 doi:10.20944/preprints202105.0255.v1

Article

Shared Data Set for Free-Text Keystroke Dynamics Authentication Algorithms

Augustin-Catalin Iapa [1,*] and Vladimir-Ioan Cretu [1]

1 Timișoara Politehnica University; Department of Computer and Information Technology; catalin.iapa@gmail.com; vladimir.cretu@cs.upt.ro

* Correspondence: catalin.iapa@gmail.com; Tel.: +40769054995

Abstract: Identifying or authenticating a computer user are necessary steps to keep systems secure on the network and to prevent fraudulent users from accessing accounts. Keystroke dynamics authentication can be used as an additional authentication method. Keystroke dynamics involves in-depth analysis of how you type on the keyboard, analysis of how long a key is pressed or the time between two consecutive keys. This field has seen a continuous growth in scientific research. In the last five years alone, about 10,000 scientific researches in this field have been published. One of the main problems facing researchers is the small number of public data sets that include how users type on the keyboard. This paper aims to provide researchers with a data set that includes how to type free text on the keyboard by 80 users. The data were collected in a single session via a web platform. The dataset contains 410,633 key-events collected in a total time interval of almost 24 hours. In similar research, most datasets are with texts written by users in English. The language in which the users wrote for this research is Romanian. This paper also provides an extensive analysis of the data set collected and presents relevant information for the analysis of the data set in future research.

Keywords: keystroke dynamics; typing pattern; keystroke data set; user authentication; user identification; free text typing; keystroke dynamics researches; keystroke analysis; biometrics; keystroke characteristics

1. Introduction

Keystroke dynamics is a research field with more and more importance in network access control and cyber security [1,2]. For now, only a few studies are about free-text keystroke dynamics, the way that the users type what text the user wants. Most of them are analyzed only fixed text, static text [1,3,4]. Fixed content and fixed length data are usernames or passwords [5]. Free text requires two phases: the user enrollment phase in the system and the user verification phase [5].

The method of continuous authentication using keystroke dynamics has several fields in which it can be successfully applied, for example, as an additional security method when a user accesses his bank account on the internet or when making a payment in a similar way [6,7]. It can be applied for e-mail accounts, or any other online platform that requires a lot of typing. The authentication process can be categorized by the number of incorporated factors: (1) something you know like a username and a password, (2) something you have, like card, token or (3) something you are, like biometrics. [8] A combination of these processes is a strong authentication [6,7].

The keystroke dynamics can also be the second factor authentication. Two-factor authentication is a large scale used approach, in some systems even mandatory, for online services [9]. The traditional password is the first factor and the second factor can be a SMS access code or a PIN generated randomly at the time of authentication [7,10].

The keystroke dynamics technique consists in capturing and analyzing the typing mode of a user. More precisely, the pressing time on one key, but also the time between

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 11 May 2021 doi:10.20944/preprints202105.0255.v1

the pressings of two consecutive keys. The rhythm along the pressure of keys plays an important role when it comes to study the cases [11]. These features are unique, as are other methods of identifying individuals such as fingerprint, facial recognition, account password, or the use of a physical card or other physical identification device [2].

Within the scientific research made about the keystroke dynamics they were identified two different branches. The first would be when a user types a default text on the keyboard, such as a user, a standard password or phrase. The second one would be the typing of a free text on the keyboard without certain conditions being imposed [12,13]. The two methods are analyzed separately by different methods in the scientific literature on this subject. Both, however, involve a phase in which the system collects data about the user, the typing times, and the typing mode, thus, creating a profile of the user that he will use later in the continuous authentication phase. The first method has been more intensively explored and the results are more successful in this direction because it is the same text entered from the keyboard each time. The second method, when the user types a free text with the help of the keyboard, without conditions, has been researched especially in recent years, and the results are increasingly improved.

Typing behavior for continuous authentication is a biometric modality proposed in [14]. The authors collected a video database from 63 users with static text and free text typing and developed computer vision algorithms to extract hand movement from the video stream.

Most studies analyze data collected in English. There are studies that research the field for texts in other languages, such as French [15], Italian [16], Japanese [17], Russian [18], Arabic [19], Korean [20] or others.

Commercial keystroke dynamic products exist. In 2003, the paper [21] presents the company BioNet Systems which patented the BioPassword authentication system [22]. In Romania, Typing DNA is a company, a start-up, that received funds of 6.2 million euros in 2020 to create a typing identity for security [23].

Other studies, like [24], incorporate the use of nonconventional typing features using free text typing dynamics. Semi-timing features along with the editing features were extracted from the users' typing flow and decision trees were used to classify each of the user data.

Algorithms of dynamic authentication can be divided into three major groups: estimation of metric distances, statistical methods and machine learning. Methods of keyboard recognition used in the literature are: distance, neural networks, statistical, probabilistic, machine learning, clustering, decision tree, evolutionary computing, fuzzy logic or other [7,18].

Some limitations of keystroke dynamics previous research were: it took a long time to train the model, data were manual preprocessed by human or large database was required [25]. The authors from [25] conclude that use of keystroke dynamics can make a more secure system.

One of the main problems facing researchers is the small number of public data sets that include how users type on the keyboard. This paper aims to provide researchers with a data set that includes how to type free text on the keyboard by 80 users. The data were collected in a single session via a web platform. The dataset contains 410,633 key-events collected in a total time interval of almost 24 hours. The language in which the users wrote for this research is Romanian.

The rest of the paper is organized as follows. Section 2 reviews the evolution of research in the field. Data acquisition methodology is presented in Section 3, where the subsections refer to the development of the platform for the acquisition the data and the acquisition and initial processing of the data. Section 4 presents the platform for data acquisition. Section 5 presents the results of continuous authentication experiments used the data set. Discussions are in the Section 6, where the subsections refer to the analysis of user information, of time and key events collected from users, of key distribution. The subsection 6.4 compares the results of this paper with results of the related works and the subsection 6.5 presents the contributions of the paper. Finally, Section 7 provides the

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 11 May 2021 doi:10.20944/preprints202105.0255.v1

conclusion and future works. The paper has also five appendices that contain details about the shared keystroke dynamics data set.

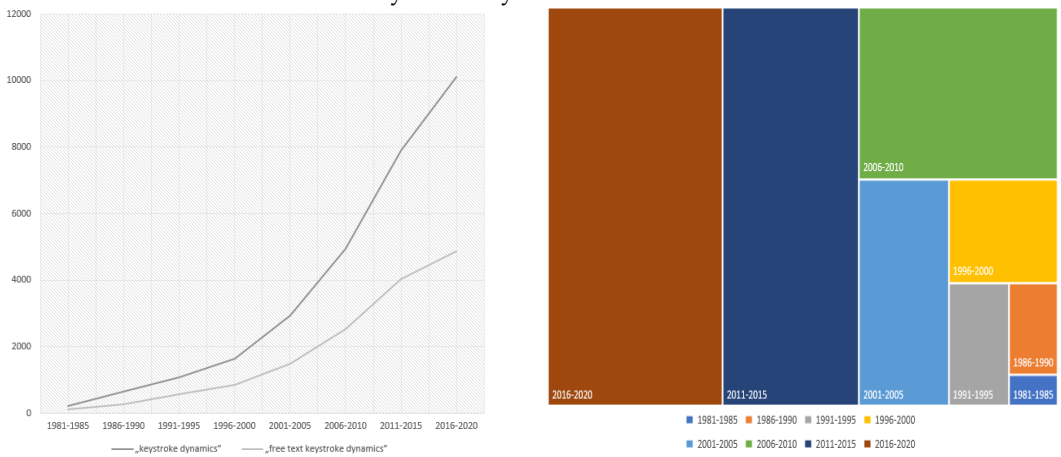
2. The evolution of research in the field

Only in the last 5 years over 10,000 scientific papers have been published about keystroke dynamics. Also, survey papers have been published as keystroke dynamics biometrics has drawn intense research interest the past couple of decades [26]. In Table 1 is the number of scientific papers in the field of "keystroke dynamics" and also in the field of "free text keystroke dynamics". The graphic represented in Figure 1(a) illustrates the growing interest in the field of "keystroke dynamics" and also in the field of "free text keystroke dynamics" [2].

Table 1. Number of scientific publication in the field [2]

Interval	„keystroke dynamics"	„free text keystroke dynamics"
1981-1985	224	108
1986-1990	643	277
1991-1995	1.080	566
1996-2000	1.630	863
2001-2005	2.950	1500
2006-2010	4.940	2520
2011-2015	7.890	4020
2016-2020	10.100	4880

The number of scientific publications in this field was counted by searching for the two text sequences on scholar.google.com, filtered on 5-year intervals [2]. It is observed that in the last 5 years over 10,000 scientific papers have been published with the topic "keystroke dynamics", and scientific papers that have addressed the branch "free text keystroke dynamics" represent about half of these, reaching about 5,000 papers published in the last 5 years [2]. In the Figure 1(b) is also a hierarchy chart with the volume of publication about "keystroke dynamics".



(**a**) (**b**)

Figure 1. (a) Evolution of publication about "keystroke dynamics" and "free text keystroke dynamics" from 1981 till 2020; **(b)** Hyerarchy chart with the volume of publication about "keystroke dynamics".

3. Data acquisition methodology

3.1 Development of the platform for the acquisition the data

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 11 May 2021 doi:10.20944/preprints202105.0255.v1

The first step in this research was to create a web platform for the acquisition of input data necessary for research. For this, the website from https://sites.google.com/view/cataliniapa was created, a form was created that would take over, besides the text typed by the users, the way of typing on the keyboard. A program in JavaScript language was written to take over the keystroke times. In order to be able to download the necessary information, a Google Sheet file was configured, and the information collected using the web form was transmitted using the platform https://api.apispreadsheets.com/. The platform for acquiring input data has been completed and functional by integrating the script written in JavaScript with the data transfer application in the Google Sheet file. The steps described can be followed in the graph in Figure 2.

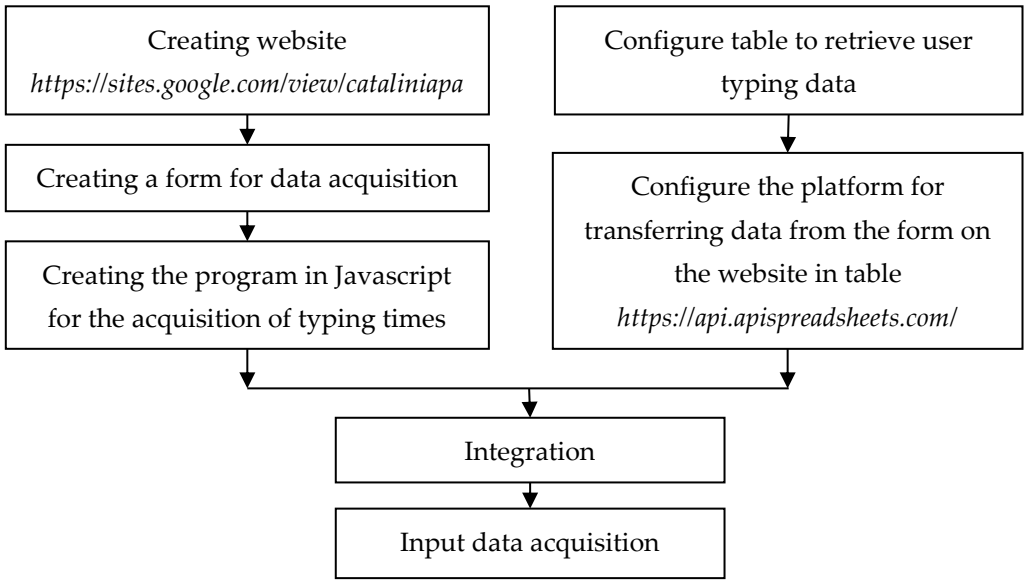


Figure 2. Steps taken to create the platform for retrieving data on how users type.

3.2 Acquisition and initial processing of the data

The acquisition and initial processing of the input data went through the following steps: Data were collected from 80 users using a web program written in JavaScript. It was collected from the 80 volunteers, through a form, the keys typed on the keyboard but also the times at which they were typed. The collected data was initially stored in a Google Sheet file via the https://api.apispreadsheets.com/ platform. With a program written in the C programming language, the data collected in the Google Sheet file was processed and transformed into key events in the following form:

68 0 123444
68 1 123555
59 0 123720
71 0 123800
59 1 123830
71 1 123992
...

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 11 May 2021 doi:10.20944/preprints202105.0255.v1

where on the first column is the key code of the pressed key, on the second column is 0 or 1, 0 represents the pressed key, and 1 represents the raised key, and the third column represents the timestamps at which the key event occurred. The file with the form presented above is the input file for the continuous authentication algorithm developed in this thesis using the keystroke dynamics method. The steps described above are summarized in the graph in Figure 3.

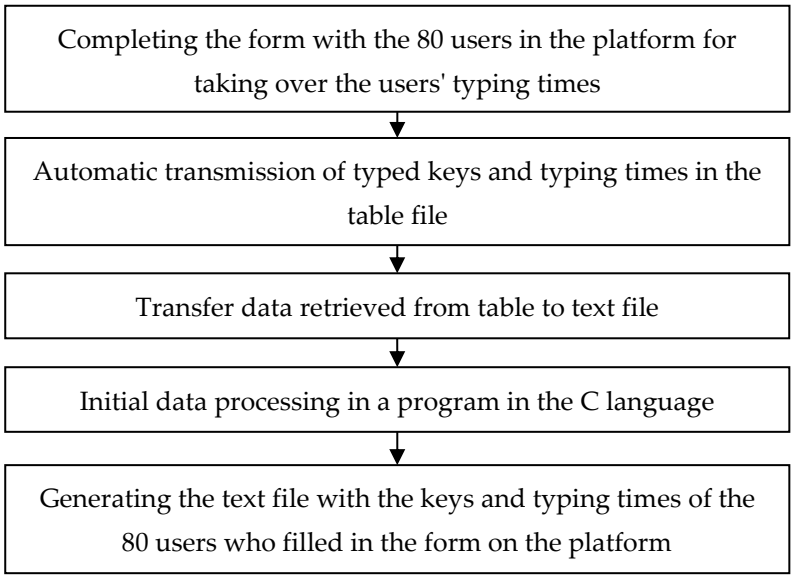


Figure 3. The steps taken for the acquisition and initial processing of key data and typing times of the 80 volunteers are in the figure.

4. The platform for data acquisition

To research in the field of keystroke dynamics biometrics the researchers need input data obtained from computer users in different real situations. The necessary data are represented by the keys typed on the keyboard but also by the times at which they are pressed. The difference between the time when a certain key is pressed, respectively the time when a certain key is raised is the keystroke time. Another important piece of information is the time between two keys. The difference between the time a key was released and the time a next key was pressed [7].

This information can only be obtained in a restrained or controlled environment, with the consent of those participating to this experiment. The agreement of the participants is necessary because it exists a possibility to form the initial text that the user typed on the keyboard with access to this data, and if, for example, a user is monitored while sending e-mails or doing other activities, the information may be confidential.

For the purpose of the research, the authors developed their own environment to obtain data from volunteers. The authors have created a web environment for taking over keys and typing times in JavaScript. A form is created that takes over the keys and typing times while completing a form on a web page [7]. The website was created on the sites.google.com platform. The web platform can be accessed at https://sites.google.com/view/cataliniapa.

To capture the keys and typing times the authors created a web form through which users were invited to answer several generic questions. The text entered from the keyboard by each user should be written freely by each user, without the need to reproduce a specific predefined text. At each text box, a series of generic questions were formulated to guide the user to a certain topic in the text he completed. The questions asked were about the weather, the ideal day or the educational system. To form the database for research is not relevant the topic of the text, but the way it is written.

The text written by users is in Romanian. Most datasets in the literature are texts captured from users who have written in English [7].

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 11 May 2021 doi:10.20944/preprints202105.0255.v1

After completing all the fields in the form, in order to send the captured data, the consent regarding the takeover for the purpose of scientific research of the participants was obtained. Two questions answered by users from the form are in Figure 4. First one is about weather and second one about the ideal day.

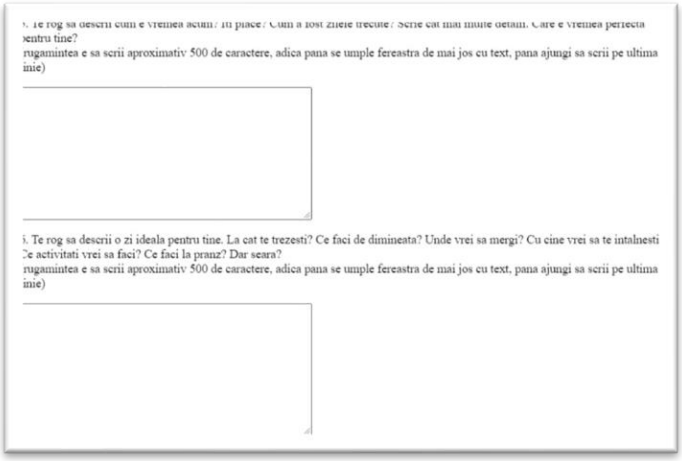


Figure 4. Two questions answered by users from the form.

Each user was instructed to pursue the following rules when filling out the form: "1. You have to write a free text about the subject managed by guidance questions; 2. You have to write a text of about 500 characters for each question (this means that all the lines in a text window should be filled); 3. Do not copy the answer from other sources; 4. You have write the answer to the questions on the spot, without consulting external sources; 5. You have to write ideas fluently, as they come to mind; 6. Do not do other activities while completing the answer to the questions. The request is to allocate about 15 minutes to complete the form; 7. The written text must be in Romanian; 8. The written text should be as generic as possible, not personal; 9. The text should be written from a physical keyboard, computer or laptop, not a touchscreen device (not a phone or tablet); 10. Please take about 15 minutes to complete the form to answer all questions without being interrupted by other activities;"

In one of the questions on the form, users were asked to describe the scene in the Figure 5, in as much detail as possible.

Figure 5. Each user has described the scene of the picture at one of the questions.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 11 May 2021 doi:10.20944/preprints202105.0255.v1

The form included statistical questions about the user's age, gender and whether he uses a computer or laptop keyboard. The questions with statistical purpose were followed by four questions to which the way of typing the answer was captured. The four questions asked in the form are presented in Table 2.

Table 2. The questions addressed to users in the form.

Questions addressed to users in the form
1. Please describe the weather. Do you like it? How were the past few days? Write as many details as possible. What is the perfect time for you?
2. Please describe an ideal day for you. What time do you wake up? What are you doing in the morning? Where do you want to go? Who do you want to meet? What activities do you want to do? What are you doing at lunch? What about the evening?
3. What is your opinion about the educational system in Romania? What is being done well? What's wrong? How is the system in other countries? What would we have to do to be better? What should students do? What about the teachers? What about the parents?
4. Describe in detail what you see in the painting. What does each character do? Where does the scene take place? How did people live then? Why were they worried? How did they spend an ordinary day? (The painting from the Figure 5)

5. Experiments and results

In order to compare the data set obtained in the present research with data sets from other research, we implemented a continuous authentication algorithm. The algorithm used as input data collected from the 80 users. We used the Equal Error Rate (EER) to quantify the performance of the algorithm. The results obtained are comparable to the results obtained in similar research.

The distance between the users was calculated using the information obtained from the di-graphs and building the user's pattern with a sample size of 1000 keys. Two distinct methods for calculating distances were used: Manhattan distance and A distance, proposed by Gunetti & Picardi in [27].

The results obtained are EER = 13.89% when using the Manhattan distance. This performance result was obtained after analyzing the most common 12 di-graphs. The distance was calculated using the total time of the di-graph.

The results obtained are EER = 6.55% in the case of using distance A. This performance result was obtained after analyzing all the collected diagrams. The distance was calculated using the total time of the di-graph.

6. Discussion

Each of us has a rhythm, a certain speed, a typing pattern, formed in time and unique while typing on a keyboard. We can differentiate the users of a computer, can identify them or authenticate them in a system only by capturing these details. To analyze a user's typing pattern, we need to capture and process it using an algorithm.

In order to be able to identify a certain user who would now be in front of a computer, using a keyboard, it is necessary, beforehand, to have his typing characteristics in a database. The database is needed in order to compare the typing mode captured live with the patterns of the users enrolled in the respective system, thus, helping to be identified. In other words, the mode of operation is similar to the username and password authentication. The computer users enter their username and their password, and the system searches them in the database to compare what the user entered with what he has previously registered, in order to make a decision.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 11 May 2021 doi:10.20944/preprints202105.0255.v1

6.1 Analysis of user information

The dataset contains keyboard typing data collected from 80 users. Of the 80 users, 35 said they were male and 44 said they were female, while one user did not report gender. The age of the users is in the range of 16-59 years. The average age of the 80 users is 28.19 years. Data was collected from users who used the keyboard from a computer or laptop. A total of 64 users used a laptop keyboard to complete the form, while only 15 used a computer keyboard and one user did not state which keyboard he used. Information about each user regarding these statistic data can be found detailed in Appendix A, Table A1.

6.2 Analysis of time and key events collected from users

The form created to purchase data sets for research purposes was completed by a number of 80 users. They handed over data for 410,633 key-events [7]. The comprise time used by all 80 users to complete the form was 23 hours, 28 minutes and 19 seconds.

The average time spent by users on the data collection platform was 17 minutes and 36 seconds. In Appendix B, the Table A2 shows the completion times of the form for each user, as well as the average and the total time spent by users to complete. In this regard, the time is expressed not only in milliseconds revealed in the second column of the table, but also in minutes show in the third column of the table. The fourth column of the table shows the total number of key events collected from each of the 80 users who filled out the form. The total number of key events collected from all users is 410,633. The average number per user is 5132 key events. Each key event contains Key Code, Down Event or Up Event and the Time Stamp.

6.3 Keys distribution analysis

A total of 100 different keys were monitored. The key that was pressed most often by users in the experiment was the SPACE key. The SPACE key has been pressed 32,387 times in total. Of the total keys, it represents the percentage of 16.17%. The next 3 frequently used keys are the vowels A, E and I. The A key was used 20,965 times and represents 10.47% of the total keys. The E key has been pressed 18,256 times and represents 9.11% of the total keys. The I key has been pressed 15,994 times and represents 7.99% of the total keys. The BACKSPACE key is also frequently pressed, which has been pressed 12,195 times.

In Table A3 from Appendix C are all the keys pressed by users in the order of their frequency in the data set collected. The most common 30 keys used by users are represented graphically in Figure 6. The first 30 keys represent 98.73% of the keys used.

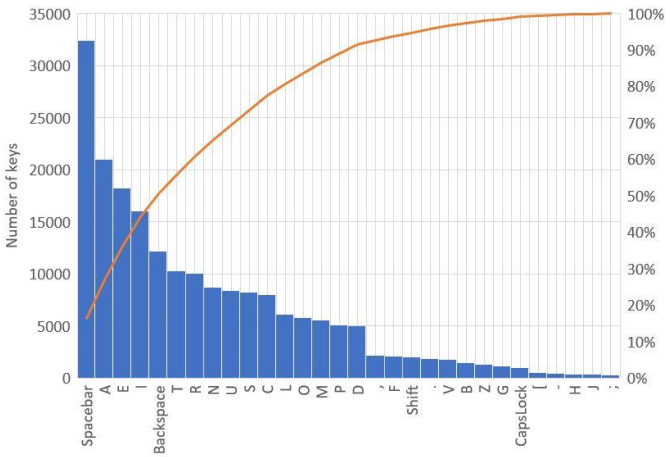


Figure 6. Graphical representation of the number of key events collected from each user, in descending order

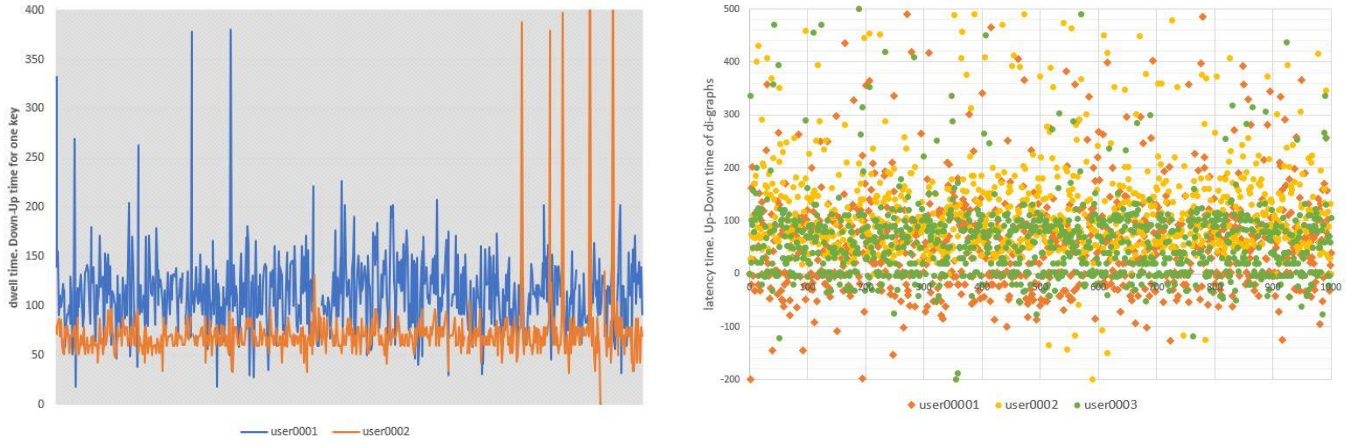
Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 11 May 2021 doi:10.20944/preprints202105.0255.v1

Analyzing studies carried out regarding the use of characters in Romanian, the conclusion is that the database collects respect the general rules, this database accurately reproduces the general characteristic of the Romanian language. According to the study conducted in [28], the most used consonants in Romanian are the consonants R and T, while the least used are X and J, except for the letters K, Q, W and Y, which are not specific to the language. The data set falls within these rules, the most used consonants being T and R, and the least used consonants being J, X, K, Q, W and Y.

The distribution of letters of the English alphabet (a-z) in the dataset is shown in Table A4 from Appendix D.

Each user has his own unique way to type text on the keyboard. This pattern is specific and does not change during a writing session or short term. The typing pattern may change over time or may differ if the same user uses different keyboards. The differences between different users, on the other hand, can be analyzed even visually, as for example in the Figure 7(a). The graph shows the typing times for two users from the database. The graph shows how the differences between the typing times for user0001 are larger, both the average of the times and the standard deviation. Most of the time intervals for user0001 are between 50 and 150 milliseconds. Instead, user0002 has a smaller difference between keystrokes. At user0002 most of the time intervals are in the range of 50-75 milliseconds [7].

The Figure 7(b) shows the first 1000 time intervals between two consecutive keys, flight time (UD time). This time interval can also have negative values, while the pressing time of a single key cannot have negative values. A negative value is taken when the second key in a di-graph is pressed before the first key is raised. The figure shows the times for three users. We can see how user0001 has the most negative time values, while user0003 has the most time values close to 0. The time value can be close to 0 when the second key is pressed exactly when the first he gets up. User0002 has the fewest negative time intervals, even their average being the highest of the time averages of the 3 users analyzed. In the analysis of the typing pattern, both the times when the keys are pressed and the times between two consecutive keys are analyzed.



(**a**) (**b**)

Figure 7. (**a**) Typing pattern from two different users [7]; (**b**) Time interval for flight time for three different users

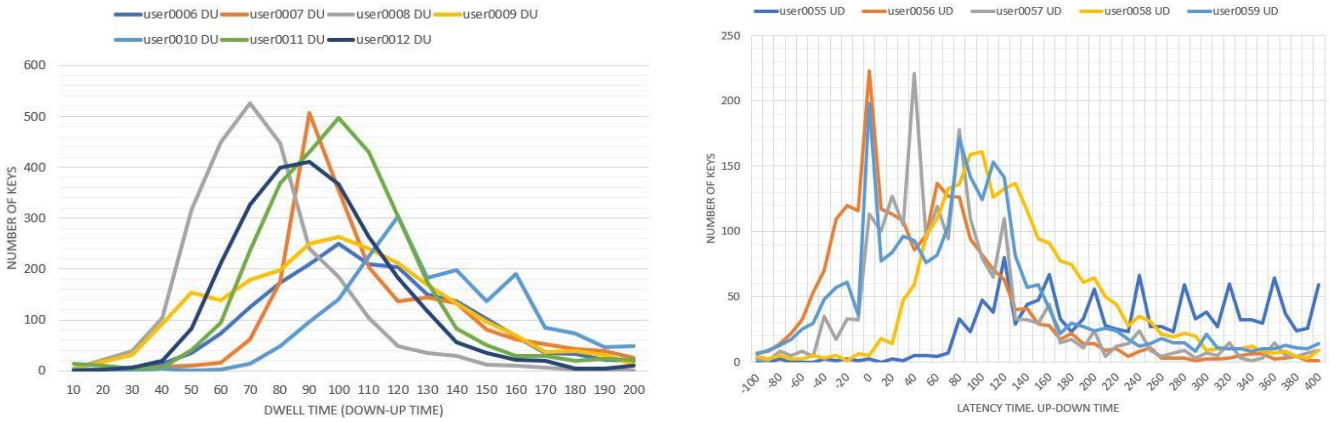
Di-graph analysis takes into account the order in which characters are typed by users. From the database collected from users, a total number of 200,227 di-graphs could be created and analyzed. The total number of unique di-graphs is 1,530. This means that there are only 1,530 unique 2-character combinations. The most used di-graphs in the text are presented in Table A5 from Appendix E. These are di-graphs that appear in texts taken from users more than 1000 times each.

A user's profile in terms of testing can be achieved based on the most frequent time intervals. The Figure 8(a) graphically represents the modes of distribution of typing time (Down-Up time) for a number of 7 users. The time distribution is a normal distribution,

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 11 May 2021 doi:10.20944/preprints202105.0255.v1

close to a Gaussian distribution or a Laplace distribution. In contrast, both the mean and the standard deviation differ from user to user.

The distribution of time intervals between two consecutive keys is represented for a total of five different users in the Figure 8(b). It is observed for two users, user0056 and user0059, a maximum of number of key intervals at the value 0 on the graph. Also, user0056 has the most negative intervals. A distribution of time intervals totally different to the other four users has user0055. Times are distributed at higher values. This means that user0055 is typing at a slower pace.



(**a**) (**b**)

Figure 8. (**a**) Key time distribution for seven different users; (**b**) The distribution of time intervals between two consecutive keys.

6.4 Comparison of the related works

The Table 3 shows the characteristics of the databases used in previous scientific research, in order to be able to compare them with the characteristics of the data set obtained in this research. The characteristics were being published and centralized in the paper [29]. In the last line of the table are the characteristics of the data set in this paper.

Table 3. Comparison of the related works [29]

Research	Year	# of users	Experimental time
Monrose & Rubin [30]	1997	42	7 Weeks
Gunetti & Picardi [27]	2005	40	1–2 Months
Villani et al. [31]	2006	40	–
Davoudi & Kabir [32]	2009	21	1–2 Months
Samura & Nishimura [17]	2009	112	–
Park & Cho [33]	2010	35	–
Messerman et al. [13]	2011	55	12 Months
Alsultan et al. [19]	2016	21	–
Alsultan et al. [34]	2017	25	–
Alsultan et al. [35]	2017	30	–
Kim et al. [36]	2018	150	–
Tsai & Shih [29]	2018	100	2 Weeks
This paper data set	**2020**	**80**	**1 Session**

6.5 Contributions of the paper and future works

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 11 May 2021 doi:10.20944/preprints202105.0255.v1

The objective of this paper was to collect a database with the typing mode from 80 users and to make it available to other interested researchers. It was created a database with typing mode from 80 users, 410.000 key events and total time of approximately 24 hours for the acquisition of the necessary data.

There are new possibilities to continue research in new directions, such as:

- Expanding the keystroke dynamics database by collecting data from a larger number of users;
- Expanding the database by collecting data from the 80 users in new sessions in order to research the evolution of the typing pattern over time

7. Conclusions

Authentication via keystroke dynamics is a topic of interest in the field of security and privacy, especially in authentication and access control. It is also a topic addressed in the field of human computer interaction (HCI), especially in interaction paradigms and interaction devices. Keystroke dynamics involves in-depth analysis of how you type on the keyboard, analysis of how long a key is pressed or the time between two consecutive keys. This field has seen a continuous growth in scientific research in the last years. One of the main problems facing researchers is the small number of public data sets that include how users type on the keyboard. The objective of this paper was to collect a database with the typing mode from 80 users and to make it available to other interested researchers. It was created a database with typing mode from 80 users, 410.000 key events and total time of approximately 24 hours for the acquisition of the necessary data. The data set is available at *https://sites.google.com/view/cataliniapa/timisoara-kd-data-set.*

Author Contributions: Conceptualization, A.C.I. and V.I.C.; methodology, A.C.I. and V.I.C.; software, A.C.I.; validation, A.C.I..; formal analysis, A.C.I.; investigation, A.C.I.; resources, A.C.I. and V.I.C.; data curation, A.C.I.; writing—original draft preparation, A.C.I.; writing—review and editing, A.C.I. and V.I.C.; visualization, A.C.I. and V.I.C.; supervision, V.I.C.. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: Publicly available datasets were analyzed in this study. This data can be found here: *https://sites.google.com/view/cataliniapa/timisoara-kd-data-set*

Acknowledgments: We thank the 80 volunteers who responded positively and filled out the data acquisition form so that we have this complete and available data set for future research in the field of keystroke dynamics authentication.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A

Table A1. Characteristics of the users who filled in the form

User	Age	Gender	Device
	Average: 28.19	**Male: 35** **Female: 44** **Unknown: 1**	**Laptop: 64** **Computer: 15** **Unknown: 1**
user0001	32	Male	Laptop
user0002	23	Famale	Laptop
user0003	29	Famale	Laptop
user0004	21	Famale	Laptop
user0005	20	Famale	Laptop
user0006			
user0007	22	Famale	Laptop

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 11 May 2021 doi:10.20944/preprints202105.0255.v1

user0008	29	Male	Laptop
user0009	18	Famale	Laptop
user0010	28	Male	Computer
user0011	19	Male	Laptop
user0012	29	Male	Laptop
user0013	19	Famale	Laptop
user0014	21	Male	Computer
user0015	23	Famale	Laptop
user0016	19	Male	Laptop
user0017	19	Male	Computer
user0018	21	Famale	Laptop
user0019	34	Famale	Laptop
user0020	24	Famale	Laptop
user0021	19	Famale	Laptop
user0022	17	Male	Laptop
user0023	51	Famale	Laptop
user0024	24	Famale	Laptop
user0025	36	Male	Computer
user0026	31	Male	Laptop
user0027	16	Famale	Laptop
user0028	26	Famale	Laptop
user0029	31	Famale	Computer
user0030	59	Male	Computer
user0031	32	Famale	Laptop
user0032	22	Famale	Laptop
user0033	42	Male	Laptop
user0034	25	Famale	Laptop
user0035	33	Male	Computer
user0036	36	Famale	Laptop
user0037	23	Famale	Laptop
user0038	29	Male	Laptop
user0039	22	Male	Laptop
user0040	25	Male	Laptop
user0041	23	Famale	Laptop
user0042	20	Famale	Laptop
user0043	24	Famale	Laptop
user0044	22	Male	Laptop
user0045	22	Famale	Laptop
user0046	21	Male	Laptop
user0047	31	Male	Laptop
user0048	23	Male	Computer
user0049	30	Male	Laptop
user0050	25	Famale	Laptop

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 11 May 2021 doi:10.20944/preprints202105.0255.v1

user0051	21	Famale	Laptop
user0052	32	Male	Laptop
user0053	21	Male	Computer
user0054	31	Famale	Computer
user0055	44	Famale	Computer
user0056	33	Famale	Laptop
user0057	22	Male	Computer
user0058	25	Famale	Laptop
user0059	32	Male	Laptop
user0060	22	Famale	Laptop
user0061	30	Famale	Laptop
user0062	30	Male	Laptop
user0063	26	Male	Laptop
user0064	36	Male	Laptop
user0065	26	Male	Laptop
user0066	21	Famale	Laptop
user0067	28	Famale	Laptop
user0068	32	Male	Computer
user0069	43	Male	Laptop
user0070	31	Famale	Laptop
user0071	41	Famale	Laptop
user0072	27	Famale	Laptop
user0073	30	Male	Laptop
user0074	34	Male	Laptop
user0075	49	Male	Laptop
user0076	37	Male	Computer
user0077	26	Famale	Laptop
user0078	32	Famale	Computer
user0079	32	Male	Laptop
user0080	43	Famale	Laptop

Appendix B

Table A2. This is a table. Tables should be placed in the main text near to the first time they are cited.

User	Time (ms)	Time (min)	Total Key Events
Total	**84501613**	**1408.36**	**410633**
Average	**1056270**	**17.60**	**5132**
user0001	1095546	18.26	3905
user0002	1177215	19.62	5715
user0003	591972	9.87	5671
user0004	1294786	21.58	5992
user0005	504168	8.40	4938
user0006	443926	7.40	3815

 doi:10.20944/preprints202105.0255.v1

user0007	1076423	17.94	4422
user0008	931092	15.52	5303
user0009	958237	15.97	4889
user0010	869952	14.50	4021
user0011	814627	13.58	6057
user0012	649341	10.82	5181
user0013	570906	9.52	4696
user0014	640486	10.67	5872
user0015	614859	10.25	5153
user0016	923821	15.40	5312
user0017	608300	10.14	4273
user0018	1038345	17.31	5311
user0019	921804	15.36	5379
user0020	570299	9.50	4521
user0021	801803	13.36	6153
user0022	512004	8.53	4612
user0023	1270468	21.17	3231
user0024	278405	4.64	1121
user0025	1324293	22.07	5083
user0026	548089	9.13	5618
user0027	1199995	20.00	6229
user0028	683331	11.39	4550
user0029	538419	8.97	3710
user0030	960423	16.01	4999
user0031	846177	14.10	4871
user0032	684500	11.41	5335
user0033	1044991	17.42	5959
user0034	1504600	25.08	4977
user0035	1298690	21.64	5682
user0036	639693	10.66	3123
user0037	1021443	17.02	4942
user0038	779141	12.99	5648
user0039	728718	12.15	5010
user0040	915012	15.25	5759
user0041	1601844	26.70	5604
user0042	1248721	20.81	5655
user0043	1248392	20.81	5766
user0044	537593	8.96	5157
user0045	1034266	17.24	6043
user0046	822826	13.71	5762
user0047	2338553	38.98	5039
user0048	1787919	29.80	5557
user0049	987686	16.46	5112

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 11 May 2021 doi:10.20944/preprints202105.0255.v1

user0050	249185	4.15	1237
user0051	562723	9.38	6111
user0052	735564	12.26	5587
user0053	829821	13.83	5451
user0054	1488714	24.81	3309
user0055	4846025	80.77	5088
user0056	553485	9.22	4757
user0057	668374	11.14	4520
user0058	1076782	17.95	5336
user0059	858249	14.30	5565
user0060	733605	12.23	4282
user0061	1218402	20.31	6606
user0062	544233	9.07	4406
user0063	996290	16.60	4725
user0064	1732492	28.87	4896
user0065	1012393	16.87	6829
user0066	1418069	23.63	6781
user0067	1208593	20.14	6465
user0068	630424	10.51	5005
user0069	854492	14.24	6034
user0070	888229	14.80	5590
user0071	2803408	46.72	5954
user0072	987762	16.46	5642
user0073	784983	13.08	5545
user0074	1601963	26.70	6263
user0075	2720502	45.34	5809
user0076	3455548	57.59	3991
user0077	965783	16.10	5523
user0078	791552	13.19	5863
user0079	786153	13.10	5427
user0080	1013715	16.90	5303

Appendix C

Table A3. This is a table. Tables should be placed in the main text near to the first time they are cited.

Key	Key Code	Total number	Percentage (%)
Total		**200299**	
Spacebar	32	32387	16,17
A	65	20965	10,47
E	69	18256	9,11
I	73	15994	7,99
Backspace	8	12195	6,09
T	84	10292	5,14

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 11 May 2021 doi:10.20944/preprints202105.0255.v1

R	82	10030	5,01
N	78	8750	4,37
U	85	8370	4,18
S	83	8210	4,1
C	67	7982	3,99
L	76	6087	3,04
O	79	5780	2,89
M	77	5556	2,77
P	80	5083	2,54
D	68	4982	2,49
,	188	2159	1,08
F	70	2108	1,05
Shift	16	2054	1,03
.	190	1848	0,92
V	86	1811	0,9
B	66	1449	0,72
Z	90	1328	0,66
G	71	1124	0,56
CapsLock	20	988	0,49
[	219	540	0,27
-	189	422	0,21
H	72	363	0,18
J	74	351	0,18
;	186	260	0,13
ArrowLeft	37	230	0,11
0	48	220	0,11
X	88	212	0,11
'	222	200	0,1
ArrowRight	39	188	0,09
]	221	162	0,08
1	49	156	0,08
\	220	98	0,05
Ctrl	17	94	0,05
Enter	13	88	0,04
Alt	18	74	0,04
2	50	74	0,04
K	75	69	0,03
9	57	67	0,03
Y	89	60	0,03
/	191	53	0,03
=	187	52	0,03
3	51	46	0,02
W	87	42	0,02

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 11 May 2021 doi:10.20944/preprints202105.0255.v1

Delete	46	38	0,02
ArrowDown	40	37	0,02
8	56	36	0,02
5	53	34	0,02
ArrowUp	38	28	0,01
7	55	28	0,01
(NumPad)-	109	28	0,01
6	54	27	0,01
-Firefox	173	20	0,01
'	192	19	0,01
NumLock	144	15	0,01
4	52	14	0,01
Tab	9	10	0,005
; Firefox	59	10	0,005
Q	81	7	0,003
(NumPad)8	104	7	0,003
(NumPad)1	97	6	0,003
= Firefox	61	4	0,002
(NumPad)7	103	4	0,002
(NumPad)/	111	4	0,002
End	35	3	0,001
(NumPad)0	96	3	0,001
PageDown	34	2	0,001
(NumPad)3	99	2	0,001
Home	36	1	0,0005
(NumPad)5	101	1	0,0005
(NumPad)6	102	1	0,0005
(NumPad)9	105	1	0,0005

Appendix D

Table A4. This is a table. Tables should be placed in the main text near to the first time they are cited.

No.	Key	KeyCode	Total number	Percentage
	TOTAL		**145261**	
1	A	65	20965	14,43%
2	E	69	18256	12,57%
3	I	73	15994	11,01%
4	T	84	10292	7,09%
5	R	82	10030	6,9%
6	N	78	8750	6,02%
7	U	85	8370	5,76%
8	S	83	8210	5,65%

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 11 May 2021 doi:10.20944/preprints202105.0255.v1

9	C	67	7982	5,49%
10	L	76	6087	4,19%
11	O	79	5780	3,98%
12	M	77	5556	3,82%
13	P	80	5083	3,5%
14	D	68	4982	3,43%
15	F	70	2108	1,45%
16	V	86	1811	1,25%
17	B	66	1449	1%
18	Z	90	1328	0,91%
19	G	71	1124	0,77%
20	H	72	363	0,25%
21	J	74	351	0,24%
22	X	88	212	0,15%
23	K	75	69	0,05%
24	Y	89	60	0,04%
25	W	87	42	0,03%
26	Q	81	7	0,005%

Appendix E

Table A5. The most used di-graphs.

No.	Key Code 1	Key Code 2	Key 1	Key 2	Occurrences	Average of di-graph time
1	8	8	Backspace	Backspace	6938	3,39
2	65	32	A	Spacebar	6663	3,39
3	69	32	E	Spacebar	6630	3,27
4	73	32	I	Spacebar	4034	6,33
5	32	83	Spacebar	S	3866	8,69
6	73	78	I	N	3121	7,31
7	32	67	Spacebar	C	3104	11,09
8	82	69	R	E	3027	6,39
9	32	80	Spacebar	P	2911	12,97
10	32	68	Spacebar	D	2846	11,58
11	32	65	Spacebar	A	2654	12,4
12	65	82	A	R	2606	9,38
13	84	69	T	E	2453	8,19
14	68	69	D	E	2343	8,45
15	67	65	C	A	2289	8,93
16	32	73	Spacebar	I	2170	16,85
17	65	84	A	T	2160	11,51
18	85	32	U	Spacebar	2034	11,61
19	188	32	,	Spacebar	2012	12,51

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 11 May 2021 doi:10.20944/preprints202105.0255.v1

20	32	77	Spacebar	M	1846	18,05
21	84	65	T	A	1812	12,01
22	78	32	N	Spacebar	1769	12,09
23	83	84	S	T	1752	13,84
24	84	73	T	I	1631	13,38
25	82	65	R	A	1540	15,28
26	83	73	S	I	1526	12,71
27	69	65	E	A	1514	17,45
28	78	84	N	T	1508	16
29	83	65	S	A	1498	13,93
30	82	73	R	I	1468	14,03
31	84	32	T	Spacebar	1457	16,65
32	69	83	E	S	1452	18,25
33	67	69	C	E	1435	14,54
34	77	65	M	A	1380	14,44
35	69	82	E	R	1373	16,97
36	32	8	Spacebar	Backspace	1372	46,92
37	85	78	U	N	1325	17,41
38	190	32	,	Spacebar	1311	28,81
39	32	70	Spacebar	F	1287	27,34
40	76	65	L	A	1272	15,25
41	85	76	U	L	1258	21,55
42	32	16	Spacebar	Shift	1210	52,88
43	80	69	P	E	1189	16,93
44	84	82	T	R	1170	17,07
45	67	85	C	U	1167	16,37
46	76	32	L	Spacebar	1154	21,6
47	32	69	Spacebar	E	1153	30,22
48	76	69	L	E	1117	16,62
49	32	76	Spacebar	L	1106	32,08
50	65	67	A	C	1103	22,21
51	80	82	P	R	1071	21,22
52	69	78	E	N	1063	20,61
53	65	78	A	N	1046	21,16
54	65	76	A	L	1043	21,54
55	32	79	Spacebar	O	1035	37,73
56	79	82	O	R	1021	22,86

References

1. Y. Zhong, Y. Deng and A. K. Jain, "Keystroke dynamics for user authentication," 2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, Providence, RI, 2012, pp. 117-123, doi: 10.1109/CVPRW.2012.6239225
2. Iapa A.C., Cretu V.I., Evaluating the performance of authentication algorithms based on keystroke dynamics used in online educational platforms, The 17th International Scientific Conference eLearning and Software for Education, Bucharest, Romania, 2021 – paper sent, unpublished

3. E. Al Solami, C. Boyd, A. Clark, and A. K. Islam, "Continuous Biometric Authentication: Can It Be More Practical?", IEEE Int'l Conf. on High Performance Computing and Communications (HPCC), pp. 647-652, 2010.
4. R. Zack, C. Tappert, and S. Cha, "Performance of a long-text-input keystroke biometric authentication system using an improved k-nearest-neighbor classification method", IEEE Int'l Conf. on Biometrics: Theory Applications and Systems (BTAS), pp. 1-6, 2010.
5. Monrose F, Reiter MK, Wetzel S (2002) Password hardening based on keystroke dynamics. Int J Inf Secur 1(2):69–83
6. Banerjee, Salil & Woodard, D.L.. (2012). Biometric Authentication and Identification Using Keystroke Dynamics: A Survey. Journal of Pattern Recognition Research. 7. 116-139. 10.13176/11.427.
7. Iapa A.C., Cretu V.I., Modified Distance Metric That Generates Better Performance For The Authentication Algorithm Based On Free-Text Keystroke Dynamics, IEEE 15th International Symposium on Applied Computational Intelligence and Informatics, Timisoara, Romania, 2021 – paper sent, unpublished
8. W. E. Burr, D. F. Dodson, and W. T. Polk. Electronic Authentication Guideline: Recommendations of the National Institute of Standards and Technology. Technical Report 800-63,National Institute of Standards and Technology (NIST), Apr. 2006.
9. Kang, Jeonil & Nyang, Daehun & Lee, KyungHee. (2014). Two-factor face authentication using matrix permutation transformation and a user password. Information Sciences. 269. 1–20. 10.1016/j.ins.2014.02.011.
10. Dasgupta, Dipankar & Roy, Arunava & Nag, Abhijit. (2016). Toward the design of adaptive selection strategies for multi-factor authentication. Computers & Security. 63. 10.1016/j.cose.2016.09.004.
11. Tsai CJ, Chang TY, Cheng PC, Lin JH (2014) Two novel biometric features in keystroke dynamics authentication systems for touch screen devices. Sec Commun Netw 7(4):750–758
12. D. Umphress and G. Williams, "Identity Verification through Keyboard Characteristics", Int'l J. Man-Machine Studies, Vol. 23, No. 3, pp. 263-273, 1985.
13. A. Messerman, T. Mustafic, S. A. Camtepe and S. Albayrak. Continuous and non-intrusive identity verification in real-time environments based on free-text keystroke dynamics. Proceedings of IEEE International Joint Conference on Biometrics. 1–8, 2011
14. J. Roth, X. Liu and D. Metaxas, "On Continuous User Authentication via Typing Behavior," in IEEE Transactions on Image Processing, vol. 23, no. 10, pp. 4611-4624, Oct. 2014, doi: 10.1109/TIP.2014.2348802.
15. P. Bours and S. Brahmanpally, "Language dependent challenge-based keystroke dynamics," 2017 International Carnahan Conference on Security Technology (ICCST), Madrid, 2017, pp. 1-6, doi: 10.1109/CCST.2017.8167838.
16. E. A. Solami, C. Boyd, A. Clark and I. Ahmed, "User-representative feature selection for keystroke dynamics," 2011 5th International Conference on Network and System Security, Milan, 2011, pp. 229-233, doi: 10.1109/ICNSS.2011.6060005.
17. T. Samura and H. Nishimura, "Keystroke timing analysis for individual identification in Japanese free text typing," 2009 ICCAS-SICE, Fukuoka, 2009, pp. 3166-3170.
18. Kochegurova, Elena & Luneva, Elena & Gorokhova, Ekaterina. (2019). On Continuous User Authentication via Hidden Free-Text Based Monitoring: Volume 2. 10.1007/978-3-030-01821-4_8.
19. Alsultan, Arwa & wei, Honq & Warwick, Kevin. (2016). Free-text Keystroke Dynamics Authentication for Arabic Language. IET Biometrics. 5. 10.1049/iet-bmt.2015.0101.
20. Junhong Kim, Pilsung Kang, Freely typed keystroke dynamics-based user authentication for mobile devices based on heterogeneous features, Pattern Recognition, Volume 108, 2020, 107556, ISSN 0031-3203
21. Ilonen, Jarmo. (2003). Keystroke dynamics. Advanced Topics in Information processing–lecture (2003).
22. Zilberman, A.G.: Security method and apparatus employing authentication by keystroke dynamics (1998) United States Patent 6,442,692
23. Stefan Koritar. (2020). Romanian startup Typing DNA raises €6.2 million in Series A funding to create 'typing identity' for security (2020).
24. Arwa Alsultan, Kevin Warwick, Hong Wei, Non-conventional keystroke dynamics for user authentication, Pattern Recognition Letters, Volume 89, 2017, Pages 53-59, ISSN 0167-8655
25. Yu, Enzhe & Cho, Sungzoon. (2004). Keystroke dynamics identity verification - Its problems and practical solutions. Computers & Security. 23. 428-440. 10.1016/j.cose.2004.02.004.
26. Zhong, Yu & Deng, Yunbin. (2015). A Survey on Keystroke Dynamics Biometrics: Approaches, Advances, and Evaluations. 10.15579/gcsr.vol2.ch1.
27. D. Gunetti and C. Picardi, "Keystroke analysis of free text," ACM Transactions on Information and System Security, vol. 8, pp. 312–347, 2005
28. Laiu-Despau Octavian, Curiozitati si amuzamente ale limbii romane introducere in ludolingvistica , Editura BrumaR, Timisoara: Brumar, 2012, ISBN 978-973-602-779-6
29. Tsai, Cheng-Jung & Shih, Kuen-Jhe. (2019). Mining a new biometrics to improve the accuracy of keystroke dynamics-based authentication system on free-text. Applied Soft Computing. 80. 10.1016/j.asoc.2019.03.033.
30. F. Monrose and A. Rubin. Authentication via keystroke dynamics. Proceedings of the 4th ACM Conference on Computer and Communications Security. 48–56, 1997.
31. M. Villani, C. Tappert, G. Ngo, J. Simone, H.S. Fort, S.H. Cha, Keystroke biometric recognition studies on long-text input under ideal and applicationoriented conditions, in: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition Workshop, June, 2006, pp. 39–46.

 doi:10.20944/preprints202105.0255.v1

32. H. Davoudi, E. Kabir, A new distance measure for free text keystroke authentication, in: Proceedings of the 14th International CSI Computer Conference, October, 2009, pp. 570–575.
33. S. Park, J.P. Cho, User Authentication based on keystroke analysis of long free texts with a reduced number of features, in: Proceedings of IEEE International Conference on Communication Systems, Networks and Applications, July, 2010, pp. 433–435.
34. A. Alsultan, K. Warwick, H. Wei, Improving the performance of free-text keystroke dynamics authentication by fusion, Appl. Soft Comput. 70 (2018) 1024–1033.
35. A. Alsultan, K. Warwick, H. Wei, Non-conventional keystroke dynamics for user authentication, Pattern Recognit. Lett. 89 (2017) 53–59.
36. J. Kim, H. Kim, P. Kang, Keystroke dynamics-based user authentication using freely typed text based on user-adaptive feature extraction and novelty detection, Appl. Soft Comput. 62 (2018) 1077–1087.