Napis/ Inscription. An anthology 3/ 2025
doi: 10.18318/napis.2025.en.3.7

Scholarly editing in the face of technological progress illustrated with the example of Wacław Potocki's *Moralia*[*]

Joanna Hałaczkiewicz
orcid: 0000-0001-7143-789x
(Jagiellonian University in Kraków)

Bumper crop. Editing in the face of technological progress

When, at the end of the last century, the Deep Blue computer played a series of winning chess games against grandmaster Garry Kasparov, the world watched with carefree curiosity. Here was a machine created by humans, fed by the experience of many outstanding chess players who provided it with ready-made moves, and it had begun to surpass human intelligence. This event from the world of elite sports and equally inaccessible Silicon Valley laboratories – a world far removed from the perspective of the lesser mortals – was the harbinger of radical technological changes that have accumulated in recent years and, in the opinion of many, have begun to threaten the established order.[1]

ChatGPT was launched as a free online tool at the end of November 2022. Netizens could not ignore this debut. Finally, there was a bot that could be talked to about any topic at any time and place, and could also be used for various unexciting tasks – in other words, a virtual companion, personal assistant and patient advisor. The brainchild of the OpenAI engineers was a huge success and surpassed the previous leaders such as TikTok or Instagram in terms of how quickly it gained new users.

[*] First printed as "Zastosowanie sztucznej inteligencji w edytorstwie naukowym – przykład *Moraliów* Wacława Potockiego" in: *Napis* issue 30 (2024), pp. 215–232.

[1] According to data provided by GWI for the Digital 2024 Global Overview Report, 71% of respondents believe that the development of artificial intelligence is too fast and therefore a source of concern. DataReportal, Digital 2024 Global Overview Report, https://datareportal.com/reports/digital-2024-global-overview-report, p. 31 (all Internet sources cited in the text are as of 20 September 2024).

The impact of the new chatbot was unprecedented,[2] but the initial enthusiasm quickly gave way to deep concern: what if a versatile and super-efficient artificial intelligence replaces us all? Not only chess players, but also doctors, lawyers, journalists, artists, programmers, office workers, economists, translators, editors and teachers? As long as expensive AI projects remained the domain of high-tech, the fear of losing one's job or a reduction in salary did not seem to exist, or at least was not widespread. When AI tools suddenly became widely available, many people began to fear for their future. Judging by the topic of the nineteenth conference of The European Society for Textual Scholarship in Budapest (*Textual Scholarship, Artificial Intelligence, Corpora and Intelligent Editions*),[3] a similar curiosity tinged with uncertainty also accompanies editors. This is true even though it would seem that this particular profession has a long tradition of awareness of and participation in technological progress, dating back to the 1940s, when the Jesuit Roberto Busa authored the digital corpus of the works of St. Thomas Aquinas.[4]

What does the future hold for scholarly editing? How will we work on texts of works and will we even do it the same way as we used to? Or maybe artificial intelligence will first search through all available online library catalogues, access the texts stored in digital libraries, automatically read their text layer, collate the collected material, select the basis for the edition, and then – having at its disposal the knowledge of the entire Internet, including metadata in the form of articles on the practice of scholarly editorship – introduce conjectures, comment on questionable passages, explain difficult words, attempt an erudite foreword with an exhaustive characterization of the transmission of the text? Will it generate its own edition by using already published digital scientific editions – especially their open code? As long as we are making fun of the absurd clumsiness of the supposedly powerful chatbot, which cannot count the number of letters 'r' in the word *strawberry*,[5] this vision seems distant; on the other hand, artificial intelligence engineers are already looking for solutions to create new models capable of conducting multi-stage re-

[2] The Google Trends graph, which records the interest of Internet users in various topics, clearly shows how the term 'artificial intelligence' saw a surge in popularity at the end of 2022. According to another report, in 2023 'ChatGPT' was the most read Wikipedia entry (ibidem, p. 108). It is worth mentioning that OpenAI is not the only artificial intelligence provider – China is developing its own project called Wu Dao, smaller start-ups such as Cohere are trying to find their place, and existing IT giants such as Microsoft (Copilot), Google (Gemini) and Facebook (Meta AI) are also working on the technology. In addition to them, non-profit organizations such as EleutherAI and, most recently, Mozilla AI are democratizing access to new technology.

[3] *Program of the ESTS Conference 2024*, https://elte-dh.hu/ests-2024-program/.

[4] Corpus Thomisticum, https://www.corpusthomisticum.org/.

[5] A. Silberling, "Why AI Can't Spell 'Strawberry'", https://techcrunch.com/2024/08/27/why-ai-cant-spell-strawberry/.

search.[6] The fact that the 'artificial' is becoming more and more 'human' and realistic is demonstrated by the following: most Internet users doubt their cognitive abilities when asked on social media to distinguish a bot from a human being or a deepfake generated by Sora[7] from a real video recording.

The ESTS conference program shows that the editing community is currently exchanging ideas on how to use AI to create scholarly editions. No longer digital scholarly editions, but intelligent editions, AI editions or at least AI-driven editions are becoming the strongest stimulus for the philological imagination.[8] This approach does not seem to deviate from the general trend in other professions 'threatened' by competition with intelligent computers: instead of sticking to a losing position, it is better to start using AI for one's own purposes, to become its operator. Interest in artificial intelligence among humanists is growing, especially since it is supported by a strong current of quantitative literature research known as distant reading, which treats literary texts as large data sets to be processed (big data) – after all, generative artificial intelligence works in a similar way: it draws its 'wisdom' from enormous data resources. Digital source editing (born-digital) is also becoming increasingly popular, which should come as no surprise given that humanity has been creating various electronic documents on a large scale for over half a century and that the service known as the World Wide Web has been in existence for over thirty years. Countless digital works are stored on the web and on computer hard drives. Among these are works that are important for literary researchers and that require proper handling. However, it may be impossible to access these works without the help of an artificial intelligence assistant.

This situation has led to a veritable flood of information at universities and research institutes. Tools and platforms are being developed independently in many places. They are often imperfect, 'under construction', 'currently being transferred to cloud servers', 'only partially accessible'. These solutions are impossible to keep up with and generally not easy to implement in one's own editorial projects because they are not versatile or user-friendly enough. The Social Sciences and Humanities

[6] A. Tong, K. Paul, "Exclusive: OpenAI Working on New Reasoning Technology under Code Name 'Strawberry'", https://www.reuters.com/technology/artificial-intelligence/openai-working-new-reasoning-technology-under-code-name-strawberry-2024-07-12.

[7] Sora is another OpenAI engineering project, capable of generating realistic films that are very difficult to distinguish from traditionally recorded material.

[8] During the DSE Communities conference at the Institute of Biology of the Polish Academy of Sciences in Warsaw (25-27 September 2024), Michael Pidd presented the C21 Editions project, which involves the extensive use of artificial intelligence to create a digital scholarly edition of *The Canterbury Tales*. Cf: C21 Editions. Scholarly Editing and Publishing in the Digital Age, https://www.c21editions.org/.

Open Marketplace catalogue[9] gives an idea of how numerous and diverse these resources are.

With so many tools at their disposal, editors are trying to reorganize their workshop and build work standards in a digital environment. This would involve carefully selecting existing services, applications or even useful scripts – for example, for automatic transcription, collation or annotation – which can then be accessed as needed. Such attempts result in case studies in which the authors describe the use of specific tools for creating editions.

In view of this dynamic development, it is also important to consider the long-term viability of digital editing projects (future-proof editing), as there is a real risk of overloading the editing process with expensive, custom-made original solutions[10] that no longer display correctly in browser windows after a few years. Some believe that the answer to these challenges is so-called minimal processing, which involves using the most standard programming and coding languages that are as independent as possible from changing internet standards;[11] research data repositories are also being created where the source code of the edition can be deposited.

The absence of a single convenient method of digital editing, which discourages many researchers, is sometimes compared to the early days of printing – at that time there was also no catalogue of good practices for creating incunabula, or even rules for writing vernacular languages, but despite this, Gutenberg's invention gradually gained acceptance, changing the social relations and intellectual culture of subsequent generations of readers.[12] As Peter L. Shillingsburg wrote:

> » It is easy to get lost or discouraged in the field of electronic texts. Every new whoop-tee-doo in these areas soon becomes last week's news in the face of even newer ones. We are tempted to wait out the turmoil, perhaps hoping to come in at the home stretch with the winners, like one who cheats in marathon races by joining for the last

[9] SSH Open Marketplace, https://marketplace.sshopencloud.eu/.

[10] Cf: E. Pierazzo, "What Future for Digital Scholarly Editions? From Haute Couture to Prêt-à-Porter", in: *International Journal of Digital Humanities* no. 13 (2019), pp. 209–220.

[11] R. Viglianti, G. del Rio Riande, "Against Infrastructure. Global Approaches to Digital Scholarly Editing", in: *C21 Digital Editions* (2023); idem, N. Hernández, R. De Léon, "Otwarte, minimalne i promujące równe szanse. Jak uczy się tworzenia naukowych edycji cyfrowych na Globalnej Północy i Globalnym Południu" ["Open, equitable, and minimal: teaching digital scholarly editing North and South"], transl. by J. Wełniak, in: *Sztuka Edycji* [*The Art of Editing*] no. 1 (2013), pp. 71–83. [Quotation in English translated from the Polish edition].

[12] H. Hollender, "Czy świat czeka przyszłość średniowiecza?" [Is the world facing a medieval future?], in: E.L. Eisenstein, *Rewolucja Gutenberga*, transl. by H. Hollender (Warsaw: 2004), passim. First published as: E.L. Eisenstein, *The printing revolution in early modern Europe* (New York: 1983).

> mile or two. The finish line, however, seems, like the horizon, to recede.[13]

It seems that the emergence of a new factor in the form of artificial intelligence significantly changes the existing rules of the game, and that is precisely why the temptation to wait it out is something that should not be given in to too much.

AI in practice, or how to teach a computer to read baroque manuscripts

The team of the Digital Editing Laboratory (LabEdyt),[14] founded at the beginning of 2023 at the Jagiellonian University, has set itself several tasks: experimenting with available tools to organize workflow systems, exchanging experiences as scientific digital editions are developed, constantly monitoring technological innovations, teaching students, and supporting scientists in the implementation of digital projects. LabEdyt is currently working on several pilot projects,[15] one of which – devoted to *Moralia* of Wacław Potocki – explores the possibilities of using machine learning to automatically create transliterations, transcriptions and XML semantic tagging.

The 17th century in the Polish-Lithuanian Commonwealth was, according to researchers, the 'age of manuscripts'. As Radosław Grześkowiak wrote: 'the most interesting works of the era were entrusted to manuscripts and reproduced in an informal circulation'. This applied to the 'works not only of such luminaries as Jan Andrzej Morsztyn, Wacław Potocki or Stanisław Herakliusz Lubomirski, but also of second-tier figures important for the history of our literature, such as Daniel Naborowski or Hieronim and Zbigniew Morsztyn'.[16] Wacław Potocki, a nobleman from Łużna who was extremely prolific in his literary output, and in this respect has been compared to Józef Ignacy Kraszewski, left virtually all his works in manuscript form – including his most famous epic poem *Transakcja wojny chocimskiej*

[13] P.L. Shillingsburg, *From Gutenberg to Google. Electronic Representations of Literary Texts* (Cambridge: 2006), p. 11.

[14] The team coordinated by Magdalena Komorowska operates within the framework of the Digital Humanities Lab flagship project, see: https://dhlab.id.uj.edu.pl/labedyt; https://labedyt.dhlab.uj.edu.pl.

[15] These are: revitalization of the Library of Old Polish and New Latin Literature "Neolatina" (https://neolatina.dhlab.uj.edu.pl), a digital scholarly edition of Wacław Potocki's *Moralia*, and an edition of Szymon Laks's correspondence with Krystyna and Czesław Bednarczyk, owners of Oficyna Poetów i Malarzy [Poets and Painters Publishers] in London.

[16] R. Grześkowiak, "Stary druk jako podstawa edycji krytycznej. Preliminaria" [Early printed books as the basis for critical editions: preliminary notes], in: *Jak wydawać teksty dawne* [How to edit old texts], eds. K. Borowiec et al. (Poznan: 2017), p. 11.

[The transaction of the Chocim war]. Only one major work was published at the end of his life: *Poczet herbów szlachty Korony Polskiej i Wielkiego Księstwa Litewskiego* [Coats of arms of the nobility of the Polish Crown and the Grand Duchy of Lithuania]. It was published in 1696 in the Cracow printing house of Mikołaj Aleksander Schedel. Other texts, including *Moralia abo rzeczy do obyczajów nauk i przestróg w każdym stanie żywota ludzkiego z łacińskich i z polskich przypowieści ojczystym krótko napisane wierszem* [Moralia, or things pertaining unto manners, lessons and admonitions for each estate of man's life, briefly set down in native verse from Latin and Polish proverbs], remained unpublished until the following centuries. The abundant work of the Old Polish writer did not attract interest until the turn of the 20th century, and Aleksander Brückner made the greatest contribution to its popularization at that time.[17]

From around 1688 until his death, Potocki worked on his most extensive work – *Moralia* – using a 1551 edition of Erasmus of Rotterdam's *Adages* printed by Froben in Basel,[18] which is a collection of short poems composed around ancient maxims selected and commented on by Erasmus. The fair copy of *Moralia*, made by Potocki, stored in the National Library,[19] has grown to an impressive size of 712 sheets, or 1424 pages, while a copy of *Adages*, which the writer used, as evidenced by his handwritten notes in the margins, was found among the duplicates of the Jagiellonian Library and subsequently transferred to the Warsaw Scientific Society (today the Library of the Institute of Literary Research of the Polish Academy of Sciences).[20]

The work has only been published once in its entirety: it was edited by Tadeusz Grabowski and Jan Łoś and published in three volumes between 1915 and 1918 in the series Biblioteka Pisarzów Polskich [Library of Polish Writers]. This edition, which is now over a century old and therefore considered a historical document by contemporary readers, has been criticized from the outset for being inaccurate and too sparse in its explanations. Aleksander Brückner pointed out many

[17] Cf: "Potocki Wacław (1621–1696)", in: *Bibliografia literatury polskiej "Nowy Korbut"* [The "New Korbut" bibliography of Polish literature], ed. K. Budzyk, vol. 3: *Piśmiennictwo staropolskie. Hasła osobowe N–Ż* [Old Polish literature. Personal entries N–Ż], compiled by R. Pollak et al. (Warsaw: 1965), pp. 119–130.

[18] L. Kukulski, *Prolegomena filologiczne do twórczości Wacława Potockiego* [Philological Prelegomena to the Works of Wacław Potocki] (Wrocław: 1962), p. 7.

[19] W. Potocki, *Moralia*, manuscript, ca. 1688–1696, The National Library, manuscript 3049 III, Polona.pl.

[20] E. Roterodamus, *Adagiorvm Chiliades Des. Erasmi Roterodami Qvatvor Cvm Dimidia Ex Postrema Avtoris Recognitione: In hac aeditione, prioribus tribus Indicibus subiunctus est quartus nouus* […], https://rcin.org.pl/dlibra/publication/74578/edition/66983/content. Cf: L. Kukulski, *Prolegomena filologiczne…*, p. 7, footnote 10.

shortcomings,[21] and Leszek Kukulski, an expert on the work of the Sarmatian poet, added to the list of errors.[22] Brückner, as if he had the gift of clairvoyance, pessimistically predicted that Potocki's work would not be published in a revised edition anytime soon: 'so the excess of frugality has been achieved at the expense of comprehensibility, which is very regrettable, because *Moralia* will probably not see a better, more careful edition'[23] – he wrote, and he was not wrong. Fragments of this collection appeared later only in selections, including the third volume of the extensive edition of Potocki's writings.[24]

Re-editing *Moralia* is a thankless task for many reasons. First of all, it is not an 'unweeded garden', to paraphrase the title of another collection by Potocki. Certainly, there is a greater temptation to break new ground and deal with unpublished texts. *Moralia*, having already had 'some' edition, lose the competition with works still awaiting publication. Besides, in order to edit them, one would have to refer to the manuscript, which editors, as diagnosed by Radosław Grześkowiak, are clearly not fond of.[25] If we add to this the extraordinary size of the text (the question immediately arises: how to convince grant committees to finance the printing?), the necessity to examine the connections with *Adages* and the fact that most of the editorial work was done by Leszek Kukulski before his death (he was also the one who made the most interesting discoveries), it is easy to come to the pragmatic conclusion that in times of scholarly haste it would be difficult to devote so much time to studying one work.[26]

However, digital editing tools may hold some hope for *Moralia*. They make it possible to create editions faster and more accurately, at a lower cost. Such an e-edition further expands the possibilities of printing, as it can include different

[21] A. Brückner, "Wacława Potockiego *Moralia* (1688), wyd. Tadeusz Grabowski, Jan Łoś, [Cracow] 1915–1918" [Wacław Potocki's *Moralia* (1688), eds. Tadeusz Grabowski, Jan Łoś, [Cracow] 1915–1918] [review] in: *Pamiętnik Literacki* [*Literary Memoir*] vol. 17/18 (1920), nos. 1–4, pp. 159–164.

[22] See, e.g.: L. Kukulski, *Prolegomena filologiczne…*, p. 14, footnote 28.

[23] A. Brückner, "Wacława Potockiego *Moralia…*", p. 161.

[24] W. Potocki, *Dzieła*, vol. 3: *Moralia i inne utwory z lat 1688–1696* [Works, vol. 3: Moralia and other works from 1688–1696], compiled by L. Kukulski (Warsaw: 1987). For a more detailed review of the latest editing accomplishments in the area of Potocki's works, see: J. Gruchała, "Wacław Potocki – problem edytorski" [Wacław Potocki – an editorial problem], in: *Wacław Potocki. W 400-lecie urodzin poety* [Wacław Potocki: on the 400th anniversary of the poet's birth], eds. K. Koehler, D. Chemperek (Warsaw: 2023), pp. 15–33.

[25] R. Grześkowiak, *Stary druk…*, p. 12.

[26] S. Grzeszczuk put it plainly: 'Potocki has a hopeless advantage over an individual researcher, no matter how hardworking and devoted to the cause. He can only be tackled by a team […]' – "O potrzebie i programie badań nad twórczością Wacława Potockiego" [On the need and program for research on the works of Wacław Potocki], in: *Wśród zagadnień polskiej literatury barokowej*, cz. 2: *Motywy – inspiracje – recepcja* [Problems of Polish baroque literature, Part 2: Motifs – inspirations – reception], ed. Z.J. Nowak (Katowice: 1980), p. 16.

versions of the text (transliteration, modernizing transcription, text with editorial commentary, facsimile of the manuscript, perhaps also images of the Froben's edition of the *Adages*), create an interactive 'key to *Moralia*',[27] provide thematic indexes, an advanced search module or even a frequency list – all without having to worry about printing sheets. Tagging the text and integrating it with existing databases (such as WikiData) also links it to the Linked Open Data network, which is a collection of open, linked data on the Internet. Creating these links enables further research, especially with the help of AI tools. Chatbots that use large language models, such as the Polish Bielik AI, are also able to read the edition and support its users. Since tagged text fragments have a semantic surplus in the form of metadata, they can be processed by a computer in a more advanced way than in the case of plain text. The editors of language corpora such as KorBa[28] could benefit from such a version of *Moralia*, for example. Even though the corpus does contain excerpts from Potocki's works, they are presented in a modernized form, as Janusz Gruchała stated: 'popular scientific rather than anything else'.[29] However, before we can start thinking about the benefits of digital editions, we first need to source the text.

Transliteration, also known as 'diplomatic transcription', is the basis for any editing of texts 'born' before the digital age – and will probably also be used in the future for many works that were written on a computer but not saved in digital form. Many have written about its indisputable significance. Peter Shillingsburg described transliteration as a form of 'reincarnation' – the subsequent embodiment of a certain intangible idea, i.e. a text. 'Reincarnation' is therefore the adoption of a new 'body' by the text (however puzzling it may sound in the context of digital space).[30] In editing, according to the Platonic concept, this embodiment of the perfect idea becomes its corruption at the same time, because matter always remains imperfect. In the process of 'reincarnation', mistakes are inevitable – and everyone makes them: copyists, editors, proofreaders, typesetters and their modern counterparts – DTP graphic designers, printers, bookbinders and even the author, making unintentional slips of the pen.

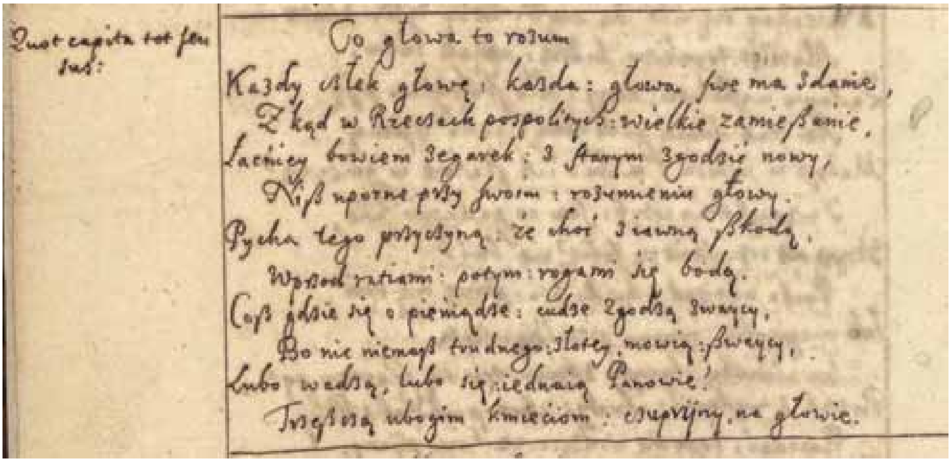
In the case of *Moralia*, it is not difficult to make an error when rewriting. The enormity of the rather monotonous material is conducive to mistakes. This is all the more likely, the more people are involved in copying a single work – there exists a risk that not everyone will be able to adhere to the accepted arrangements

[27] Cf: L. Kukulski, "Klucz do *Moraliów*" [The Key to *Moralia*], in: idem, *Prolegomena filologiczne…*, pp. 64–68.

[28] "KorBa" electronic corpus of Polish texts from the 17th and 18th century, https://korba.edu.pl.

[29] J.S. Gruchała, *Wacław Potocki…*, p. 19.

[30] P.L. Shillingsburg, *From Gutenberg…*, p. 27.

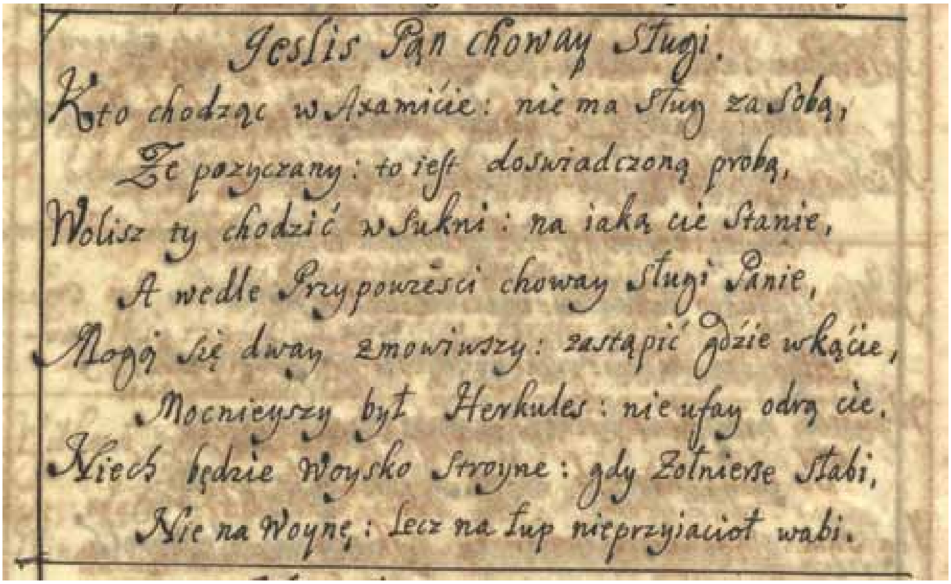


Quot capita tot ſenſus:

Co głowa to rozum

Każdy człek głowę: każda: głowa ſwe ma zdanie,
Z kąd w Rzeczach pospolitych: wielkie zamieſzanie.
Łacniey bowiem zegarek: z ſtarym zgodzić nowy,
Niſz uporne przy ſwoim: rozumieniu głowy.
Pycha tego przyczyną: że choć z iawną ſzkodą.
Wprzod ratiami: potym: rogami ſię bodą.
Cyſz idzie ſię o pieniądze: cudze zgodzą zwaycy.
Bo nie niemaſz trudnego: złotey, mowią: ſwięcy.
Lubo wadzą, lubo ſię: iednaią Panowie!
Trzęſą ubogim kmieciom: czupryny, na głowie.

Fig. 1. Sample of handwriting from the first pages of the *Moralia* manuscript. Source: Polona (scan 4v–5r).



Ieſliś Pan choway ſługi.

Kto chodząc w Axamicie: nie ma ſługi za ſobą,
Że pożyczany: to ieſt doświadczoną probą,
Wolisz ty chodzić w ſukni: na iaką cię ſtanie,
A wedle Przypowieści choway ſługi Panie,
Mogą ſię dway zmowiwszy: zastąpić gdzie w kącie,
Mocnieyszy był Herkules: nie ufay odrą cię.
Niech będzie Woysko ſtroyne: gdy Zołnierze ſłabi,
Nie na Woynę: lecz na łup nieprzyiacioł wabi.

Fig. 2. Sample of handwriting from the last pages of the *Moralia* manuscript – fragments written by a 'boy copyist' come from the *Ogród fraszek*[31] [Garden of epigrams]. Source: Polona (scan 641v–642r).

despite their best intentions. In addition, *Moralia* is a text that is convoluted from the perspective of today's reader, full of archaisms, and therefore incomprehensible in places – a great deal of this is due to baroque poetics. This was already Aleksander Brückner's opinion about the work a hundred years ago,[32] so what can a 21st-centu-

[31] Cf: J. Łoś, "Wstęp" [Preface], in: *Wacława Potockiego "Moralia" 1688*, vol. 3, ed T. Grabowski, J. Łoś (Cracow: 1918), p. XXVI.

[32] A. Brückner, "Wacława Potockiego *Moralia*…", pp. 159–161.

ry Polish speaker say? Failure to understand the text can lead to erroneous, hasty readings, and the form of the message – a manuscript – increases the difficulty of the task, although it must be emphasized that Potocki's handwriting is legible.

With these reservations in mind, the Digital Editing Laboratory team decided to perform the transliteration using an automatic handwriting recognition tool – the Transkribus application. This program (although in its current state of development it should probably be called a SaaS – Software as a Service – tool, as the so-called desktop client has already been discontinued) uses user-prepared samples of images and their associated transcriptions to train specialized text reading models. The 'training' itself consists of applying a basic artificial intelligence function, namely machine learning, to the provided training set. The trained model can be saved in a private library assigned to the account or shared with the Transkribus community.

The application offers a number of ready-made models that recognize both handwritten and printed texts, but none of them were suitable for transliterating *Moralia*. When selecting a model, several parameters must be taken into account. Of course, the most important is the close visual similarity between the text we want to automatically transliterate and the text used to train the model. In practice, this means that a model for English cursive (Copperplate) will not work on a sample written in uncial – similarly, if the author's handwriting differs even slightly from that of the model, the results of automatic transliteration will not be satisfactory. Apart from this obvious issue, it is also worth remembering the CER (Character Error Rate), which determines how often the computer makes mistakes when reading characters; the higher the CER (above 5%), the less accurate the reading. A CER of 15% means about fifteen incorrectly recognized characters per hundred, or almost fifteen corrections per standard line of 12-point Times New Roman text in Microsoft Word – this is a lot, so correcting such a transcription can take more time than creating it from scratch.

When considering a ready-made solution, one should also take into account the language of the text on which the AI was trained. A model trained on English-language documents will certainly not recognize Polish diacritical marks and will also make more mistakes because it has not learned the character combinations that are typical for Polish but absent in English. What is more, observations of transcriptions generated in Transkribus show that the model learns not only characters, but also the shape of entire words, which is why it is able to correct scanned text to a certain extent (!).[33] Few public models in Transkribus are suitable for use

[33] This is not always desirable. For example, the Polish Schwabacher model, designed to generate transliterations of Polish Schwabacher, consistently adds a stroke to the letter 'a' in the word 'náprzod' ('naprzód'), even when the scan clearly shows the letter 'a' without a stroke. An editor who wants to preserve this unstable spelling in the transliteration must also remain vigilant.

by Polish editors. In the gallery of ready-made solutions containing over two hundred models – from Church Slavonic to Tibetan cursive – only three are intended for Polish.[34] Two of these are large models that are constantly being developed and fed with large data sets – one for print, the other for manuscripts (multilingual Transkribus Print M1 – CER 2.2%, Transkribus Polish M2 – CER 4.1%). The third model, The Polish Schwabacher (CER 0.87%), was developed in early 2024 by editing students at the Jagiellonian University as part of a course on creating digital scholarly editions and is used to automatically transliterate Schwabacher.

To achieve the best possible results, a new model specifically designed to read Potocki's manuscripts had to be trained. This involved preparing a sample of real data (ground truth) that the algorithm would treat as an ideal model to follow – in other words, even if we wanted to automate the work, we first had to do a considerable amount of it ourselves. The transliteration of the one first hundred pages of the manuscript was done by Lidia Nowak, a graduate of editing and currently a doctoral student at the Jagiellonian University Doctoral School in the Humanities. The manuscript had to be read as accurately and unambiguously as possible, because in the ground truth sample, each character in the manuscript should have a fixed equivalent in the transliteration – otherwise, the AI training would not be as effective. For a computer, every character is equally abstract and meaningless: if we consistently show it that the capital letter A on the scan is actually a lowercase g, it will begin to recognize it as such.

In accordance with previously adopted rules, the transliteration of *Moralia*, among other things:

– retained the layout and breaks in pages, lines, and marginalia,

– retained punctuation marks appearing in the original (without retaining the inconsistent spacing preceding these marks),

– decomposed *æ, œ*, & ligatures into *ae*, *oe* and *et*,

– standardized the three variants of the grapheme z, ʒ, ȝ to z (analogously to the sign ż – if the variants had a dot above),

– rendered long s as *ſ*,

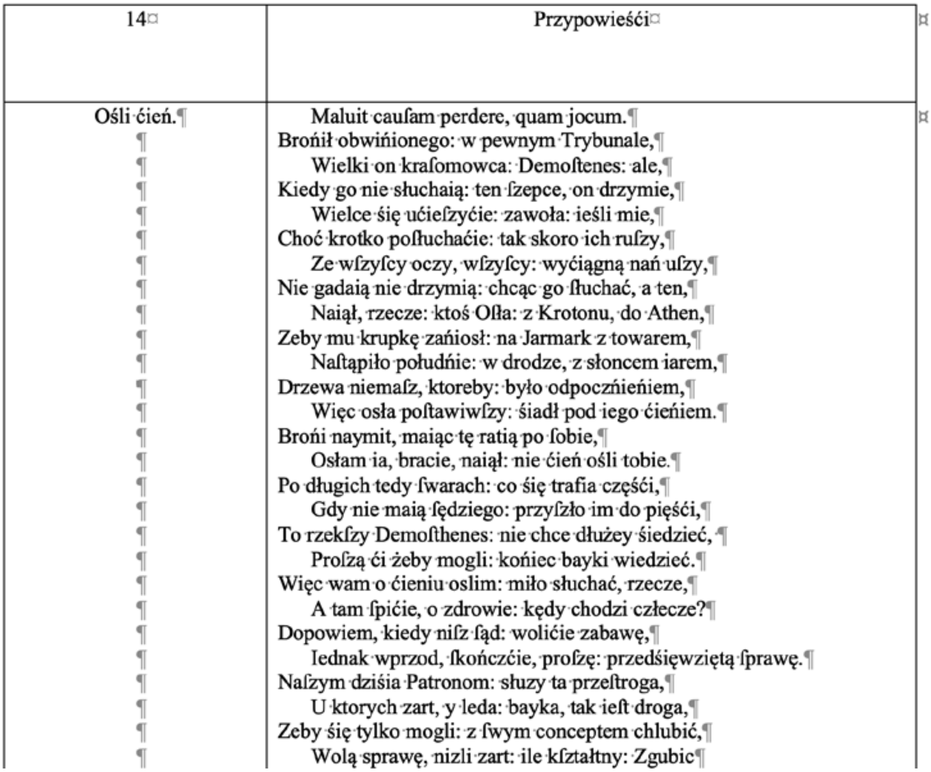
– retained the author's decisions regarding the use of capital letters (including the inconsistent but clear distinction between capital letters *I* and *J*),

– expanded abbreviations,

– used < > brackets to mark damaged and illegible places.

The transliteration was done in Word, so it had to be transferred to Transkribus and linked to the scans of the document. This procedure was carried out in several stages. First, scans of *Moralia* were downloaded from the Polona Digital

[34] As of December 2024.

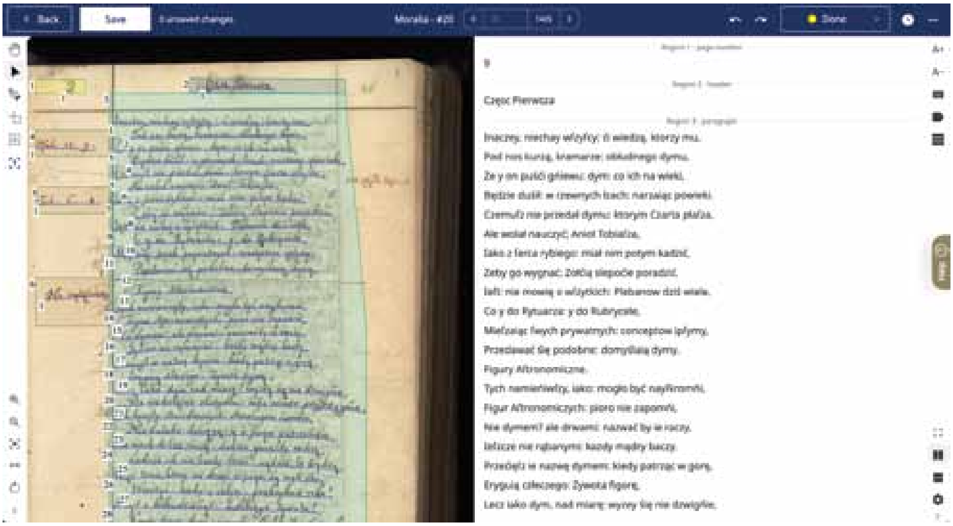


14¤	Przypowieśći¤
Ośli ćień.¶	Maluit cauſam perdere, quam jocum.¶
¶	Brońił obwińionego: w pewnym Trybunale,¶
¶	Wielki on kraſomowca: Demoſtenes: ale,¶
¶	Kiedy go nie słuchaią: ten ſzepce, on drzymie,¶
¶	Wielce śię ućieſzyćie: zawoła: ieśli mie,¶
¶	Choć krotko poſłuchaćie: tak skoro ich ruſzy,¶
¶	Ze wſzyſcy oczy, wſzyſcy: wyćiągną nań uſzy,¶
¶	Nie gadaią nie drzymią: chcąc go ſłuchać, a ten,¶
¶	Naiął, rzecze: ktoś Oſła: z Krotonu, do Athen,¶
¶	Zeby mu krupkę zańiosł: na Jarmark z towarem,¶
¶	Naſtąpiło połudńie: w drodze, z słoncem iarem,¶
¶	Drzewa niemaſz, ktoreby: było odpocząieńiem,¶
¶	Więc osła poſtawiwſzy: śiadł pod iego ćieńiem.¶
¶	Brońi naymit, maiąc tę ratią po ſobie,¶
¶	Osłam ia, bracie, naiął: nie ćień ośli tobie.¶
¶	Po długich tedy ſwarach: co śię trafia częśći,¶
¶	Gdy nie maią ſędziego: przyſzło im do pięśći,¶
¶	To rzekſzy Demoſthenes: nie chce dłużey śiedzieć, ¶
¶	Proſzą ći żeby mogli: końiec bayki wiedzieć.¶
¶	Więc wam o ćieniu oslim: miło słuchać, rzecze,¶
¶	A tam ſpićie, o zdrowie: kędy chodzi człecze?¶
¶	Dopowiem, kiedy niſz ſąd: wolićie zabawę,¶
¶	Iednak wprzod, ſkończćie, proſzę: przedśięwziętą ſprawę.¶
¶	Naſzym dziśia Patronom: służy ta przeſtroga,¶
¶	U ktorych zart, y leda: bayka, tak ieſt droga,¶
¶	Zeby śię tylko mogli: z ſwym conceptem chlubić,¶
¶	Wolą sprawę, nizli zart: ile kſztałtny: Zgubic¶

Fig. 3. Transliteration made in Microsoft Word.

National Library, then, using a batch function (a feature of Photoshop), the scans were reduced in size and divided into separate files with individual pages. The edited graphics were transferred to Transkribus[35] servers, where the next step was to build and organize the page layout. The program offers an automatic layout recognition feature and detects columns and lines of text on its own, but the editor must check this process because automatic recognition is not perfect. Often lines of text are out of order, baselines are broken, or a single line is interpreted as two separate lines (for example, when the author used larger than usual spacing between words). It is particularly important to ensure that the baselines reach the end of the text lines, as characters not on the baseline will not be read later. At the layout

[35] The activities described took place in the first half of 2023, when Transkribus was still operating as a program that could be downloaded and run on a computer. Even then, there were concerns the security of storing research data on third-party servers (in this case, those belonging to the READ–COOP association based in Innsbruck). While *Moralia* is in the public domain, the results of LabEdyt's work have not yet been made available under an open license, although this is planned. An even bigger problem is posed by works with unexpired copyrights. Currently, researchers who wish to use such legally protected materials for scholarly purposes have the option of creating a Transkribus instance on their own server (Transkribus On-Prem).



9

Część Pierwsza

Inaczey, niechay wßyſcy: ô wiedzą, ktorzy mu.
Pod nos kurzą, kramarze: obkadnego dymu.
Że y on puśći gęstego: dym: co ich na wieki,
Będzie duśił w trzewnych łbach: narzaiąc powieki.
Czemuż nie przedał dymu: ktorym Czarta płaſza.
Ale wolał nauczyć; Anioł Tobiasza.
Iako z serca rybiego: miał nim potym kadźić.
Żeby go wygnać: Żołćią ślepoćie poradźić.
Iaſz: nie mowię o wßytkich: Plebanow dźiś wiele.
Co y do Rytuarza: y do Rubrycele.
Mieſzaiąc ſwych prywatnych: conceptow ięfymy.
Przedawać się podobno: domyślaią dymy.
Figury Aſtronomiczne.
Tych namieśniwtry, iako: mogło być nayſkromni,
Figur Aſtronomiczych: pioro nie zapomni,
Nie dymem? ale drwami: nazwać by w roczy.
Ieszcze nie rąbanymi: każdy mądry baczy.
Przecięż ie nazwę dymem: kiedy patrząc w gorę,
Erygują człeczego: Żywota figurę,
Lecz iako dym, nad miarę wyżey się nie dźwignie.

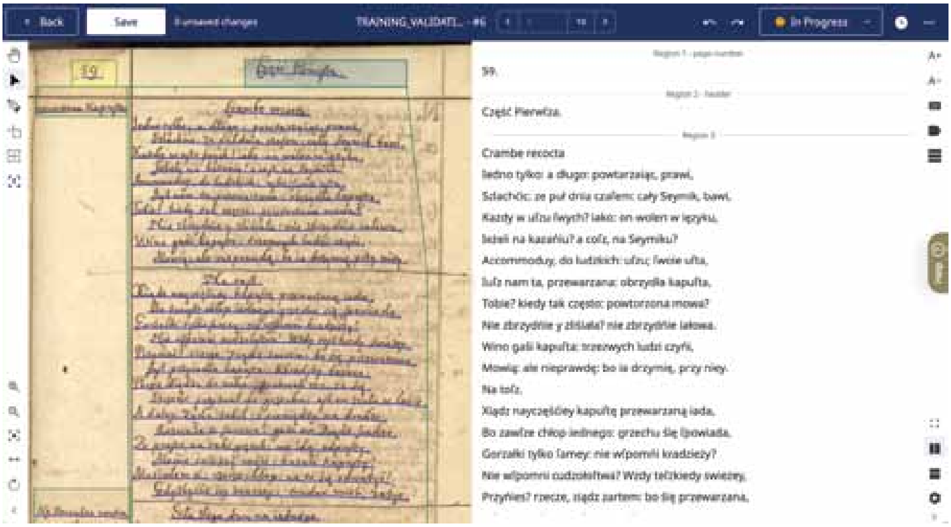
Fig. 4. Transkribus browser application – preview of the facsimile and transliteration. On the pages of the manuscript, apart from the main text, text fields with pagination, marginalia and a running header have been highlighted.

stage, much depends on the quality of the scan. For example, if the digitized book did not open fully, the inner part of the column on the scan will 'curl' towards the spine. Such curling causes two problems for Transkribus: first when recognizing the layout, and later when reading the characters (because they are distorted and tilted – it is not without reason that CAPTCHA images displayed on some websites and in electronic forms as protection against bots contain 'wavy' letters that humans can recognize without any problems, but computers cannot... at least until recently).

The finished transliteration of *Moralia* was assigned to previously mapped text fields and lines of text. Of course, Transkribus allows users to enter transliterations directly in the browser window, but preparing a sample in Word first has its advantages, primarily related to the more advanced text editing options available in this editor (searching, replacing, using regular expressions); its stability, which still is an issue in Transkribus, is also important.

The creators of the service suggest in the documentation that a sample of five to fifteen thousand words is sufficient to train an effective model.[36] Printed text recognition models usually perform well even on small samples, while manuscripts require much more material. Originally, according to the instructions, the LabEdyt team intended to use fifty pages of the manuscript (approximately fifteen thousand words), but the model obtained from this sample had a fairly high CER of 5.9%.

[36] Transkribus Help Center, "2. Data Preparation", https://help.transkribus.org/data-preparation.



59.

Część Pierwsza.

Crambe recocta
Iedno tylko: a długo: powtarzaiąc, prawi,
Szlachcic: ze puł dnia czaſem: cały Seymik, bawi.
Każdy w uſtu ſwych? Iako: on wołem w ięzyku.
Ieżeli na kazaniu? a coſz, na Seymiku?
Accommoduy, do ludzkich: uſzu; ſwoie uſta,
Iuż nam ta, przewarzana: obrzydła kapuſta.
Tobie? kiedy tak często: powtorzona mowa?
Nie zbrzydłnie y złśiała? nie zbrzydłnie iakowa.
Wino gali kapuſta: trzeźwych ludzi czyni,
Mowię: ale nieprawdę: bo ia drzymię, przy niey.
Na toſz.
Xiądz nayczęśćiey kapuſtę przewarzaną iada,
Bo zawſze chłop iednego: grzechu ſię ſpowiada.
Gorzałki tylko ſamey: nie wſpomni kradzieży?
Nie wſpomni cudzołoſtwa? Wżdy teżkiedy swieżey,
Przyńieś? rzecze, xiądz zartem: bo ſię przewarzana.

Fig. 5. Validation set for the Potocki-2 model. On the right, text read by a computer, no editor intervention.

It was only after adding an additional fifty pages of transliteration that the CER was reduced to a satisfactory level of 2.6%.

What conclusions can be drawn from this experiment? First and foremost, it is important to consider when it is really worth investing time in training an AI model to recognize text. Had the *Moralia* manuscript been 300 pages long, it might have been quicker, easier, and more accurate to transcribe it manually into a text file. After all, the use of artificial intelligence does not relieve the scholarly publisher of the obligation to carefully proofread the text several times; the preparation of the training sample itself is also time-consuming. Nevertheless, the procedure presented here has a significant advantage if we intend to create a digital scholarly edition in the long term. Transkribus' user-friendly graphic interface is equipped with useful tagging functions, thanks to which it is possible, already at the transcription stage (or transliteration, depending on the user's needs), to mark illegible, damaged, deleted, supplemented or corrected places, and also to indicate elements of the source text's structure, such as pagination, running headers, catchwords, signatures, marginalia, titles, and the like. Recently, it has become possible to use the <persName> and <placeName> tags, as well as to define one's tags. All this can be achieved without entering a single line of code, which is an undoubted advantage for editors who are reluctant to use markup languages. The text obtained in this way is not actually plain text (although it can be downloaded from Transkribus, as the application allows export in many different formats), but has an additional layer of data that will later be included in the exported Page XML file converted to TEI XML.[37] This greatly facilitates work on digital scholarly editions, as it allows one to obtain pre-tagged and mapped files in a semi-automatic way, in which lines of text or even individual words are assigned coordinates on the digital facsimile.

Currently, the correction of the transliteration of the entire *Moralia* is nearing completion. The next stage of work will begin soon – the semi-automatic creation of a transcription modernizing the spelling. To this end, the team intends to use a word frequency list obtained by processing a clean text file using a simple script written in Python. While work was underway on reading Potocki's manuscript (the Potocki-2 model was created on 16 June 2023), significant progress was made in this field. Transkribus itself has undergone a huge transformation, evolving from a simple, virtually free tool offering unrivalled functionality at the time to a commercial service operating in a freemium model.[38] However, it is worth remembering that digital scholarly editing has exactly the same goal as traditional editing – the scholarly preparation of a text in accordance with accepted guidelines. Therefore, one should not overly fetishize digital humanities tools or become particularly attached to them. If a program that better fulfills its purpose appears today or in the near future, and there are already opinions[39] that it could be made available under an open license such as eScriptorium, then nothing will prevent subsequent transcriptions of literary works from being created in it.

*

In the 1960s, Swiss typographer Adrian Frutiger collaborated with the European Computer Manufacturers Association (ECMA) to develop a typeface that would be recognizable by optical readers. This resulted in the 'standardized Latin alphabet' OCR-B (Optical Character Recognition – font B), which was a compromise between the requirements of the digital machines of the time and the centuries-old tradition of typography.[40] The designer tried to draw the individual characters of

[37] Transkribus allows data to be exported directly to a TEI file, but this option is only available to users who subscribe to the premium plan.

[38] Dynamic changes in access to software pose a significant threat to projects financed under the grant model. They force project managers to take into account not only currency exchange rates, but also potential increases in license fees.

[39] The developers of TEI Publisher, a tool for publishing digital scholarly editions, have announced plans to integrate it with eScriptorium. However, this software has a significant drawback: because it is developed in an open source model, it is not as user-friendly as its competitor, Transkribus.

[40] A. Frutiger, "OCR-B. Znormalizowane pismo o czytelności optycznej" [OCR-B: A standardized character for optical recognition], transl. by R. Tomaszewski, in: *Litera* [suplement *Poligrafiki*] [Letter (supplement to Poligraphics)] Y. 3 (1968), nos. 23–25, pp. 91–93 [completed in no. 25, pp. 97–102]. [All quotations in English translated from the Polish edition].

the numbers and alphabet so that none of them (for example, I and l or B and 8) were similar to each other, as 'flawed' computers could not cope with small differences in shape. Even then, Frutiger had no doubt that 'one day, a reading machine will become so advanced that it will be able to read the characters and forms of any style of contemporary alphabets without error'.[41] The typographer considered his work on the OCR-B typeface, which was not particularly attractive, to be a 'success in the field of ethics', because 'it is not the machine that forces man to use a "mechanized" style of characters; it is man who tries to "teach" the machine to read the writing that is in common use, the writing that has developed over centuries from stone hieroglyphs, through the pen and parchment, the engraver's chisel, to the current methods of graphic designers, typesetters, and printers of our time'.[42] Is that so?

The artificial intelligence revolution we are witnessing today is the future Frutiger dreamed of. Its course raises many ethical and existential questions: about the future of existing professions, about ecology (supercomputers consume enormous amounts of energy), about data privacy. Scholarly editors are also asking themselves these questions. Automatic transcription or transliteration using AI seems like a fairly innocent procedure. It is obvious that in the end the text will be read and corrected by a 'human instance'. But what about attempts to create commentaries with the use of chatbots? Who will be the author of such an edition, who will take responsibility for it – the editor writing the prompts or the computer processing the knowledge of all humanity? Is a text written at our request our text? Will the editor continue to be an editor – the one who knows the work best – or perhaps a specialized machine operator?

Translated by Maja Jaros,
verified by Marek Pąkciński

Bibliography

Brückner A., *Wacława Potockiego „Moralia" (1688), wyd. Tadeusz Grabowski, Jan Łoś, [Kraków] 1915–1918* [recenzja], „Pamiętnik Literacki" 1920, t. 17/18, nr 1–4.

C21 Editions. Scholarly Editing and Publishing in the Digital Age, https://www.c21editions.org/ (wszystkie przywoływane źródła internetowe – stan z 20 września 2024 r.).

Corpus Thomisticum, https://www.corpusthomisticum.org/.

DataReportal, *Digital 2024 Global Overview Report*, https://datareportal.com/reports/digital-2024-global-overview-report.

[41] Ibidem, p. 102.

[42] Ibidem.

Digital Humanities Lab, https://dhlab.id.uj.edu.pl/.

Eisenstein E.L., *Rewolucja Gutenberga*, tłum. H. Hollender, Warszawa 2004.

Elektroniczny korpus tekstów polskich XVII i XVIII w. „KorBa", https://korba.edu.pl.

Frutiger A., *OCR-B. Znormalizowane pismo o czytelności optycznej*, tłum. R. Tomaszewski, „Litera" [dod. „Poligrafiki"] 1968, R. 3, nr 23–25 [dokończ. w nr. 25].

Gruchała J., *Wacław Potocki – problem edytorski*, w: *Wacław Potocki. W 400-lecie urodzin poety*, red. K. Koehler, D. Chemperek, Warszawa 2023.

Grzeszczuk S., *O potrzebie i programie badań nad twórczością Wacława Potockiego*, w: *Wśród zagadnień polskiej literatury barokowej*, cz. 2: *Motywy – inspiracje – recepcja*, red. Z.J. Nowak, Katowice 1980.

Grześkowiak R., *Stary druk jako podstawa edycji krytycznej. Preliminaria*, w: *Jak wydawać teksty dawne*, red. K. Borowiec et al., Poznań 2017.

Hollender H., *Czy świat czeka przyszłość średniowiecza?*, w: E.L. Eisenstein, *Rewolucja Gutenberga*, tłum. H. Hollender, Warszawa 2004.

Kukulski L., *Prolegomena filologiczne do twórczości Wacława Potockiego*, Wrocław 1962.

Pierazzo E., *What Future for Digital Scholarly Editions? From Haute Couture to Prêt-à-Porter*, „International Journal of Digital Humanities" 2019, nr 13.

Potocki W., *Dzieła*, t. 3: *Moralia i inne utwory z lat 1688–1696*, oprac. L. Kukulski, Warszawa 1987.

Potocki W., *Moralia*, rękopis, ok. 1688–1696, Biblioteka Narodowa, rps 3049 III, Polona.pl.

Potocki W., *Wacława Potockiego „Moralia" 1688*, t. 1–3, wyd. T. Grabowski, J. Łoś, Kraków 1915–1918.

Potocki Wacław (1621–1696), w: *Bibliografia literatury polskiej. „Nowy Korbut"*, red. K. Budzyk, t. 3: *Piśmiennictwo staropolskie. Hasła osobowe N–Ż*, oprac. R. Pollak et al., Warszawa 1965.

Program of the ESTS Conference 2024, https://elte-dh.hu/ests-2024-program/.

Roterodamus E., *Adagiorvm Chiliades Des. Erasmi Roterodami Qvatvor Cvm Dimidia Ex Postrema Avtoris Recognitione: In hac aeditione, prioribus tribus Indicibus subiunctus est quartus nouus* […], https://rcin.org.pl/dlibra/publication/74578/edition/66983/content.

Shillingsburg P.L., *From Gutenberg to Google. Electronic Representations of Literary Texts*, Cambridge 2006.

Silberling A., *Why AI Can't Spell 'Strawberry'*, https://techcrunch.com/2024/08/27/why-ai-cant-spell-strawberry/.

Tong A., Paul K., *Exclusive: OpenAI Working on New Reasoning Technology under Code Name 'Strawberry'*, https://www.reuters.com/technology/artificial-intelligence/openai-working-new-reasoning-technology-under-code-name-strawberry-2024-07-12.

Transkribus – Unlocking the Past with AI, https://www.transkribus.org/.

Viglianti R., del Rio Riande G., *Against Infrastructure. Global Approaches to Digital Scholarly Editing*, „C21 Digital Editions" 2023.

Viglianti R., del Rio Riande G., Hernández N., De Léon R., *Otwarte, minimalne i promujące równe szanse. Jak uczy się tworzenia naukowych edycji cyfrowych na Globalnej Północy i Globalnym Południu*, tłum. J. Wełniak, „Sztuka Edycji" 2023, nr 1.

Abstract

In the early 2020s, Artificial Intelligence became extremely popular. Due to new technology, more processes from various fields are subject to automation every day.

Editing is no different, as concepts of utilising the feats of engineering to create scholarly editions are voiced ever more confidently. In the international community of Humanists, there is an ongoing debate on the possibilities for creative uses of generative artificial intelligence and machine learning in work with text.

The first part of the article portrays the current state of scholarly editing in the digital age. The second part is devoted to a case study of utilising machine learning for the automatic generation of a transliteration of the seventeenth-century manuscript of *Moralia* by Wacław Potocki. The text concludes with a discussion on the ethics of editing aided by AI. The endeavour is carried out at the Jagiellonian University as part of the Digital Humanities Lab project.

Keywords: Wacław Potocki (1621–1696), digital scholarly edition, the application of AI in the Humanities, Digital Humanities, OCR and HTR