Studia UBB Digitalia, Volume 65 (LXV), 2020, June, Issue 1, 37-52

Published Online: 2020-12-10

DOI:10.24193/subbdigitalia.2020.1.03

Preliminary Research on Computer-Assisted Transcription of Medieval Scripts in the Latin Alphabet using AI Computer Vision techniques and Machine Learning. A Romanian Exploratory Initiative

Adinel C. Dincă and Emil Ștețco
Babeș-Bolyai University
Zetta Cloud, Cluj-Napoca

Abstract: The objective of the present paper is to introduce to a wider audience, at a very early stage of development, the initial results of a Romanian joint initiative of AI software engineers and palaeographers in an experimental project aiming to assist and improve the transcription effort of medieval texts with AI software solutions, uniquely designed and trained for the task. Our description will start by summarizing the previous attempts and the mixed-results achieved in *e-palaeography* so far, a continuously growing field of combined scholarship at an international level. The second part of the study describes the specific project, developed by Zetta Cloud, with the aim of demonstrating that, by applying state of the art AI Computer Vision algorithms, it is possible to automatically binarize and segment text images with the final scope of intelligently extracting the content from a sample set of medieval handwritten text pages.

Keywords: Middle Ages, Latin writing, palaeography, Artificial Intelligence, Computer Vision, automatic transcription.

A. Introduction

In 1971 György Granasztói (1938-2016), a Hungarian historian and demographer, teamed up with a Russian technician, Valentin A. Ustinov (1938-2015), and together used the computer – inputting data via punched cards – to analyse from a quantitative perspective (Granasztói 1971) a tax register compiled in 1475 in Brașov (an account book regarding the tax revenue of *Portica* neighbourhood in Brașov,

preserved at Serviciul Județean Brașov al Arhivelor Naționale [Brașov County Service of the National Archives], „Primăria orașului Brașov", Oficiul impozitelor oraș Brașov, seria III Da, no. 1/1, 1475). The authors were following a fashionable ideological trend of that moment, which pointed out that, by using mathematical and computer-oriented tools, historians could move toward a fundamental scientific objective: the development of a common language for interdisciplinary work that could solve complex social problems (Putnam 1971, 24). Beyond the political principles that motivated this endeavour, some elements still valid 50 years later need to be highlighted: first, the research was based on an **experiment**, which was conducted with the use of an **innovative** machine at that time. Secondly, it involved a Hungarian historian, a Russian computer analyst and a Romanian archival item which reflected information concerning a medieval Saxon town in Transylvania – thus, the operation did not only have an **international** character, but was also validated by a **cross-field** level investigation. These key-words: "Experimenting" – "Innovating" – "International" – "Cross-field level" still define today the interaction of humanities and computer sciences, even though the results have moved greatly from a simple quantitative report produced half a century ago towards the next phase, simply put as the transcription of medieval writing with the help of an intelligent machine.

The history of handwriting, **palaeography,** is a specific field within medieval studies. According to the definition:

> "this discipline studies the appearance and development of different types of script and their uses by diverse social groups across time and in diverse documents (books, records, charters, etc.). Further, it analyses the transmission and staging of a written message, attending not only to the text but also to its form, script, layout and support." (Stutzmann 2016)

In these circumstances, the transcription of medieval manuscript texts sets itself apart as a complex, highly specialized, and time-consuming activity. Experts in palaeography need to be trained in the language of the text, the historical usages of various styles of handwriting, common writing customs, and scribal/notarial abbreviations. Thus, one may spend many hours transcribing rather short medieval original documents to offer researchers ways of indexing, finding, and consulting such evidence. Complete editions of extended works usually take years to complete, a reason why modern scholars, confronted with the issue of "short-termism", avoid engaging (too often) in such endeavours or prefer to do partial editions (Bertrand 2005, Słoń 2015). Some religious texts such as sermons or saints' lives, for example, were completely neglected by editorial projects, due to their large number of individual copies transmitted in variants copied over centuries in a plethora of writing solutions and contexts.

More than any other disciplines, palaeography is most dependent on visual evidence. Thus, the invention of photography in the 19th century compelled Ludwig Traube (1861-1907), the palaeographer who held the first chair of Medieval Latin at the "Ludwig Maximilian University" of Munich, to address the 1900s as 'the age of photography' in the history of this particular field of study (Lehmann 1909, 57). Visual contexts provided palaeographers with ways to broadcast and compare manuscripts

while slowly linking researchers into interconnected networks around the digitized images (Widner 2017) that became valuable tools for consulting important manuscripts to which access could be hard to obtain, for reasons of conservation or value (Wakelin). The conversion of texts and pictures into a digital form that can be processed by a computer produced a *digital representation* or, more specifically, ***a digital image*** in the form of binary numbers, which facilitated computer processing and other operations. This process of **digitalization** has set a milestone in the study of automated transcription, yet, in order to be "read" by either the human eye or the artificial intelligence, the image had to be further "prepared" for browsing.

The classical problem in computer vision, image processing and machine vision has been that of **recognition**, from the first **Computer Vision** projects that were meant to mimic the human visual system. Experimenting began in the late 1960s, at universities in the West, as a task connected to artificial intelligence, and the pioneering studies in the next decades formed the foundations for many of the computer vision algorithms that exist today. The next phase, bringing further life to the field of computer vision, was the so-called deep-representation learning based on neural networks. **Deep Learning techniques** have currently the accuracy and performance close to that of humans, relying on **convolutional neural networks (CNNs).** This evolution, not only in terms of sophistication of the learning process, has been paralleled with the progression of the performed tasks: while

> "the first wave of digital humanities' work, from the late 1990s to early 2000s, was described as quantitative, mobilizing the search and retrieval powers of the database, the second wave emphasizes the keywords: qualitative, interpretive, experiential, emotive, generative. [...] It harnesses the digital toolkits in the service of the Humanities' core methodological strengths: attention to complexity, medium specificity, historical context, analytical depth, critique and interpretation." (*Digital Humanities Manifesto 2.0*)

Experts in palaeography or the "auxiliary sciences of history" – such as epigraphy, codicology, numismatics, sphragistics to name just a few – were always a scarce resource, and currently the changes in the educational system make them even more difficult to find. As it has been noted, "palaeographic skills can be acquired only through intensive training and prolonged exposure to rare artifacts that can be difficult to access" (Kestemont et al. 2017, S87). Recent opinions even suggested that such experts are obsolete due to their subjective, dogmatic and authoritarian perspective (Stokes 2009, 313-317), and plead for objective criteria in palaeography, by generating a set of measurements which can ultimately be statistically analysed and compared by computers. The result of these divergent perspectives may lead to a paradoxical situation: humanity will soon have an exhaustive digital memory of its medieval evolution, but it will not have access to it, due to the missing corpus of trained specialists able to decode old forms of writing.

The variability of the handwriting, the complexity of the vocabulary and styles, the multi-linguistical and multi-graphical aspects, the difficulty of automatically isolating the characters and segmenting the text lines make the currently systems based on recognition, the Optical Character Recognition (**OCR**) and the Handwritten Text

Recognition (**HTR**), unable to automatically recognize and transcribe medieval texts. While OCR is considered a closed problem in computer vision, handwritten text recognition still presents an open challenge. Yet, the need to provide efficient solutions to time-consuming and laborious palaeographic tasks pushes all academic fields towards multi-disciplinary collaboration in the search for new options. The creation of the Text Encoding Initiative **(TEI)** (https://tei-c.org/), developed in the 1990s, has offered humanities scholars common standards for encoding electronic texts and representing texts in digital form, TEI **guidelines** being used around the world by libraries, museums, publishers, and individual scholars for online research, teaching, and preservation.

B. Computer-assisted palaeography. An overview of European projects

Medieval scripts developed in the Latin cultural area are characterized by different handwriting styles from diverse places and periods of time over one thousand years, from the 6th to the 15th century. Moreover, the typological diversity of such graphical solutions for human communication (Derolez 2003) is increased by the intention laying behind a certain text. For example, a notarial document will never be similar to a liturgical text, as much as a university handwritten textbook could never be mistaken for a formal papal bull or a royal charter. Different sets of letterforms were specifically designed for individual contexts of use and destinations, each context being closely connected to the issuer and the beneficiary, respectively with their intention regarding the text. Therefore, notions like ***hierarchical structure of information, formality***, ***standardization*** or ***level of execution*** play an important role in classifying and describing a certain sample of medieval handwriting. All these factors create the uniqueness of every handwritten record produced during the Middle Ages.

The current scholarly field of “computer-assisted” or “artificial palaeography” represents the small tip of an iceberg: its massive base assembles the digitized heritage of “European” handwriting, that is, based on the Latin alphabet, with initiatives that also take into consideration Greek and Coptic (https://d-scribes.philhist.unibas.ch/en/home/), or Hebrew (http://www.erabbinica.org/), with all their quirks and features.

A large part of the medieval texts preserved in archives or libraries all over the world has been intensively digitized during the last two decades. Considerable institutional or private funding was invested in preserving the textual memory of the medieval past, aiming at the same time an improved accessibility to the source material for international scholarship, as usually repositories with medieval books or documents are in various countries across the continent(s). Today, virtually every single larger library and archival collection from Austria, France, Germany, Italy (with the Vatican state), Switzerland, to name only a few, has its own digitization project, on various levels of development but comprising tens of thousands of scanned units with uncountable number of pages. However, even though such digital libraries – BVMM (https://bvmm.irht.cnrs.fr/), Gallica (https://gallica.bnf.fr/), e-Codices (https://www.e-codices.unifr.ch/en), Manuscripta Mediaevalia (http://www.manuscripta-mediaevalia.de/#|1), DigiVatLib (https://digi.vatlib.it/) – and archives – Monasterium (http://monasterium.net:8181/mom/home) – are amassing reproductions of medieval manuscripts and archives, they offer scarce metadata. The

construction of such impressive digital libraries required so far technical solutions and skills and usually neglected or undervalued the traditional scholarship based on palaeography and textual criticism.

As briefly stated, digitization projects represent at this point an accomplished task for most European archives, Romania included (see, for instance, the digitization project of the Romanian National Archives, http://arhivamedievala.ro/). Database creation and management is also a fruitful ongoing activity; however, this approach does not cover the specific needs of digital palaeography. Beyond the first stage of acquiring images, the interaction of palaeography with computerized tools seeks to provide efficient solutions to the consuming palaeographic tasks. Various attempts have produced mixed-results in **e-palaeography** over the last two decades, to cite just a few: the **System for Paleographic Inspections (SPI)** was the first software dedicated to digital palaeography developed in 1999 by a group of researchers in the History Department, University of Pisa, for the inspection of ancient Roman manuscripts – from this project derived in 2004 the first attempts at automatically clustering scripts, which led to Arianna Ciula's coining of the term "digital palaeography" (Aiolli et al. 1999; Ciula 2005; Aiolli & Ciula 2009; Aussems & Brink 2009).

The **Monk** system (https://www.ai.rug.nl/~mrolarik/MLS; /http://monkweb.nl/) is a continuous project, developed at the University of Groningen in 2005 by a research group at the Institution of Artificial Intelligence and Cognitive Engineering (ALICE), under the supervision of Lambert Schomaker. The **GRAPHEM** (*Grapheme based retrieval and analysis for palaeographic expertise of medieval manuscripts*) research project was funded from 2007 to 2011 by the French National Agency for Research, under the supervision of Dominique Stutzmann, on the basis of automated image analysis without the need to select individual letters (Cloppet et al. 2007; Muzerelle & Gurado 2011).

The research programme **ORIFLAMMS** (*Ontology research, image feature, letterform analysis on multilingual medieval scripts* –http://oriflamms.teklia.com/), 2013-2016, tackled the same issues with new methods that combined letterform analysis with Computer Vision for script classification. ORIFLAMMS, coordinated by Dominique Stutzmann, aimed at analysing the evolution of writing systems and graphical forms during the Middle Ages and according to their production contexts (informal, documentary, book scripts) and languages (Latin or vernacular) (Stutzmann 2016; Oriflamms 2017).

DigiPal (*Digital Resource for Palaeography* – http://www.digipal.eu/) developed between 2010-2014 under the supervision of Peter A. Stokes, a scholar at King's College, London (Stokes 2009; Stokes 2011). The project's aim was to catalogue, describe and, where possible, source digital images of about 1200 scribal hands. Data was organized with the help of **Archetype** (http://archetype.ink/), an integrated suite of web-based tools for the study of medieval handwriting, art and iconography. Peter A. Stokes also proposed a different approach (Stokes 2007), consisting of two stages, the first known as "feature extraction", which involves generating the numerical measurements, and the second, "data mining", finding similarities and classifying handwriting based on these measurements.

tranScriptorium project (http://transcriptorium.eu/) was active between 2013-2015; its aim was to develop innovative, efficient and cost-effective solutions for the indexing, search and full transcription of historical handwritten document images, using an enhanced version of Handwritten Text Recognition (HTR) technology and Layout Analysis. The result, the **Transkribus** software (https://transkribus.eu/Transkribus/), is currently hosted by the "Digitisation and Digital Preservation group" (DEA) at the University of Innsbruck. It is also used by the EU-funded e-infrastructure project **READ** (Recognition and Enrichment of Archival Documents), coordinated by Günter Mühlberger of the University of Innsbruck (https://read.transkribus.eu/), an undertaking that concluded just some months ago.

Another recent project was **HIMANIS** (*HIstorical MANuscript Indexing for user-controlled Search* – https://www.himanis.org/), again under the coordination of Dominique Stutzmann, aimed at developing cost-effective solutions for querying large sets of handwritten document images, more precisely the textual heritage of the French Royal Chancery compiled in the 14th and 15th centuries: charters, registries and formularies. To this end, innovative keyword spotting, indexing and search methods have been developed, tested, adapted and/or scaled up to meet the challenges of more than 83.000 digitized pages (Stutzmann 2018).

The **Scripta-PSL: The History and Practices of Writing** programme at Université Paris Sciences et Lettres (https://escripta.hypotheses.org/) aims at integrating the fundamental sciences that deal with written artefacts (palaeography, codicology, epigraphy, history of the book, etc.), with other disciplines in the humanities and social sciences (linguistics, history, anthropology, etc.), together with digital and computational humanities, around the study of writing. The digital component, **eScripta**, intends to mash software programmes (such as **Archetype**, developed by Peter Stokes and the team at King's College London, which allows deep annotation and extensive palaeographic study of writing, an open-source OCR software called **kraken**, developed by Benjamin Kiessling (PSL research engineer) and **Pyrrha**, a tool for post-correction of POS tagging and lemmatization with the help of CNNs to isolate objects from their background and distinguish between main writing, decoration (illuminations, drop capitals, etc.), and interlineal or marginal annotations (Stokes et al. 2019).

The **Digital forensics for historical documents** project, funded by the Royal Netherlands Academy of Arts and Sciences (https://en.huygens.knaw.nl/projecten/digital-forensics-for-historical-documents/?noredirect=en_GB), is currently ongoing (2018-2021) under the supervision of Mariken Teeuwen. The research is developing the results of previous projects that have explored automatic methods for handwritten text analysis: **Monk**, **Transkribus**, **DigiPal** and **Artificial Palaeography** (developed recently by Dominique Stutzmann e.a., see Kestemont et al. 2017). The project Digital Forensics aims at creating a bridge in between two different ways of handwriting analyses: the forensic method (graphanalysis) and the palaeographical method (the study of the development of scripts through space and time), combining the two methods in a single 'deep learning system'.

Before concluding this brief overview of the blooming **Digital Palaeography** arena, it is our intention to share some thoughts regarding how Artificial Intelligence, Computer Vision techniques and Machine Learning processes could contribute to a more efficient and accurate transliteration of historical texts issued within the cultural area of the Latin Middle Ages. First, the technology is not yet mature, and it poses a series of obvious issues. Difficulties in fluid communication between palaeographers and computer scientists is a prevailing problem, another is restriction of property: licensed vs. open source software (Hassner et al. 2012, 15, 24). Reiterating the opening paragraph of this presentation, digital humanities work best when a fruitful collaboration is established between scholars, computer scientists and, last but not least, cultural heritage institutions – in other words, when the roles and competence of each party have been identified and acknowledged. Integration of user feedback (participatory engagement) is another aspect that needs to be taken into consideration when discussing the software development strategies. Our view on the matter is therefore not to replace the expert's eye and hard acquired training (as it is sometimes casually suggested and expected), but to equip palaeographers with new tools and solutions that would help them to face with improved efficiency the overwhelming mass of medieval manuscripts available now in the digital environment (see a parallel experiment in Fischer et al. 2009). Interpretation of the results would have to derive from an interdisciplinary discussion across fields of expertise, elaborated by scholars of both Humanities and Computer Science disciplines. After all, the final goal of computer scientists is to develop an expert system that will emulate human expertise, in this particular case that of palaeographers' competences, methodologies, taxonomies and "ground-truth" (Stutzmann 2015).

C. A computer-assisted transcription project for medieval text images

In our specific project, developed by Zetta Cloud (a SME entity), the goal was to demonstrate that, by applying state of the art AI Computer Vision algorithms, we are able to automatically binarize and segment text images with the final scope of automatically extracting text from a sample set of medieval handwritten text pages. For applying AI algorithms used in Computer Vision to the handwriting transcription task, we trained, in an experimental manner, a Recurrent Neural Network (RNN) to recognize text by feeding it with enough previously humanly transcribed text image segments. For this purpose, a sample was selected from the 12th century handwritten Latin manuscript from Engelberg, Stiftsbibliothek: *Vitae Sanctorum et passiones Martyrum*. *Pars aestivalis* (https://www.e-codices.unifr.ch/de/list/one/bke/0002/). The selected manuscript was copied in the second half of the 12th century within the premises of Engelberg Abbey, Switzerland, a Benedictine foundation of 1120. The script engaged for the transmission of the widely spread text (Lives of the Saints, organized according to the liturgical calendar, the summer part) is a Caroline Minuscule in its late development before the spreading of Gothic script (http://carolinenetwork.weebly.com/). This particular form of Caroline Minuscule reflects accurately the calligraphic mastery of the Swiss Benedictine Convent

(Bruckner 1950), the script achieving at that time an almost typographic regularity. There is virtually no variation of the individual letter forms and the use of elements of abbreviation, as well as the diacritical signs, are remarkably constant. All these features (further details in Dincă 2011), combined with the exceptionally good digital copy at hand and an almost optimal state of preservation, advocate the above-mentioned manuscript as a suitable sample for our research. It must be added that the book comprises hundreds of folios copied by the same hand, thus providing enough reliable material to be worked with. Minor, subjective variations of forms (dimensions, rightward slanting, marginalia etc.), if needed, can be inserted into recognizing and identification patterns used in Computer Vision.

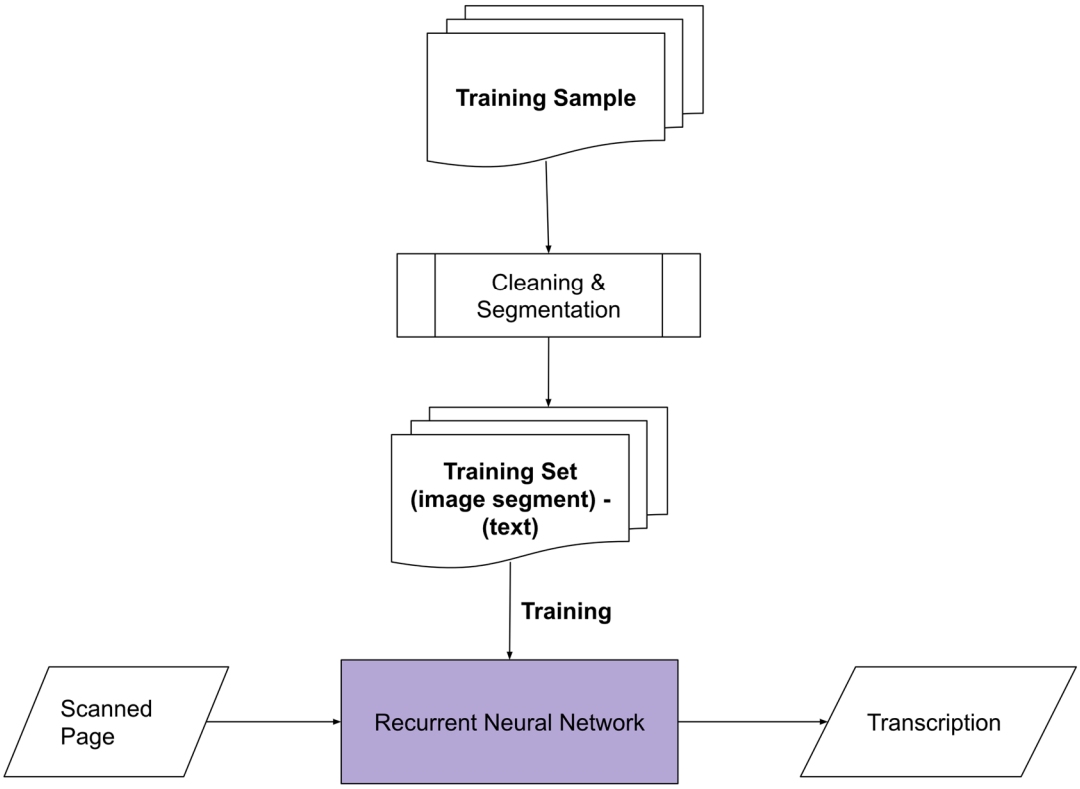
We have done extensive research to find the best implementation of an RNN for our purposes. Also, previous research carried out by well-known universities were analysed in order to gain as much knowledge as possible. In search for the best RNN implementation, our goal was to find an Open Source library that allows the specification and use of different Recurrent Networks architectures with ease and, if possible, with a no-code approach. Kraken implements a dialect of the Variable-size Graph Specification Language (VGSL), enabling the specification of different network architectures for image processing purposes using a short definition string. For our project purposes, the need arose for a library that could be used to train lightweight AI Computer Vision models where the recognition is done not on the character level but rather on image pattern level, this to allow expansions of abbreviations used by the medieval writer. On the other hand, Kraken implements the full Handwritten Recognition Pipeline one would expect, meaning: Page Segmentation followed by Line Segmentation and Handwriting recognition. At the end of our research, we have decided to use the Kraken Open Source Library as the foundation for our implementation.

We applied three training phases, each followed by testing phases carried out using previously humanly transcribed text image segments that were not used during the training phase to assess the achieved model accuracy.

Once the image acquisition issue was solved, the first training stage, consisting of a set of examples or training samples, could be put together. In order to obtain this training set, a pre-process of preparation was required: images were cleaned from noises but still preserved the text and its features, while an automatic binarization process was applied on all training sample pages. The set was then split into (1). a training set and (2). a testing set (typically, the ratio was 80% of the set for training and 20% of the set for testing), followed by training sessions of the Machine Learning Algorithm with this set; accuracy was automatically determined using the testing set. Essentially, the run experiments used intuition and statistics to make the machine able to understand what needs to be done to solve the specific handwriting transcription problem.

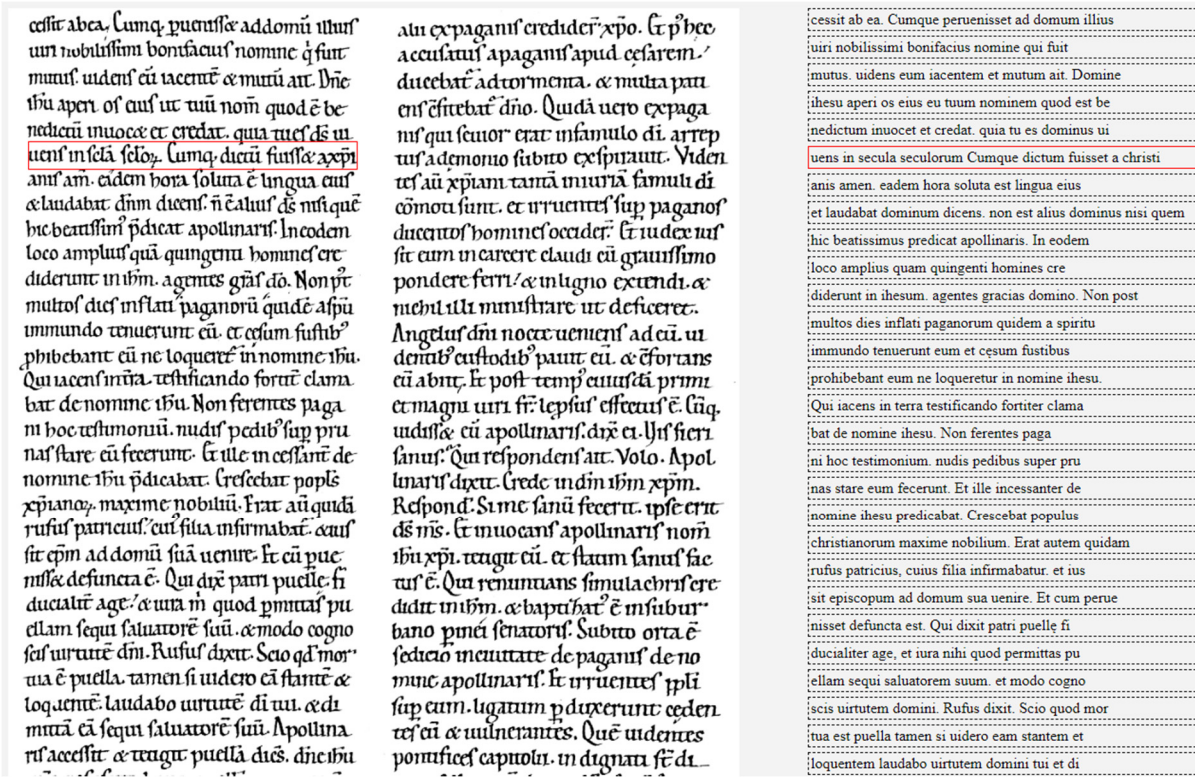
In addition to the automatic validation at the end of a training phase, human expert evaluation was absolutely needed during both training and experiment phases. The training corpus consisted of pairs of images and their “caption” transcriptions, and

the initial aim was to prepare a training data set comprising at least 50 text images with their corresponding transcriptions. Computer scientists have tailored and then used algorithms for automatic binarization of original text images. The same procedure was followed with implementing and using AI algorithms for automatic segmentation of the original text images. The result of this phase was used by the expert in palaeography during the transcription phase. Segmented versions of the original documents were manually transcribed for the purpose of creating the training and testing sets used at training the AI computer vision algorithm. The manual transcription has been undertaken by a qualified expert in Latin palaeography, to ensure the accuracy of the data set.



A typical Machine Learning Flow

For the first training phase, 2972 segments were used, extracted from the first 38 pages (1r to 19v) of the sample manuscript. Training the RNN AI algorithm with these 2972 segments (keeping just 80% of the set for training and 20% for testing and control purposes) generated an automatically computed accuracy of 0.9461962161730662 (94.61%). This is an excellent result for a first phase of training. We have also analyzed automatic transcriptions of pages not seen by the AI algorithm by now (any pages starting from 20r outwards) and the results were extremely positive.



cessit ab ea. Cumque peruenisset ad domum illius
uiri nobilissimi bonifacius nomine qui fuit
mutus. uidens eum iacentem et mutum ait. Domine
ihesu aperi os eius eu tuum nominem quod est be
nedictum inuocet et credat. quia tu es dominus ui
uens in secula seculorum Cumque dictum fuisset a christi
anis amen. eadem hora soluta est lingua eius
et laudabat dominum dicens. non est alius dominus nisi quem
hic beatissimus predicat apollinaris. In eodem
loco amplius quam quingenti homines cre
diderunt in ihesum. agentes gracias domino. Non post
multos dies inflati paganorum quidem a spiritu
immundo tenuerunt eum et cesum fustibus
prohibebant eum ne loqueretur in nomine ihesu.
Qui iacens in terra testificando fortiter clama
bat de nomine ihesu. Non ferentes paga
ni hoc testimonium. nudis pedibus super pru
nas stare eum fecerunt. Et ille incessanter de
nomine ihesu predicabat. Crescebat populus
christianorum maxime nobilium. Erat autem quidam
rufus patricius, cuius filia infirmabatur. et ius
sit episcopum ad domum sua uenire. Et cum perue
nisset defuncta est. Qui dixit patri puelle fi
ducialiter age, et iura nihi quod permittas pu
ellam sequi saluatorem suum. et modo cogno
scis uirtutem domini. Rufus dixit. Scio quod mor
tua est puella tamen si uidero eam stantem et
loquentem laudabo uirtutem domini tui et di

As planned, the goal was to understand how would the accuracy of the algorithm evolve when the training set was increased to 5427 segments, extracted from approximately 68 pages (1r to 34v). At this point, a second RNN Training phase was conducted, starting with an increased effort to transcribe further pages in a traditional manner, up to page 34v. By segmenting all these pages, a total of 5427 segments were generated for this second training phase. The computed achieved accuracy was of 0.9509186465708205 (95.09%), consequently the accuracy gain was of only 0.48%.

Therefore, by almost doubling the manual transcription effort a gain of only 0.48% of computed accuracy was achieved. Nonetheless, by conducting expert analyses of the automatic transcriptions generated by this second trained AI model, we have concluded that the 0.48% accuracy gain was worth the effort since the AI algorithm seemed to improve the recognition of abbreviations and spaces between words. This second experimental phase allowed for a consolidation of information both from the perspective of palaeography and computer science: that is to better understand how the AI algorithm managed to differentiate and recognize on an improved level various details of palaeographical and linguistic nature: abbreviating solutions, words as linguistic units, double letter forms (“s”-long, “s”-round, varieties of “a”), or letter similarities (“c” vs. “t” confusions). Such details are not always easy to determine even for the human reader at a medium level of expertise in this field. It can be concluded that an increased effort of traditional transcription may not bring a significant advance in accuracy of automatic reading but could lead to a sensible improvement of the transliteration quality. More experiments should be conducted to verify such hypotheses.

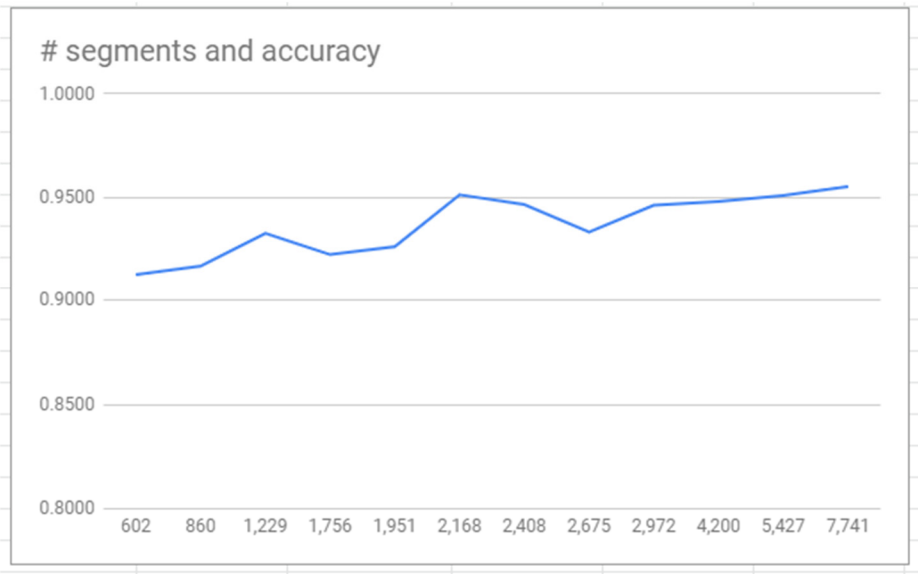
Even if it seemed at this point that the results were indicative of the Pareto principle (Newman 2005) applied to machine learning, the fact that the microscopic

accuracy gain had an abundance of meaning for the preciseness of the transcription, made the Zetta Cloud team decide to continue increasing the set of segments for the last RNN training phase.

# pages	# segments	accuracy	
8	602	0.9128	Experiment 8
11	860	0.9168	Experiment 7
16	1,229	0.9327	Experiment 6
22	1,756	0.9224	Experiment 5
25	1,951	0.9263	Experiment 4
28	**2,168**	**0.9512**	**Experiment 3**
31	2,408	0.9466	Experiment 2
34	2,675	0.9334	Experiment 1
38	**2,972**	**0.9462**	**Phase I**
53	4,200	0.9481	Experiment 9
68	**5,427**	**0.9509**	**Phase II**
100	**7,741**	**0.9553**	**Phase III**

To find the **effort inflection point**, the minimum number of manually transcribed pages (number of segments) needed to achieve sufficient transcription accuracy, we have operated several training and testing phases using sets of reduced number of segments, starting from the number of segments used by the Phase I - 2972 segments. Experiments were conducted with 4200 segments, as well. This is to plot the accuracy obtained by training the RNN AI algorithm with a number of segments between the number used for Phase I training and the Phase II training.

Here are all observations plotted on a simple graph:



Effort Inflection Point is at 2168 segments (28 pages)

Experiment 8, conducted by using just 602 segments (approximately 8 pages, manually transcribed) shows that 91.28% accuracy could have been achieved with a minimum effort. To conclude, the **effort inflection point** is at **2168 segments** (approximately **28 pages**), meaning that the algorithm could have been trained with **just 28 pages** of manual transcription to gain an accuracy value that is very close to what was achieved by manually transcribing **100 pages** and training the algorithm with them.

For the purpose of testing the transcription AI algorithm that was trained during the project (the result of Phase III – 100 manually transcribed pages for training), we've developed and deployed a very simple UI interface that one could use to upload one or more page images to obtain an automatically generated transcription. The interface is available here: https://catmeti.intellidockers.com/.

We have also put together a quick video tutorial that shows how to use the interface to obtain automatically generated transcription for your page images. The tutorial is available http://bit.ly/catmetipoc. (NOTE: You should use pages starting from 51r of the manuscript. These pages were never seen by the algorithm and they will show you how the AI computer vision algorithm is doing when put in front with completely new pages to be transcribed).

D. Conclusion

An innovative collaboration between two apparently opposing fields of study, palaeography and computer science, has led to a pioneering Romanian experiment aiming to demonstrate that by applying state of the art AI Computer Vision algorithms, text images can be automatically binarized and segmented for the systematic extraction of content from a sample set of medieval handwritten text pages. Beyond the theoretical challenges pertaining to the computational aspects of the endeavour, the Zetta Cloud computer scientists auspiciously tackled the larger scope of such an enterprise: to respond to the complex needs of the end-users, from experienced palaeographers interested in maximising the effort of transcribing a larger volume of handwritten text to the young enthusiasts observing the contents of a historical source that had been, so far, mediated by printed editions. However, making palaeography accessible to non-experts does not directly imply that the assessment of experienced scholars – perceived as academics existing in an ivory tower – is overrated and inconsequential, but rather that a highly restricted field of study can open up the scientific enquiry to young researchers, by facilitating access to the raw historical source and speeding up the formative stage of a specialised work force and academic career.

A handful of factors have contributed to the favourable outcome of this experiment: by carefully selecting the writing samples to be transcribed (i.e. Carolingian minuscule for continental manuscripts from the 9th/10th to the 12th/early 13th, as well as for the Humanistic calligraphic iterations from the 15th to 16th century) an immense chronological and geographical area of employment for this highly advanced palaeographical tool can be achieved. In this context, the next logical step

would be to apply the same workflow for obtaining AI computer vision models to other handwriting styles and complexities.

Previous and ongoing developments in the field of e-palaeography make the present project even more viable: potential collaborations and exchange of experience are currently heading towards an international community of practice, suitable for open dialogue and scientific progress. Leveraging crossover skills in today's academic environment means, more than any other time, pushing one's versatility and adaptability in the direction of new interdisciplinary projects and, last but not least, of prospective funding opportunities from both the public and the private sectors. Romanian memory institutions and their holdings may be the first to benefit by such computational advancement, while researchers could explore the historical material in new ways, from putting together editions of previously unpublished documents / account books / letter collections – to name just some of the potential archival records that would make suitable candidates for such an approach – to solving larger "puzzles" arising from the emerging field of "fragmentology" (as applied to the situation in Romania, where the study of medieval manuscript fragments has an explored potential, see Dincă 2011 and Dincă 2017). Thus, the long-term goal of this collaborative team is to develop a platform aiming to assist the experts in the transcription process, automating the tasks, speeding up the rendition of documents and improving its capabilities along the way. The automatic, Computer Vision based, handwriting recognition system should cooperate with the human expert to generate the final, highly qualitative text transcription.

This platform, with the implementation of a full Machine Learning Loop, would become smarter with every human expert intervention. Machine Learning Loop or Human-in-the-loop (HITL) is the process where the machine is unable to solve a problem with great accuracy initially starting with a small amount of annotated samples and it needs human intervention for creating a continuous feedback loop allowing the algorithm to give every time better results. Mainly HITL approach is used, when there is not much data available initially so that human experts can produce just enough annotated samples to obtain the best accuracy possible. The AI Computer Vision models produced using this approach can then be applied to the high volume of manuscripts to obtain the best accurate transcription. In this respect, our experiment shows that by initially annotating just 602 segments (approximately 8 pages, manually transcribed) we could obtain a model with 91.28% accuracy that is very near to the 95.53% accuracy that we obtained with training the same RNN neural network with 100 annotated paged. Our envisioned system would allow human experts to start with a very few annotated samples and reach a just enough accuracy level in a quick as guided manner through the HITL process.

Together with further AI technology like **Named Entities Extraction**, **Automatic Summarization** and **Forensic Author Identification**, the collaborative team will be able to do content extraction and various summaries which will allow any researcher to focus on the most relevant parts of the verified texts. Eventually, any sort of text in any documentary form will be available for further research purposes with a moderate transcription effort.

Acknowledgment

This work was supported by a grant from the Romanian National Authority for Scientific Research, CNDI–UEFISCDI, project PN-III-P4-ID-PCCF-2016-0064: "The Rise of an Intellectual Elite in Central Europe: Making Professors at the University of Vienna, 1389-1450" (https://rise-ubb.com/). We thank our two anonymous reviewers for their comments.

Works cited

Aiolli, Fabio & Ciula, Arianna. "A case study on the System for Paleographic Inspections (SPI): challenges and new developments". *Proceedings of the 2009 Conference on Computational Intelligence and Bioengineering: Essays in Memory of Antonina Starita*, IOS Press, 2009, pp. 53-66.

Aiolli, Fabio & Simi, Maria & Sona, Diego & Sperduti, Alessandro & Starita, Antonina & Zaccagnini, Gabriele. "SPI: A System for Paleographic Inspections". *AIIA Notizie*, vol. 4, 1999, pp. 34-48.

Aussems, Mark & Brink, Axel. "Digital palaeography". *Codicology and palaeography in the digital age 2*. Edited by Rehbein, Malte & Sahle, Patrick & Schassan, Torsten, BoD, 2009, pp. 293-308.

Bertrand, Paul. "La numérisation des actes: evolutions, révolutions. Vers une nouvelle forme d'édition de textes diplomatiques?". *Vom Nutzen des Edierens. Akten des internationalen Kongresses zum 150-jährigen Bestehen des Instituts für Österreichische Geschichtsforschung, Wien, 3.-5. Juni 2004*. Merta, Brigitte & Sommerlechner, Andrea & Weigl, Herwig Böhlau, 2005, pp. 171-176.

Bruckner, Albert. *Scriptoria medii aevi Helvetica: Denkmäler schweizerischer Schreibkunst des Mittelalters*, vol. VIII. *Schreibschulen der Diözese Konstanz, Stift Engelberg*, Genf, 1950.

Ciula, Arianna. "Digital palaeography: Using the digital representation of medieval script to support palaeographic analysis". *Digital Medievalist*, 1, 2005. www.digitalmedievalist.org/journal/1.1/ciula/. [20.06.2020].

Cloppet, Florence & Daher, Hani & Églin, Véronique & Emptoz, Hubert & Exbrayat, Matthieu & Joutel, Guillaume & Lebourgeois, Frank & Martin, Lionel & Moalla, Ikram & Siddiqi, Imran & Vincent, Nicole. "New Tools for Exploring, Analysing and Categorising Medieval Scripts". *Digital Medievalist*, vol. 7, 2011. journal.digitalmedievalist.org/articles/10.16995/dm.44/. [20.06.2020].

Derolez, Albert. *The palaeography of Gothic manuscript books from the twelfth to the early sixteenth century*. Cambridge University Press, 2003.

Dincă, Adinel C. "Datarea manuscriselor medievale latineşti. Evaluari metodologice", *Anuarul Institutului de Istorie «George Barițiu» din Cluj-Napoca, Series Historica*, tome L, 2011, pp. 295-306.

Dincă, Adinel C. "The Medieval Book in Early Modern Transylvania. Preliminary Assessments", *Studia UBB, Historia*, vol. 62, Issue 1, 2017, pp. 23-34.

Fischer, Andreas & Wüthrich, Markus & Liwicki, Marcus & Frinken, Volkmar & Bunke, Horst & Viehhauser, Gabriel & Stolz, Michael. *Automatic Transcription of Handwritten Medieval Documents*. Conference paper at Proc. 15th Int. Conf. on Virtual Systems and Multimedia (VSMM'09), 2009. DOI: 10.1109/VSMM.2009.26.

Granasztói, György. "Computerized Analysis of a Medieval Tax Roll. *Acta Historica Academiae Scientiarum Hungaricae*, vol. 17, no. 1/2, 1971, pp. 13-25.

Hassner, Tal & Rehbein, Malte & Stokes, Peter A. & Wolf, Lior. "Manifesto from Dagstuhl Perspectives Workshop 12382. Computation and Palaeography: Potentials and Limits". *Dagstuhl Manifestos*, vol. 2, issue 1, 2012, pp. 14-35, drops.dagstuhl.de/opus/volltexte/2013/4167/pdf/dagman-v002-i001-p014-12382.pdf. [20.06.2020].

Hassner, Tal & Sablatnig, Robert & Stutzmann, Dominique & Tarte, Ségolène. "Report from Dagstuhl Seminar 14302. Digital Palaeography: New Machines and Old Texts". *Dagstuhl Reports*, vol. 4, issue 7, 2014, pp. 112-134. www.researchgate.net/publication/269168418_Digital_Palaeography_New_Machines_and_Old_Texts_Dagstuhl_Seminar_14302 [20.06.2020].

Kestemont, Mike & Christlein, Vincent & Stutzmann, Dominique. "Artificial Paleography: Computational Approaches to Identifying Script Types in Medieval Manuscripts". *Speculum*, vol. 92 (S1), 2017, pp. S86-S109. DOI: 10.1086/694112.hal-01854939. [20.06.2020]

Kiessling, Benjamin. *Kraken – a Universal Text Recognizer for the Humanities*. Paper presented at Digital Humanities Conference 2019 (DH2019), Utrecht, the Netherlands. doi.org/10.34894/Z9G2EX, dev.clariah.nl/files/dh2019/boa/0673.html [20.06.2020]

Lehmann, Paul (ed.). *Ludwig Traube, Zur Paläographie und Handschriftenkunde*, Beck, 1909.

Muzerelle, Denis & Gurrado, Maria, (eds.), *Analyse d'image et paléographie systématique : travaux du programme "Graphem": communications présentées au colloque international "Paléographie fondamentale, paléographie expérimentale: l'écriture entre histoire et science" (Institut de recherche et d'histoire des textes (CNRS), Paris, 14-15 avril 2011)*. Association Gazette du livre médiéval, 2011.

Newman, M. E. J. "Power laws, Pareto distributions and Zipf's law". *Contemporary Physics*, vol. 46, no. 5, 2005. DOI: 10.1080/00107510500052444.arxiv.org/PS_cache/cond-mat/pdf/0412/0412004v3.pdf [20.06.2020]

Oriflamms. *Compte-rendu final du projet ORIFLAMMS / ORIFLAMMS Final report*. 2017. oriflamms.hypotheses.org/files/2017/04/Oriflamms-Compte-rendu-final.pdf. [20.06.2020].

Putnam, George F. "Soviet historians, quantitative methods, and digital computers", Computers *and the Humanities*, vol. 6, Issue 1, September 1971, pp. 23-29.

Schnapp, Jeffrey & Presner, Todd & Lunenfeld, Peter & Drucker, Johanna. *Digital Humanities Manifesto 2.0*, jeffreyschnapp.com/wp-content/uploads/2011/10/Manifesto_V2.pdf. [20.06.2020]

Słoń, Marek. “Pryncypia edytorstwa źródeł historycznych w dobie rewolucji cyfrowej [Principles of Editing Historical Sources at the Time of the Digital Revolution]”, *Studia Źródłoznawcze/Studies in Historical Sources*, vol. LIII, 2015, pp. 155-161.

Stokes, Peter A. & Kiessling, Benjamin & Tissot, Robin & Stökl Ben Ezra, Daniel. *EScripta: A New Digital Platform for the Study of Historical Texts and Writing*, paper presented at Digital Humanities Conference 2019 (DH2019), Utrecht, the Netherlands. hal-02310781. dev.clariah.nl/files/dh2019/boa/0322.html [20.06.2020].

Stokes, Peter A. “Computer-Aided Palaeography, Present and Future”. *Kodikologie und Paläographie im digitalen Zeitalter / Codicology and Palaeography in the Digital Age*, BoD, 2009, pp. 309-338, kups.ub.uni-koeln.de/2978/. [20.06.2020].

Stokes, Peter A. “Computer-Aided Palaeography, Present and Future”. *Kodikologie und Paläographie im digitalen Zeitalter / Codicology and Palaeography in the Digital Age.* Edited by Rehbein, Malte & Sahle, Patrick & Schaßan, Torsten, BoD, 2009, pp. 309-38.

Stokes, Peter A. “Digital Resource and Database for Palaeography, Manuscripts and Diplomatic”. *Gazette du livre médiéval*, vol. 56-57, 2011, pp. 141-142; www.persee.fr/doc/galim_0753-5015_2011_num_56_1_1991. [20.06.2020].

Stokes, Peter A., “Palaeography and Image-Processing: Some Solutions and Problems”. *Digital Medievalist*, vol. 3, 2007. http://doi.org/10.16995/dm.15. [20.06.2020].

Stutzmann, Dominique & Kermorvant, Christopher & Vidal, Enrique & Chanda, Sukalpa & Hamel, Sébastien & Puigcerver Pérez, Joan & Schomaker, Lambert & Toselli, Alejandro H. “Handwritten Text Recognition, Keyword Indexing, And Plain Text Search In Medieval Manuscripts”. Conference paper at *Digital Humanities 2018 Puentes-Bridges. Book of Abstracts*, p. 298-302. dh2018.adho.org/wp-content/uploads/2018/06/dh2018_abstracts.pdf. [20.06.2020].

Stutzmann, Dominique. “Clustering of medieval scripts through computer image analysis: Towards an evaluation protocol”. *Digital Medievalist*, vol. 10, 2016. DOI: http://doi.org/10.16995/dm.61. [20.06.2020]

Wakelin, Daniel. “«An anthology of images»: DIY digital photography in manuscript studies”, *DIY Digitization*. diydigitization.org/contributed-papers/wakelin/ [20.06.2020]

Widner, Michael. “Toward Text-Mining the Middle Ages: Digital Scriptoria and Networks of Labor”. *The Routledge research companion to digital medieval literature*. Edited by Boyle, Jennifer & Burgess, Helen J., Routledge, 2017, pp. 131-144.