

Improving automated transcription to compile large and diverse historical corpora

Tanja Säily, Ari Vesalainen, Samuli Kaislaniemi, David Denison, Nuria Yáñez-Bouza, Pete Morris, Akseli Kettunen, Inga Kokkonen

Introduction

- Historical published texts:
 - Readily available in large quantities BUT socially unrepresentative
- Manuscript sources like personal letters:
 - More ‘speech-like’ (Culpeper & Kytö 2010: 17), could be written by anyone literate
- Transcription: expert work, requires time + resources
 - manuscript-based corpora are small
 - research limited to frequent phenomena
- We need corpora that are both large and socially diverse!
 - **more efficient transcription methods needed**

State of the art in Handwritten Text Recognition (HTR)

- Current best practice: transformer- and attention-based HTR models (Alkendi et al. 2024) (super models in Transkribus)
 - Main strength: accurate transcription when **trained on suitable** historical data
 - Challenges: degraded pages, complex layouts, marginal notes, abbreviations, and **low-resource languages/varieties**
 - Pipeline: layout > line segmentation > reading order > recognition
 - Sometimes followed by LLM-based post-correction
- Future: multimodal LLMs (mLLMs) that read and interpret document images directly (Humphreys et al. 2025, Crosilla et al. 2025)

Pilot study: Late Modern English correspondence

- Fine-tuning a mid-sized open source mLLM (Qwen2.5-7B-Instruct, Bai et al. 2023) to transcribe LModE letters
 - Part of the project “New Methods for Developing Diverse Corpora” (DEDICO, University of Helsinki, 2025-)
- Research questions:
 - How does transcription performance vary across different test datasets and social groups?
 - How does the transcription performance of our model compare with Transkribus on letters from the lower social ranks?
 - What are the takeaways for future research?

Datasets for training and testing

Training dataset

Subset	Images	Source
Clift 1	25	Clift family letters, sample re-edited by Samuli
MHP	918	Mary Hamilton Papers, gold standard sample curated by Nuria+David
EMCO	2808	Elizabeth Montagu Correspondence Online

Test dataset

Subset	Images	Note
Clift 2	48	Different letters from Clift 1
MHP	127	Random, no overlap
EMCO	154	Random, no overlap
CEEC = Corpora of Early English Correspondence , sample re-edited by Samuli	40	Socially representative sample

You will oblige me by sending the books
 when with the magazines to Mr. Abrey, directed for me & so
 soon as you give me notice I will send for them from there
 If it is not convenient for you to convey them by this
 means, be please to let me know in immediately & I
 will think of some other method.

Sweet Puchinotta called here
 this morning to take leave, he only waits for post here
 to set off to Italy - I cannot express how sorry
 am to see him go! the taking leave of him was quite
 an operation, but yet a cordial leave taking is
 always a total parting from a lost body, so it is
 the more so in this instance, he seems so much
 to mind himself - he left his best respects &
 Compliments to you, & a request that you would
 be so kind to him as well as my Father, & should

But when you wrote your letter I suppose you
 thought to tease me as you did John yet I hope
 you dont think me altogether of my Brothers
 Disposition tho when first I Peruse it over
 I could not help thinking it a pretty smart
 Punishment for my neglect & respect which I did

who gave me my life, & thought I must be content to live my last day
 of it & it is long enough. We are not in great confusion upon the
 enquiry that are about & there is generally a damp upon the spirits of
 all men that wish will to their Country, Or war against France last
 year carried on with great success & now we are almost an
 ally with us ~~to~~ to receive of fruit of it it is now in the
 of our hands by our own hands, & we are in a great measure
 of us now 7 years since at least such a peace is will be happily
 England. we have in our country at least don't mean to France the
 when we have don assist them with all our strength and we are at the

Humble beg^g you will be pleased to write to me
 of if you think I may be desired to go into your
 you will assure if I am under so much sorrow
 that is my condition to be expert or can support my
 self for you declare I have thought I have
 might since he hath been there I have another way
 to desire you to let me know that you would be pleased to consider
 his condition & that if place is very changeable &
 we persons that he does owe but diligence to but
 it is better to be in the hands of the law

you have miss'd the truth by a wide
 bye & more - it was not neglect
 I am too proud for that - it was not
 forgetfulness - sir I am out so
 forgetful - it was not Idleness,
 the excuse of fools - nor Henry of
 Walsingham - the defence of Traitors -

cost me in Surgeons Apothecary Quises Mr. Street
 four or five hundred pounds. I have bin to receive
 of my friend as not to demand satisfaction for my
 retainer, do pray the favor of the said Quises have
 a Bill made out for my ~~retainer~~ what you think -
 Government standing bin out of my money now 36 or 37

CEEC test dataset: 13(+) letters by diff writers, 1690-1784 (3 women, 10 men; from ex-slave to nobility). Non-professional photographs ("DIY digitization")

Dear brother
 the pen is put into my hand, that I
 may inform you how affairs are here. first then you
 may remember I had great expectations of a bear,
 that provided herself a nest in the stable, the first instance
 of bringing me in chickens, as I vainly imagined she would
 present me with only 4 or 5 these so small, I thought I should have
 been oblig'd to have borrow'd Master Rowles's spectacles to have
 seen them. I wish you had staid a few days longer, & you would
 see them.

a thousand thanks to the Duke
 of Newcastle & yourself for the exceeding good
 news I received from you last night, & the
 confirmation I had of it this morning. it has
 given us all the greatest pleasure, but not
 the least surprise to me, tho I know the

Dear Sir!
 Newcastle Sept. 10. 1748.
 I rec'd your kind letter & had enquir'd
 the former had not I been out when it came - I shall
 be oblig'd for the translation of the Greek history, my
 Greek was having now with it - the volume you mention
 of Modestius Adams works is the very volume I have not
 having those of the Lawyers, Philosophers & Physicians but
 not of Divines - I bought for my self about 50 volumes, &

The principal affair you men-
 tion is under examination, & will
 yet is over, I am not inform'd
 sufficiently to make any other
 judgement of ye matter yet

This, Madam, is the short, & true state of my
 Case. They that make their Court to y^e Ministers, &
 not their Majestys. succeed better. If my Case deserves
 some consideration, & you can serve me in it, I
 humbly hope, & believe, you will. I shall therefore
 trouble you no farther, but beg leave to subscribe my
 self, with most Respect, & Gratitude

as in our reflections upon the behaviour of the Clergy,
 Arncliffe, & told me, that, being upon good terms, as I
 know, in St. Gilberts family, who is one of the
 Carleton House prints as it is call'd, & what is more
 to the purpose, Treasurer of the Navy, & he could,

with I am sure y^e company would do a
 great way towards my recovery, for I
 assure you no body has a greater value
 for you than I have, & hope I shall
 have the good fortune to see you before

Data processing

- One page per image
- Resize images to a fixed maximum width and height while preserving the original aspect ratio
- Remove images with ambiguous page structure or inconsistent text orientation
- TEI to plaintext conversion
 - Normalize whitespace
 - Preserve line breaks, superscript
 - Remove notes, stamps, ...

Method: Fine-tuning Qwen2.5-7B-Instruct for HTR

Goal: Transcribe handwritten English letters diplomatically from image-text pairs.

Training data: Pairs of handwritten letter images and manually prepared ground-truth transcriptions.

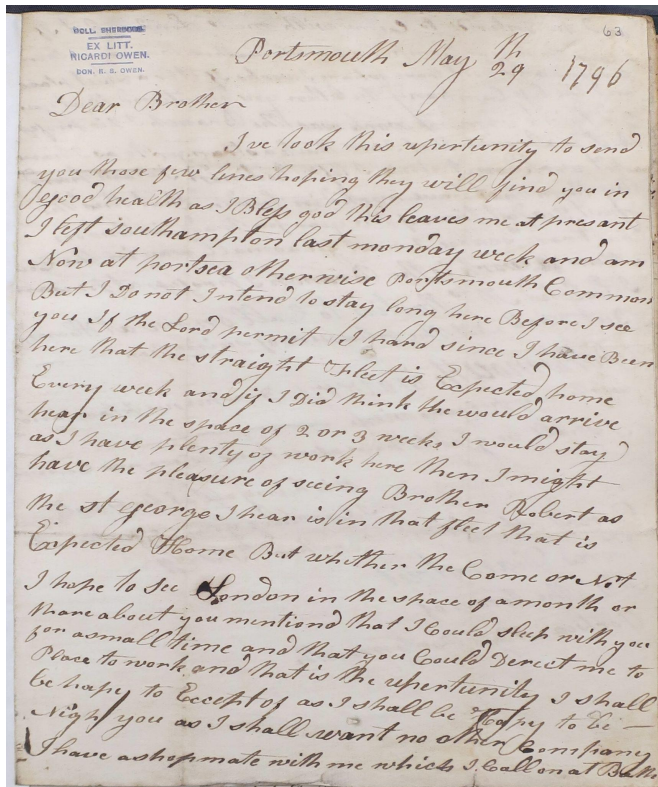
Prompt design: The model was instructed to preserve the document as written:

“Transcribe this handwritten English letter diplomatically. Preserve original spelling, capitalization, punctuation, and line breaks exactly as written. Do not normalize or modernize the text. Output only the transcription.”

Output target: The model learns to generate only the diplomatic transcription, without commentary, correction, modernization, or normalization.

Results

Example of good results (Character Error Rate = CER 5%)



Portsmouth May 29th 1796

Dear Brother

[D] I [-] ve took this oportunity to send
you those few lines hoping they will find you in
[R] G ood health as I Bless god [R] t his leaves me at pres [R] a nt
I left southampton last monday week and am
Now at portsea otherwise Portsmouth Common
But I Do not Intend to stay long here Before I see
you If the Lord permit [-] I ha [D] r d since I have Been
here that the straight [R] F leet is Expected home
Every week and if I [R] D id think [R] t he would arrive
h [D] e a [R] r in the space of 2 or 3 weeks I would stay
as I have plenty of work here then I might
have the pleasure of seeing Brother Robert as
the st [R] G eorge I hear [D] is in that fleet that is
Expected Home But whether the Come or Not
I hope to See London in the space of a month or
[R] tha re about you mentiond that I Could sleep with you
for a [-] small time and that you Could [R] De rect me to
[R] P lace to work and [R] t hat is the [R] u p [R] e rtunity I shall
be ha [-] py to Except of as I shall be [R] H a [-] py to be [-]
[R] N i [R] g h you as I shall want no other Company
I have a [R] sh o [-] p [D] m a [R] te with me which I C [-] all on at B [R] ath

Example of good results (Character Error Rate = CER 5%)

HOLL. SHERBOURNE
EX LITT.
RICARDI OWEN.
DON. R. S. OWEN.

Portsmouth May 29th 1796

Dear Brother

I've took this upertunity to send
you those few lines hoping they will find you in
Deveod health as I Bless god this leaves me at presant
I left southampton last monoday week and am
Now at portsea otherwise Portsmouth Common
But I Do not Intend to stay long here Before I see
you If the Lord permit I shand since I have Been
here that the straight Fleet is Expected home
Every week and if I did think the would arrive
heaf in the space of 2 or 3 weeks I would stay
as I have plenty of work here then I might
have the pleasure of seeing Brother Robert as
the St George I hear in that fleet that is

Portsmouth May 29th 1796

Dear Brother

[D] I [-] ve took this upertunity to send
you those few lines hoping they will find you in
[R] G ood health as I Bless god [R] t his leaves me at pres [R]
I left southampton last monday week and am
Now at portsea otherwise Portsmouth Common
But I Do not Intend to stay long here Before I see
you If the Lord permit [-] I ha [D] r d since I have Been
here that the straight [R] F leet is Expected home
Every week and if I [R] D id think [R] t he would arrive
h [D] e a [R] r in the space of 2 or 3 weeks I would stay
as I have plenty of work here then I might
have the pleasure of seeing Brother Robert as
the st [R] G eorge I hear [D] is in that fleet that is
Expected Home But whether the Come or Not

Overall statistics, 4 test sets

CER evaluation results

Subset	CER% (case sensitive)	CER% (all lowercase)*)	Size (characters)
EMCO	8.60	8.10	160,549
MHP	9.45	8.91	89,169
Clift 2	14.11	12.75	38,244
CEEC	15.08	14.16	29,774

*) All text in lowercase, long-s as normal s, multiple hyphens collapsed to one

Overall statistics: Summary of results

- More training data → better results
 - EMC0: also the least variation, only letters by Montagu
- More social representativeness → worse results?
 - Clift 2, CEEC - but also less training data
- Surprising: lower-class letters (Clift 2) perform better than a more balanced sample (CEEC)
 - Clift: some training data, CEEC: none
 - CEEC: more variation in hands, styles
- More variation → more challenging to model

CEEC: Variation across social groups

Small sample, but indications of transcription performance:

- Women < men
- Lower < higher social ranks
- Earlier < later

Letter ID	Sender name	Sender gender	Sender rank	Year	CER%
BENTHAJ_016	Jeremy Bentham	Male	Professional	1772	3.39
DUKES_089	Charles Lennox	Male	Nobility	1746	6.32
TWINING_005	Thomas Twining	Male	Clergy (Lower)	1764	6.45
YOUNG_026	Edward Young	Male	Clergy (Lower)	1730	6.81
SWIFT_033	Mary Butler née Somerset	Female	Nobility	1723	8.34
BLOMEFI_034	Francis Blomefield	Male	Clergy (Lower)	1748	9.35
SANCHO_034	Ignatius Sancho	Male	Other non-gentry	1779	11.01
BURNEY_040	Charles Burney	Male	Professional	1784	11.63
HATTON2_048	Elizabeth Hatton née Scroggs	Female	Gentry (Lower)	1690	14.36
PRIDEA2_035	Humphrey Prideaux	Male	Clergy (Upper)	1710	16.00
CARTER_015	Elizabeth Carter	Female	Clergy (Lower)	1739	20.55
PINNEY_056	Nathaniel Pinney	Male	Merchant	1706	34.57
HADDOC2_008	Richard Haddock	Male	Professional	1709	41.60

Qwen vs. Transkribus

CER and WER (Word Error Rate) results,
Clift 2 test dataset (48 images)

Model	CER%	WER% (all lowercase)
The Text Titan I ter (Transkribus)	10.26	24.25
Our model (Qwen)	14.11	27.43

Qwen vs. Transkribus: Token-level error analysis

Error category	Qwen	Titan
Correct	5887	6157
Real-word substitution	813	750
Deletion	552	310
Capitalization	352	372
Non-word error	322	345
Real-word insertion	166	124
Non-word insertion	35	54
Long s or apostrophe	8	0

Qwen vs. Transkribus: Summary of results

- Transcription performance: Qwen < Transkribus
- Qwen
 - More **real-word errors** (as expected from an LLM) → harder to detect
 - Example: ground truth **house**, prediction **horse**
 - More deletions
- Transkribus
 - More **non-word errors** (compared to a historical glossary)
 - Example: ground truth **Brother**, prediction **Brolher**
 - More capitalization errors

Conclusions

Discussion

- Training data matters
 - Training data is not neutral: transcription conventions shape model behaviour
 - Granularity matters: line-level data may improve alignment between image and text
 - Quality beats volume when sources are heterogeneous or inconsistent
 - Should separate model limits from data effects
- What is the ground truth?
 - Include what is visible and transcribable; exclude editorial or structural TEI markup not present in the image?

Discussion (cont.)

- Training data matters → deep-dive approach to future corpus compilation
 - Lots of data per source, hand-corrected training data for each source, standardized conventions
- Pros of Qwen over Transkribus:
 - We have control over the process, can keep improving
 - Transkribus is commercial, Qwen free for us
- Cons of Qwen:
 - Use of resources? (Supercomputer at CSC – IT Center for Science)
 - Identifying error source more difficult (page-level transcription)

Summary of results

- More training data → better results
- More variation → more challenging to model
- Social variation in transcription performance
 - Women < men
 - Lower < higher social ranks
 - Earlier < later
- Multimodal LLMs are a promising alternative to earlier methods but more work needed
 - Transkribus still beats our model (but only a smallish sample compared)

Conclusion

- Future work
 - More data; try to use more of the CEEC?
 - How training data scale and curation choices influence the error structure of mLLMs in historical HTR
 - Benchmarking, try other models besides Qwen
- mLLMs significantly facilitate semi-automated transcription already
 - Mary Hamilton project: Llama
 - Potential experiment: how much do different models speed up human transcribers' work (experts vs. student assistants)?
 - Issue of social variation remains but training data helps

References

- Alkendi, W., Gechter, F., Heyberger, L., & Guyeux, C. 2024. Advancements and Challenges in Handwritten Text Recognition: A Comprehensive Survey. *Journal of Imaging*. <https://doi.org/10.3390/jimaging10010018>
- Bai, J., Bai, S., Chu, Y., Cui, Z., Dang, K., Deng, X., Fan, Y., Ge, W., Han, Y., Huang, F., et al. 2023. Qwen technical report. arXiv preprint arXiv:2309.16609. <https://arxiv.org/abs/2309.16609>
- CEEC = *Corpora of Early English Correspondence*. Compiled by T. Nevalainen, H. Raumolin-Brunberg, S. Kaislaniemi, J. Keränen, M. Laitinen, M. Nevala, A. Nurmi, M. Palander-Collin, T. Säily, & A. Sairio at the Department of Languages, University of Helsinki. <https://varieng.helsinki.fi/CoRD/corpora/CEEC/>
- Crosilla, G., Klic, L., & Colavizza, G. 2025. Benchmarking Large Language Models for Handwritten Text Recognition. *J. Documentation*, 81, 334-354. <https://doi.org/10.48550/arxiv.2503.15195>.
- Culpeper, J. & Kytö, M. 2010. *Early Modern English Dialogues: Spoken Interaction as Writing*. CUP.
- EMCO = *Elizabeth Montagu Correspondence Online*. Registered Charity no. 1174697 in collaboration with the University of Swansea and Oxford Brookes University. <https://emco.swansea.ac.uk>
- Humphries, M., Leddy, L., Downton, Q., Legace, M., McConnell, J., Murray, I., & Spence, E. 2025. Unlocking the archives: Using large language models to transcribe handwritten historical documents. *Historical Methods: A Journal of Quantitative and Interdisciplinary History*, 58, 175-193. <https://doi.org/10.1080/01615440.2025.2500309>
- MHP = *The Mary Hamilton Papers (c.1740-c.1850)*. Compiled by D. Denison, N. Yáñez-Bouza, T. Oudesluijs, C. Ulph, C. Wallis, H. Barker, & S. Coulombeau, University of Manchester, 2019-2023. <https://doi.org/10.48420/21687809>
- Transkribus. Innsbruck: READ-COOP SCE. <https://transkribus.eu>