Using Genetic Algorithms to Design an Optimized Keyboard Layout for Brazilian Portuguese

Gustavo Pacheco[1], **Eduardo Palmeira**[1], **Keiji Yamanaka**[1]

[1]Faculdade de Engenharia Elétrica – Universidade Federal de Uberlândia (UFU)
Uberlândia – MG – Brazil

{pacheco.gustavo.alves,egqpalmeira}@gmail.com, keiji@ufu.br

Abstract. *Currently, keyboards are the most common means of communicating with computers. Despite being the most commonly used keyboard layout, QWERTY has had various issues raised concerning its effectiveness, as it is not efficient in English (target language) or in fact other languages. Therefore, this paper presents the development process of a Genetic Algorithm with the intention of generating a more adequate and coherent layout proposal for Brazilian Portuguese, which has its focus on ergonomics and user productivity. Using five ergonomic criteria and a statistical analysis of the characters and sequences of most frequently used pairs in Brazilian Portuguese, a layout approximately 53% better than QWERTY was obtained.*

1. Introduction

Currently, keyboards are the most common means of communicating with computers. The electronic era has strengthened the unique and dominant position of the QWERTY layout keyboard. This layout had its design conceived in 1874, for typewriters, but it was only in the first decade of the 20th century that it became established as the international keyboard standard. At that time, writers commonly used only their index fingers for typing. The first typists searched for the character and, after finding it, pressed it to make its impression on the paper. It was a slower and more cautious interaction. Only after the 1930s, the typing speed was improved when all ten fingers started to be used as a conventional manner of interaction. Researchers and keyboard designers have raised various concerns with the reasoning behind the choice of the QWERTY keyboard layout. However, there seems to be no logical reason for the arrangement of the characters. Thus, despite some attempts at explanations, doubts regarding its origin remain [Noyes 1983, Noyes 1998].

Also in the 1930s, the efficiency of the QWERTY layout began to be questioned and received several criticisms. The layout has problems distributing workload concerning hands, fingers, and key rows when typing using the English language itself. Some believed that the QWERTY layout would no longer be used in the future as it is not suitable for the English language, and it was not designed with a user-centered design regarding the hands of the typist. Still in the 20th century, design proposals for new layouts began (e.g., Dvorak keyboard [Dvorak and Dealey 1936]), but although some did appear to be superior, none managed acceptance as a standard and overthrow the QWERTY supremacy. Thus, although the QWERTY layout remains unchanged, its design has remained questionable [Noyes 1983, Noyes 1998].

If the QWERTY layout, which was developed based on the English language, is inconsistent with it, then regarding other languages, it does not occur differently. This

limitation encouraged research that both investigated and developed new keyboard layout proposals suitable for other languages [Liao and Choe 2013, Deshwal and Deb 2003, Khorshid et al. 2010]. Such investigations are based on concepts of ergonomics and contemplate the use of Genetic Algorithms (GA) to find an optimized keyboard layout solution.

GA offers good solutions to complex problems. These algorithms are inspired on elements of the theory of evolution, proposed by Charles Darwin, which is to adapt a data set to specific conditions. These are classified as search and optimization algorithms, and their origins can be found at the University of Michigan, through the work of John Holland, along with his colleagues and students. This group researched complex artificial systems, capable of eliminating non-beneficial or harmful characteristics of individuals, while reinforcing the positive ones [Tomassini 1995, Goldberg 1989]. Therefore, this paper presents the development process of a GA with the intention of generating a more adequate and coherent layout proposal for Brazilian Portuguese, which aims at ergonomics and user productivity.

To this end, the remainder of this paper is organized as follows: Section 2 presents other works that attempted to improve the typing, by using a variety of techniques and aiming at different aspects of the keyboard. Section 3 is divided into eight parts, and describes the GA implemented. Worth noting that subsection 3.2 unravels every value used to classify each keyboard, explaining the statistical analysis performed (3.2.1), the functions utilized to express the criteria (3.2.2 to 3.2.7) and, finally, the overall fitness score (3.2.8). The results obtained are presented in subsection 3.8, which are discussed in Section 4, concluding this paper.

2. Related Works

Despite being the most commonly used keyboard layout, a number of issues have been raised concerning QWERTY and its applied use. Dvorak developed one of the most notable alternatives [Dvorak and Dealey 1936]. In 1936, he presented his proposed layout, the Dvorak Simplified Keyboard (DSK), and even today, the model has its quota of users. In developing the layout, Dvorak aimed to reduce typing errors and increase operational speed. Thus, reducing unnatural hand movements and reducing writer's fatigue, while modifying the writing flow, in order to adapt it to the most frequently used English sequences.

Moreover, several other authors have evaluated the keyboard design in order to improve the interaction between the user and the keyboard. Some efforts have been directed towards ambiguous keyboards [Garbe 2000, Oommen and Zgierski 1991], which have a limited set of buttons, less than the number of letters, requiring the user to press the same key several times (e.g., telephone keypad). However, the most significant highlight is in research related to conventional devices, which have a specific button for each letter of the alphabet. For these, two approaches are proposed: concerning the change in physical characteristics, and the other regarding the variation of the keyboard key arrangement.

The first focuses on the physical modification of the equipment, and seeks to improve the user experience by improving the user's posture concerning the device. An example of this aspect is [Heidner 1915], who, in his patent, divided the typewriter keyboard into two parts, providing better visualization of the keys and greater comfort during

typing, by increasing the distance between the hands.

The other approach, in which this work is placed, seeks to modify the arrangement of the keys and optimize the keyboard layout. Most publications in this research area are based on statistical data, assessing the frequency of letters and sequences of each language, and then applying some optimization process, based on specific factors and criteria. Some examples of studies were found in the literature that optimized the keyboard layout for Chinese [Liao and Choe 2013], Hindi [Deshwal and Deb 2003], and Arabic [Khorshid et al. 2010]. However, was not found in literature an attempt to apply this methodology for Brazilian Portuguese, being one of the objectives of this paper to fill this gap.

Furthermore, such related works used GA to find a keyboard layout proposal, and followed the ergonomics criteria set out by [Wagner et al. 2003]. In addition to GA, other optimization strategies have been identified in the literature. Such a case is found in [Eggers et al. 2003], which used an ant colony optimization, and [Light and Anderson 1993], which used simulated annealing. However, although this paper makes use of GA and [Wagner et al. 2003] criteria, which will be explored in the next section, it adapts this methodology according to the peculiarities of Brazilian Portuguese. Also, it adds a new criterion to the fitness function and proposes an execution strategy aiming at statistical reliability.

3. Genetic Algorithm

This section describes the steps incorporated for developing the algorithm. The determination of a keyboard layout, adequate to certain factors, is conditioned to the search for a specific arrangement, within a set of possible key combinations. Due to this solution space size, GA stands out as a viable strategy for performing this search. Thus, each step of this study was in accordance with the ergonomics criteria and the specifications of the Brazilian Portuguese language.

The algorithm process was divided into seven steps. The first, initialization, in addition to the random generation of individuals, covers the data modeling process and representation of the keyboard. The second, evaluation, uses a statistical analysis of Brazilian Portuguese and five of the six ergonomics criteria [Wagner et al. 2003, Eggers et al. 2003], checking letter frequencies and letter pairs in texts from the public domain, in order to evaluate each keyboard layout in the set. Step numbers three to six are responsible for the composition of a new population. Thus, the third, elitism, guarantees the perpetuation of the best individuals. The fourth, selection, selects the most suitable individuals, based on probabilistic processes, so that they undergo genetic operations in steps five and six, crossover and mutation, respectively. Such steps induce crossover between individuals and random mutations. The seventh and last step, end of execution, is responsible for assessing whether the stop condition has been satisfied, or whether the process will continue to run, from the second step, now with a more mature population. Finally, the resulting optimized layout is presented.

3.1. Initialization

GA are developed as iterative programs. In these, the populations of individuals are implemented as data sets, and each repetition step represents a generation of this population.

Table 1. Weight of ergonomic indicators. Adapted from [Wagner et al. 2003]

Criterion	Weight
Tapping workload distribution	0.45
Hand alternation	1.00
Finger alternation	0.80
Avoidance of big steps	0.70
Hit direction	0.60

The group size is constant over the generations, and individuals are designed to abstract the element to be optimized, in which any one of them presents an answer to the optimization problem.

One of the initial challenges of implementing a GA is to represent, in a generic way, this answer in the form of computational data. Individuals, also known as chromosomes, must be designed in such a way that they are susceptible to modifications caused by the genetic operators. Most of these are implemented from some sequences, such as strings or lists. Each element of this structure is called a gene.

For this work, lists of characters represented individuals. A list is a structure of abstract data, in which each individual has a reference to the next element, ensuring ordination. In this structure, there are 27 items. Each gene represents a letter of the alphabet (including 'Ç'). The initial population, composed of 200 individuals, was implemented as a list of randomly generated individuals, starting from a process that shuffles the 27 letters.

3.2. Evaluation

Here, each individual is evaluated, measuring its fitness concerning the established environment. The fitness function must have an individual as input, and an appropriate score as an output. In this study, for each keyboard layout in the set, a number was associated. This value was obtained through the mathematical analysis of five ergonomics criteria. These are tapping workload distribution, hand alternation, finger alternation, avoidance of big steps, and hit direction. Unlike [Wagner et al. 2003], the criterion related to the number of keys per word was not used, as this refers to ambiguous keyboards.

These criteria aim at simulating the typing effort. As this research is directed to Brazilian Portuguese, it was necessary to perform a study concerning the frequency of letters, individual and in pairs, of the language. So, the workload was distributed according to the periodicity, and the alternations and directions favorable to the typing were guaranteed.

Finally, the overall fitness corresponds to the weighted average of the five indicators. The weights of each coefficient are shown on Table 1, obtained from [Wagner et al. 2003]. Low total values represent better individuals. The equations used are similar to those of [Liao and Choe 2013]. Also, another penalty (standard deviation) was added to the current fitness, a procedure that was not found in the literature. This was added to ensure that there were no resolutions with a great overall fitness, but with poor results on specific indicators.

Table 2. Relative Frequency (RF) of the characters

Character	A	E	O	S	R	I	N	D	M
RF (%)	13.47	12.68	10.70	8.16	6.84	6.02	5.11	5.06	4.90

Character	U	T	C	L	P	V	Q	H	G
RF (%)	4.58	4.47	3.28	2.87	2.62	1.75	1.31	1.29	1.23

Character	F	B	Ç	Z	J	X	W	Y	K
RF (%)	1.06	1.04	0.49	0.44	0.35	0.25	0.01	0.01	0.01

3.2.1. Statistical Analysis

The ergonomics criteria used to measure the fitness of an individual require a database regarding the frequencies of letters, individual and in pairs. Thus, identifying which are the most frequently used characters and the most typed sequences.

For this analysis, 42 works were chosen and evaluated from the public domain[1], in Brazilian Portuguese, with relevance placed on the number of times it was accessed. Among these works, there are copies of Machado de Assis, José de Alencar, Fernando Pessoa, among others. From these files, the relative frequencies of each character were collected, ordered in descending order on Table 2. The 26 letters of the alphabet were considered, plus 'Ç' and the relative frequencies of the letter pairs, shown on Table 3.

3.2.2. Tapping Workload Distribution

Considering a constant workload, shared between the hands, it is natural that the division amongst the fingers is proportional to the strength of each one of them. Besides, the keyboard home row, which begins with the character 'A' in Figure 1, should be given priority. The study by [Wagner et al. 2003] proposed an ideal workload distribution, in which weights were allocated to the rows and the columns, relative to the finger responsible for these. Based on this distribution, this paper proposes a new desirable workload distribution, considering the ABNT keyboard specifications [ABNT/CB-021 1991] (Figure 1) regarding the letters. In this proposal, the home row and the following order of fingers are prioritized as index, middle, little and ring.

Unlike the aforementioned research, the objectives of this paper is not to include characters other than letters (e.g., numbers, diacritics and special characters). By doing so, the general organization of groups in the keyboard remains the same, while affecting only the distribution of the one composed by letters. Consequently, there is no reason to consider the thumbs in this distribution, usually designated to the space bar. Therefore, the relative coefficients of the rows and columns are multiplied, determining the desirable workload for each key (Figure 2). Noteworthy here is that the tapping workload is equally distributed between the hands, even if the right one is responsible for a smaller number of keys, according to the division shown in Figure 1.

Equation (1) represents the evaluation function for indicator I_1. It represents how far from the ideal the workload distribution on a keyboard is. I_1 is calculated from the dif-

[1] www.dominiopublico.gov.br

Table 3. Relative frequencies of letter pairs. The lines correspond to the first letter of the pair, and the columns, the second.

	A	B	C	Ç	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z
A	0,0000	0,0027	0,0031	0,0027	0,0099	0,0001	0,0007	0,0022	0,0001	0,0047	0,0003	0,0000	0,0092	0,0090	0,0129	0,0021	0,0027	0,0013	0,0168	0,0200	0,0038	0,0011	0,0055	0,0000	0,0000	0,0000	0,0015
B	0,0028	0,0000	0,0000	0,0000	0,0000	0,0029	0,0000	0,0000	0,0000	0,0011	0,0001	0,0000	0,0002	0,0000	0,0000	0,0018	0,0000	0,0000	0,0033	0,0003	0,0000	0,0005	0,0000	0,0000	0,0000	0,0000	0,0000
C	0,0102	0,0001	0,0000	0,0000	0,0000	0,0045	0,0000	0,0000	0,0025	0,0052	0,0000	0,0000	0,0007	0,0000	0,0000	0,0149	0,0000	0,0000	0,0016	0,0000	0,0002	0,0020	0,0000	0,0000	0,0000	0,0000	0,0000
Ç	0,0024	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0013	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
D	0,0142	0,0000	0,0000	0,0000	0,0000	0,0240	0,0000	0,0000	0,0000	0,0064	0,0000	0,0000	0,0000	0,0001	0,0000	0,0189	0,0000	0,0000	0,0006	0,0000	0,0000	0,0013	0,0001	0,0000	0,0000	0,0000	0,0000
E	0,0009	0,0008	0,0038	0,0006	0,0022	0,0002	0,0007	0,0026	0,0000	0,0079	0,0008	0,0000	0,0085	0,0110	0,0159	0,0003	0,0015	0,0004	0,0161	0,0209	0,0025	0,0050	0,0018	0,0000	0,0014	0,0000	0,0018
F	0,0026	0,0000	0,0000	0,0000	0,0000	0,0025	0,0000	0,0000	0,0000	0,0030	0,0000	0,0000	0,0006	0,0000	0,0000	0,0028	0,0000	0,0000	0,0010	0,0000	0,0000	0,0009	0,0000	0,0000	0,0000	0,0000	0,0000
G	0,0032	0,0000	0,0000	0,0000	0,0000	0,0019	0,0000	0,0000	0,0000	0,0012	0,0000	0,0000	0,0002	0,0000	0,0003	0,0029	0,0000	0,0000	0,0023	0,0000	0,0000	0,0036	0,0000	0,0000	0,0000	0,0000	0,0000
H	0,0057	0,0000	0,0000	0,0000	0,0000	0,0041	0,0000	0,0000	0,0000	0,0005	0,0000	0,0000	0,0000	0,0000	0,0000	0,0049	0,0000	0,0000	0,0000	0,0000	0,0001	0,0004	0,0000	0,0000	0,0000	0,0000	0,0000
I	0,0105	0,0004	0,0032	0,0008	0,0050	0,0008	0,0005	0,0022	0,0000	0,0002	0,0002	0,0000	0,0025	0,0051	0,0107	0,0049	0,0006	0,0002	0,0062	0,0091	0,0054	0,0010	0,0020	0,0000	0,0011	0,0000	0,0014
J	0,0015	0,0000	0,0000	0,0000	0,0000	0,0005	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0010	0,0000	0,0000	0,0000	0,0000	0,0000	0,0008	0,0000	0,0000	0,0000	0,0000	0,0000
K	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
L	0,0070	0,0001	0,0003	0,0001	0,0003	0,0053	0,0001	0,0009	0,0059	0,0039	0,0000	0,0000	0,0000	0,0009	0,0000	0,0042	0,0002	0,0001	0,0000	0,0002	0,0015	0,0013	0,0007	0,0000	0,0000	0,0000	0,0000
M	0,0131	0,0017	0,0000	0,0000	0,0001	0,0112	0,0000	0,0000	0,0000	0,0039	0,0000	0,0000	0,0000	0,0000	0,0000	0,0074	0,0030	0,0000	0,0001	0,0001	0,0000	0,0021	0,0000	0,0000	0,0000	0,0000	0,0000
N	0,0067	0,0000	0,0038	0,0011	0,0094	0,0031	0,0009	0,0015	0,0053	0,0023	0,0002	0,0000	0,0000	0,0000	0,0000	0,0066	0,0000	0,0004	0,0001	0,0033	0,0153	0,0013	0,0006	0,0000	0,0001	0,0000	0,0001
O	0,0008	0,0019	0,0017	0,0006	0,0027	0,0004	0,0006	0,0008	0,0001	0,0027	0,0002	0,0000	0,0036	0,0084	0,0074	0,0000	0,0006	0,0001	0,0133	0,0200	0,0010	0,0078	0,0011	0,0000	0,0001	0,0000	0,0004
P	0,0083	0,0000	0,0000	0,0000	0,0000	0,0071	0,0000	0,0000	0,0000	0,0016	0,0000	0,0000	0,0010	0,0000	0,0000	0,0083	0,0000	0,0000	0,0056	0,0000	0,0001	0,0008	0,0000	0,0000	0,0000	0,0000	0,0000
Q	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0173	0,0000	0,0000	0,0000	0,0000	0,0000
R	0,0221	0,0003	0,0010	0,0004	0,0019	0,0148	0,0002	0,0012	0,0000	0,0093	0,0000	0,0000	0,0001	0,0017	0,0011	0,0085	0,0003	0,0005	0,0033	0,0006	0,0038	0,0012	0,0007	0,0000	0,0000	0,0000	0,0000
S	0,0081	0,0001	0,0025	0,0000	0,0002	0,0159	0,0004	0,0002	0,0000	0,0037	0,0000	0,0000	0,0001	0,0011	0,0000	0,0061	0,0026	0,0003	0,0001	0,0076	0,0097	0,0026	0,0001	0,0000	0,0000	0,0000	0,0000
T	0,0124	0,0000	0,0000	0,0000	0,0000	0,0144	0,0000	0,0000	0,0002	0,0062	0,0000	0,0000	0,0000	0,0001	0,0000	0,0119	0,0000	0,0000	0,0070	0,0000	0,0000	0,0032	0,0000	0,0000	0,0000	0,0000	0,0000
U	0,0048	0,0005	0,0008	0,0002	0,0015	0,0148	0,0001	0,0006	0,0000	0,0031	0,0002	0,0000	0,0023	0,0076	0,0025	0,0003	0,0006	0,0001	0,0034	0,0029	0,0021	0,0000	0,0007	0,0000	0,0002	0,0000	0,0004
V	0,0067	0,0000	0,0000	0,0000	0,0000	0,0069	0,0000	0,0000	0,0000	0,0050	0,0000	0,0000	0,0000	0,0000	0,0000	0,0029	0,0000	0,0000	0,0006	0,0000	0,0000	0,0001	0,0000	0,0000	0,0000	0,0000	0,0000
W	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0000	0,0000	0,0000
X	0,0009	0,0000	0,0002	0,0000	0,0000	0,0004	0,0000	0,0000	0,0000	0,0004	0,0000	0,0000	0,0000	0,0000	0,0000	0,0004	0,0004	0,0000	0,0000	0,0000	0,0002	0,0001	0,0000	0,0000	0,0001	0,0000	0,0000
Y	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
Z	0,0010	0,0000	0,0000	0,0000	0,0000	0,0016	0,0000	0,0000	0,0000	0,0008	0,0000	0,0000	0,0000	0,0000	0,0000	0,0002	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0000	0,0000	0,0000	0,0000	0,0000

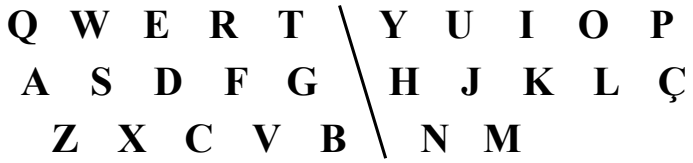

Figure 1. QWERTY keyboard layout and the division of keys by hand.

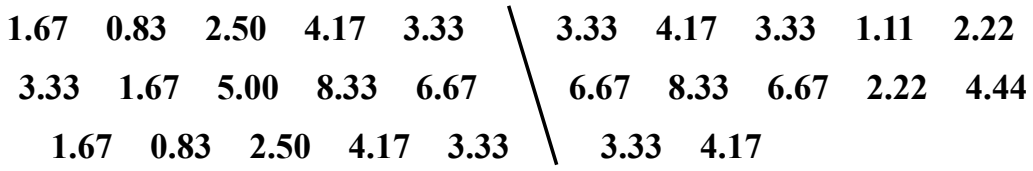

Figure 2. Desirable distribution proposed for ABNT keyboard.

ference between the frequency $\mathrm{f_S}(i)$ of each letter i, going from A to Z in the summation, and the desirable workload distribution $\mathrm{distr}_I(x)$ where this letter is positioned, $\mathrm{pos}(i)$, according to Table 2. Finally, this result is squared.

$$I_1 = \sum_{i=A}^{Z} \left(\mathrm{f_S}(i) - \mathrm{distr}_I(\mathrm{pos}(i))\right)^2 \tag{1}$$

3.2.3. Hand Alternation

Ideally, the hands should alternate in the tapping of consecutive letters, ensuring speed and comfort during writing. Thus, (2) sweeps the statistical basis, checking if each letter pair is typed using the same hand on the generated keyboard. If so, the relative frequency of that pair, $\mathrm{f_P}(i,j)$, is added to I_2. In this equation, $\mathrm{sh}(i,j)$ returns 1 if each letter from the pair is typed by the same hand, otherwise returns 0.

$$I_2 = \sum_{i=A}^{Z} \sum_{j=A}^{Z} \mathrm{f_P}(i,j) * \mathrm{sh}(i,j) \tag{2}$$

3.2.4. Finger Alternation

Typing consecutive keys using the same finger is also detrimental to the writing flow. The indicator I_3, presented in (3), adds a penalty if this occurs. In (3), $\mathrm{sh}(i,j)$ checks if the letter pair is typed by the same hand, and $\mathrm{sf}(i,j)$, by the same finger. The relative frequency of the letter pair, $\mathrm{f_P}(i,j)$, is multiplied by the distance, $\mathrm{d}(i,j)$, calculated by summing the absolute difference of i and j columns, $\mathrm{col}(x)$, with the difference between rows, $\mathrm{row}(x)$, see (4), before being added to I_3.

$$I_3 = \sum_{i=A}^{Z} \sum_{j=A}^{Z} \mathrm{f_P}(i,j) * \mathrm{d}(i,j) * \mathrm{sf}(i,j) * \mathrm{sh}(i,j) \tag{3}$$

Table 4. Penalty coefficients for big steps. Adapted from [Wagner et al. 2003]

		Second Finger			
		Index	Middle	Ring	Little
First Finger	Index	0	5	8	6
	Middle	5	0	9	7
	Ring	8	9	0	10
	Little	6	7	10	0

$$\mathrm{d}(x, y) = |\,\mathrm{col}(x) - \mathrm{col}(y)| + |\,\mathrm{row}(x) - \mathrm{row}(y)| \tag{4}$$

3.2.5. Avoidance of Big Steps

When the same hand types consecutive letters, it is important that the fingers involved provide a natural posture to the hand. It is known that certain combinations of finger movements are preferable than others. For example, the index-middle sequence results in a more comfortable movement than the index-ring one (Table 4).

This criterion is therefore considered in I_4. Thus, (5) initially checks if only one hand is typing the pair. If so, the relative frequency of that pair, $\mathrm{f_P}(i, j)$, is multiplied by a coefficient, $\mathrm{steps}(x, y)$, concerning the combination of fingers used, obtained through Table 4.

$$I_4 = \sum_{i=A}^{Z} \sum_{j=A}^{Z} \mathrm{f_P}(i, j) * \mathrm{steps}(i, j) * \mathrm{sh}(i, j) \tag{5}$$

3.2.6. Hit Direction

It is more comfortable for most people that the typing flow happens while using the same hand, and that this occurs from the extremities to the center of the keyboard, that is, from the little finger to the index [Wagner et al. 2003]. Thus, I_5 represents the penalty for this occurrence. According to (6), the relative frequency of the letter pair, $\mathrm{f_P}(i, j)$, is multiplied by $\mathrm{dir}(x, y)$, which returns 1 if the aforementioned case occurs, and 0 if not.

$$I_5 = \sum_{i=A}^{Z} \sum_{j=A}^{Z} \mathrm{f_P}(i, j) * \mathrm{dir}(i, j) * \mathrm{sh}(i, j) \tag{6}$$

3.2.7. Standard Deviation

The standard deviation represents how far the measurements move away from the average value of the set. In this research study, which seeks a better solution than that presented by QWERTY, significant improvements are expected across all indicators. Therefore, the ergonomics indicators obtained previously are compared to those of the QWERTY,

for the same statistical database. The percentage improvement for the target keyboard are obtained and the standard deviation for these measures are calculated. The value S_I is added to the final fitness, in order to penalize individuals that present the mentioned characteristics. Equation (7) encapsulates this process, where $I_{\%_i}$ is the percentage of improvement over each indicator and $\overline{I_\%}$ the average over all the improvements.

$$S_I = \sqrt{\frac{\sum_{i=1}^{5} \left(I_{\%_i} - \overline{I_\%}\right)^2}{4}} \tag{7}$$

3.2.8. Overall Fitness

The information collected regarding each ergonomics indicator for a specific keyboard is condensed into a single value, I_T, by (8). In this, the standard deviation S_I of the improvement percentage is added to the weighted sum of the ergonomics indicators, in which the weights $\mathrm{weight}(I_i)$ of each indicator are obtained as seen on Table 1.

$$I_T = S_I + \sum_{i=1}^{5} \mathrm{weight}(I_i) * I_i \tag{8}$$

3.3. Elitism

Elitism ensures that the best individuals are not lost over the generations. After a value is assigned for each keyboard model, following the ergonomic criteria, the population of the next generation is initiated. In this paper, the 20 fittest individuals, representing 10% of the total population [Rani et al. 2019], were copied directly to the next generation, without changes. This is necessary, as the main stages of composing a new population, the crossover and mutation, which will be discussed later, both present random elements, which may modify individuals and, perhaps, generate worse individuals than the previous ones.

3.4. Selection

In the selection process, some individuals are chosen to pass their characteristics on to the next population. However, it is preferable to have a random factor over this technique, preventing only the best individuals from prospering and the solution converging to a local maximum (or minimum, in this case). To this end, a method called tournament was used. In this study, three individuals, from the overall population, were chosen randomly. From among these, the element with the lowest score (most suitable keyboard layout) was selected. This process is repeated 90 times per generation, forming enough pairs to fill the 180 remaining individuals in the population, after the participation of all of them in the next steps: crossover and mutation.

3.5. Crossover

During crossover, the previously selected individuals combine their characteristics, generating new individuals, called offspring. The Partially Mapped Crossover (PMX) technique is responsible for merging such chromosomes, without repeating part of the genetic

Z G S C V E A W U J
D N M P T K Y Ç H I
L B F R Q X O

Figure 3. Optimized keyboard layout proposed.

material in each individual [Goldberg and Lingle 1985]. The central part of the individuals is switched. Then, the PMX evaluates all the other genes of this new chromosome, in order to avoid any duplication in the individual, which would cause a letter to repeat or disappear. For each gene, its presence is verified across the rest of the representation. If found, it is replaced by the character that is in the same position, in the other individual. In this research there was a 40% probability of ignoring this step, leading the unchanged individuals to the mutation process.

3.6. Mutation

The mutation changes individuals randomly. This step is responsible for introducing new options to the search space. After crossover, each individual is analyzed separately. In this research, each gene was put to the test, for which there was a 1% probability to swap places with another one.

3.7. End of Execution

Randomness is one of the main characteristics of a GA. In light of this, the results can present considerable variations. In this study, unlike related works, a strategy based on multiple executions of the GA was implemented, in order to obtain a statistically better solution. Each execution ended after 70 generations, and this process was repeated 20 times, independently, with different seeds for each run. Thus, the initial population was never the same, avoiding biasing the results. Then, a final run was made, using the best individual from each of the 20 executions as the initial population, randomly generating the remaining 180 individuals. Thanks to elitism, only better, or equal, results were obtained from this new run. Consequently, the fittest individual from these multiple executions could be selected.

3.8. Optimized Layout

By executing the algorithm, a better layout than QWERTY was obtained, with 52.99% of enhancement, see Figure 3. The fitness comparison between this layout and the QWERTY can be seen on Table 5. There were improvements across all indicators. The lowest of these was in tapping workload distribution, with 17.58%. This is due to the addition of the standard deviation. Without standard deviation, a result of 50.36% was reached, but at the expense of other indicators, which presented worse values than those of QWERTY.

4. Conclusion

This paper aimed at implementing a GA intended for the optimization of a more ergonomic and efficient keyboard layout for typing in Brazilian Portuguese. The approach described is capable of providing mathematically reliable solutions, since the optimization guidelines are all based on equations. In addition, the GA allows for a large number

Table 5. Fitness comparison between keyboard layouts

Indicators	QWERTY	Optimized	Improvement (%)
I_1	0.0621	0.0512	17.5822
I_2	0.4949	0.2238	54.7847
I_3	0.1254	0.0578	53.8999
I_4	2.7851	1.2860	53.8248
I_5	0.1920	0.1172	38.9787
I_T	2.6880	1.2636	52.9911

of possible solutions to be analyzed, providing a layout with a good fitness. Thus, the optimized layout found by this research is in accordance with initial expectations. It assured an improvement of approximately 53% when compared to QWERTY, according to the ergonomic indicators.

Besides, this paper shows contributions in four aspects. First, it presented a statistical analysis of the characters and sequences of pairs most commonly used in Brazilian Portuguese. Second, it proposed an desirable workload distribution suitable for the keyboard with a layout for the mentioned language, with the scope reduced only to the region of the letter on the keyboard. Third, the addition of a penalty for the standard deviation of percentage improvements ensured a steady improvement across all indicators. Finally, the execution of the algorithm several times, using the best individuals as the initial population, guaranteed statistical reliability to the GA.

Due to the great combinations of GA's parameters, it is not possible to prove that the proposed design is the best. Far from it, it is known that with small changes in the GA configuration, other optimized results could be obtained, which would develop some aspect to the detriment of another. One of the difficulties of this approach is to balance individual gains with collective ones, while trying to configure the parameters of the GA effectively. Moreover, due to the penalty for standard deviation, indicators in which the improvement would be more significant end up truncated to the average values.

For individuals who have never been exposed to keyboards, an optimized keyboard layout would be a great solution. However, for those who already have some experience with the QWERTY layout, training would be necessary. So, although the QWERTY layout is not the most efficient designated layout, experienced users are more skilled at this layout than with others. Thus, so that the optimized layout found through this investigation does not remain just an academic exercise, as a future work user studies will be conducted that evaluate human factors and ergonomics, as well as analyzing the resistance of users to the implementation of this layout as a primary way of interaction. Moreover, comparing the layout obtained, and the method used, with others from the literature. Furthermore, the search will be made for an optimized keyboard that includes numbers, special characters, and diacritics, which recur in Brazilian Portuguese.

References

ABNT/CB-021 (1991). Information tecnology - keyboards for data processing equipaments - standardization. Technical Report NBR 10346, Associação Brasileira de Normas Técnicas (ABNT).

Deshwal, P. S. and Deb, K. (2003). Design of an optimal hindi keyboard for convenient and efficient use. Technical Report KanGAL 2003005, Indian Institute of Technology, Kanpur.

Dvorak, A. and Dealey, W. L. (1936). Typewriter keyboard. US Patent 2,040,248.

Eggers, J., Feillet, D., Kehl, S., Wagner, M. O., and Yannou, B. (2003). Optimization of the keyboard arrangement problem using an ant colony algorithm. *European Journal of Operational Research*, 148(3):672–686.

Garbe, J. U. (2000). Optimizing the layout of an ambiguous keyboard using a genetic algorithm. Master's thesis, Universität Koblenz-Landau, Department of Computer Science, Germany.

Goldberg, D. E. (1989). *Genetic Algorithms in Search, Optimization and Machine Learning*. Addison-Wesley Longman Publishing Co. Inc., Boston, MA, United States.

Goldberg, D. E. and Lingle, R. (1985). Alleles, loci and the travelling salesman problem. In *Proceedings of the First International Conference on Genetic Algorithms and Their Applications*, pages 154–159, Pittsburgh, PA, United States. Psychology Press.

Heidner, F. (1915). Type-writing machine. US Patent 1,138,474.

Khorshid, E., Alfadli, A., and Majeed, M. (2010). A new optimal arabic keyboard layout using genetic algorithm. *International Journal of Design Engineering*, 3(1):25–40.

Liao, C. and Choe, P. (2013). Chinese keyboard layout design based on polyphone disambiguation and a genetic algorithm. *International Journal of Human-Computer Interaction*, 29(6):391–403.

Light, L. W. and Anderson, P. G. (1993). Typewriter keyboards via simulated annealing. *AI Expert*, September:1–11.

Noyes, J. (1983). The qwerty keyboard: A review. *International Journal of Man-Machine Studies*, 18(3):265–281.

Noyes, J. (1998). Qwerty-the immortal keyboard. *Computing & Control Engineering Journal*, 9(3):117–122.

Oommen, B. J. and Zgierski, J. R. (1991). Keyboard optimization using genetic techniques. In *1991 Tenth Annual International Phoenix Conference on Computers and Communications*, pages 726–732, Scottsdale, AZ, USA. IEEE.

Rani, S., Suri, B., and Goyal, R. (2019). On the effectiveness of using elitist genetic algorithm in mutation testing. *Symmetry*, 11(9).

Tomassini, M. (1995). A survey of genetic algorithms. *Annual reviews of computational physics*, 3(1):87–118.

Wagner, M. O., Yannou, B., Kehl, S., Feillet, D., and Eggers, J. (2003). Ergonomic modelling and optimization of the keyboard arrangement with an ant colony algorithm. *Journal of Engineering Design*, 14(2):187–208.