Advancements in Dating Undated Manuscripts through Dual Methodologies[*]

Maksim Iavich[1,*,†], Maia Ninidze[2,†]

[1] *Caucasus University, School of Technology, 1 Paata Saakadze St, Tbilisi 0102, Georgia*
[2] *Ivane Javakhishvili Tbilisi State University, 1 Chavchavadze av., Tbilisi 0179, Georgia*

Abstract

Manuscript dating, particularly in the context of analyzing handwritten materials, poses a distinct challenge compared to author identification in anonymous holographs. The intricacy arises from the broader spectrum of differences in handwriting styles among various authors, overshadowing the subtler variations within the handwriting of a single author across different years. To address this complexity, our research explores diverse methodologies and technologies for accurately dating the undated holographs of Galaktion Tabidze, a prominent Georgian poet of the 20th century. This article delineates two distinct approaches employed in our study, presenting experiments conducted to assess the efficacy of the proposed dating method. By delving into these methodologies, we aim to contribute valuable insights and enhance the accuracy of dating historical manuscripts.

Keywords

Automatic dating approach; Neural network; Automated handwriting dating, Georgian handwriting, Georgian studies

1. Introduction

Cursive manuscripts, imbued with the fluidity of handcrafted strokes and the enigma of tied letter pairs, serve as intricate windows into linguistic evolution and historical context. In our exploration, we undertake a dual-methodological journey, blending traditional graphematic studies with the innovative capabilities of neural networks [1-3]. This study centers on the poetic manuscripts of Galaktion Tabidze, where the continuous flow of the pen and the distinctive artistry of tied letters pose unique challenges for deciphering temporal nuances.

We offer dual methodologies for the dating of manuscripts. The first one is the manual, traditional one and another one uses machine learning techniques.

The manual approach, as detailed in the preliminary stages of this research, involves a meticulous analysis of tied letter pairs, recognizing their interconnectedness as not only a temporal signifier but also an expressive element in the poet's work. We present a template, crafted with special tables encompassing all 33 letters of the Georgian alphabet, cataloging 561 tied letter pairs observed in manuscripts spanning the years 1907 to 1959. This traditional method, while invaluable in providing qualitative insights, faces challenges in scalability and objectivity. In tandem with the manual methodology, we introduce a revolutionary neural network-based approach. Leveraging a diverse dataset of cursive writing samples, our neural network model is trained to autonomously decipher evolving grapheme forms and temporal patterns within tied letter pairs. This computational framework adds a layer of efficiency and objectivity, complementing the rich qualitative data obtained through the manual method.

Our experiments aim to seamlessly integrate the outputs from both approaches, offering a holistic understanding of the temporal evolution of Galaktion Tabidze's manuscripts. The neural network,

[*] *IVUS2024: Information Society and University Studies 2024, May 17, Kaunas, Lithuania*
[1,*] Corresponding author
[†] These author contributed equally.

✉ miavich@cu.edu.ge (M. Iavich); maia.ninidze@tsu.ge (M. Ninidze)
iD 0000-0002-3109-7971 (M. Iavich); 0000-0002-6552-2163 (M. Ninidze)

functioning as a time-traveling computational companion, collaborates with the manual template to systematically organize and analyze the tied letter pairs. By navigating the interplay between manual and automated methods, we anticipate uncovering latent temporal trends and refining the dating process.

Our work doesn't only help us learn about Georgian manuscripts. It also adds to the bigger conversation about using both old and new ways to study history and language. We think that by combining looking closely and using computers, we can better understand the details of cursive writing over time, like a woven story across the years.

2. Literature review

The literature in grapheme-to-phoneme conversion presents diverse approaches to tackle the complexities of this linguistic process. Weingarten [1] delves into comparative graphematics, offering valuable insights into the representation of graphemes. Andersen et al. [2] contribute by comparing tree-structured approaches, providing a nuanced understanding of grapheme-to-phoneme conversion strategies. Kheang et al. [4] propose a two-stage neural network-based solution, addressing conflicts in phoneme conversion.

In the realm of handwritten Arabic grapheme segmentation, Elkhayati et al. [3] employ a directed convolutional neural network and mathematical morphology operations, showcasing advancements in the understanding and application of segmentation techniques. Turning to the field of manuscript analysis, researchers have explored various aspects, such as dating, localization, and preservation. Wahlberg et al. [6] focus on large-scale style-based dating of medieval manuscripts, providing valuable insights into dating methodologies. Legendre [7] introduces tools for dating and localizing manuscripts, contributing to the broader discussion on manuscript analysis. Omayio et al. [8] offer an overview of traditional and modern trends in historical manuscript dating, enriching the understanding of evolving practices. Karlsson [9] addresses the localization and dating of medieval Icelandic manuscripts, adding a unique perspective to the broader discourse.

Conservation and restoration efforts for historical manuscripts are explored by Hajji et al. [10], who present a multi-analytical approach for evaluating the efficiency of conservation-restoration treatments, underscoring the interdisciplinary nature of preserving historical documents.

Deep learning techniques have found application in historical manuscript analysis. Hamid et al. [11] propose a deep learning-based approach for historical manuscript dating, reflecting the increasing integration of advanced technologies in traditional disciplines. Boudraa and Bennour [12] combine local features and deep learning for historical manuscripts dating, contributing to the growing body of literature on the subject. Assael et al. [13] focus on restoring and attributing ancient texts using deep neural networks, showcasing the potential of AI in historical text restoration.

The automated dating of handwritten texts is a notable focus, with Tvalavadze et al. [14] presenting an approach for Galaktion Tabidze's handwritten texts. He et al. [15] contribute with a multiple-label guided clustering algorithm for historical document dating and localization, offering innovative methods in the automated analysis of historical texts.

3. Manual approach

The most challenging problem at the initial stage of graphematic studies of cursive manuscripts is the following: writing without lifting a pen and tying letters to each other causes specific changes in the grapheme forms. In case we split them, we can’t get the same forms that those graphemes would have if they were written separately. Therefore, we decided not to extract graphematic components of the tied letters but to study them as whole units. The style of their tying might be no less meaningful for graphematic studies than the forms of the graphemes. As our aim was to date Galaktion Tabidze’s manuscripts, we created a template with special tables for all 33 letters of Georgian alphabet and all the possible letter pairs that might be tied to each other. This particular author’s manuscripts observed by us revealed 561 such pairs. Then we started to extract images of the graphematic units from the manuscripts of different years and to put them in the corresponding tables. Figure 1 illustrates the images of a pair of Georgian letters ან [an] from a manuscript dated back to 1914.

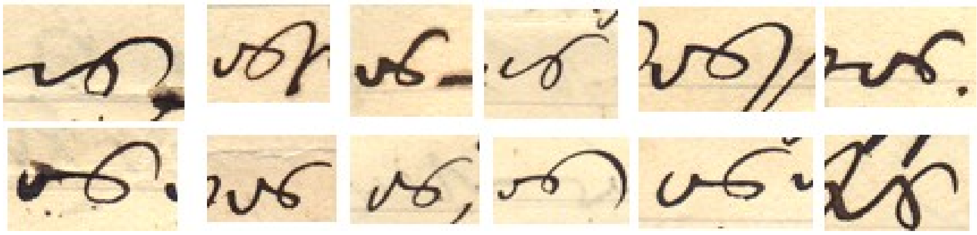


Fig. 1. Unit ან in 1914

On the whole we analyzed 130 handwritten documents of the years 1907 - 1959 i. e. two or three manuscripts per year. Thus, we created a database of possible forms for every graphematic unit (a letter or a pair of tied letters) represented in the documents of each year. In cases when several letters were tied, we grouped them in the following way: first two, second and third, third and fourth, fourth and fifth etc., so that both – the left and the right sides of all the letters from the graphematic chain were represented and considered.

Galaktion Tabidze, as many other authors, used to go back to his earlier manuscripts to resume creative work on them. In some texts there are passages with a great number of crossed out and respelled words. In the cases when there is a difference in ink or pencil color, it is easy to understand that the interventions are made later but we should know that any correction that is not in line with the initial text may be made later. If the words are struck out and the alternative ones are overwritten, written between lines or on the margins, this may be done even years after the first layer was created. Therefore, it was decided not to include letters from such passages in the database of the graphematic units. The data of every dated document should represent the images of the indicated year and not of the later layers of the manuscript.

While creating a chronological database of the dated documents, one more specific case was considered. It is obvious that at the end of the text authors indicate the time of their creation but what is indicated by the date written in the beginning of a literary work, a section with multiple texts or on the notebook cover. As experience shows, if there are only digits, without comments, they also indicate the time of the text creation and not the time of making their copy in the particular substrate. Galaktion Tabidze's notebook (GTDA 476) with the date 1909 at the top of the first page, includes his poems written in the indicated year but they are copied much later. When we know the date of the text creation, the only thing that we can say by sure about any of its clear copies is that they should be written later. As any misdated data included in the training set, may have negative impact on the results of the research we have to be very careful about it.

At this stage of investigation, the data of the grapheme forms were grouped in 130 files each describing a particular manuscript of a particular year. Such organization of the data was convenient for the overview of all the units of one and the same year but not for the detection of the changes undergone by particular graphematic units in course of time. To organize the data in a more convenient way for our survey, we grouped images of each of those 594 graphematic units detected in the observed documents in the chronological order. This made it easier to look through the changes undergone by each unit year by year. It became clear that in most cases changes were made not only in the whole units but in their particular parts, their smaller elements. It was decided to make these small elements of graphematic units the main focus of our investigation. So, it was necessary to identify them. As the number of units was almost six hundred, their parts or smaller elements that used to change year by year, would be much more. Therefore, we decided to start our research with the analysis of only those 33 Georgian Graphemes that were written separately.

In order to group and organize the changes undergone by the graphemes written separately, we used cardinal numbers. We indicated by numbers not only changing elements of the graphematic units but also all the detected types of those elements. Figure 2 illustrates eight types of the first element (upper part) of the letter ე[e] revealed in the whole database of the observed documents dated back from 1907 to 1959.

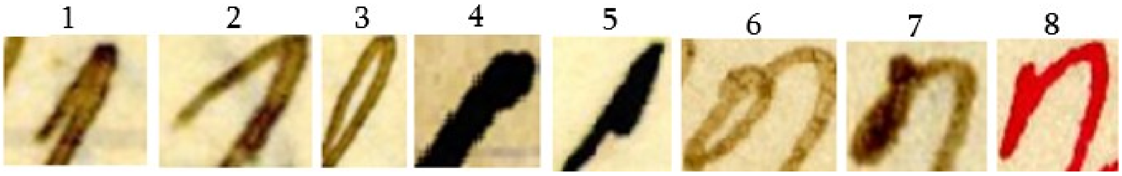
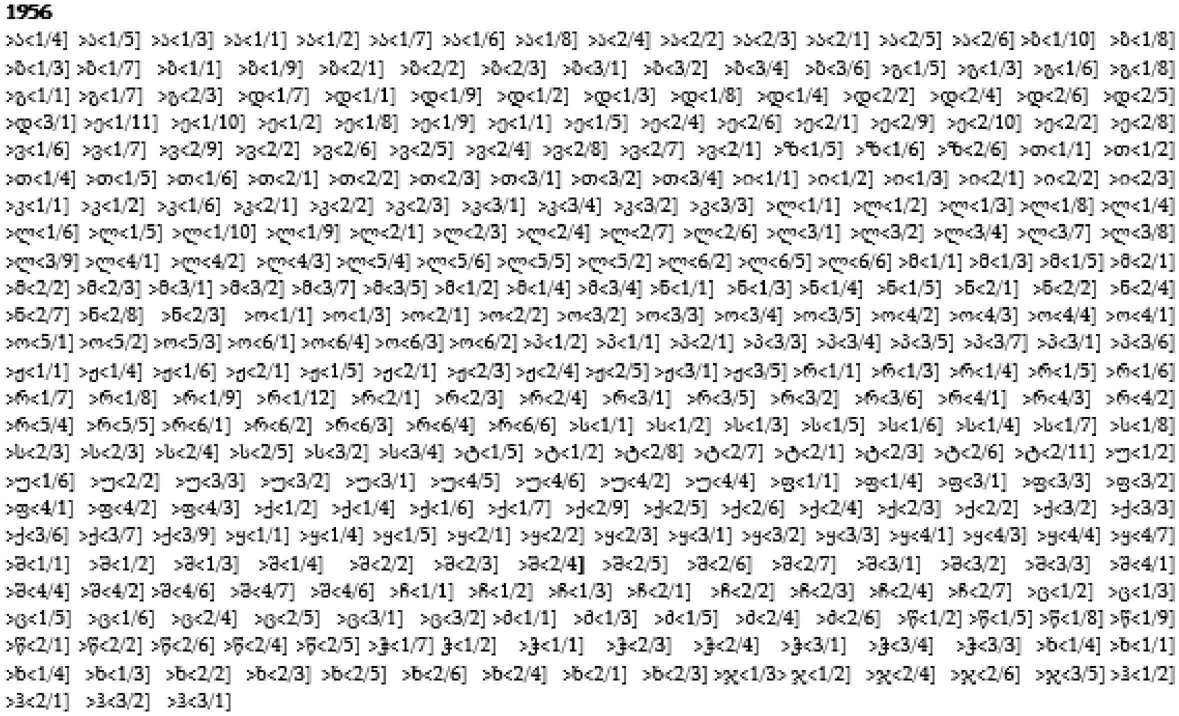


Fig. 2. Types of the upper part of the letter ე

We identified from 2 to 6 changing elements in graphemes and from 2 to 12 types of those elements. In the letters that have simple forms like ა [a], we observed the remarkable changes of only two elements – the upper and the lower parts of the letter but in the graphemes that have more complicated forms we indicated up to six elements. E. g. in the letter ლ [l] changes occurred in upper left, upper middle, upper right, lower left and lower right parts and in the horizontal line that the poet often added above this grapheme. On the whole, we identified 436 Graphematic Element Types (GET) for all the letters of Georgian alphabet written separately in Galaktion Tabidze's 130 documents. These Graphematic Element Types were considered to be the basic date-informative features in the investigation.

We constructed the encoding system from three variables: the unit name (it is put between the angle brackets), cardinal number given by us to the changing element of the letter (it is written before the slash) and a cardinal number given by us to the particular type of this element (is written after the slash). A square bracket was used as an ending mark. So, the code of the eighth type of the first element (upper part) of the letter ე [e] given in Fig. 2, is >ე<1/8. Having encoded all the 436 GETs and having created a codebook, we started to identify all types of all graphemes, characteristic for every particular year of the period 1907-1959 and created a chronological database. Figure 3 illustrates the data of only one year:



1956

>ა<1/4] >ა<1/5] >ა<1/3] >ა<1/1] >ა<1/2] >ა<1/7] >ა<1/6] >ა<1/8] >ა<2/4] >ა<2/2] >ა<2/3] >ა<2/1] >ა<2/5] >ა<2/6] >ბ<1/10] >ბ<1/8]
>ბ<1/3] >ბ<1/7] >ბ<1/1] >ბ<1/9] >ბ<2/1] >ბ<2/2] >ბ<2/3] >ბ<3/1] >ბ<3/2] >ბ<3/4] >ბ<3/6] >გ<1/5] >გ<1/3] >გ<1/6] >გ<1/8]
>გ<1/1] >გ<1/7] >გ<2/3] >დ<1/7] >დ<1/1] >დ<1/9] >დ<1/2] >დ<1/3] >დ<1/8] >დ<1/4] >დ<2/2] >დ<2/4] >დ<2/6] >დ<2/5]
>დ<3/1] >ე<1/11] >ე<1/10] >ე<1/2] >ე<1/8] >ე<1/9] >ე<1/1] >ე<1/5] >ე<2/4] >ე<2/6] >ე<2/1] >ე<2/9] >ე<2/10] >ე<2/2] >ე<2/8]
>ვ<1/6] >ვ<1/7] >ვ<2/9] >ვ<2/2] >ვ<2/6] >ვ<2/5] >ვ<2/4] >ვ<2/8] >ვ<2/7] >ვ<2/1] >ზ<1/5] >ზ<1/6] >ზ<2/6] >თ<1/1] >თ<1/2]
>თ<1/4] >თ<1/5] >თ<1/6] >თ<2/1] >თ<2/2] >თ<2/3] >თ<3/1] >თ<3/2] >თ<3/4] >ი<1/1] >ი<1/2] >ი<1/3] >ი<2/1] >ი<2/2] >ი<2/3]
>კ<1/1] >კ<1/2] >კ<1/6] >კ<2/1] >კ<2/2] >კ<2/3] >კ<3/1] >კ<3/4] >კ<3/2] >კ<3/3] >ლ<1/1] >ლ<1/2] >ლ<1/3] >ლ<1/8] >ლ<1/4]
>ლ<1/6] >ლ<1/5] >ლ<1/10] >ლ<1/9] >ლ<2/1] >ლ<2/3] >ლ<2/4] >ლ<2/7] >ლ<2/6] >ლ<3/1] >ლ<3/2] >ლ<3/4] >ლ<3/7] >ლ<3/8]
>ლ<3/9] >ლ<4/1] >ლ<4/2] >ლ<4/3] >ლ<5/4] >ლ<5/6] >ლ<5/5] >ლ<5/2] >ლ<6/2] >ლ<6/5] >ლ<6/6] >მ<1/1] >მ<1/3] >მ<1/5] >მ<2/1]
>მ<2/2] >მ<2/3] >მ<3/1] >მ<3/2] >მ<3/7] >მ<3/5] >მ<1/2] >მ<1/4] >მ<3/4] >ნ<1/1] >ნ<1/3] >ნ<1/4] >ნ<1/5] >ნ<2/1] >ნ<2/2] >ნ<2/4]
>ნ<2/7] >ნ<2/8] >ნ<2/3] >ო<1/1] >ო<1/3] >ო<2/1] >ო<2/2] >ო<3/2] >ო<3/3] >ო<3/4] >ო<3/5] >ო<4/2] >ო<4/3] >ო<4/4] >ო<4/1]
>ო<5/1] >ო<5/2] >ო<5/3] >ო<6/1] >ო<6/4] >ო<6/3] >ო<6/2] >პ<1/2] >პ<1/1] >პ<2/1] >პ<3/3] >პ<3/4] >პ<3/5] >პ<3/7] >პ<3/1] >პ<3/6]
>ჟ<1/1] >ჟ<1/4] >ჟ<1/6] >ჟ<2/1] >ჟ<1/5] >ჟ<2/1] >ჟ<2/3] >ჟ<2/4] >ჟ<2/5] >ჟ<3/1] >ჟ<3/5] >რ<1/1] >რ<1/3] >რ<1/4] >რ<1/5] >რ<1/6]
>რ<1/7] >რ<1/8] >რ<1/9] >რ<1/12] >რ<2/1] >რ<2/3] >რ<2/4] >რ<3/1] >რ<3/5] >რ<3/2] >რ<3/6] >რ<4/1] >რ<4/3] >რ<4/2]
>რ<5/4] >რ<5/5] >რ<6/1] >რ<6/2] >რ<6/3] >რ<6/4] >რ<6/6] >ს<1/1] >ს<1/2] >ს<1/3] >ს<1/5] >ს<1/6] >ს<1/4] >ს<1/7] >ს<1/8]
>ს<2/3] >ს<2/3] >ს<2/4] >ს<2/5] >ს<3/2] >ს<3/4] >ტ<1/5] >ტ<1/2] >ტ<2/8] >ტ<2/7] >ტ<2/1] >ტ<2/3] >ტ<2/6] >ტ<2/11] >უ<1/2]
>უ<1/6] >უ<2/2] >უ<3/3] >უ<3/2] >უ<3/1] >უ<4/5] >უ<4/6] >უ<4/2] >უ<4/4] >ფ<1/1] >ფ<1/4] >ფ<3/1] >ფ<3/3] >ფ<3/2]
>ფ<4/1] >ფ<4/2] >ფ<4/3] >ქ<1/2] >ქ<1/4] >ქ<1/6] >ქ<1/7] >ქ<2/9] >ქ<2/5] >ქ<2/6] >ქ<2/4] >ქ<2/3] >ქ<2/2] >ქ<3/2] >ქ<3/3]
>ქ<3/6] >ქ<3/7] >ქ<3/9] >ყ<1/1] >ყ<1/4] >ყ<1/5] >ყ<2/1] >ყ<2/2] >ყ<2/3] >ყ<3/1] >ყ<3/2] >ყ<3/3] >ყ<4/1] >ყ<4/3] >ყ<4/4] >ყ<4/7]
>შ<1/1] >შ<1/2] >შ<1/3] >შ<1/4] >შ<2/2] >შ<2/3] >შ<2/4] >შ<2/5] >შ<2/6] >შ<2/7] >შ<3/1] >შ<3/2] >შ<3/3] >შ<4/1]
>შ<4/4] >შ<4/2] >შ<4/6] >შ<4/7] >შ<4/6] >ჩ<1/1] >ჩ<1/2] >ჩ<1/3] >ჩ<2/1] >ჩ<2/2] >ჩ<2/3] >ჩ<2/4] >ჩ<2/7] >ც<1/2] >ც<1/3]
>ც<1/5] >ც<1/6] >ც<2/4] >ც<2/5] >ც<3/1] >ც<3/2] >ძ<1/1] >ძ<1/3] >ძ<1/5] >ძ<2/4] >ძ<2/6] >წ<1/2] >წ<1/5] >წ<1/8] >წ<1/9]
>წ<2/1] >წ<2/2] >წ<2/6] >წ<2/4] >წ<2/5] >ჭ<1/7] ჭ<1/2] >ჭ<1/1] >ჭ<2/3] >ჭ<2/4] >ჭ<3/1] >ჭ<3/4] >ჭ<3/3] >ხ<1/4] >ხ<1/1]
>ხ<1/4] >ხ<1/3] >ხ<2/2] >ხ<2/3] >ხ<2/5] >ხ<2/6] >ხ<2/4] >ხ<2/1] >ხ<2/3] >ჯ<1/3> ჯ<1/2] >ჯ<2/4] >ჯ<2/6] >ჯ<3/5] >ჰ<1/2]
>ჰ<2/1] >ჰ<3/2] >ჰ<3/1]

Fig. 3. Encoded data

At the validation stage we extracted images from 5 dated manuscripts that were not included in the training set, identified GETs, and using the "search" engine highlighted the coincidences in each year of the codebook data, hoping that the greater number of similarities would point at the year of the document creation. As a result, the greater number of coincidences was revealed with those years the documents of which were greater in volume and accordingly had a greater number of detected GETs. There might be one more reason for the incorrect results: while identifying GET codes of particular years we generally made an account of different forms revealed in the document, not paying attention to the quantity of each. Those with a greater number (sometimes tens) might be a more essential feature for the year and those with only a single or a pair of samples – random. Such random forms of letters cannot have the same value in the investigation as those that are frequently used and very peculiar for the particular year. It became clear that for the better representation of the date-informative features this difference should have been considered.

In order to avoid the above-mentioned confounding factors, the next experiment was carried out with the database of only two years (one of which was the year of the document's creation), volumes of the observed manuscripts of which were equal in size. Besides, in order to differentiate the value of each GET for the particular manuscript or frequency of its use, we calculated percentage of the number of particular GETs in the document from the total number of this particular letter in it and used this digit as a coefficient. For example, if there were 50 letters ბ [b] in the whole document and only 15 of them were of the type >ბ<3/2], we added coefficient 30 to the code – GET 30>ბ<3/2]. After highlighting all the coincidences in the data of the two years, we summed up their coefficients separately and compared to each other but again without success – the bigger number did not point at the correct year. Analyzing the reasons, we concluded that the failure might be caused by rarely used Georgian letters. When this or that grapheme is used only once in the document, its GETs, even if they are not highly date-informative, take coefficient 1 from 1, i. e. 100, while very specific GET of some other letter that is used 9 times, if the total number of this letter in the document is 10, takes as a coefficient smaller digit – 90. At present we are looking for other ways of avoiding all the confounding factors revealed in the previous experiments and are continuing our work on graphematic pairs.

While analyzing different values of GETs in different years, we noticed that some of them are used in a rather short period of time and this might be efficiently used for the identification of the documents belonging to those years. Our attention was attracted by a very unusual for Georgian script form. This is a second type of the second element (lower part) of the grapheme გ [g] >გ<2/2]. See Figure 4.

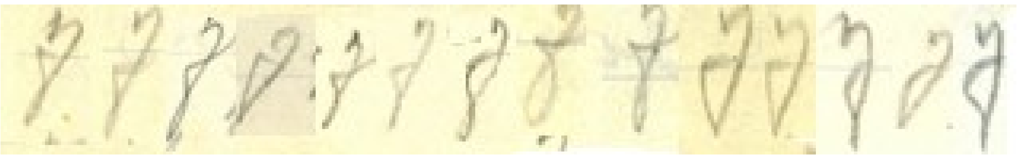


Fig. 4. GET >გ<2/2]

Its lines are screwed like in the digit 8. This form is applied in Galaktion Tabidze's manuscripts written only in the years 1908-1910. One voluminous handwritten text (GTDA 1040) started by the author in 1907 and continued in 1908 with corresponding dates at the end of each section, helped us to comprehend the changes in the author's handwriting of the period. It was revealed that there is not a single case of using the above mentioned GET in 1907 while they are quite numerous in the part of the document written in 1908. We found 120 undated documents with the similar GET in the poet's digital archive but there were dated ones as well and according to those dates, this GET should be used by the author from April 1908 to December 1910.

In order to identify the dates with an accuracy of a year we carried out comparative analysis of the encoded data of these three years and got sure that it was rather an easy task to pick out the documents belonging to the year 1908 as in that year the author used very specific forms of the graphemes ნ [n] and ზ [z] and there was not a single case of applying two particular GETs of the graphemes ო [o] and ლ [l] that are widely used in the documents of all the following years. These date-informative features were quite enough for us to belong the great Georgian poet's one personal letter, eight poems and a translated story to the year 1908 (GTDA 20, 52, 182, 416, 784, 791, 731). There were four documents in the poet's archive with similar ნ [n], ზ [z], ო [o] and ლ [l] but without single case of GET >გ<2/2]. As Galaktion Tabidze started to use this form since April 1908, we dated those four documents: GTDA 415, 694, 778 and 1130 back to the period before April 1908.

Analyzing GETs of other graphemes, used in the rest 113 undated documents with the specific type of the grapheme გ [g], we revealed that there are 21 manuscripts with the particular form of the grapheme დ [d]. Having analyzed the dated documents of the period, we found out that these two specific GETs should be used together from September to December in 1910. So, the manuscripts:

GTDA 146, 189, 365, 380, 414, 425, 460, 468, 503 (p. 9-11), 531, 564, 633, 725, 777, 866, 895, 911, 967, 1116, 1181 and 1201 that include several poems, dramatic poems, New Year rhymes and documentary texts, were dated back to September-December 1910. All the rest 92 manuscripts (GTDA 3, 4, 16, 19, 23, 30, 40, 46, 58, 133, 149, 152, 166, 172, 174, 181, 184, 185, 197, 207, 209, 261, 280, 287, 289, 339, 340, 344, 353, 363, 374, 379, 387, 390, 392, 393, 404, 405, 423, 456, 475, 501, 530, 534, 551, 563, 566, 588, 593, 613, 632, 639, 640, 706, 721, 725, 728, 733, 738, 742, 754, 770, 785, 786, 894, 899, 904, 931, 934, 942, 955-957, 965, 970, 1028, 1119, 1127, 1131, 1150, 1161, 1165, 1166, 1168, 1169, 1182, 1183, 1192, 1213, 1214, 1217 and 1219) are given approximate date – from 1909 to September 1910.

4. Automated approach

For the automated approach we have designed the neural network. We have used the labeled images consisting of fifty-four different years for the dataset. We followed a systematic process to train a neural network for image recognition within the context of a mentioned dataset.

We initiated the process by curating a dataset laying the groundwork for our program to understand various image categories. To enhance its versatility, we introduced data augmentation techniques, such as flipping images horizontally or vertically.

The dataset was then divided into a training set, where the computer learned, and a validation set, enabling us to assess the model's learning progress. Specifically, 75% of the data was allocated for training, while 25% was reserved for validation.

For constructing the neural network, TensorFlow and Keras were employed to create a layered architecture, mimicking the brain's ability to recognize features like shapes, colors, and patterns. Training involved exposing the program to numerous images from the training set, guiding it to discern patterns and fine-tune its 'brain' for image comprehension.

We evaluated the program's performance by testing it with unseen images, gauging accuracy and confidence in classifications.

The acquired knowledge of the neural network was saved for future use, streamlining subsequent analyses.

Here is offered the pseudo code for the system:

```
 Load necessary libraries
Load required libraries: NumPy, PIL, TensorFlow, Keras, Matplotlib, and others
 Set up dataset
Define the path to the dataset directory
Create an image folder dataset using TensorFlow datasets
Print dataset information

 Configure batch size and image dimensions
Set batch size for processing images
Define image height and width
 Prepare training and validation datasets
Create training and validation datasets using image_dataset_from_directory:
  - 25% of the data for validation (validation_split=0.25)
  - Subset set to "training" for training dataset
  - Subset set to "validation" for validation dataset
  - Image size and batch size are specified
 Explore class names
Print the number of class names in the dataset
 Visualize sample images
Display a 3x3 grid of sample images with corresponding class names
 Normalize and cache datasets
Apply normalization to pixel values
Cache and prefetch training and validation datasets for improved performance
 Data augmentation
Implement data augmentation using random horizontal and vertical flips
```

```
 Define the neural network model
Build a convolutional neural network model with three convolutional layers, max-pooling, and dense layers
Compile the model with Adam optimizer and sparse categorical cross-entropy loss
 Train the model
Fit the model to the training dataset, validating on the validation dataset
Train for 30 epochs, with early stopping and model checkpointing
 Save and load the model
Save the trained model to Google Drive
Load the saved model for further evaluation
 Evaluate on test images
Load images from a specified directory for testing
Predict class labels for each image and store the top three predictions in a dictionary
 Display results
Print the dictionary containing the top three predicted class labels for each test image
```

We have trained the model and received the maximal accuracy score on 24-th epoch.
The validation accuracy score of our model was 75.11 %.

5. Methodology

This methodology section provides a comprehensive overview of the procedures followed in both the manual and automated approaches for analyzing handwritten manuscripts of Galaktion Tabidze.

For the manual approach the following procedures where implemented:

Data Collection and Organization: Handwritten manuscripts attributed to Galaktion Tabidze were collected for analysis. A template with special tables for all 33 letters of the Georgian alphabet and possible letter pairs was created. Images of graphematic units from manuscripts dated between 1907 and 1959 were extracted and organized into corresponding tables, grouped by year.

Analysis: Each manuscript was meticulously examined to identify grapheme forms. Graphematic units were analyzed as whole units without splitting tied letters. Changing elements within graphemes were identified and encoded using cardinal numbers. A database of graphematic element types (GETs) was constructed, representing the basic date-informative features.

Validation: The encoded data were validated using separate manuscripts. Coincidences between the encoded data and the manuscripts were highlighted to assess accuracy.

Refinement: Various techniques were explored to avoid confounding factors. Specific attention was given to grapheme forms unique to certain years to improve accuracy in identifying document origins.

For the automated approach the following techniques were used:

Data Preparation: A labeled dataset comprising images of handwritten manuscripts from different years was curated. Data augmentation techniques, including horizontal and vertical flips, were applied to enhance dataset variability.

Model Building: A convolutional neural network (CNN) model was constructed using TensorFlow and Keras. The model architecture mimicked human visual perception, enabling it to recognize features in images. The model was trained on the training set and validated on the validation set.

Model Evaluation: The trained model was evaluated using unseen images to assess its accuracy and confidence in classifying handwritten manuscripts.

Model Deployment: The trained model was saved for future use and implemented to facilitate further analyses of handwritten manuscripts.

Validation and Comparison: The accuracy and effectiveness of both manual and automated approaches were evaluated and compared. Metrics such as accuracy, precision, recall, and F1 score were calculated to assess the performance of each approach in identifying the origin and date of handwritten manuscripts.

6. Experiments

We have tested our system using the images, which were not in our set. We took into consideration three years, predicted with the highest probability.

Table 1 illustrates the received results:

Table 1
Experiments table

Title	Correct year	Predicted years
14	1928	1925, 1921, 1927
15.1	1912	1912, 1913,1914
13	1929	1931,1928,1929
10	1928	1930, 1931, 1935
18	1912	1958,1908,1943
17.1	1912	1941,1955,1949
16.3	1912	1925,1943,1939
16.2	1912	1911,1912,1947
16.1	1912	1907,1908,1911
15.2	1912	1913,1929,1915
22.1	1922	1908,1921,1909
21.2	1915	1913, 1909,1958
21.1	1915	1913,1926,1958
19	1955	1908,1948,1954
24.2	1914	1911,1912,1913
24.1	1914	1914,1941,1908
22.2	1928	1950, 1955, 1930
27.4	1937	1935, 1930, 1936
27.3	1937	1935, 1940, 1937
27.2	1937	1930, 1935, 1937
27.1	1937	1935,1931,1937
29.2	1948	1948,1957,1955
29.1	1948	1948,1957,1955
28	1948	1947, 1956, 1948
7	1910	1941, 1914, 1935
30	1956	1956, 1954, 1949
9.1	1915	1914, 1915, 1908
8	1915	1914, 1907, 1905
9.2	1915	1915, 1914, 1930

In our analysis of predicting the creation years of artworks using the provided neural network model, we observed notable outcomes. Among the 29 titles considered, the predictive model exhibited diverse levels of accuracy.

Firstly, the model performed well in predicting the correct year as the top choice for a subset of titles, achieving a success rate of 17.2%. Additionally, the second prediction demonstrated reasonable accuracy, correctly identifying the creation year in the second position for 6.9% of the titles. Furthermore, the third prediction added another layer of reliability, successfully forecasting the correct year for 24.1% of the titles in the third position of predictions.

Remarkably, a significant portion of the predictions fell within a ±3 year range of the actual creation year. Approximately 48.3% of the titles had predictions that were closely related to the correct year, highlighting the model's ability to capture temporal proximity.

7. Conclusions and Future Plans

While the model showed promise in predicting creation years, it's essential to acknowledge the challenges associated with the inherent complexity and subjectivity of handwritings. It must be mentioned that all the handwritings were written by the same person. This factor can contribute to variations in the predicted years.

In conclusion, our neural network-based approach, despite its inherent challenges, provides valuable insights into predicting creation years of artworks. The combination of accurate top predictions and the model's ability to identify closely related years enhances its utility in the nuanced task of dating artworks. As we continue refining and expanding our dataset, we anticipate further improvements in the model's performance. Additionally, our examination of handwritten documents spanning 1907 to 1959 revealed intricate challenges in graphematic analysis. Despite these challenges, we refined our approach, leveraging distinctive grapheme forms for precise dating.

The combination of automate approach together with the manual approach described in the paper can bring better results.

8. Acknowledgements

This work was supported by The Association for Textual and Editorial Studies and Digital Humanities under grant [No. ATESDH-6] – Creation of an algorythm and a working web-platform for the graphematic studies.

9. References

[1] Weingarten, Rüdiger. "Comparative graphematics." Written Language & Literacy 14.1 (2011): 12-38.

[2] Andersen, Ove, et al. "Comparison of two tree-structured approaches for grapheme-to-phoneme conversion." Proceeding of Fourth International Conference on Spoken Language Processing. ICSLP'96. Vol. 3. IEEE, 1996.

[3] Elkhayati, Mohsine, Youssfi Elkettani, and Mohammed Mourchid. "Segmentation of handwritten arabic graphemes using a directed convolutional neural network and mathematical morphology operations." Pattern Recognition 122 (2022): 108288.

[4] Kheang, Seng, et al. "Solving the phoneme conflict in grapheme-to-phoneme conversion using a two-stage neural network-based approach." IEICE TRANSACTIONS on Information and Systems 97.4 (2014): 901-910.

[5] Eyben, Florian, et al. "From speech to letters-using a novel neural network architecture for grapheme based ASR." 2009 IEEE Workshop on Automatic Speech Recognition & Understanding. IEEE, 2009.

[6] Wahlberg, Fredrik, Lasse Mårtensson, and Anders Brun. "Large scale style based dating of medieval manuscripts." Proceedings of the 3rd International Workshop on Historical Document Imaging and Processing. 2015.

[7] Legendre, Olivier. "Some tools for dating and localizing manuscripts." The Journal of the Early Book Society for the Study of Manuscripts and Printing History 11 (2008): 181-197.

[8] Omayio, Enock Osoro, Sreedevi Indu, and Jeebananda Panda. "Historical manuscript dating: traditional and current trends." Multimedia Tools and Applications 81.22 (2022): 31573-31602.

[9] Karlsson, Stefán. "The localisation and dating of medieval Icelandic manuscripts." Saga-book 25 (1998): 138-158.

[10] Hajji, Latifa, et al. "A multi-analytical approach for the evaluation of the efficiency of the conservation–restoration treatment of Moroccan historical manuscripts dating to the 16th, 17th, and 18th centuries." Applied spectroscopy 69.8 (2015): 920-938.

[11] Hamid, Anmol, et al. "Deep learning based approach for historical manuscript dating." 2019 International Conference on Document Analysis and Recognition (ICDAR). IEEE, 2019.

[12] Boudraa, Merouane, and Akram Bennour. "Combination of local features and deep learning to historical manuscripts dating." International Conference on Intelligent Systems and Pattern Recognition. Cham: Springer Nature Switzerland, 2023.

[13] Assael, Yannis, et al. "Restoring and attributing ancient texts using deep neural networks." Nature 603.7900 (2022): 280-283.

[14] Tvalavadze, Tea, et al. "Automated Dating of Galaktion Tabidze's Handwritten Texts." International Conference on Computer Science, Engineering and Education Applications. Cham: Springer Nature Switzerland, 2023.

[15] He, Sheng, et al. "A multiple-label guided clustering algorithm for historical document dating and localization." IEEE Transactions on Image Processing 25.11 (2016): 5252-5265.