heritage

MDPI

Article

Experimenting with Training a Neural Network in Transkribus to Recognise Text in a Multilingual and Multi-Authored Manuscript Collection

Carlotta Capurro [1,*], Vera Provatorova [2] and Evangelos Kanoulas [2]

[1] Department of History and Art History, Utrecht University, Drift 15, 3512 BR Utrecht, The Netherlands
[2] Informatics Institute, University of Amsterdam, Science Park 900, 1012 WX Amsterdam, The Netherlands; v.provatorova@uva.nl (V.P.); e.kanoulas@uva.nl (E.K.)
* Correspondence: c.capurro@uu.nl

Abstract: This work aims at developing an optimal strategy to automatically transcribe a large quantity of uncategorised, digitised archival documents when resources include handwritten text by multiple authors and in several languages. We present a comparative study to establish the efficiency of a single multilingual handwritten text recognition (HTR) model trained on multiple handwriting styles instead of using a separate model for every language. When successful, this approach allows us to automate the transcription of the archive, reducing manual annotation efforts and facilitating information retrieval. To train the model, we used the material from the personal archive of the Dutch glass artist Sybren Valkema (1916–1996), processing it with Transkribus.

Keywords: handwritten text recognition (HTR); neural networks; digital archives; digital cultural heritage; Transkribus

check for updates

Citation: Capurro, C.; Provatorova, V.; Kanoulas, E. Experimenting with Training a Neural Network in Transkribus to Recognise Text in a Multilingual and Multi-Authored Manuscript Collection. *Heritage* **2023**, *6*, 7482–7494. https://doi.org/10.3390/heritage6120392

Academic Editor: Andreas Aristidou

Received: 4 September 2023
Revised: 10 November 2023
Accepted: 20 November 2023
Published: 29 November 2023

1. Introduction

The work described in this contribution aims at developing an optimal strategy to automatically transcribe a large quantity of uncategorised, digitised archival documents when resources include handwritten text by multiple authors and in several languages. This research question has emerged following the digitisation of the personal archive of the Dutch glass artist Sybren Valkema (1916–1996). The archive covers the entire lifetime of the artist, including notebooks from his childhood, personal annotations on his work and artistic practice, and correspondence documents drafted by Valkema or received from his colleagues and international acquaintances. In order to make the digitised resources searchable, we have developed a methodology that allows us to comparatively evaluate the efficiency of several handwritten text recognition (HTR) models using Transkribus, a platform enabling AI-powered text recognition [1].

The digitisation of Valkema's personal archive represents an outstanding new source of data for the study of modern glass production in Europe. Sybren Valkema was one of the founders of the Studio Glass Movement's European branch, known in the Netherlands as Vrij Glas [2]. Studio Glass spread internationally from the US in the early 1960s and represented a revolution in the world of glass art, marking a shift from the traditional method of glass production in factories to a more individualised approach that allowed artists to create unique pieces. The Movement was pioneered by the American artist Harvey Littleton (1922–20) and the glass research scientist Dominick Labino (1910–1987), who introduced the first glassblowing workshop at the Toledo Museum of Art in 1962. This workshop approach was made possible by the innovative studio furnace they prototyped. Such a furnace allowed artists to become independent from industrial facilities and experiment with the form, colour, and texture of glass in their workshops [3].

One of the Studio Glass Movement's main strengths was the high interconnection of its members. Contrary to the secrecy surrounding glass production in traditional centres such as Venice, Studio Glass artists were carrying out research on new recipes for producing glass and exchanging their findings with their colleagues and students. In promoting this open practice, the Studio Glass Movement also significantly impacted art education and academia. Many of the Movement's pioneers were educators who believed in the importance of teaching glassblowing as a form of art to stimulate the free circulation of knowledge about glass and glassmaking techniques [4]. Therefore, they established several courses on glassblowing in universities and art academies in the US, Europe, and worldwide.

During his life, Sybren Valkema combined glassblowing with an intense career as a teacher. Employed as an aesthetic design teacher by the Glass School of the Royal Leerdam glass factory in 1943, Valkema then became a drawing instructor at the Dutch Institute for Art and Crafts (Instituut voor Kunstnijverheidsonderwijs), later renamed Gerrit Rietveld Academie. During the 1950s, he collaborated with the glass masters working in the Leerdam glass factory, producing his first designs for glassware and becoming interested in glass as a means of creative expression. When he attended the inaugural meeting of the World Crafts Council in New York City in 1964, Valkema was particularly impressed by the presentation of the portable oven developed by Labino and Littleton and established a close friendship with many of the exponents of the American Studio Glass Movement. Once back in the Netherlands, he built the first studio glass furnace in Europe at the Rietveld Academy. In 1969, he inaugurated a new teaching curriculum around glassmaking, organised as a workgroup where people from the different departments of the Rietveld Academy could learn to work with glass [5].

After Valkema's passing, his family donated his archive to the Netherlands Institute for Art History (RKD) [6]. The collection is a valuable resource for studying the development of contemporary glassmaking and the Studio Glass movement both in the Netherlands and worldwide. It contains over 103,000 pages documenting Valkema's career, including teaching materials, letters, designs, sketches, descriptions of processes, and many glass recipes. Between 2013 and 2015, the archive was digitised in collaboration with the Rakow Research Library of the Corning Museum of Glass [7]. In 2018, the Art DATIS project was inaugurated with the aim of making the Sybren Valkema archive publicly accessible and easily searchable [8].

2. Handwritten Text Recognition on the Archival Material

The archive contains documents that have been typewritten and pages of handwritten text. In order to extract the text from images and make content searchable, images need to be converted into machine-encoded text [9]. This process, called Text Recognition, requires different tools and procedures depending on the type of document to be processed. This contribution focuses on how we have dealt with the problem of automatically transcribing handwritten documents.

In computer vision, the process of recognising and interpreting handwritten text by a machine and transforming it into machine-encoded text is called handwritten text recognition (HTR). Due to the complexity of dealing with handwriting, which varies significantly from author to author, HTR has developed as an autonomous research area [10]. To support our work, we used the platform Transkribus, developed in the framework of the recognition and enrichment of archival documents (READ) European Union Horizon 2020 project [11]. The platform, designed to support researchers dealing with historical manuscripts, is based on open technology and the scientific community's contribution [12].

Transkribus is based on machine learning principles and uses artificial neural networks to support researchers in training models capable of reading and automatically transcribing their corpora. To train an efficient model, researchers need to manually transcribe a number of documents that serve as examples for the neural network to learn to identify the specific writing sample. These manually transcribed documents form an objective basis of information called ground truth (GT). Transkribus usually requires between 5000 and 15,000 words (around 25–75 pages) of transcribed material to start training a model [13]. Once adequately trained, the model can be applied to a larger collection of similar documents to automate their transcription successfuly.

However, performing HTR on the Valkema archive is particularly complex due to two main issues. First, the archive includes not only documents written by Valkema at any stage of his life, therefore presenting a variation in his handwriting, but also documents written by other people and in several languages. Current HTR methods require extensive manual annotation efforts to create training data: every language and every handwriting have unique features and ideally require the use of a separate model. Trankribus recommends using at least 10,000 words for each distinct writer in order to train an efficient model [13]. Second, digital resources in the RKD digital archive are largely undocumented. Due to the lack of metadata about the language and author, it is impossible to automatically sort out the material and process it with a specific model. In theory, each page should be processed with the appropriate language-specific model. However, this approach is not feasible, requiring an enormous amount of human labour to select the best model to transcribe the page manually.

To overcome these challenges, we needed to train a model capable of interpreting all the documents in the archive regardless of their language and author, providing an adequate automatic transcription[1]. Therefore, we aimed to develop a multilingual model whose performance could be comparable to that of specific monolingual models. In pursuing this objective, we were conscious of the tradeoff between low resources and high accuracy involved in the process. For this reason, we aimed to develop a model with the lowest possible error rate to make the documents searchable for further research without aiming at obtaining zero errors.

Therefore, we created a multilingual, multi-author model trained on sample documents from the Valkema archive. We identified Dutch (NL), English (EN), and German (DE) as the most frequently used languages in handwritten documents. Therefore, the main requirement of the multilingual model we trained was to be flexible enough to generate an acceptable level of character error rate (CER) in transcribing these three languages. We use CER as the evaluation metric to ensure a fair comparison with the models already available on Transkribus, which report CER (but not word error rate) on training and validation datasets.

The study we present in this contribution aims to answer the following research questions:

1. Can a multilingual HTR model be a viable alternative to language-specific HTR models for automatically transcribing a set of handwritten documents in multiple languages? We consider the multilingual model to be usable if the performance drop (an increase in character error rate) compared to using a language-specific model is not larger than the threshold value of 10%. This threshold value is chosen arbitrarily since, to the best of our knowledge, no guidelines on choosing it are available from prior research.
2. Does applying OCR postcorrection to the output of a multilingual HTR model improve its performance? At the time of the experiment, no HTR postcorrection software was publicly available, with research on this topic being scarce [15,16]. As implementing an HTR postcorrection algorithm from scratch is out of the scope of this study, OCR postcorrection was used instead. We applied automatic postcorrection to the output generated using the multilingual, multi-author model and evaluated the results (with and without postcorrection) on a separate subset of manually annotated documents sampled from the same archive. A comparison of our empirical results shows that OCR postcorrection is not applicable to HTR models.

Due to the rapid digitisation of archives, many projects have investigated the best practices to develop HTR models capable of automatically transcribing large collections of historical manuscripts (such as [17–22]). Most of these models prove to be very performative, having been trained on large corpora of coherent documents. The innovative aspect of our work, therefore, is not the process of model creation per se, which will be anyhow addressed in Section 4 of this article, but the attempt to develop an agile model capable of automating transcribing an incoherent corpus of handwritten documents, and the methodology we developed to test its quality.

The following part of this contribution describes the methodology we developed to assess the quality of our multilingual and multi-author model compared to using a separate model for each language.

3. Methodology

In order to test the quality of the multilingual, multi-author model that we trained using the pages from the Valkema archive, we developed a four-step experiment using the most common languages in the archive: Dutch, English, and German.

First, for each of the three languages, we manually selected and annotated a set of 50 documents. These pages, all written between the 1940s and 1980s, were carefully sampled in the archive in order to offer a representative variety of handwriting styles and layouts. This process could not be automated since Transkribus does not offer any language detection support, nor was the language of the files previously documented. The manual annotation of these 150 documents, the details of which are listed in Table 1, represented the GT on which we calculated the amount of error of each of the trained models.

Table 1. Composition of the ground truth (GT) test set used to evaluate the performance of the trained models.

Test Documents (GT)	NL	50 pages (letters, postcards, notes) 42 handwriting styles	Nr. of lines transcribed: 1204 Nr. of words transcribed: 6933
	EN	50 pages (letters, postcards) 46 handwriting styles	Nr. of lines transcribed: 1229 Nr. of words transcribed: 6440
	DE	50 pages (letters, postcards) 37 handwriting styles	Nr. of lines transcribed: 954 Nr. of words transcribed: 4551
Total		**150 pages** **121 handwriting styles ***	**Nr. of lines transcribed: 3387** **Nr. of words transcribed: 17,924**

* The difference between the number of handwriting styles and the total number of pages manually transcribed is explained by the presence of multiple documents from the same authors. In particular, we selected more samples of documents handwritten by Sybren Valkema in all three languages.

Second, every document was automatically transcribed using two different models:

- a monolingual model specific to the language of the document, which is expected to offer high-quality results;
- a multilingual model, which is expected to provide lower quality yet still usable results in all the languages.

Third, the transcriptions produced using the multilingual model were exported from Transkribus and processed with an OCR postcorrection algorithm to fix the transcription mistakes. Last, the accuracy of the results of the multilingual model (with and without OCR postcorrection) is compared with the automatic transcription obtained using the corresponding monolingual models. Figure 1 shows an overview of the experiment.

The results of this experiment allow us to devise an agile strategy to automate the datafication of the archival collection, reducing the amount of human labour involved in the process. Furthermore, a good multilingual model has several critical advantages for cultural heritage institutions. First, it can be re-used on similar collections, sensibly reducing the investment of resources in training new ad-hoc models. These resources include not only the labour time of personnel manually transcribing pages to feed the AI engine but also the environmental costs behind the training [23]. As the study by Strubell et al. demonstrated, training a single deep-learning natural language process model can produce approximately 0.3 tonnes of carbon dioxide [24]. Therefore, the environmental impact of producing new ad-hoc models for each documentary collection is hefty in the long run. Second, when automatic HTR quality is accurate, archival resources can be documented using entities already in the text [25]. This approach substantially reduces the interpretative efforts of heritage documentation, thus granting more equitable access to resources [26].

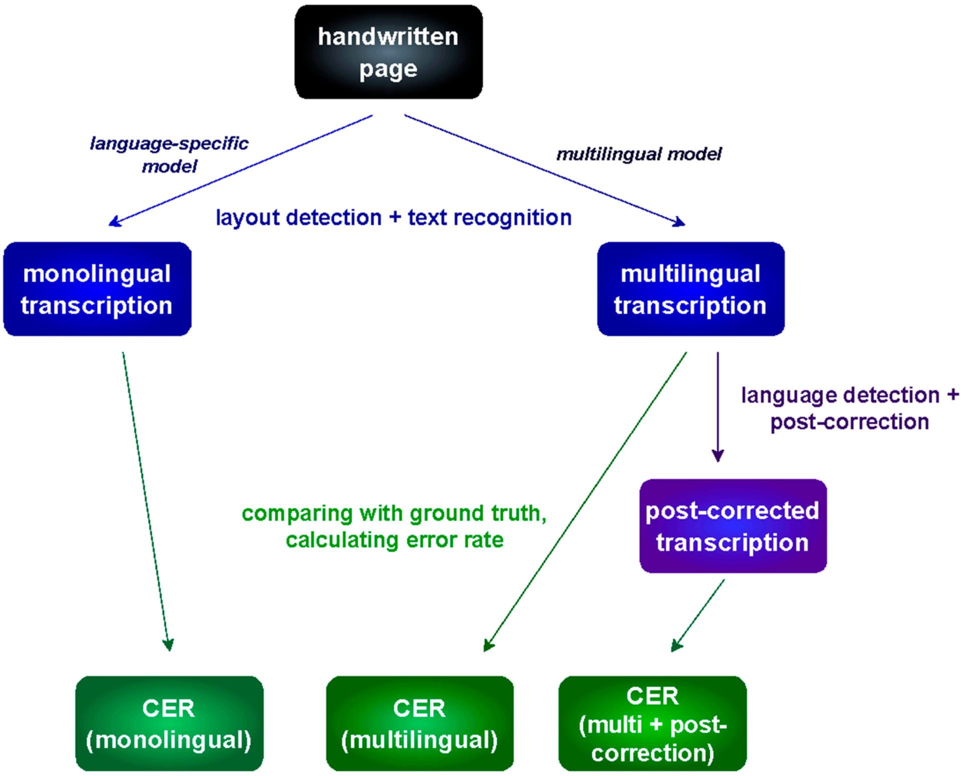
handwritten page
language-specific model
multilingual model
layout detection + text recognition
monolingual transcription
multilingual transcription
language detection + post-correction
comparing with ground truth, calculating error rate
post-corrected transcription
CER (monolingual)
CER (multilingual)
CER (multi + post-correction)

Figure 1. The pipeline of the experiment.

4. Training the HTR Models

Before carrying out the experiment described above, it was necessary to train the models based on the main languages included in the archive. We identified Dutch, English, and German as the most recurring languages in handwritten documents. Therefore, the main requirement of the multilingual model we trained was to be flexible enough to generate an acceptable level of CER in transcribing any of these three languages.

For each of the three languages, Transkribus offers public models trained on documents provided by its community of users. We tested them on the 150 documents we used as the ground truth for our experiment to verify their usability for our scopes. Table 2 shows that the CER of each of these models calculated on our GT drastically increased in all the languages, proving inadequate for our scopes. Therefore, we needed to train our own models. This process included selecting and manually transcribing an adequate number of documents in each language. The Transkribus platform offers the environment for easily producing manual transcriptions of digitised documents. After importing the image files, each document can be automatically analysed, and the page layout is detected. This means that each line of text is automatically identified in the image. The transcriber then has the capability to correct any possible mistake in the detected layout and manually type in the text as it appears in each line. In creating a dataset for training a model, it is crucial to pay attention to correctly reporting capital letters, signs, spaces between words and letters, and typos. The neural network will then use this information to make sense of the signs detected on the page and learn to identify them in other documents.

Table 2. Performance of monolingual models available on Transkribus calculated for our collection. Training and validation datasets here were selected by the Transkribus team, and the test datasets were extracted from our collection. High variance is explained by the diverse handwriting styles and layouts.

Language	Transkribus Model	Performance of the Model Calculated by Transkribus	Performance of the Model Calculated on Our Collection
EN	Transkribus English Handwriting M3 (ID 37646)[2]	CER on training: 6.5% CER on validation: 5%	CER on the test: 18.7% (STD 13.2%)
NL	Transkribus Dutch Handwriting M2 (ID 45422)[3]	CER on training: 6.1% CER on validation: 4.9%	CER on the test: 21.7% (STD 10.7%)
DE	Transkribus German handwriting M1 (ID 35909)[4]	CER on training: 6.2% CER on validation: 4.7%	CER on the test: 24.6% (STD 18%)

In total, we manually transcribed 344 documents, of which 244 were NL (six had some uncertainties), 87 EN, and seven DE documents. These documents were used as the training set and did not include any of the 150 pages forming the test set illustrated in Table 1. The disparity in these numbers testifies to the linguistic composition of the archive. As a personal archive of a Dutch artist, many of the documents drafted by Valkema (including notes, drafts, and articles) are in Dutch. The international correspondence was mainly in English or German, with a small number of documents in other languages such as French and Swedish. Still, these documents form only a small portion of the archive. The documents used to create our training set were randomly selected from the archive, with attention to include in the sample a diverse number of handwriting styles and document types (e.g., letters, postcards, drafts, and notes, all of which display different and irregular text layouts). When our starting corpus was ready, we proceeded with training the models.

When we started working on our corpus, we planned to compare the performance of the two AI engines available on Transkribus: HTR+ and PyLaia. The former was especially recommended for heterogeneous documents since it was deemed more capable of recognising irregular layouts [27]. Unfortunately, Transkribus decided to discontinue it, leaving us with the only option of using PyLaia to train our models [28]. Table 3 illustrates the descriptions, training details, and evaluation results of the models we created during this experiment.

First, we trained two monolingual models, ArtDATIS_English_latest and ArtDATIS_Dutch_latest, using the respective monolingual models listed in Table 2 as base models. No monolingual model was trained for German since Transkribus recommends using at least 20 pages of training data, and we only had seven pages available. The evaluation results in Table 3 show considerable performance improvements compared to using the existing models without fine-tuning them on our data: 11.1% CER compared to 18.7% for English and 18.9% CER compared to 21.7% for Dutch.

Next, we compared two approaches for training a multilingual model. The first model, ArtDATIS_Multilanguage_V2, was trained from scratch on all the available training data. For the second model, ArtDATIS_Dutch_English, we excluded all non-ground truth pages from the training set (that is, all the pages with uncertainties in the transcription) and used Transkribus Dutch Handwriting M2 as the base model. The reasoning behind choosing a Dutch base model is to optimise the resulting model for our archive; since a vast majority of the documents are in Dutch, the overall performance is expected to be improved. Evaluation results show that ArtDATIS_Multilanguage_V2 is likely overfitted to the training data (since its error rate on the validation set is much higher than the error rate on the train set), while ArtDATIS_Dutch_English appears to be free from this problem.

Table 3. Overview of the HTR models trained during the experiment.

Language	Training Details	Title & Transkribus ID	Performance during Training (Average CER)	Performance of the Model Calculated on Our Collection (Average CER)
EN	Base model: Transkribus English Handwriting M3 (ID 37646) Training data: 87 pages (10% used for validation)	ArtDATIS_English_latest (53449)	CER on training: 4.7% CER on validation: 2.1%	CER on the test: 11.1% (STD 12.8%)
NL	Base model: Transkribus Dutch Handwriting M2 (ID 45422) Training data: 215 pages (10% used for validation)	ArtDATIS_Dutch_latest (53462)	CER on training: 9.8% CER on validation: 6.1%	CER on the test: 18.9% (STD 9.5%)
multi	No base model training data: 338 pages (14 used for validation) DE—7 pages EN—87 pages NL—244 pages	ArtDATIS_Multilanguage_V2 (ID: 48828)	CER on training: 1.9% CER on validation: 12.8%	CER on the test: 22.8% on average for all languages (STD 15.1%) 13.8% EN (STD 14.1%) 33.6% DE (STD 14.1%) 21% NL (STD 9.5%)
multi	Base model: Transkribus Dutch Handwriting M2 (ID 45422 Training data: EN—87 pages NL—215 pages (10% in each language is used for validation)	ArtDATIS_Dutch_English (53444)	CER on training: 6.6% CER on validation: 3.2%	CER on the test: 20.1% all languages (STD 14.8%) 12.5% EN (STD 13.1%) 30.7% DE (STD 14.9%) 17.1% NL (STD 9.5%)

All models were trained for 100 epochs with early stopping possible after 20 epochs, using default parameters provided by Transkribus (learning rate 0.0003, batch size 24, normalised height 64).

5. Testing the Models

Once we created the models, we used the documents in our GT to test their performances. The documents in EN, NL, and DE were first transcribed using their language-specific HTR model (respectively, ArtDATIS_English_latest, ArtDATIS_Dutch_latest, and Transkribus German handwriting M1) and with the multilingual model ArtDATIS_Dutch_English. In Table 3, the column ‘Performance of the model calculated on our collection’ illustrates the evaluation results. First, results confirm that using a Dutch base model improves the general performances of our multilingual model: the average CER on all documents in the test set is 20.1% for ArtDATIS_Dutch_English compared to 22.8% for ArtDATIS_Multilanguage_V2, with unsurprisingly better performance on Dutch (17.1% vs. 21%) and slight improvements in English (12.5% vs. 13.8%). An unexpected observation is that performance on the German documents has also improved, even though the German pages were excluded from the training data: 30.7% CER compared to 33.6%. We hypothesise that the parametric knowledge of the large Dutch base model allows us to achieve improvements on the other two languages for two reasons: firstly, both German and English share linguistic similarities with Dutch, and secondly, a wide variety of handwritings and layouts seen by the base model during training allows it to better generalise to noisy data regardless of the language. Since ArtDATIS_Dutch_English performs better than ArtDATIS_Multilanguage_V2 in all languages, it has been chosen as the multilingual model for further evaluation and postcorrection experiments.

Since the aim of the experiment was to test the quality of the multilingual model to use it for streamlining the transcription process of the entire collection, we decided to closely follow and evaluate Transkribus behaviour on our test sets, breaking down the text recognition process, and documenting any potential pitfalls.

First, we tested the layout detection on our GT dataset. To automatically detect the layout of our documents, we used the Transkribus Layout Analysis tool, preserving the preset parameters defined by Read Coop (Transkribus LA 0.0.5). Such an experiment—performed on a relatively small dataset but representative enough of the whole collection—offered an idea of the quantity and quality of errors or missing information that a fully automated document transcription process might generate. At the same time, it also offers an idea of the amount of manual work required to fine-tune the results of the automatised process, which is essential information when dealing with the datafication of large document collections.

After performing the layout analysis on the Dutch-language collection, nine out of 50 documents (almost one out of five) had no layout detected. It was, therefore, necessary to run the layout analysis on them a second time. In the English-language collection, the layout consistently failed to be recognised in one document. In the German-language collection, the layout failed to be properly detected on eight documents. After being processed a second time, the layout of three documents still required some adjustments, and one document consistently failed to be recognised. Overall, we identified some recurring errors in the process, especially in detecting text areas oriented differently than the main document. This issue could be especially problematic in documents, such as postcards or annotated texts, where the layout is not regular.

Second, we tested the quality of the layout analysis embedded in Transkribus' text recognition tool and proceeded to transcribe the documents using the monolingual models we trained. The tool we first used was the PyLaia decoding 0.9.1. When processing the Dutch-language collection, it emerged that 16 documents had the layout wrongly detected, resulting in major transcription mistakes. In particular, two errors occurred: the layout analysis detected two or more text areas, some of which were interpreted and transcribed as if the orientation of the text was upside down. This represented 15 of the cases, six of which with minor errors. Second, the text written in the opposite orientation was not detected with the layout analysis, remaining untranscribed. The English-language collection presented the same issues, with 22 documents that had to be reprocessed. Concerning the German-language collection, nine documents presented issues with layout detection and text orientation. Among them, one document presented important issues with the text area consistently failing to be detected. Table 4 offers an overview of the issues detected. Once reprocessed one-by-one and with the updated version of the tool (PyLaia decoding 1.2.0), many issues with text orientation and transcription were solved. It is important to note that the difference in performance between the Transkribus Layout Analysis tool and the layout detection tool embedded in the Text Recognition tool can be explained with the support of the language model that is present in the second tool. This results in better performances, especially in self-trained models (Transkribus Support 2023). The transcriptions resulting from this phase were not manually corrected and were exported as the first subset of data in our experiment and used to evaluate the performance of the monolingual models.

Thirdly, we ran the ArtDATIS_Dutch_English multilingual model on the three datasets. The resulting transcriptions were then exported without any manual adjustment, constituting our experiment's second subset of data to evaluate the model's performance.

Lastly, to find out whether the output of the multilingual model could be improved with OCR postcorrection, we ran an additional experiment: for every document in the collection, we automatically detected its language using the LangDetect library [29] and then applied a language-specific, state-of-the-art OCR postcorrection model [30].

Table 4. Layout analysis. Overview of the issues detected.

Documents' Language	Errors in Layout Detection (Document Nr.)
NL	P. 18, 19, 21, 23, 33 (minor), 35, 36 (minor), 38 (minor), 41, 42, 43 (minor), 44 (minor), 48, 50 (minor)—Several text regions detected, and wrong interpretation of text orientation. P. 30—text in different orientations is not detected. P. 46—second text region detected, but no text in it.
EN	P. 4, 12 (minor), 16–17, 19 (minor), 21 (minor)–25, 27, 30, 33–34, 38 (minor), 40–43, 45, 47–50—Several text regions detected, and wrong interpretation of text orientation.
DE	P. 4, 7, 10—text in different orientations is not detected. P. 20—text region detected, but wrong interpretation of text orientation. P. 21—text in different orientations is not detected, text region detected, but wrong interpretation of text orientation. P. 26, 42—a large portion of text not detected. P. 41—text lines partially detected. P. 25 (postcard)—failed layout analysis. The text areas had to be manually added.

6. Results and Discussion

Following the evaluation practice of Transkribus, we used the character error rate (CER) as a metric for comparing our models in different setups. The metric was calculated in a Python script using the FastWER library by comparing the output of the models with the ground truth and calculating the character error rate [31].

Figure 2 shows the detailed results of our experiment. The outlier pages include the most unusual handwritings (with the letters tilted to the left) and the most challenging layouts (for example, a postcard with typed text in multiple languages alongside the actual handwritten content). The high variance in all models is explained by the high diversity of the material; there are handwriting variations even for the pages written by a single author, as well as diverse layouts of the documents.

The figure also shows that postcorrection does not lead to any improvements. On the contrary, it brings a slight increase in character error rate. This result is not surprising, given that HTR mistakes and OCR mistakes have a different nature, and attempting to perform OCR postcorrection on HTR data ends up introducing extra noise.

Considering the performance of the different models, two observations can be made. Firstly, fine-tuning a monolingual base model improves the results both for Dutch and for English, which means that adapting to the writing styles present in a specific collection is important. Secondly, using a large Dutch HTR model as the base for fine-tuning a multilingual model leads to performance improvements not only on Dutch but also on English texts, with a noticeable yet acceptable increase in error rate on German. Table 5 presents the details of the performance tradeoff when using the multilingual model on in our collection. Table 6 shows the results of statistical significance testing: we report p-values obtained by the pairwise Mann–Whitney–Wilcoxon test. With Bonferroni correction for six hypotheses at $\alpha = 0.01$, a result is statistically significant when $p < \alpha/6 = 0.0017$.

These results allow us to draw some conclusions about the research questions outlined in the introduction of this paper. First, the experiments in this study show that a multilingual model is indeed usable for automatically transcribing our multilingual collection: it works better than the existing language-specific models for the two most common languages, Dutch and English, while achieving acceptable performance on German. Second, the evaluation results confirm the assumption that automatic OCR postcorrection is not useful for HTR data, with the most likely reason being the different visual features of typed and handwritten texts. As no software specifically designed for HTR postcorrection was publicly available at the time of the experiment, we conclude that improving the output of

HTR models without developing new tools or applying extensive manual effort could only be achieved by optimising the training process.

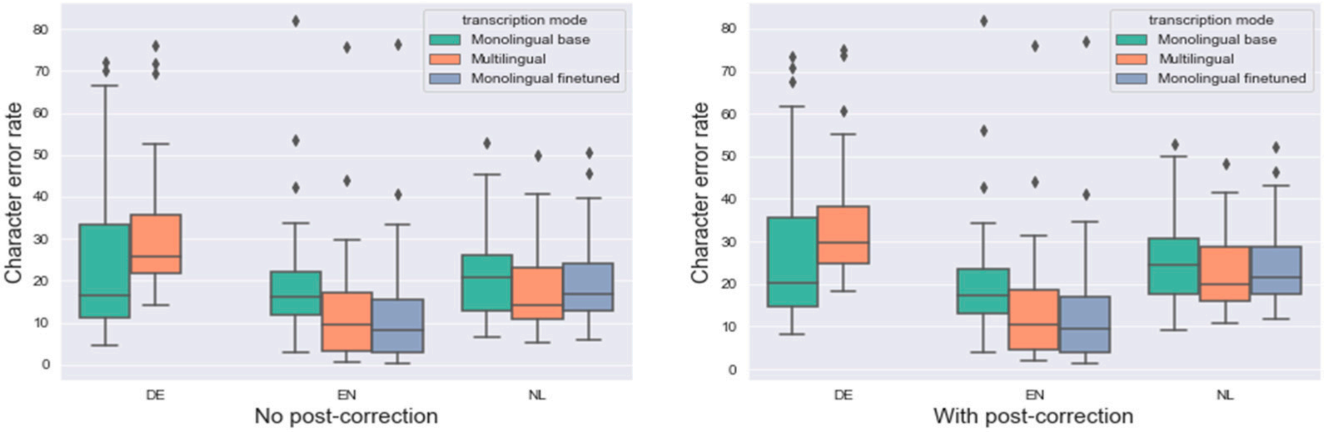

Figure 2. Results of the experiment: character error rate distribution for every setup. In the graph, the diamonds represent the pages with unusual handwritings or challenging layouts.

Table 5. Performance tradeoff of the multilingual model.

Language	Average Performance Difference, % CER	
	Compared to a Monolingual Base Model	Compared to Monolingual Fine-Tuned Model
NL	1.6% improvement (17.1% vs. 18.7%)	6% drop (17.1% vs. 11.1%)
EN	9.2% improvement (12.5% vs. 21.7%)	5.4% improvement (12.5% vs. 18.9%)
DE	6.1% drop (30.7% vs. 24.6%)	-

Table 6. Statistical significance testing results.

Hypothesis	*p*-Values Per Language		
	NL	EN	DE
Fine-tuning a monolingual model significantly improves its performance	0.0004	0.0000	–
Our multilingual model performs significantly better than a base monolingual model	0.0000	0.0000	1
Our multilingual model performs significantly worse than a fine-tuned monolingual model	1	0.0001	–

7. Conclusions and Future Work

We have trained and evaluated two monolingual and two multilingual HTR models, with the goal of determining whether a multilingual model can be a viable alternative to language-specific models in a limited resources setting when aiming at automating the transcription of a multilingual and multi-authored collection of documents. Our experiments show that a multilingual model based on a large Dutch model achieves good results on Dutch and English and an acceptable performance drop on German, making it suitable for transcribing the archive of Sybren Valkema (where Dutch is the most common language, followed by English and a smaller amount of German).

The next step of our project includes adding more German-language data to the training set of the multilingual model and analysing the results to see whether it improves HTR performance on the German text set without losing too much accuracy on English and

Dutch documents. Moreover, the results of our experiment inspire several extra questions outside of the scope of the current project. These questions are the following:

- Is language family an important feature for HTR models? I.e., would the multilingual model from our experiment perform better on another Germanic language (such as Swedish) than on a language from a different group (such as French)? Since a small number of documents in these two languages have been identified in the archive, we will be able to partially assess this matter after proceeding with the automatic transcription of the files in the archive.
- Do different handwriting styles influence HTR performance more than different languages? For example, if model A is trained on a single-authored collection of documents in English and Dutch, and model B is trained on a multi-authored collection of documents in English, which of the models performs better on a collection that contains both languages and is written by multiple authors?

In the case of the second question, we expect the writing style to be an important feature, following the findings in [32], a study on multilingual OCR, where using different fonts in training data turned out to be more beneficial than using different languages (Finnish and Swedish). However, more research is needed to find out whether this generalises to handwritten texts, as well as to other languages.

Another direction for future work involves experimenting with HTR-specific postcorrection when new tools become available since the most recent developments in the field reveal promising new directions [33].

We hope that our contribution inspires further research in the field of multilingual handwritten text recognition.

Author Contributions: Conceptualization, C.C.; methodology, C.C., V.P. and E.K.; software, V.P., READ COOP; validation, C.C. and V.P.; formal analysis, C.C. and V.P.; investigation, C.C. and V.P.; resources, RKD, Vrij Glass foundation; data curation, C.C.; writing—original draft preparation, C.C. and V.P.; writing—review and editing, C.C. and V.P.; visualization, C.C. and V.P.; supervision, E.K.; project administration, Art Datis; funding acquisition, Art Datis. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the NWO Innovational Research Incentives Scheme Vidi, (016.Vidi.189.039), grant number NWO Smart Culture—Big Data/Digital Humanities, 314-99-301. All content represents the opinion of the authors, which is not necessarily shared or endorsed by their respective employers and/or sponsors.

Data Availability Statement: Restrictions apply to the availability of these data. Data was obtained from RKD and are available at https://rkd.nl/nl/explore/collections/246 (accessed on 1 September 2023) with the permission of RKD.

Acknowledgments: We want to thank Sven Dupré and Marieke Hendriksen for providing feedback on previous versions of this article.

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

Notes

[1] In July 2023, after the completion of our work, READ Coop released a new generation of 'supermodels' working on two features: 'an optical part that processes the images and an extensive language model that tries to make sense of and improve the extracted text information' [14]. These supermodels are able to deal with more than a single language (both old and new forms of the languages). These models, which promise to improve the quality of document transcriptions radically, are currently neither fine-tunable nor trainable by users. Therefore, they cannot be used as base models to improve the performance of the models trained during the Art Datis project.

[2] More information on the materials used to train the model: https://readcoop.eu/model/english-handwriting-18th-19th-century-2/ (accessed on 1 September 2023).

[3] The general model for Dutch handwriting curated by the Transkribus team, has been trained on 4,626,201 words and 898,252 lines.

[4] More information on the materials used to train the model: https://readcoop.eu/model/transkribus-german-handwriting-m1/ (accessed on 1 September 2023).

References

1. Colutto, S.; Kahle, P.; Guenter, H.; Muehlberger, G. Transkribus. A Platform for Automated Text Recognition and Searching of Historical Documents. In Proceedings of the 2019 15th International Conference on EScience (EScience), San Diego, CA, USA, 24–27 September 2019; IEEE: San Diego, CA, USA; pp. 463–466. [CrossRef]
2. Valkema, S.; Laansma, K. *Sybren Valkema*; De Prom: Baarn, The Netherlands, 1994.
3. Byrd, J.F.; Littleton, H.K.; Harvey, K. *Littleton—A Life in Glass: Founder of America's Studio Glass Movement*; Skira Rizzoli: New York, NY, USA, 2011.
4. Frantz, S.K. *Artists and Glass: A History of International Studio Glass*; The University of Arizona: Tucson, AZ, USA, 1987.
5. Meihuizen, J.; Temminck, J. *De Wereld Volgens Valkema*; Glascahiers: Leerdam, The Netherlands; Nationaal Glas Museum Leerdam: Leerdam, The Netherlands, 2005.
6. Archief (Verzameling) Sybren Valkema. 2021. RKD Explore. 2021. Available online: https://rkd.nl/nl/explore/collections/246 (accessed on 11 November 2023).
7. Archief Sybren Valkema. RKD Nederlands Instituut voor Kunstgeschiedenis. Available online: https://rkd.nl/nl/projecten-en-publicaties/projecten/265-archief-sybren-valkema (accessed on 2 January 2023).
8. Art DATIS Project. 2018. Art DATIS. Available online: https://artdatis.nl (accessed on 8 March 2023).
9. Gupta, M.R.; Jacobson, N.P.; Garcia, E.K. OCR Binarization and Image Pre-Processing for Searching Historical Documents. *Pattern Recognit.* **2007**, *40*, 389–397. [CrossRef]
10. Leedham, C.G. Historical Perspectives of Handwriting Recognition Systems. In *IEE Colloquium on Handwriting and Pen-Based Input*; IET: Stevenage, UK, 1994; pp. 1/1–1/3.
11. European Commission. Recognition and Enrichment of Archival Documents (READ). CORDIS. 2019. Available online: https://cordis.europa.eu/project/id/674943 (accessed on 11 November 2023).
12. Muehlberger, G.; Seaward, L.; Terras, M.; Oliveira, S.A.; Bosch, V.; Bryan, M.; Colutto, S.; Dejean, H.; Diem, M.; Fiel, S.; et al. Transforming Scholarship in the Archives through Handwritten Text Recognition: Transkribus as a Case Study. *J. Doc.* **2019**, *75*, 954–976. [CrossRef]
13. READ-COOP. Training Models. Data Preparation. 2023. Available online: https://help.transkribus.org/data-preparation (accessed on 11 November 2023).
14. READ-COOP. Introducing Transkribus Super Models—Get Access to "The Text Titan I". 2023. Available online: https://readcoop.eu/introducing-transkribus-super-models-get-access-to-the-text-titan-i/ (accessed on 11 November 2023).
15. Quiniou, S.; Mohamed, C.; Eric, A. Error handling approach using characterization and correction steps for handwritten document analysis. *Int. J. Doc. Anal. Recognit. (IJDAR)* **2012**, *15*, 125–141. [CrossRef]
16. Neto, A.F.d.S.; Bezerra, B.L.D.; Toselli, A.H. Towards the natural language processing as spelling correction for offline handwritten text recognition systems. *Appl. Sci.* **2020**, *10*, 7711. [CrossRef]
17. Dunley, R. The National Archives—Machines Reading the Archive: Handwritten Text Recognition Software. Text. The National Archives Blog. The National Archives. 2018. Available online: https://blog.nationalarchives.gov.uk/machines-reading-the-archive-handwritten-text-recognition-software/ (accessed on 19 March 2018).
18. Rabus, A. Recognizing Handwritten Text in Slavic Manuscripts: A Neural-Network Approach Using Transkribus. *Scr. E-Scr.* **2019**, *19*, 9–32.
19. Philips, J.P.; Tabrizi, N. Historical Document Processing: Historical Document Processing: A Survey of Techniques, Tools, and Trends. *arXiv* **2020**, arXiv:2002.06300. [CrossRef]
20. Parziale, A.; Giuliana, C.; Angelo, M. One step is not enough: A multi-step procedure for building the training set of a query by string keyword spotting system to assist the transcription of historical document. *J. Imaging* **2020**, *6*, 109. [CrossRef] [PubMed]
21. Santoro, A.; Marcelli, A. Using keyword spotting systems as tools for the transcription of historical handwritten documents: Models and procedures for performance evaluation. *Pattern Recognit. Lett.* **2020**, *131*, 329–335. [CrossRef]
22. Schwarz-Ricci, V.I. Handwritten Text Recognition per registri notarili (secc. XV–XVI): Una sperimentazione. *Um. Digit.* **2022**, *6*, 171–181. [CrossRef]
23. Van Wynsberghe, A. Sustainable AI: AI for Sustainability and the Sustainability of AI. *AI Ethics* **2021**, *1*, 213–218. [CrossRef]
24. Strubell, E.; Ganesh, A.; McCallum, A. Energy and Policy Considerations for Deep Learning in NLP. *arXiv* **2019**, arXiv:1906.02243. [CrossRef]
25. Provatorova, V.; Vakulenko, S.; Kanoulas, E.; van Hulst, J.M. 'Named Entity Recognition and Linking on Historical Newspapers: UvA.ILPS & REL At'. In CEUR-WS, Vol. 2696. Thessaloniki, 2020. Available online: https://ceur-ws.org/Vol-2696/paper_209.pdf (accessed on 1 September 2023).
26. Capurro, C.; Plets, G. Europeana, EDM, and the Europeanisation of Cultural Heritage Institutions. *Digit. Cult. Soc.* **2020**, *6*, 163–189. [CrossRef]
27. READ-COOP. HTR+. 2021. Available online: https://readcoop.eu/glossary/htr-plus/ (accessed on 9 March 2023).
28. READ-COOP. PyLaia. 2021. Available online: https://readcoop.eu/glossary/pylaia/ (accessed on 9 March 2023).

29. Danilak, M.M. Langdetect: Language Detection Library Ported from Google's Language-Detection. Python. 2014. Available online: https://github.com/Mimino666/langdetect (accessed on 11 November 2023).

30. Ramirez-Orta, J.A.; Xamena, E.; Maguitman, A.; Milios, E.; Soto, A.J. Post-OCR Document Correction with Large Ensembles of Character Sequence-to-Sequence Models. In Proceedings of the AAAI Conference on Artificial Intelligence, Virtual, 22 February–1 March 2022; Volume 36, pp. 11192–11199.

31. Wang, C. Fastwer: A PyPI Package for Fast Word/Character Error Rate (WER/CER) Calculation. Python. 2020. Available online: https://github.com/kahne/fastwer (accessed on 11 November 2023).

32. Drobac, S.; Lindén, K. Optical Character Recognition with Neural Networks and Post-Correction with Finite State Methods. *Int. J. Doc. Anal. Recognit. (IJDAR)* **2020**, *23*, 279–295. [CrossRef]

33. Pavlopoulos, J.; Kougia, V.; Platanou, P.; Shabalin, S.; Liagkou, K.; Papadatos, E.; Essler, H.; Camps, J.-B.; Fischer, F. Error Correcting HTR'ed Byzantine Text. 2023. Available online: https://www.researchsquare.com/article/rs-2921088/v1 (accessed on 11 November 2023).

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.