An Approach to Recognizing and Digitizing Old Handwritten Documents with Mathematical Heterogeneous Content in Cyrillic script

Olesea Caftanatov, Valentina Demidova, Tatiana Verlan

Abstract

This paper describes the steps through which the authors passed during the process of digitization of manually written mathematical texts with formulas and figures. Some difficulties met are also discussed. Our project highlighted the challenges associated with working with handwritten, non-homogeneous texts stored on outdated records, but it also demonstrated the effectiveness of combining modern technology with traditional manual methods.

Keywords: digitization, handwritten texts, Cyrillic script, mathematical texts, punch cards, tools for digitization.

MSC 2020: 68T50, 68U15.

1 Introduction

Nowadays, the process of text digitization in different domains is very actual and popular. We try to restore and keep rare documents and books existing only on paper, or to keep the important information with the purpose of being able to work with it in the future. The goals for digitization are different, for example, our colleagues developed the HeDy platform [1] for digitization historical Cyrillic texts from XVIII-XX century, or even digitization of musical scores [2]. Each case requires different steps in the process of digitization, as well as different tools for realization are selected or even elaborated.

It is worth mentioning that digitizing historical, particularly, mathematical texts is a complex process that requires careful attention to

doi:10.56415/csjm.v33.03

details and the use of specialized tools. We describe our way of solving this problem. Our task was to digitize mathematical texts containing such objects as formulas and figures. Some other features of our task are the following:

- the texts and formulas are written manually;
- geometric figures are hand drawn;
- the information carrier is old punch cards;
- the texts are written in Cyrillic script;
- the texts are not homogeneous and structured;
- as a result of the digitization process we had to obtain texts in LaTeX format.

On the other hand, these features represent the difficulties or challenges the authors had to face during their work. Because of manually written and painted information, in many places, the handwriting is illegible. Some words or designations have been crossed out, and a corrected version has been written in. Often, there are the following cases that complicate the recognition process: the insertions are written between the lines; the lines are crooked or written diagonally; the ink has faded. Not always (especially when it comes to solutions and hints) information or formulas are arranged sequentially but in the form of sketches or ideas. Surely, all this complicates the recognition process.

2 Tools selection and steps of the working process

The features described in Section 1 defined the tools that were selected by the authors for the task implementation.

2.1 Tools that were rejected

At first, we would like to mention some existing tools for digitizing that were not selected for our work and the reasons of this.

There are a lot of online services, programs, and applications for digitizing text, including those. with free access (in many cases –

limited in volume). A couple of the most popular ones are the following: NewOCR, OCR Convert, Яндекс OCR, ABBYY FineReader, CuneiForm, WinScan2PDF, Transkribus. They are good for different purposes but not for our one. We tried some of them.

For example:

– *NewOCR* (https://www.newocr.com/). It can recognize files in JPEG, PNG, GIF, BMP, TIFF, PDF, DjVu formats for free without any restrictions, and you can upload several pages at a time with no limits on the number of files you can upload. It can recognize texts from images in DOC, DOCX, RTF, and ODT files. It supports 58 languages and can translate text using Google Translate. You can save the obtained recognition results in TXT, DOC, ODT, RTF, PDF, and HTML formats.

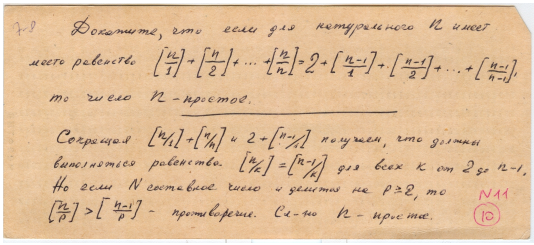
– *OCR Convert* (https://www.ocrconvert.com/). Free online text recognition service that does not require registration. Supports PDF, GIF, BMP, JPEG formats. This OCR service allows you to convert PDF to Text, JPEG to Text, and scanned images into editable text documents.

– *Transkribus* (https://www.transkribus.org/): "Transkribus is your AI-powered ally designed to simplify your time-consuming and laborious work with historical documents".

We tried to use these tools and obtained results that didn't satisfy us (see Figs. 1 and 2 with the recognition results of the same punch card 'Task_T_1' containing text in Russian and formulas). Also, we tried to run Transcribus with content containing only formulas. The result of recognition didn't satisfy us as well (see Fig. 3). Moreover, we tended to finally obtain the result as the text and formulas in LaTeX format. No one of these tools didn't provide us with such a possibility. That is why, we refused to use them in our work.

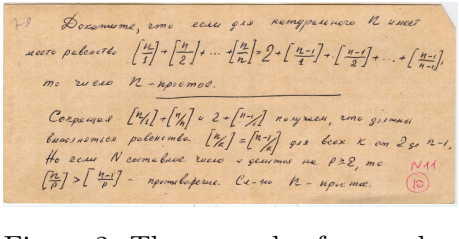
2.2 Brief description of the initial task

To better understand the initial task for the authors of this paper, it is necessary to describe briefly the source material for the work and the main goal. In the family scientific archive of mathematician Boris Cinic, who was a scientific researcher at the Institute of Mathematics



+. ' © Орви 7 , 2.227 г вече, $" -Р етим есзе #о, ишее/Гр, >. _ °27 с реверс 7 (= = = 'а, 4-2 ['\|, и-у1 2 я 4 [72 /"... + =| р й- —1/*, #772 277-2757) #2. — ро сто. . .НИо. Сори сле. 1%] +/47 и 2 -/* Мо .2 9 24т 2+7 с фа. мене. г _: ей #/)/ ./И- ||| |И И=*е с 4. 1% / =" 9+ беж Ш в?» 2 5 12. -/,© 2е.4е Фот еле. биение <' 22 ЗРталу .„- >=) [4=17 " ме РА, те №44>. _ о 2 и тх у[27 > Юр + прозы воофельие. С4 - 40 А. = Арье: Я «е, (5 и. . . . о. ОИ . _ . . —— (и

Figure 1. The example of a punch card 'Task_T_1' and the result of its recognition with NewOCR tool

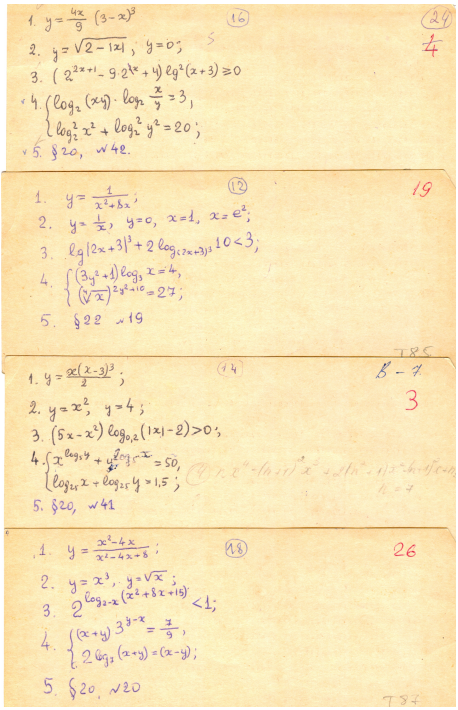


"„/ЛЮ ‘„ «2/9 : „. #74") 9/07 „ ага «_а 99 ,г у ,астму/,734, 14— „««/;и _-: сйхгазм"% 1/2- .^""5- ' ' ' . `‘ёьа/Ъ /)шС@И4/ _сг' "'—"'] ‹: "" "` >;"?! [[О]—;"- ii'7/ _: __;1',.и а ,11 ! 9 ' ' ‘в +. „ .`“ „ ";,#7712 ЫЁМ & 9/%3421›Ёт_/1ё:_ССЭ/' 9.5.3 ";/ ЦАР). [&'/,],г/У/ # 2 т/i/y/ );д 7 мы“! 24770 (у: ,“"/“&'!` ‘ " в '.47 , $6 5/ "; {д` ‘ 11 — - ,] .{&'/16,1111}, {($ („11/7' " ' ‹ 7 ‘ [% "`/›)- *а; _? (::(11 1/5: (7,77 %() 5: 71, "`/‘а !#0 "36 Щ Л' ," аш 771%." — (,ФО/‹ ЁА/см' « 5241: ;,ъщши А", /i' 5; [7 „" (; „.... , „ —„. _ „ „ ,_i__1,.; ;; , ,[? _;1/>[р * ir/6'д'д'щ (&&/&'441 -41 Сч". “ “О ,41 «› И/Л-,д таз; Ё ЁЁ...» ,

Figure 2. The example of a punch card 'Task_T_1' and the result of its recognition with OCR Convert tool

and Computer Science of the Academy of Sciences of Moldova since its foundation, there is a wealth of mathematical and geometrical problems he developed for lyceum students. When he prepared for the lessons, he wrote these problems on the punch cards [3]; in such a way he formulated the individual tasks for each student in the class. Thus, these problems have found their application in the real educational process, thereby confirming the usefulness of this educational material. Many punch cards also contain solutions or hints for these problems. They were written manually in the 90s of the last century and today are also of interest. So, it would be a good idea to keep this material in electronic form and make a manual for students with it.

Therefore, to begin with, these materials must be digitized.



(3-x) 16 1. 9= 9 y= v2-121, 9=0 9227 9.2 +4 lq (a+3)
20 a xy). log y 23 log2 Clog x + log ye =20, 5. 820,
W42. 19 x+bx Y= E, 4-0, x-1 2-e 2. lg12x +31 +
2 log. 2213 1031 3y &log, x-4 4. 225-10-24 22 19
14 B-7 1.4=22-3 2. y22, 7=4 3. (6x-xd l092
(121-2)0 logs +403 =50 4. logis 2 + l0g2s4 =1.5
5. 820 W42 2-42 18 9= 42-4X+3 26 1 y= 2, y=
VE. 2. log.- (x? +82 +15) L1;
-x - (x+9 5 4 269, (2+9) = (x-9) N20 520 27 4

Figure 3. The example of punch cards 'Tasks_T84_87' and the result of its recognition with Transcribus tool

2.3 Brief description of the working steps and selected tools

Below is a brief description of the steps we took in processing the source materials.

a) The existing materials were thoroughly analyzed from the context point of view, because, depending on the context and its complexity, the respective processing tool was selected for work:

   - punch cards containing only text information, which may contain short letter designations of variables. These could be, for example, problem statements and, in some cases, their solutions (see Fig. 4);

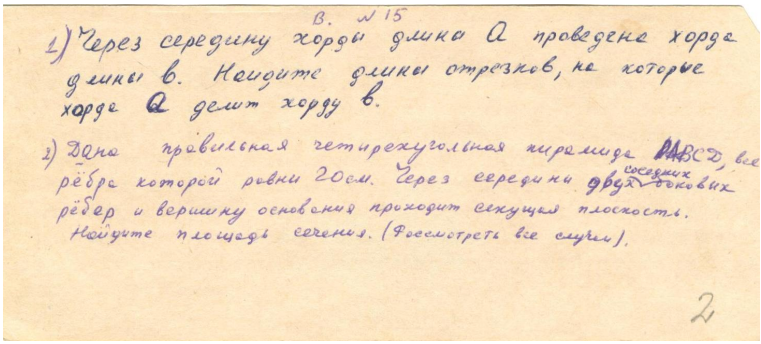


Figure 4. Punch card containing only manually written Cyrillic text

- punch cards containing only formulas (see Fig. 5);
- punch cards containing text and complex formulas (see Fig.6);
- punch cards containing figures (see Fig. 7).

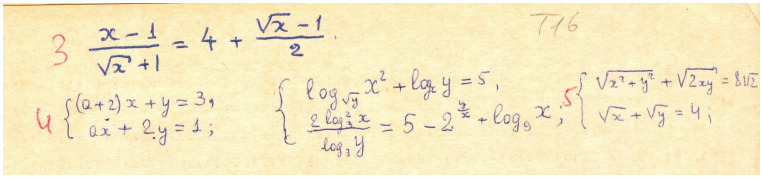


Figure 5. Punch card containing only manually written formulas

b) The existing tools for the digitization of handwritten texts were analyzed with the aim of preparing an electronic problem book with these materials. It was decided to work with the following tools: MathPix, ChatGPT, Google Docs, and LaTeX.

- The Google Docs tool [4] is simple and easy to use and copes well with recognizing handwritten text in Russian, which is extremely important for our task. However, it does not cope with recognizing formulas. Therefore, this tool was chosen to work with purely text punch cards. Its simplicity and

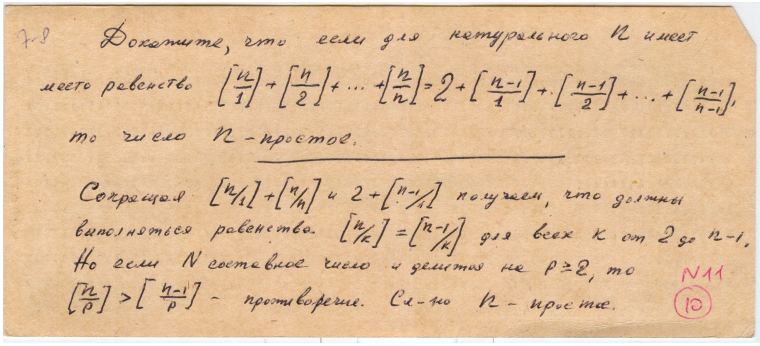


Figure 6. Punch card with manually written Cyrillic texts and formulas

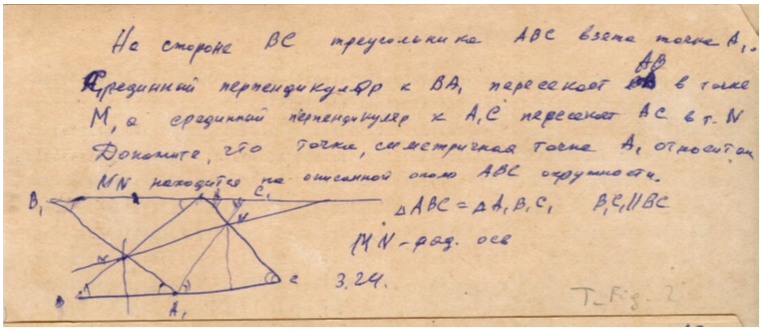


Figure 7. Punch card with manually written Cyrillic text, formulas, and manually painted geometric figure

accuracy in text recognition made it the ideal choice for these cards.

The recognition system runs automatically when the scanned image of the document is opened. The scanned punch card is loaded into Google Docs and opened with the help of Google Documents; after that, the conversion starts, and the result is opened in a Microsoft Word environment to continue the edition process. The quality of recognition is good.

- The MathPix tool [5] perfectly recognizes formulas and generates the corresponding LaTeX code, which greatly simpli-

fies and speeds up the process of further layout of the already recognized material. However, it "does not understand" the Russian language (see Fig. 8 with the original punch card and Fig. 9 with the respective recognized text).

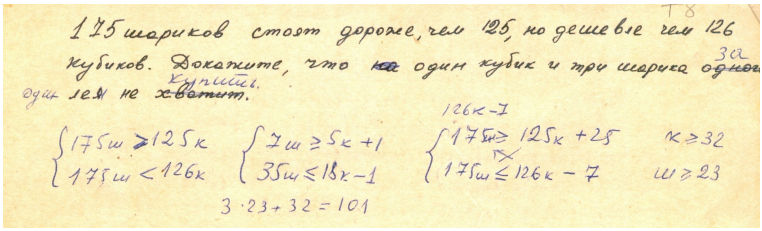


Figure 8. Original punch card T08 featuring a text problem in Russian and its corresponding formulas for the solution

175·шарикоb·стоят·дороже,·чем·125·,·но·дешевле·чем·126·кубиков.·Докажите,·что·на·один·кубик·и·три·шарика·одного·купишь,·один·лей·не·хватит.¶

$$\begin{cases} 175m \geqslant 125k \\ 175m < 126k \end{cases} \begin{cases} 7m \geqslant 5k+1 \\ 35m \leqslant 18x-1 \\ 3 \cdot 23 + 32 = 101 \end{cases} \begin{cases} 175k \geqslant 125k + 25 \\ 175 \lll 126k - 7 \end{cases} \quad k \geqslant 32, \quad u \geqslant 23 \P$$

Figure 9. Recognized formulas and "Russian" text from punch card T08

Another problem we had to face while working with MathPix was related to the especially bad handwriting. Some texts were written in more clumsy symbols (see Fig. 10). MathPix simply did not even "see" such punch cards, absolutely refusing to recognize them.

- The ChatGPT tool [6] does a great job of recognizing both formulas and text in Russian and also generates the corresponding LaTeX code. However, it is too creative in its approach to the task and often offers distorted text. Despite this, this creative approach sometimes helps to understand the idea of a written problem and, which is not less important, the idea of its solution. Thus, in some cases, we

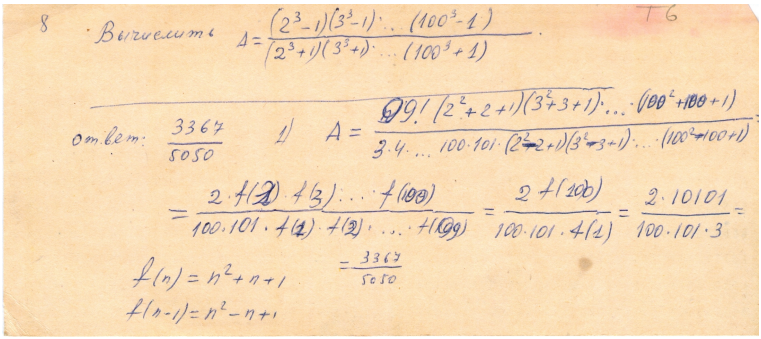


Figure 10. Punch card with bad handwriting

decided to retain these GPT's creative hints with the respective annotations.

However, it should be noted that the results of the work of all the applied tools require careful verification and correction of both texts and formulas to ensure the integrity and accuracy of the final digitized material (see Fig. 11).

Card T08

175 шариков стоят дороже, чем 125 кубиков, но дешевле, чем 126 кубиков. Докажите, что один кубик и три шарика за один лей не купить.

Подсказка:

$$\begin{cases} 175\ \text{ш} > 125\ \text{к} \\ 175\ \text{ш} < 126\ \text{к} \end{cases} \quad \begin{cases} 7\ \text{ш} \geq 5\ \text{к} + 1 \\ 35\ \text{ш} \leq 18\ \text{к} - 1 \end{cases} \quad \begin{cases} 175\ \text{ш} \geq 125\ \text{к} + 25 \\ 175\ \text{ш} \leq 126\ \text{к} - 7 \end{cases} \quad \begin{matrix} \text{к} \geq 32 \\ \text{ш} \geq 23 \end{matrix}$$

$3 \cdot 23 + 32 = 101$

Figure 11. Verified, corrected, and formatted punch card T08 in .pdf file. English translation of the Russian text is: *"175 marbles are more expensive than 125, but cheaper than 126 cubes. Prove that one cube and three marbles cannot be bought for one lei. The Clue:"*

- The LaTeX tool [7] was chosen as a system for the convenient finalizing of the layout of the obtained material including formulas and figures into the required format. It provided a

convenient and consistent way to organize and present the digitized content.

c) Thus, having the analyzed and classified source material and the chosen instruments for work, we proceeded in the following way:

- Scanning of handwritten texts into .jpg or .pdf files;
- Processing of images obtained by scanning;
- Recognizing handwritten texts in Russian using the Google Docs online service and obtaining the respective LaTeX code;
- Recognition of mathematical formulas with the help of the MathPix online service and obtaining the respective LaTeX code;
- Recognition of handwritten texts in Russian and mathematical formulas using the GPT online service and obtaining the respective LaTeX code;
- Recognition of geometrical drawings. Sometimes this stage included scanning and separate graphic processing of the images.
- Recognized resources have been manually checked, corrected, and validated in the LaTeX environment.

It should be noted that the results of the work of all the applied tools require careful verification and correction of both texts and formulas to ensure the integrity and accuracy of the final digitized material.

2.4 Work with geometrical drawings

The work with geometrical drawings is also a specific task and deserves special attention. As we mentioned earlier, the drawings on punchcards with geometrical problems were also made by hand. So, the lines are often crooked and the circles are irregular. Moreover, the drawings were usually made schematically, only as sketches. That is why, sometimes the lines of circles are unfinished or do not converge; the lines connecting some points are not straight, and the cross-points are not points but there is a distance between the crossing lines. Another reason of the conventionality of the drawings is that Prof. B.Cinic was a

rather aged person when he made them; so, in some cases, they were done by a faltering hand. But on the other side he always liked to say that "the correctly made drawing according to the condition of the problem is 50% of its solution". So, he tried to be maximally precise in his "conditionalities".

While working with drawings on the punchcards, we wanted to adjust some lines and denotations, make them more clear and straight, remove the dirt, and, at the same time, try to keep the conventional nature of the drawings and the character and unique handwriting of their author (see Figs. 12 – 14).

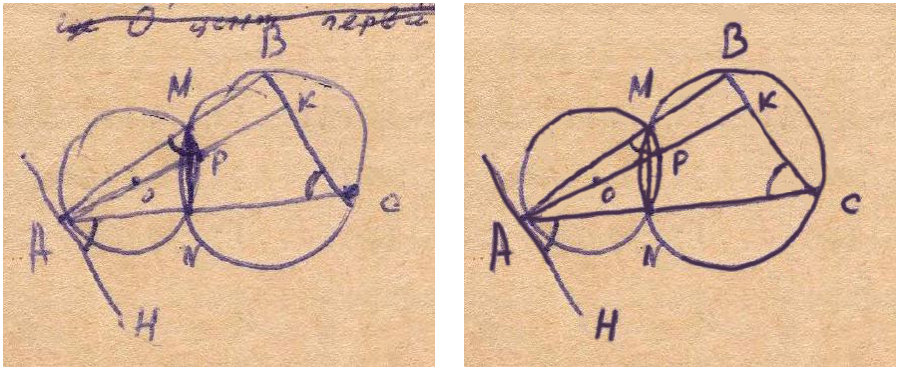


Figure 12. Drawing from punchcard T_Fig_15 before and after adjustments

All corrections and adjustments were made with the help of Paint, FastStone Image Viewer, and Adobe Photoshop. Correction/adjusting/drawing of the lines and designations was made also by hand.

3 Conclusion and Future Work

At the current moment, more than 300 handwritten punch cards were processed. This is only a part of the whole archive. The information from 240 processed punch cards has been thoroughly checked, and the LaTeX code generated by MathPix and GPT online services has been manually verified and corrected.

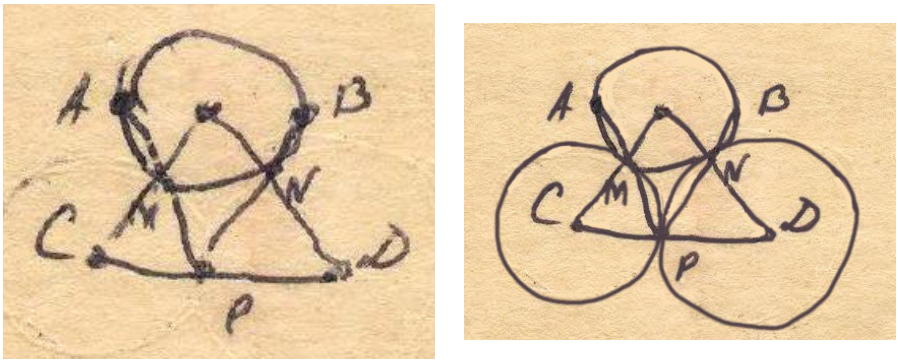


Figure 13. Drawing from punchcard T_Fig_22 before and after adjustments

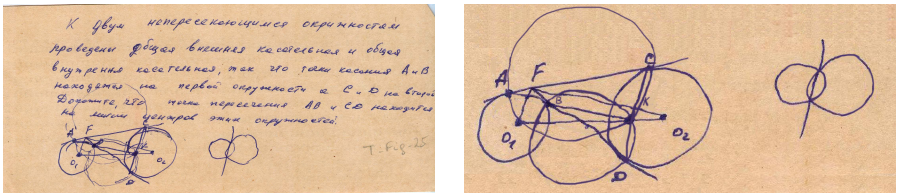


Figure 14. Original punchcard Card_T_Fig_25 and the drawing from it after adjustments

These statistics allow us to assert that at this stage we have chosen successful tools to solve the task at hand and have developed a successful mechanism for continuing the work. Through this careful selection and application of tools, we were able to efficiently digitize the handwritten texts and formulas, preparing them for inclusion in an electronic problem book.

In summary, the digitization of handwritten mathematical materials containing texts in Cyrillic script from punched cards was successfully achieved using a carefully structured process involving multiple tools and techniques. By scanning the original punched cards into digital formats and processing these images, we were able to use various services to recognize and convert handwritten Russian texts, mathematical formulas, and geometric drawings into LaTeX code. Using Google Docs contributed to the accurate recognition of handwritten

Russian text, while MathPix proved effective in translating mathematical formulas into LaTeX. ChatGPT provided additional support for recognizing both Russian text and formulas. Specialized tools were used to digitize geometric figures, ensuring their accuracy.

Manual verification and correction within the LaTeX environment played a decisive role in the finalization of the digitized materials.

In addition, we will work closely with our colleagues to integrate digitized texts, mathematical formulas, and geometric images into the AI Tutoring system [8]. This integration will involve matching the digitized content with the system's algorithms to ensure that it can effectively help students solve problems.

This paper is the extended and revised version of the conference paper [9] presented at IMCS-60.

Acknowledgments. Project SIBIA – 011301, "Information systems based on Artificial Intelligence" has supported part of the research for this paper.

References

[1] T. Bumbu, L. Burţeva, S. Cojocaru, A. Colesnicov, and L. Malahov, "Distinctive features of recognition for documents printed in the Romanian transitional alphabets," *Computer Science Journal of Moldova,* vol. 31, no. 3(93), pp. 340–350, 2023. DOI: `https://doi.org/10.56415/csjm.v31.17`, `http://www.math.md/en/publications/csjm/issues/v31-n3/13848/`.

[2] A. Colesnicov, S. Cojocaru, M. Luca, and L. Malahov, "On Digitization of Documents with Script Presentable Content," in *Proceedings of the Fifth Conference of Mathematical Society of Moldova IMCS-55, September 28 - October 1, 2019*, (Chisinau, Republic of Moldova), 2019, pp. 321–324. `https://ibn.idsi.md/sites/default/files/imag_file/321-324_7.pdf`.

[3] S. Lubar, "'Do not Fold, Spindle or Mutilate': A Cultural History of the Punch Card," *Journal of American Culture*, vol. 15, no. 4, pp. 43–55, 1992, DOI: `https://doi.org/10.1111/j.1542-734X.1992.1504_43.x`.

[4] Google Workspace, "Online, collaborative documents," Google Docs, https://www.google.com/intl/en-GB/docs/about/.

[5] "MathPix, AI-powered document automation," https://mathpix.com/.

[6] "ChatGPT 4o mini," https://chatgpt.com/.

[7] "LaTeX – A document preparation system," https://www.latex-project.org/.

[8] O. Caftanatov and A. Parahonco, "The Virtual GPT Assistant: Emulating the Teaching Style of a Real Professor," in *Proceedings of International Conference dedicated to the 60th anniversary of the foundation of Vladimir Andrunachievici Institute of Mathematics and Computer Science, MSU, Octomber 10-13, 2024*, (Chisinau, Republic of Moldova), 2024, pp. 247–258, https://www.math.md/imcs60/combined_IMCS_v6_with_preface.pdf.

[9] O. Caftanatov, V. Demidova, and T. Verlan, "Our Approach to Digitizing Handwritten Mathematical Text in Cyrillic Containing Formulas and Drawings," in *International Conference dedicated to the 60th anniversary of the foundation of Vladimir Andrunachievici Institute of Mathematics and Computer Science, MSU, October 10 – 13, 2024*, (Chisinau, Republic of Moldova), 2024, pp. 237–246, https://www.math.md/imcs60/combined_IMCS_v6_with_preface.pdf.

Olesea Caftanatov[1],
Valentina Demidova[2], Tatiana Verlan[3]

Received December 30, 2024
Revised February 12, 2025
Accepted February 14, 2025

[1,2,3] Vladimir Andrunachievici Institute of Mathematics and Computer Science, Moldova State University

[1] E–mail: olesea.caftanatov@math.md
ORCID: https://orcid.org/0000-0003-1482-9701
[2] E–mail: valentina.demidova@math.md
ORCID: https://orcid.org/0009-0006-7260-8375
[3] E–mail: tatiana.verlan@math.md
ORCID: https://orcid.org/0009-0006-4519-1105