]. The usage of eight fingers for typing and the invention of touch-typing has made the design of better layouts a competitive area. The two main objectives in developing a new KL design are to increase typing speed and reduce repetitive strain injury (RSI). Also, the tremendous evolution in electronic data and devices has made the keyboard design domain worthy of even greater interest and has led to numerous studies on various methods of HMI.

The main contribution of this study is the proposal for an improved alternative Arabic KL for the standard Arabic (101) KL, which involved determining how the letters should be placed on the KL in order to minimize the total time required to type a certain amount of text by using statistical metrics. This study conducted an in-depth analysis of the current standard Arabic (101) KL according to finger and hand loads, jumps and alternations, KL rows, Arabic letter distribution, bi-grams, and word frequencies using a large corpora. To improve the Arabic KL, it was necessary to adhere to the following five principles: (i) Letters should be typed by alternating between hands, (ii) for maximum speed and efficiency, the most common letters and bi-grams should be the easiest to type, which means that they should be on the home row, which is where the fingers rest, and under the strongest fingers, (iii) the least common letters should be on the bottom row, which is the hardest row to reach, (iv) the right hand should do more of the typing because most people are right-handed, and (v) bi-grams should not be typed with adjacent fingers [7, 9].

The rest of the paper is organized as follows. Section 2 reviews in detail the most popular KLs for both the English and Arabic languages. Section 3 provides a detailed description of the steps taken to design and propose a new improved Arabic KL. Section 4 presents and discusses the results of the comparisons and evaluations of the improved and current Arabic KLs according to some metrics. Finally, section 5 makes some conclusions and suggestions for future work.

2. Literature Review

The term keyboard layout or KL refers to the way in which letters, numbers, punctuation marks, and symbols are mapped on a keyboard. The English KL was inherited from the mechanical typewriter developed in 1870 by Christopher Latham Sholes [3, 9, 10]. Sholes' KL was alphabetically ordered but it jammed easily because after the user had pressed a key, the corresponding type-bar retracted relatively slowly, and if a second key was pressed quickly thereafter and it was near to the first, it would stick to the first and jam [11]. Jamming could be reduced if the most common two-letter sequences (known as bi-grams) were far apart in the layout, or if the typist slowed down. For example, a layout that encouraged pressing a pair of keys with the same finger [11]. Sholes developed a KL to solve the jamming issue by experimenting with bi-grams and assigning them to opposite sides of the KL [9]. This resulted in the QWERTY layout, which was optimal in avoiding typewriter keys jamming together. As the main aim of the QWERTY layout was to avoid jamming it was mapped to be a slow-typing layout [10, 12]. Later, another layout was developed after several years of intensive research by August Dvorak, which was known as the Dvorak Simplified Keyboard (DSK) [1]. In the DSK, the seven most used letters according to linguistic analyses and measurements were placed under the fingers in the resting

position and the result was optimal in terms of greater speed, reduced fatigue, and easier learning [1, 9]. However, despite these claims, the DSK failed to find widespread acceptance and it failed to replace the QWERTY keyboard [2].

In 1999, Mackenzie and Zhang [13] designed a new, optimized layout, called OPTI. They first placed the 10 most frequent letters in the center of the keyboard, then assigned the 10 most frequent digraphs to the top 10 keys. The placement of all the letters and digraphs was done by trial and error. They later made a further improved 5 × 6 layout, called the 5 × 6 layout or OPTI II. As in their QWERTY estimation, Mackenzie and Zhang used the character–space–character tri-graph approach to handle the multiple space keys. Therefore, they made the same probability miscalculation for the tri-graphs on the OPTIs as they did on the QWERTY. Another KL was designed by Smith and Zhai [14] based on alphabetical ordering

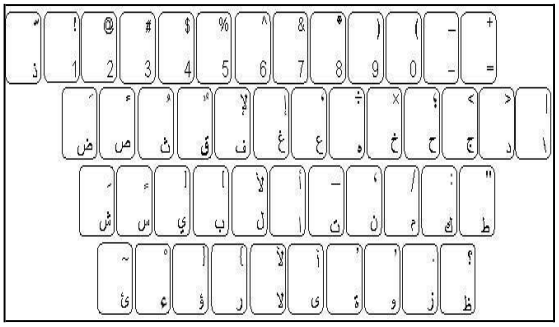


Figure 1. Current standard Arabic (101) KL

tendency and with a little movement efficiency. In addition, a new approach for improving the speed of computer-based writing was proposed by Zhai and Kristensson [15]. This approach was named SHARK (shorthand-aided rapid keyboarding) and was developed to improve stylus keyboarding through the use of shorthand gesturing. An experiment showed that their KL improved novice users' performance and was used by most participants in the study.

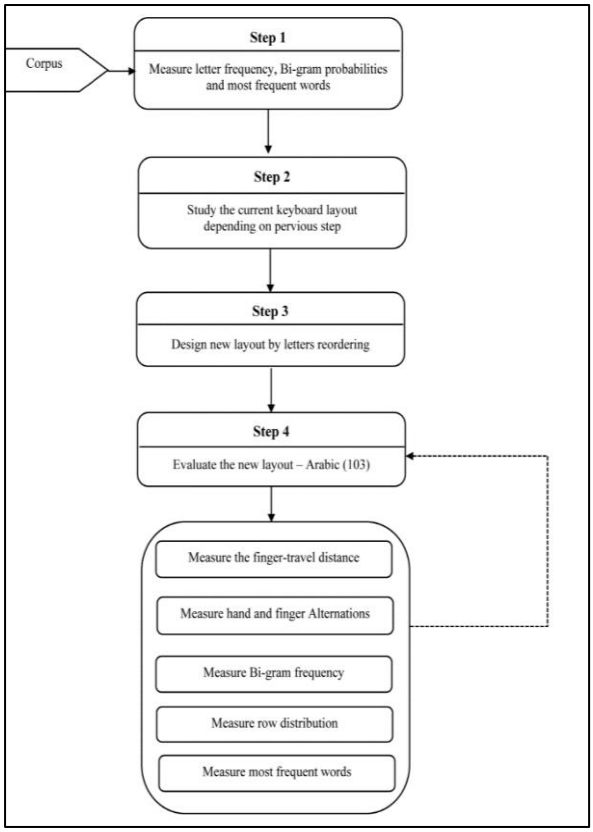
In 2005, Hartmut Goebel [16] proposed a KL named the NEO or ergonomic KL, which was developed in 2004. The NEO layout was established for the German linguistic context. By the cryptographic statistics of the letter frequency distributions in the German language, NEO layout is designed, which results the arrangement of the keys are more or less. Goebel matched German words on the QWERTY, Dvorak, and NEO KLs and scores of 75, 1400, and 3600 words, respectively [17].

The above works illustrate that the optimum position of the characters on a KL is a complex matter. To arrive at an optimum layout requires the consideration of many variables including motion economy principles related to hand and finger movements; finger strength and flexibility, i.e., the human neuromuscular structure. In addition to all of these factors, other issues need to be taken into account such as language constraints, such as letter confusions which result in common spelling errors and then appear as common keying errors, and allowance for statistical frequency of letters, single and combinations of di- and tri-graphs, especially those in the commonest words [5]. Furthermore, for accurate keying and for ease of learning, the keyboard letter layout should take account of the cybernetic requirements of the specific language. The highest source of error in reading and in spelling occurs in relation to vowels and vowel graphemes. On the QWERTY KL the highest source of error is in the vowels "e" and "i" [7]. The Maltron letter layout [18] was developed to solve the above issues.

In respect of the Arabic KL, the order of the letters on the current layout has remained the same as that on the KL that was first designed for the Arabic typewriter in 1914. This means that the current layout has the same problem as the QWERTY KL i.e., slow speed of typing. Currently, there are two types of Arabic keyboard in use: the Arabic (101) and the Azerty (102) standard KLs. The Arabic (101) KL, which was proposed and designed by Microsoft, is considered to be the most commonly used KL (see Figure 1). The only difference between the Arabic (101) and Azerty (102) layouts is the position of the letter (ذ) (*Thal*). However, to the best knowledge of the authors, there is no firm proof as to whether the currently used standard Arabic KL is truly optimal and it is not clear what optimization methods were used in its development. For this reason and because there are other possible ergonomically optimized layouts, the authors were motivated to investigate the possibility of designing an alternative optimal Arabic KL based on ergonomic standards [19]. Another motivating factor was the lack of studies on the usage of letter frequencies in Arabic KL design, despite the existence of some studies on letter frequencies in other areas [17].

3. Methodology Steps

This section describes in detail the steps that were taken to design the proposed new improved KL. As mentioned above, the current Arabic KL was inherited from the typewriter layout, and it is still in use, with an unchanged layout, today. It was therefore important to fully study and analyze the current KL in order to gain a better understanding of how it could be improved. In the first step, two main metrics were computed on a corpus, the letter frequencies and the bi-gram probabilities. The goal of these two computations was to determine the best arrangement of Arabic letters based on the most used letters and the least used letters. In the second step, the letter positions and the letter load distribution among the fingers as well as the hand alternations of the current KL were also analyzed. In the third step, a new layout was constructed based on the results of steps one and two in order to arrange the letters in such a way so as to maximize typing speed and balance the load among the fingers. Then in the fourth and final step, the current and the new improved layout were evaluated according to typing speed by computing the letter frequency and the finger-travel distance, hand and finger balance, bi-gram frequency, row distribution, and most frequent words. Figure 2 shows the methodology steps that were followed in this study in order to design an improved KL.



2. Methodology steps

3.1 Hand rest position

As explained earlier, the mapping of letters onto the keys of a keyboard is a complex matter and it is necessary to consider many variables in order to develop an optimal layout. These variables include the nature of the language, the distance traveled by fingers during typing and finger movements. Moreover, in order to achieve keying at high speed, it is necessary to balance the load between the two hands as well as the fingers while at the same time making some allowance for right-hand dominance and reducing finger motions to a minimum [1, 7].

A standard KL has three main rows, the upper, the home, and the bottom row. The most important row is the home row as this is where the hands are held in a resting position, as shown in Figure 3 and

Figure 4. This position is standard regardless of the KL language or key mapping. The fingers of the hand are generally known as the thumb, index, middle, ring, and pinkie. Each finger is responsible for pressing a certain number of keys.

For example, the right index finger is responsible for typing *Ain Ghain*, *Ta'a*, *Aleph*, *Ta'a marboota*, and *Aleph maqsora*, and another three phonetics of *Aleph*.

Figure 3. Hand rest position on standard KL

The index fingers of both hands are the strongest and the most flexible. They are responsible for typing more letters than the other fingers. Another point to note in respect of the anatomy of the hands in the resting position is that the outside fingers have more freedom to move further so they can reach more keys.

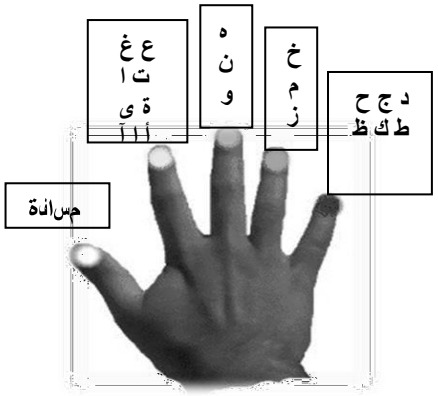


Figure 4. Right-hand letters on standard KL

3.2 Arabic corpus

The Arabic corpus in this study was compiled from several sources, We used the corpus of Arabic text studied and assessed by Goweder and Roeck [20], which is an electronic archive derived from the international Arabic newspaper, *Al-Sharq Al-Awsat*. We also used the Culture, Economics, Science, Industry, Medical, Politics, Religion, Technologies, and Blogs pages from the Maktoob social network webpage as well as the content of six different Arabic dictionaries, namely, *Alfarahidi*, *Alein Dictionary*, *Taj-Alaroos*, *Lesan Alarab*, *Almujam Alwaseet*, and *Almunjed*. Last but not least, The Holy Qur'an was also included in the Arabic corpus. Hence, the corpus contained roughly five million words. For measurement, this paper uses Zipf's law [21], which states that "for a representative sample the graph should be a straight line with slope-1.

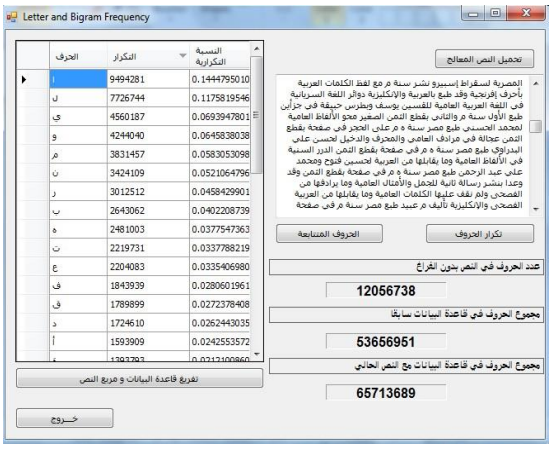


Figure 5. Frequency of single letters

The graph improves as the size of the text increases." (p. 1). A sample of words from the corpus, sorted and ranked, is shown in Figure 5 in the next subsection.

3.3 Letter frequency

In [1], the researchers used a large corpora to obtain the frequencies of English single letters by conducting the initial letter accounts. Similar to [22], we used a large corpora for the Arabic alphabet consisting of 28 main letters. However, there are 36 letters on the Arabic (101)

KL; the eight extra letters are ئ ء ؤ أ آ ى إ ة.

We built a new letter frequency software solution using VB.NET and MS-Access which can deal with a large corpus. The developed software consists of two stages. The first stage of the software takes the input text and processes it by taking only the Arabic letters and eliminating the other characters. This is done by a function that was created to replace each non-Arabic letter with a null value. In addition, unneeded punctuation marks are eliminated and then each multiple space or carriage return is converted into a single space that separates the words in the text. Then, the processed text is saved to a text file. This text file is then loaded into the second stage of the software, which counts each occurrence of each Arabic letter and saves this information into a database file. The resulting database file is accumulative, which means that the letter frequencies of the Arabic letters in first processed text file are added to the new values obtained from the subsequent text files, as shown in Figure 5, which presents a snapshot of the output of the software.

The results of the Arabic single letter frequency analysis were used to calculate the relative frequency or the occurrence probability of each letter according to the following equation:

$$P(l1) = freq(l1) / total\ freq \quad (1)$$

where *P()* is the probability of letter *l*, *freq()* is the frequency (occurrence) of letter *l*, and *total freq* is the total occurrences of all letters in the corpus. Then, these frequencies were sorted and ranked, as shown in Table 1.

Table 1. Frequencies and Probabilities of Single Arabic Letters

Rank	Letter	Frequency	Probability
1	ا	9494281	0.144479501
2	ل	7726744	0.117581955
3	ي	4560187	0.069394780
4	و	4244040	0.064583804
5	م	3831457	0.058305310
6	ن	3424109	0.052106480
7	ر	3012512	0.045842990
8	ب	2643062	0.040220874
9	ه	2481003	0.037754736
10	ت	2219731	0.033778822
11	ع	2204083	0.033540698
12	ف	1843939	0.028060196
13	ق	1789899	0.027237841
14	د	1724610	0.026244304
15	أ	1593909	0.024255357
16	ة	1393793	0.021210086
17	س	1382634	0.021040274
18	ك	1320916	0.020101078
19	ح	1236376	0.018814588
20	ج	987276	0.015023902
21	ص	669503	0.010188182
22	ش	668087	0.010166634
23	ذ	642723	0.009780656
24	خ	541104	0.008234266
25	ى	531818	0.008092956
26	ط	531552	0.008088908
27	ض	495496	0.007540225
28	إ	464446	0.007067721
29	ث	462878	0.007043860
30	ز	443594	0.006750405
31	غ	344269	0.005238924
32	ء	327871	0.004989387
33	ئ	193228	0.002940453
34	ظ	143086	0.002177415
35	ؤ	74261	0.001130069
36	آ	65212	0.000992366

At the end of this process, the 36 letters were grouped into three classes. The first class consisted of the highest ranked 11 letters (ا ل ي و م ن ر ه ب ت ع), the second class consisted of 12 letters (ف ق د أ ة س ك ح ج ص ش ذ), and the third class consisted of the remaining 13 letters which were classified as the least frequent letters (خ ى ط ض إ ث ز غ ء ئ ظ ؤ آ). Table 2 shows the three classes together with their "action" and weight metrics, which are explained below.

Table 2. Classes of Letters for the Improved KL

Class	Letters	Action	Weight
First	ا ل ي و م ن ر ه ب ت ع	0–1	High
Second	ق ف أ د ك س ح ة ج ش ص ذ	2	Mid
Third	خ ط ض ى إ ث ز غ ء أ ظ ؤ آ	3	Low

In order to improve the current KL, the most frequent letters should be placed in the home row. Referring to Table 2, the home row of the current KL contains seven of the 11 (63%) most frequent (first-class) letters, and the remaining four letters (ط ك س ش) are located in the upper and bottom rows of the improved KL. The letters (ك س ش) in the home row of the current KL are in the second class of letters in Table 2 and the letter (ط) in the home row of the current KL is in the third class in Table 2.

To further refine the classification of the letters for the new KL, we added another metric, called action, which represented how many moves are needed to reach the letter. We also added a weight metric to denote the importance of each class (keyboard row). The upper row comes after the home row in importance because it is easier to reach than the lower row. The upper row of the current KL contains 12 letters (د ج ح خ ه ع غ ف ق ث ص ض), six (50%) of which are in the second class in Table 2. The upper row letters in the current KL were replaced by the second-class letters. The bottom row (the weakest row for typing) of the current KL contains 10 letters (ظ ز و ة ى آ ر ؤ ء ئ) and these letters were replaced by the third-class letters in Table 2. After this process had been completed, the bi-gram frequency was measured as in [1, 22].

3.4 Bi-gram frequency

In [1, 16], the researchers used the bi-gram to improve the English and the German KL, respectively. A bi-gram (also called a digraph) is a sequence of two letters. It is used very commonly as a basis for the simple statistical analysis of text using the following equation:

$$P(l_1l_2) = freq\ of\ l_1l_2 / total\ freq\ of\ bi\text{-}grams \quad (2)$$

where *P()* is the probability of letter l_1 coming before letter l_2, *freq()* is the frequency (occurrence) of l_1 l_2, and *total freq* is the total occurrence of all bi-grams in the text.

In this study, the bi-gram measure was used to check the bi-gram combinations that occurred most frequently and then arrange the letters on the new improved KL in accordance with the results, and also to avoid placing bi-grams on the same finger or on successive (adjacent) fingers. The typing speed increases when a bi-gram combination can be typed as quickly as possible, but this does not mean that bi-gram combinations should be placed to each other. For example, the bi-gram (في) should be typed by different hands or different (not adjacent) fingers.

The bi-gram analysis process involved taking each Arabic letter as a combination of two letters from (ا) to (ي), which meant computing the probability of letter (ا) occurring with (ا) to computing the probability of (ا) occurring with (ي) and proceeded until the probability of (ي) occurring with (ي) was reached. This analysis was performed using the same software that we built to determine the frequencies of single letters. The results therefore consisted of 36 * 36 = 1296 states. The software receives the processed text and then for each letter, but not the space it calculates the two combinations of the single Arabic letters. Figure 6 shows a snapshot of the bi-gram results.

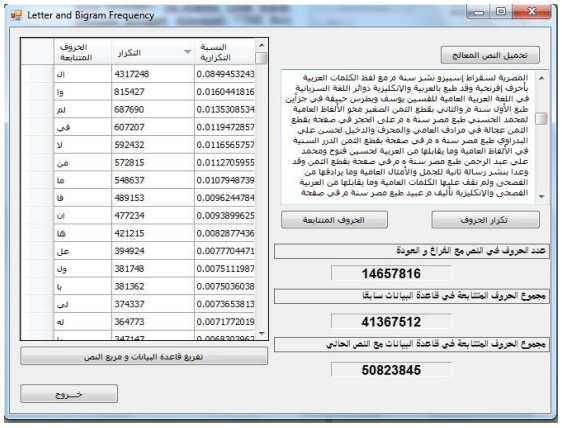


Figure 6. Bi-gram occurrence and probability snapshot

After the bi-gram analysis process was completed, the results were sorted and ranked. Table 3 shows the 50 highest ranked bi-grams. For instance, (ال) is the highest ranked bi-gram. So, its rank is 1. The column titled "Jump" contains the number of rows that need to be jumped to type the bi-grams (if both letters are on the home row the value of the jump = 0, whereas a jump from the home to upper or from the home to lower row = 1, a jump from upper row to lower row or vice versa = 2, and other than that the jump value = 3). For example, to type the bi-gram (هو) on the current KL requires a jump of 2.

Table 3. Fifty Highest Ranked Bi-grams

Rank	Bi-gram	Frequency	Probability	Jump
1	ال	4317249	0.0849453	0
2	وا	815427	0.0160442	1
3	لم	687690	0.0135309	0
4	في	607207	0.0119473	1
5	ال	592432	0.0116566	0
6	من	572815	0.0112706	0
7	ما	548637	0.0107949	0
8	قا	489153	0.0096245	1
9	ان	477234	0.0093900	0
10	ها	421215	0.0082877	1
11	عل	394924	0.0077704	1
12	ول	381748	0.0075112	1
13	با	381362	0.0075036	0
14	لي	374337	0.0073654	0
15	له	364773	0.0071772	1
16	را	347147	0.0068304	1
17	أل	326125	0.0064168	0
18	ين	320038	0.0062970	0
19	اب	314699	0.0061920	0
20	نا	309616	0.0060919	0
21	لت	300713	0.0059168	1
22	ية	294338	0.0057913	1
23	ري	292994	0.0057649	1
24	بي	290492	0.0057157	0
25	ار	290347	0.0057128	1
26	ير	290195	0.0057098	1
27	لى	285739	0.0056221	1
28	أن	280346	0.0055160	0
29	لل	271356	0.0053391	0
30	نه	264601	0.0052062	1
31	اء	258745	0.0050910	1
32	ات	253773	0.0049932	1
33	ام	252497	0.0049681	0
34	وق	250926	0.0049372	2
35	لك	248158	0.0048827	0
36	لع	246166	0.0048435	1
37	مع	238958	0.0047017	1
38	با	237353	0.0046701	0
39	بن	233086	0.0045862	0
40	ذا	230898	0.0045431	3
41	لب	230735	0.0045399	0
42	لح	226577	0.0044581	1
43	ون	225624	0.0044393	1
44	هم	221513	0.0043584	1
45	عن	220187	0.0043324	1
46	لق	216102	0.0042520	1
47	دي	211447	0.0041604	1
48	يه	210267	0.0041372	1
49	ني	206987	0.0040726	0
50	وم	205837	0.0040500	1

3.5 Finger-travel distance

According to [1], the function below is used to calculate the total time necessary to type a certain amount of text as follows:

$$f(layout,lang) = \sum_{c \in letters} (freq(c,lang) * t(c,layout)) \quad (3)$$

where *freq()* is the occurrence probability for character c in a given language and *t()* is the time to reach *c* for the given layout.

A layout analysis cannot, however, be based solely on this kind of time analysis. Even if non-statistical factors are ignored, occurrence frequencies of two- and three-letter structures as well as top row–bottom row jump frequencies should be taken into account if more precise results are desired [1]. So, analysis of the current KL is taken into consideration and all most of its drawbacks are resolved.

In addition, according to [6], the following equation for total finger-travel distance can also be used:

$$f(layout,lang) = \sum_{c \in letters} (freq(c) * dist(c)) \quad (4)$$

where *freq()* is the measured relative occurrence frequency for letter c and *dist()* is the distance to the character as defined above in column "Jump" in Table 3. Furthermore, according to [6], another equation can be used to measure the time taken to reach a certain key, as shown below.

$$T(layout,lang) = n * t * \sum_{c \in letters} (freq(c) * reach(c) * t_{reach}(c)) \quad (5)$$

where *t* is the time needed to press a key, t_{reach} is the time required to reach a certain key, *n* is the total number of different letters and *reach()* is a factor that takes individual finger abilities into account. The *reach()* function assigns reach difficulty to letters ranging from 8 to 10 from the middle to the little finger, respectively. In this case, two

additional factors are taken in consideration as follows: (i) any key press takes a certain amount of time and (ii) not all fingers are equally strong or fast: index fingers are strongest followed by middle fingers and so on.

3.6 Word frequency

For this study we also built a software solution to count the occurrences of each word in the processed text. This software works as follows: First, each word that appears in the text is entered as a single record in the first table by separating each word from another using the space as the terminator of each word. Second, the distinct words (i.e., non-repeating words) are saved in second table. Third and finally, each occurrence of each word is counted and the information is saved in a third table containing the word and its number of occurrences. Figure 7 shows a snapshot of this software solution in action. Table 4 contains the 15 most frequent Arabic words in the corpus that were identified as a result of this process.

ما11845

According to Figure 7 and Table 4 above, two metrics are important for typing speed in order to increase KL efficiency [4, 5]. These are the words most frequently used in the Arabic language and the number of words that can be typed using the home row of the current and improved KL Without any jumps.

Table 4. Top 15 Most Frequent Words in the Corpus

Word	No. of Occurrences
من	43420
أن	34695
في	32475
إلى	29295
أن	27570
أي	24275
إلى	23745
إذا	22935
قال	22785
أو	20620
أي	20520
الذي	18965
الذي	15580
هلا	15545
على	14765
إذا	14280
ان	13295
إن	12480

3.7 Drawbacks of current Arabic (101) keyboard layout

Our analysis of the current KL revealed several key drawbacks:

- Some letters are positioned alphabetically such as (ح خ ج), (س ش ص ض), (ع غ ف ق) and (م ن).
- Some adjacent letters have almost the same shape, which confuses the typist, such as (خ ﺟ ﺣ), (ص ض), (ا ت ي ب), (س ش),(ف ق ع غ),(ئ ء).
- The letters (ط ك س ش) are placed on the home row despite their low frequency of usage compared with first-class letters identified in subsection 3.3, which slows down the typing process.
- Total finger-travel distance is not organized according to frequency.
- There is an uneven balance of finger load, as illustrated in Figure 4, where the right index finger has a heavy load and the frequently used keys are not fairly distributed in respect of the finger responsible for typing them.
- The current layout has many bi-grams that need to be typed by successive fingers, the same finger, and/or the same hand, which results in a slow typing speed.

t

Figure 7. Word frequency software snapshot

- Row distribution is not applied to the home row, which ideally should contain the most used letters, followed by the upper low and finally the lower row, and there is no need to put the letter (ذ) on the number row as this makes it hard to reach this letter.
- The reserved key (ال) is not frequently used because the typist can type the bi-gram (ال) on the home row faster than pressing it on the bottom row.

3.8 An Improved Arabic keyboard layout

The improved Arabic KL is proposed as a way to solve the drawbacks of the current KL. Figure 8 shows the layout of the improved Arabic keyboard. This layout represents the result of an in-depth study and analysis of the occurrence of Arabic letters, the occurrence of bi-grams, the frequencies of Arabic words, and the best way to achieve hand balance when typing.

Figure 8. Improved layout of Arabic keyboard

4. RESULTS AND DISCUSSION

This study compared the improved Arabic KL, with the current Arabic KL, Arabic (101), based on the following metrics:

- Letter frequency and finger-travel distance
- Hand and finger balance
- Bi-gram frequency
- Row distribution
- Most frequent words.

4.1 Letter frequency and finger-travel distance

Using equation (4) in section 2.5, the total finger-travel distance (single letter frequency) of the current KL was calculated and compared with that of the improved KL.

Table 5 shows the comparison of the 36 letters on the current and the improved KL based on finger dist(), travel ratio, reach(), and typing time (sec). According to Table 5, the comparisons show that in terms of the results for total finger-travel distance the improved KL outperforms the current KL (0.759 versus 1.306). Likewise the improved

Table 5. Comparison between Current KL and Improved KL for the 36 Arabic letters based on dist(), travel ratio, reach(), and typing time

Order	Rank	Letter	Frequency	Relative Frequency	Current Layout	Improved Layout	Current Layout	Improved Layout	Current Layout	Improved Layout	Current Layout	Improved Layout
					Dist()	Dist()	Finger-travel Ratio	Finger-travel Ratio	Reach()	Reach()	Typing Time (sec)	Typing Time (sec)
1	1	ا	9510962	0.1445	1	0	0.1445	0.000	8	8	1.156	0.0000
23	2	ل	7726744	0.1176	1	0	0.1176	0.000	8	8	0.941	0.0000
28	3	ي	4560187	0.0694	0	0	0.0000	0.000	8	9	0.000	0.0000
27	4	و	4244040	0.0646	3	0	0.1938	0.000	8	10	1.550	0.0000
24	5	م	3831457	0.0583	0	0	0.0000	0.000	9	10	0.000	0.0000
25	6	ن	3424109	0.0521	0	0	0.0000	0.000	8	9	0.000	0.0000
10	7	ر	3012512	0.0458	3	0	0.1375	0.000	8	10	1.100	0.0000
26	8	ب	2643062	0.0402	0	0	0.0000	0.000	8	8	0.000	0.0000
2	9	ه	2481003	0.0378	2	1	0.0755	0.038	8	10	0.604	0.3775
3	10	ت	2219731	0.0338	0	1	0.0000	0.034	8	8	0.000	0.2702
18	11	ع	2204083	0.0335	2	0	0.0671	0.000	8	8	0.537	0.0000
20	12	ف	1843939	0.0281	2	2	0.0561	0.056	8	8	0.449	0.4490
21	13	ق	1789899	0.0272	2	2	0.0545	0.054	8	10	0.436	0.5448
8	14	د	1724610	0.0262	2	2	0.0525	0.052	10	8	0.525	0.4199
30	15	أ	1593909	0.0243	2	3	0.0485	0.073	8	8	0.388	0.5821
12	16	ة	1393793	0.0212	3	2	0.0636	0.042	9	9	0.573	0.3818
29	17	س	1382634	0.0210	0	2	0.0000	0.042	8	8	0.000	0.3366
22	18	ك	1320916	0.0201	0	2	0.0000	0.040	10	8	0.000	0.3216
6	19	ح	1236376	0.0188	2	2	0.0376	0.038	10	9	0.376	0.3387
5	20	ج	987276	0.0150	2	2	0.0300	0.030	10	10	0.300	0.3005
13	21	ص	669503	0.0102	2	2	0.0204	0.020	10	8	0.204	0.1630
14	22	ش	668087	0.0102	0	2	0.0000	0.020	9	10	0.000	0.2033
9	23	ذ	642723	0.0098	4	2	0.0391	0.020	10	8	0.391	0.1565
7	24	خ	541104	0.0082	2	3	0.0165	0.025	9	10	0.148	0.2470
16	25	ى	531818	0.0081	3	2	0.0243	0.016	10	10	0.243	0.1619
33	26	ط	531552	0.0081	1	3	0.0081	0.024	8	8	0.065	0.1941
15	27	ض	495496	0.0075	2	3	0.0151	0.023	10	10	0.151	0.2262
4	28	إ	464446	0.0071	3	3	0.0212	0.021	8	8	0.170	0.1696
11	29	ث	462878	0.0070	2	3	0.0141	0.021	9	9	0.127	0.1902
31	30	ز	443594	0.0068	3	3	0.0203	0.020	8	9	0.162	0.1823
36	31	غ	344269	0.0052	2	3	0.0105	0.016	8	8	0.084	0.1257
19	32	ء	327871	0.0050	3	3	0.0150	0.015	8	8	0.120	0.1197
35	33	ئ	193228	0.0029	3	3	0.0088	0.009	10	10	0.088	0.0882
17	34	ظ	143086	0.0022	3	3	0.0065	0.007	10	8	0.065	0.0523
34	35	ؤ	74261	0.0011	3	1	0.0034	0.001	8	10	0.027	0.0113
32	36	آ	65212	0.0010	4	1	0.0040	0.001	8	8	0.032	0.0079
Total			65730370	1			**1.3060**	**0.759**			**11.011**	**6.6220**

KL showed better performance compared to the current KL in respect of typing time (6.6220 seconds versus 11.011 seconds).

4.2 Hand and finger balance

Balancing the usage of the fingers on the same hand for typing is a major metric for KL efficiency; the more balanced layout is the more flexible one. The load on each finger was calculated by summing the finger-travel ratio of each letter for a specific finger. Tables 6 and 7 show the loads on the fingers of the right hand and the left hand, respectively. The load on a finger is determined by computing the sum (Distance to reach letters * probability of letters), "letters that a finger is responsible to type".

According to Tables 6 and 7, it took about half the time to reach the letters using the improved KL as compared to the current KL. Specifically, the improved KL was 45% faster than the current KL. The results in Tables 6 and 7 also reveal that the current KL lacks balance among the fingers, which increases the loads on fingers.

The load on the right hand in the current KL is focused on the index and the middle fingers, which increases the risk of RSI in those fingers. As for the left hand, the load is concentrated on the index finger only, which makes the typing process exhausting for this finger. However, in the improved KL the need to achieve a balance between the fingers and between the hands was taken into consideration.

Hand balance is the cumulative result of the balance achieved among the fingers on each hand. In order to increase the typing speed and reduce the chance of RSI, hand balance must be achieved. Figure 9 shows the differences between the two hands, the fingers of the two hands, and the cumulative load when using the current and improved KL. It is evident from Figure 9 that the improved layout decreased the load and achieved an acceptable balance between the hands and the fingers of each hand. In the case of the right hand, the results showed that the improved KL outperformed the current KL by 49% (0.382 versus 0.784). As for the left hand, the results showed that the improved KL also outperformed the current KL by 53% (0.325 versus 0.615).

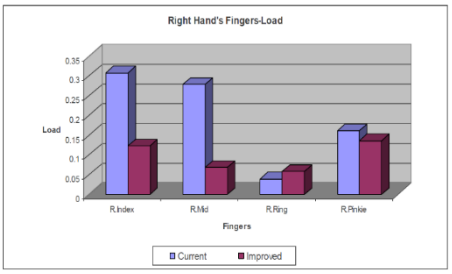
Table 6. Right-hand Finger Loads

Finger	R. Index	R. Mid	R. Ring	R. Pinkie	Total
Current Layout	ع غ ا ت أ إ آ ى ة	و ه ن	ز خ م	ظ ط ك د ج ح	0.784
	0.307	0.278	0.039	0.160	
Improved Layout	ك ا ت ء إ ص أ آ	س ع ط	ح ي ث	ج ش ذ ه م ئ	0.382
	0.122	0.067	0.057	0.136	

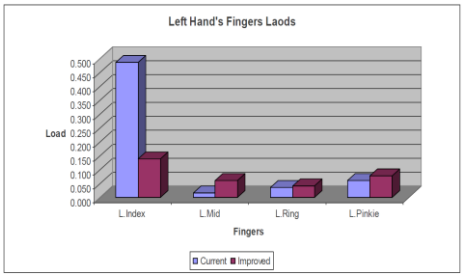
Table 7. Left-hand Finger Loads

Finger	L. Index	L. Mid	L. Ring	L. Pinkie	Total
Current Layout	ق ف ب ل لا ر	ؤ ي ث	ص س ء	ذ ض ش ئ	0.615
	0.489	0.017	0.036	0.073	
Improved Layout	ف د ب ل غ ظ	ة ن ز	ى و ؤ خ	ق ر ض	0.325
	0.141	0.062	0.043	0.079	

a)



b)



c)

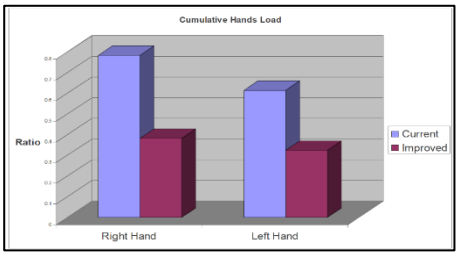


Figure 9. (a) Right-hand finger loads, (b) left-hand finger loads, (c) cumulative hand loads

Bi-gram frequency

The typing performance of the improved KL and the current KL was also compared for the highest 50 bi-grams. The comparisons of the two layouts are shown in Table 8, which shows the bi-gram distribution on the KL according to the use of successive fingers, the same finger and the same hand. As shown in Table 8, the improved layout distributes the bi-grams in a significant way and is therefore much better than the current layout. Furthermore, as shown in Table 9, the highest 50 bi-grams and the time ratio needed to type each bi-gram is satisfied by adding the ratio of each letter in a single bi-gram, i.e., for the (ال) bi-gram, the time ratio was computed by adding the ratio of (ا) to the ratio of (ل) for the two layouts.

Table 8. Bi-gram Distribution

	Successive Fingers					Same Finger				Same Hand					
Current Layout	وا	ها	لي	من	ر ي	نه	ات	ون	وق	وا	ها	لي	من	ري	في
	نا	ان	ير	بي	وم					نا	ان	ير	بي	وم	ما
	هم	يل								ام	ات	ون	وق	نه	لق
															يل
Improved Layout	وق	ون	اس			هم		ات		ات	هم	اس	ون	وق	ذا
										ام	ها	يا	يه	لق	ما
Total (Current:: Improved)	**12::3**					**4::2**				**19::12**					

Table 9. Time Needed to Type Bi-Grams

Improved	Current	Bi-gram
0.000	0.262	ال
0.000	0.338	وا
0.000	0.262	لا
0.000	0.118	لم
0.056	0.056	في
0.000	0.144	ما
0.054	0.199	قا
0.000	0.000	من
0.000	0.144	ان
0.000	0.144	ها
0.000	0.311	ول
0.038	0.220	با
0.000	0.185	عل
0.000	0.118	لي
0.000	0.282	را
0.073	0.166	لأ
0.000	0.118	له
0.038	0.220	اب
0.000	0.144	نا
0.000	0.282	ار
0.016	0.155	اء
0.000	0.000	ين
0.024	0.126	لى
0.000	0.138	ير
0.000	0.144	ام
0.000	0.235	لل
0.034	0.118	لت
0.038	0.076	بي
0.054	0.248	وق
0.042	0.000	ية
0. 00	0.000	نه
0.000	0.138	ري
0.000	0.185	لع
0.034	0.144	ات
0.040	0.118	لك
0.020	0.184	ذا
0.000	0.194	ون
0.038	0.155	لح
0.000	0.144	يا
0.000	0.000	يه
0.054	0.172	لق
0.038	0.193	لب
0.000	0.118	لن
0.054	0.054	يق
0.042	0.208	اس
0.000	0.067	مع
0. 00	0.194	وم
0.052	0.052	دي
0.000	0.000	هم
0.000	0.118	يل
0.840	**7.391**	**Total**

Figure 10 shows the difference between the current KL and the improved KL in terms of time needed to type the 50 highest ranked bi-grams. The improved KL was confirmed to be eight times better than the current KL (0.840 vs. 7.391).

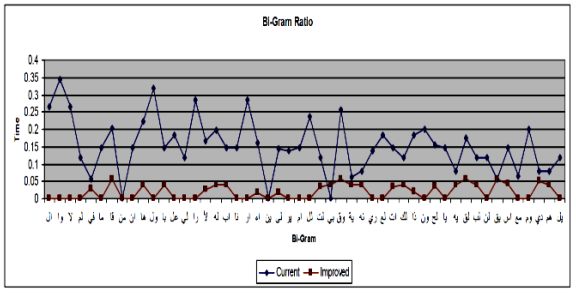


Figure 10. Time needed to type top 50 most frequent bi-grams

Finally, we built a software solution that connected to the database and used the distinct words of the corpus as input. The software solution was used to calculated how many real words could be written by using only the home row of both layouts.

4.3 Row distribution

The rows of a keyboard should be distributed according to the importance of the rows in terms of use. The home row should have the highest usage, then the upper row and finally the lower row. Figure 11 and Figure 12 show the row distribution of the improved KL and current KL, respectively. Note that the numeric row in the current KL includes the letter (ذ).

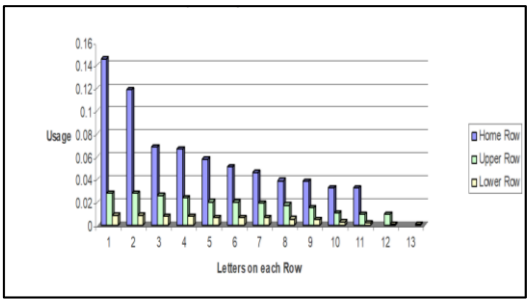


Figure 11. Row distribution of improved Arabic keyboard layout

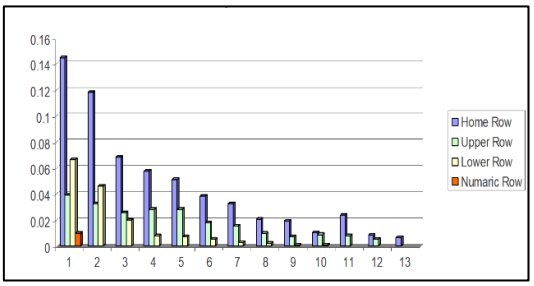


Figure 12. Row distribution of current Arabic (101) KL.

Table 10: Time Needed to Type Top 50 Most Frequent Words

إلى	0.162468624	0.042602511
الى	0.287888726	0.016151925
بين	0.000000000	0.038574512
كما	0.145258041	0.038360878
كل	0.118402797	0.038360878
إلى	0.162468624	0.042602511
مع	0.065984342	0.000000000
حتى	0.060385043	0.085407306

بعد	0.143829134	0.09047104
مثل	0.13244889	0.021069139
أنه	0.126222821	0.06311141
اسم	0.145258041	0.041743489
أيضاً	0.208575234	0.047087343
إلا	0.283498778	0.026450586
إليه	0.21657511	0.065617773
إن	0.01983794	0.026450586
او	0.344839162	0.000000000
محمد	0.114001948	0.088053684
تحت	0.036157156	0.102353607
ذات	0.184246182	0.052592296
أكثر	0.200785852	0.08337424
أنا	0.193146488	0.023944224
أحد	0.161890395	0.111997908
ألا	0.311549285	0.023944224
أمر	0.186739759	0.023944224
أم	0.047888447	0.023944224
أما	0.193146488	0.023944224
أعلم	0.232275587	0.023944224
Total	**7.218041791**	**1.737886774**

Word	Current Layout	Improved Layout
في	0.056725677	0.028362838
من	0.000000000	0.000000000
قال	0.319815239	0.056154401
على	0.208615027	0.016151925
ما	0.145258041	0.000000000
أي	0.047888447	0.023944224
عن	0.065984342	0.000000000
ابن	0.145258041	0.038574512
هو	0.277915495	0.039167187
أن	0.047888447	0.023944224
هذا	0.262580556	0.058661257
له	0.196737171	0.039167187
ان	0.145258041	0.000000000
أبو	0.247469569	0.062518736
أو	0.247469569	0.023944224
لم	0.118402797	0.000000000
ذلك	0.157390939	0.057854949
إذا	0.204084122	0.045944657

4.4 Most frequent words

As for the bi-gram, the time ratio needed to type each word is satisfied by adding the ratio of each letter in a single word. For example, for the word (ابن), the time ratio was computed by adding the ratio of (ا) to the ratio of (ب) and the ratio of (ن) for the two layouts. Table 10 shows the top 50 most frequent words and the time ratio needed to type each word when using the two layouts (current KL, improved KL).

As shown in Table 10 above, the improved layout takes significantly less travel time in typing the most frequent words and is therefore much better than the current layout. In addition, the last row of Table 10 shows that the improved KL was five times better than the current KL (1.74 vs. 7.22). Figure 13 above graphically illustrates the difference between the current KL and the improved KL.

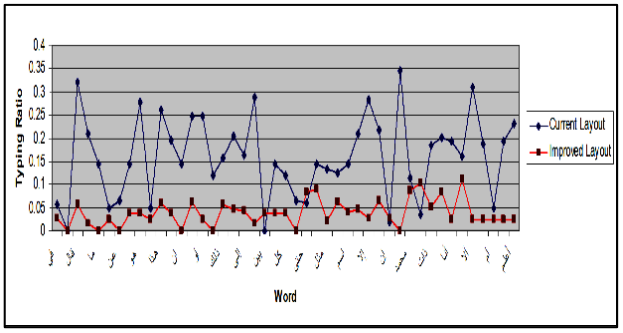


Figure 13. Comparison of time needed to type most frequent words using current and improved KL for the highest 50 words

4.5 Training and feedback on improved keyboard layout

Eight persons volunteered to type text using the new KL. They were divided into two groups. The first group consisted of five persons who had already experienced Arabic typing using the current layout. The second group consisted of three persons who were used to using the "pick and strike" approach in typing in Arabic. The initial feedback from the first group was promising, as illustrated by the following excerpt: "In the beginning it was difficult to learn the new improved KL because of being accustomed to the current layout, but later the new improved KL became easier". The feedback from the second group initial was very optimistic, as typified by the following statement in which they declared that "the new positions of the letters make it easier to pick the letters and there is no confusing of similar letters on the same row."

5. CONCLUSION AND FUTURE WORK

In this study, the current Arabic standard keyboard was analyzed in depth and a lot of problems were identified that affected the efficiency of this legacy keyboard. These findings informed the development of a new and improved Arabic KL, which was designed based on a deep analysis of the letter distances and the frequencies and probabilities of Arabic letters and bi-grams. A large corpus was created for the keyboard development process. This corpus consisted of five million words. The improved KL was evaluated and compared with the current KL based on letter frequency and finger-travel distance, hand and finger load, bi-gram frequency, row distribution, and most frequent words. The comparisons showed that the improved KL was more efficient than current KL.

In future research, the use of a genetic algorithm may need to be considered in order to calculate the word frequency accurately and thereby further reduce typing time and improve the Arabic KL. Finally, it is hoped that the proposed Arabic keyboard will help commercial organizations and individuals in typing Arabic words more efficiently. The improved layout may also serve as a foundation upon which researchers can build to enhance the computer keyboard layout as a part of HMI research and the work undertaken in this study could be extended into new research areas.

REFERENCES

1. NakiC-AlfireviC, T. and M. Durek. "The Dvorak keyboard layout and possibilities of its regional adaptation". IEEE. i*n 26th International Conference on Information Technology Interfaces*, (2004)

2. Pradeepmon, T., V.V. Panicker, and R. Sridharan, "Hybrid estimation of distribution algorithms for solving a keyboard layout problem". *Journal of Industrial and Production Engineering*, **35**(6): p. 352-367(2018).

3. Yamada, H., "A historical study of typewriters and typing methods, from the position of planning Japanese parallels".*Journal of Information Processing*, (1980)

4. Hobday, S. "Keyboard Designed to Fit Hands & Reduce Postural Stress" . *9th Congress of the IEA*, (1985)

5. MIP&I., S.W.H.L.A. "Computer Related Upper Limb Disorder. A Keyboard to Eliminate the Stress & the Pain An Interim Success Report. 1994 [20/03/2018]; Available from:https://www.maltron.com/computer-related-upper-limb-disorder.html.

6. Malt, L.G. "Keyboard design in the electronic era". *in Printing Industry Research Association. Symposium Paper*. (1977)

7. Hobday, S.W., "Keyboard to increase productivity and reduce postural stress", *Elsevier*, (1988)

8. Brewbaker, C.R., "Optimizing stylus keyboard layouts with a genetic algorithm: customization and internationalization". *Dept. of Computer Science, Iowa State University*, (2008)

9. Buzing, P., "Comparing different keyboard layouts: aspects of qwerty, dvorak and alphabetical keyboards". *Delft University of Technology Articles*, (2003)

10. Onsorodi, A.H.H. and O. Korhan, "Application of a genetic algorithm to the keyboard layout problem". *PloS one*, **15**(1): p. e0226611. (2020)

11. Piepgrass, D. "Why QWERTY, And What's Better?" 2006 25/03/2018]; Available from: http://millikeys.sourceforge.net/asset/why-qwerty.pdf.

12. Alswaidan, N., M.I. Hosny, and A.B. Najjar. "A genetic algorithm approach for optimizing a single-finger arabic keyboard layout". in *Science and Information Conference, Springer*, (2014)

13. MacKenzie, I.S. and S.X. Zhang. "The design and evaluation of a high-performance soft keyboard". *in Proceedings of the SIGCHI conference on Human Factors in Computing Systems*. (1999)

14. Smith, B.A. and S. Zhai. "Optimised Virtual Keyboards with and without Alphabetical Ordering-A Novice User Study". *in INTERACT*. (2001)

15. Zhai, S. and P.-O. Kristensson. "Shorthand writing on stylus keyboard". i*n Proceedings of the SIGCHI conference on Human factors in computing systems*. (2003).

16. H, G. "Ergonomic layout of a standard keyboard "NEO" ". 2005 18/08/2018]; Available from: http://pebbles.schattenlauf.de/layout/index_us.html.

17. Sawalha, N. and M. Sawalha, "A Study of Arabic Keyboard". *New Trends in Information Technology (NTIT)*, p. 100, (2017)

18. Hosken, M., "An introduction to keyboard design theory: What goes where? Implementing writing systems: an introduction", edited by Melinda Lyons. Dallas, Texas: SIL, Non-Roman Script Initiative, p. 121-137, (2001)

19. Khorshid, E., A. Alfadli, and M. Majeed, "A new optimal Arabic keyboard layout using genetic algorithm". *International Journal of Design Engineering*, **3**(1): p. 25, (2010)

20. Goweder, A. and A. De Roeck. "Assessment of a significant Arabic corpus". *in Arabic NLP Workshop at ACL/EACL*. (2001)

21. Ha, L.Q., et al. "Extension of Zipf's law to words and phrases". i*n Proceedings of the 19th international conference on Computational linguistics*,**1**,(2002). Association for Computational Linguistics.

22. Jones, M.N. and D.J. Mewhort, "Case-sensitive letter and bigram frequency counts from large-scale English corpora". *Behavior research methods, instruments, & computers*. **36**(3): p. 388-396 (2004)