Handwriting recognition for Scottish Gaelic

Proceedings of the CLTW 4 @ LREC2022 , pages 60–70
Marseille, 20-25 June 2022

Handwriting Recognition for Scottish Gaelic

Mark Sinclair, William Lamb, Beatrice Alex
Quorate Technology Ltd, University of Edinburgh, University of Edinburgh
mark.s.sinclair@gmail.com, w.lamb@ed.ac.uk, b.alex@ed.ac.uk

Abstract

Like most other minority languages, Scottish Gaelic has limited tools and resources available for Natural Language Processing research and applications. These limitations restrict the potential of the language to participate in modern speech technology, while also restricting research in fields such as corpus linguistics and the Digital Humanities. At the same time, Gaelic has a long written history, is well-described linguistically, and is unusually well-supported in terms of *potential* NLP training data. For instance, archives such as the School of Scottish Studies hold thousands of digitised recordings of vernacular speech, many of which have been transcribed as paper-based, handwritten manuscripts. In this paper, we describe a project to digitise and recognise a corpus of handwritten narrative transcriptions, with the intention of re-purposing it to develop a Gaelic speech recognition system.

Keywords: Scottish Gaelic, Handwriting Recognition, minority languages, Low-Resource NLP, Digital Humanities

1. Introduction

Few minority languages have progressed beyond an inchoate developmental stage in language technology and Natural Language Processing (NLP). As the emphasis in these fields has shifted from rule-based approaches to deep-learning, the challenges for most minority languages have intensified. For many, the requisite training data do not exist. For some, the data are available, but must be digitised – a less imposing, but still significant barrier. In this latter category is Scottish Gaelic, a minority language spoken by 57,000 people in Scotland (National Records of Scotland, 2015).[1] A wealth of transcribed spontaneous speech and corresponding audio exist in Gaelic, but these transcriptions generally occur as handwritten manuscripts. Thus, to use these data for training an automatic speech recognition (ASR) system, for instance, one must first convert them to digital text.

Most of the transcriptions of natural language available in Gaelic are paper-based and stem from linguistic and ethnological fieldwork carried out in the mid-20th century by the School of Scottish Studies (University of Edinburgh).[2] Although some of these documents are typed, the majority are handwritten.[3]

Optical character recognition (OCR) for roman type is considered less challenging than handwriting recognition (HWR) due to language-specific parameters, variability in handwriting styles and the character-touching problem (Chen et al., 2021). If a robust HWR tool could be developed for Gaelic, it would unlock a vast trove of data useful both to the Digital Humanities and NLP research.

This paper reports on a one-year pilot study[4] to develop such a tool, by utilising the configurable HWR platform, Transkribus (Kahle et al., 2017), which is described further below. A Scottish Gaelic HWR resulting from our work is publicly available on the Transkribus site.[5]

Central to the effort were three research questions:

1. Given that most of the transcriptions were from one hand, to what extent would models developed using that hand alone generalise to the other hands in the dataset?

2. Manual annotation is by nature costly: How much data is required to produce a model that is accurate enough to allow a semi-supervised or unsupervised approach (i.e. one requiring little further editing)?

3. What impact do other resources (e.g. a lexicon and language model) have on error rates vis-à-vis training data alone (i.e. what is the most efficient combination of parameters to produce a usable model quickly)?

[1] https://www.scotlandscensus.gov.uk/census-results/at-a-glance/languages/

[2] https://www.ed.ac.uk/information-services/library-museum-gallery/cultural-heritage-collections/school-scottish-studies-archives

[3] A recent survey of the transcriptions held by the School of Scottish Studies' Tale Archive indicates that 77% are handwritten and 23% are typed

[4] We gratefully acknowledge funding from the University of Edinburgh's Challenge Investment Fund towards this project.

[5] https://readcoop.eu/model/scottish-gaelic-1949-1979/

2. Related Work

2.1. Speech and Language Processing for Scottish Gaelic

Given the lack of available electronic data for Scottish Gaelic, speech and language processing research for the language remains fairly limited. However, there has been recent work to develop: a Scottish Gaelic part-of-speech tagger (Lamb and Danso, 2014); an online linguistic analyser (Boizou and Lamb, 2020); a dependency treebank and parser (Batchelor, 2019); an automatic speech recognition system (Rasipuram et al., 2013); machine translation from Gaelic to Irish (Murchú, 2019);[6] an embedding model for Scottish Gaelic (Lamb and Sinclair, 2016); a derivation of a categorical grammar (Batchelor, 2016; Batchelor, 2019); a wordnet (Bella et al., 2020) and a text-to-speech system (developed by Cereproc).[7] Akhmetov et al. (2020) have also included Scottish Gaelic in their experiments on language-independent word lemmatisation.

Aside from existing speech and language processing work, there are digital corpora and lexical resources for Scottish Gaelic, including the Digital Archive of Scottish Gaelic (DASG) (O Maolalaigh, 2016)[8], the Annotated Representative Corpus of Scottish Gaelic (ARCOSG)[9] and the online dictionary, Am Faclair Beag (Bauer and MacDhonnchaidh,)[10].

2.2. Handwriting Recognition

Methods for HWR, also referred to as Handwritten Text Recognition, were first developed in the 1950s (Dimond, 1957). Since then, HWR has developed into an extremely active research field in computer science, which has been covered by a series of surveys and reviews (Hull, 1994; Plamondon and Srihari, 2000; Santosh and Nattee, 2010; Tagougui et al., 2012; Parvez and Mahmoud, 2013; Pal et al., 2012; Manoj et al., 2016; Al-Salman and Alyahya, 2017; Choudhary et al., 2017; Kumbhar and Kunjir, 2017; Das et al., 2018; Ramzan et al., 2018; Wang et al., 2021). A number of approaches including machine learning (Xu et al., 1992; Marti and Bunke, 2001) and neural network based learning (Graves et al., 2009; Boquera et al., 2011; Bluche, 2015; Wu et al., 2017; Naz et al., 2015; Voigtlaender et al., 2016; Chowdhury and Vig, 2018; Pham et al., 2014), or combinations thereof, have been applied to this task. The state-of-the-art is driven by regular international competitions on HWR and the availability of public datasets to compare performance of systems developed by different research groups (Menasri et al., 2011; Yin et al., 2013; Sánchez et al., 2014; Sánchez et al., 2017; Nguyen et al., 2018; Potanin et al., 2021). While HWR tended to be applied for financial or commercial purposes (Pal et al., 2012; Dimauro et al., 1997; Hafemann et al., 2017), with the increasing availability of digitised manuscript collections made available by libraries and archives, it has more recently been applied to historical manuscripts (Terras, 2006; Fischer et al., 2009; Fischer et al., 2014; Bhunia et al., 2019; Firmani et al., 2018; Chammas et al., 2018). There is also related work on applying HWR to different languages (Alipour et al., 2016; Zhang et al., 2018; Altwaijry and Al-Turaiki, 2020) or devising methods which work for multiple languages (Mondal et al., 2010; Keysers et al., 2017; Carbune et al., 2020). Carbune et al. (2020) are the only group we are aware of with a system that supports Scottish Gaelic alongside 101 other languages. They found that, compared to their previous segment-and-decode method, their Long Short-Term Memory (LSTM) based algorithm reduced the character error rate by between 20-40%, but they reported only for languages for which they had sufficient evaluation data (Figure 7 in Carbune et al. (2020)). They did not provide evaluation results for Scottish Gaelic. To the best of our knowledge, our paper is the first to report performance of HWR models applied to Scottish Gaelic text.

Transkribus (see Section 5) uses a deep neural network based algorithm for HWR (Muehlberger et al., 2019) and currently provides access to over 80 publicly available HWR models for different languages, each time reporting their character error rates against a validation set.[11] The platform has been used for training models for a series of languages, including low resource languages and scripts such as dialectal Finnish (Blokland et al., 2019), South Tyrolean (König et al., 2020), Low Saxon (Siewert et al., 2021), Evenki and Russian (Arkhipov et al., 2021), Greek, Slavic and Latin (Thompson and others, 2021), 16th century Romanian (Burlacu and Rabus, 2021) and Croatian Glagolitic (Rabus, 2022), to name but a few. Terras (2022) surveyed the registered users of Transkribus in early 2019 and examined how HWR had been by adopted libraries, archives and academia. Her work clearly shows that most of the documents processed by Transkribus projects were in German, Latin, English and French. A lot less material in other languages was processed at that point. Since the survey was conducted the user base has more than doubled and many more languages have been included, showing the potential and demand of HWR technology.

3. Digitisation and Correction of the Corpus

The training data for the current study came from a subset of the School of Scottish Studies Archives (University of Edinburgh) known as the Tale Archive. The

[6] NB: Gaelic also was added to Google Translate in 2016.

[7] https://www.cereproc.com

[8] https://dasg.ac.uk/

[9] https://github.com/Gaelic-Algorithmic-Research-Group/ARCOSG

[10] https://www.faclair.com/

[11] https://readcoop.eu/transkribus/public-models/

Tale Archive is an extensive collection of traditional narrative texts (c30k pages), most of which are entirely or partly in Scottish Gaelic.[12] Together, they represent the largest collection of transcribed Scottish Gaelic in the world. Although most of the participants from whom they were recorded lived in areas that continue to be Gaelic-speaking at the time this paper was written, many participants spoke regional variants that are now moribund or no longer extant. Thus, these data are uniquely valuable for their linguistic and ethnological content, as much as their potential to provide robust speech data for language technology applications.

We began the project by manually recording key metadata about the transcripts. Following this, we randomly selected documents totalling 2724 pages for digitisation. The transcripts were originally gathered between 1949 and 1979 and the distribution across that time period is shown in Figure 1. Here we can see some spikes in frequency corresponding to periods of increased activity for the project.

Despite spanning several decades, the narratives were predominantly transcribed by a single principal hand (approximately 85%) with the remaining portion (approximately 15%) distributed across 10 other hands. Given the over representation of this single hand in the data, a particular theme of our research was to examine how this imbalance would affect the potential generalisation of the HWR system.

The digitisation process involved converting the paper texts to a multi-page PDF format using a feed-based scanner and single-page scanning booth, depending on whether the source was an original or photocopy. Subsequently, the texts were uploaded to Transkribus for manual editing by a Domain Expert and, eventually, automatic recognition. The following section outlines the segmentation and transcription process in detail.

4. Handwriting Recognition

The task of Handwriting Recognition (HWR) involves automatically transcribing handwritten text into a digital form. HWR is similar to the task of Optical Character Recognition (OCR). The main difference is that the latter involves the recognition of printed text which, due to its uniformity, is typically less challenging to recognise automatically than handwriting.
Before carrying out HWR, handwritten documents must be captured as digital images, typically using digital imaging technologies such as scanners or cameras. Generally, modern HWR systems will them process these images in two main stages: *Segmentation* and *Transcription*.

[12]Roughly 75% are primarily in Gaelic, with another 25% mainly in Scots, English or Irish.

Segmentation is the task of removing non-relevant information from an image. This is typically achieved by defining tight geometric boundaries around areas of the image that are hypothesised to contain handwritten text. The purpose of this stage is to reduce noise in the input as well as to reduce the search space of any recognition algorithm in order to increase efficiency. An example is shown in Figure 2a.

Transcription is the task of estimating the text within each text segment and providing the results as standard digitised text. An example is shown in Figure 2b.

5. Transkribus

Transkribus is a software platform that helps to facilitate both manual and automatic transcription of historical written documents, as well as providing tools for searching and archiving. The main components of Transkribus include:

- An editing tool for manual and automatic segmentation, transcription and searching of documents.
- Cloud services that provide compute and storage resources for automated system components, including training HWR models.
- Web-based documentation and 'how-to' guides

5.1. Automatic Text Segmentation

The Transkribus platform provides an automatic text segmentation tool that is limited to Latin character sets, but is otherwise language-independent. This means that the tool is able to automatically find the boundaries of any text regions within Gaelic handwritten documents without the need for a specialised model. An example of fully automatic Transkribus segmentation on Scottish Gaelic is shown in Figure 3.

The text segmentation system component is not guaranteed to be error free and may require manual edits to be regarded as 'gold standard'. On the other hand, it is likely that such efforts will be minimal.

5.2. Manual HWR

Transkribus provides functionality for manually transcribing documents by means of an editor tool. This is a graphical interface that focuses an image viewer on each text segment and allows a human transcriber to easily type in the correct matching transcript.

5.3. Automatic HWR

Transkribus also facilitates automatic HWR. This system, however, relies on language-dependent neural network models in order to function accurately. Models are provided for a limited set of languages, including English and German, but no known Transkribus model for Gaelic existed before the current study. In order to

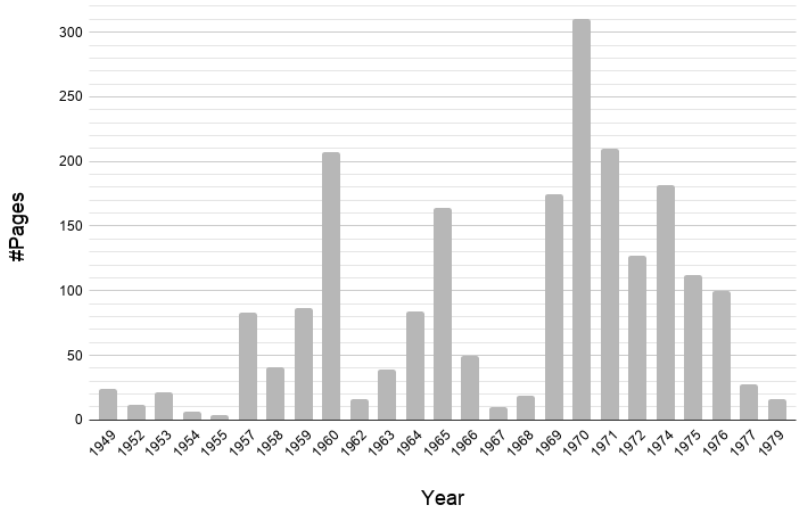
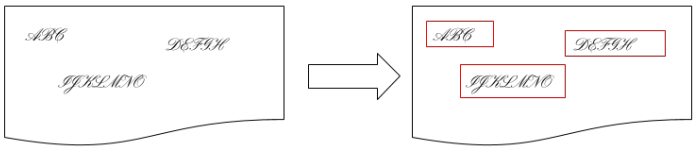
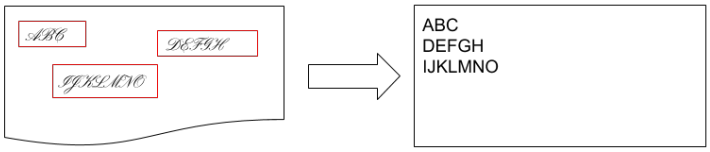


Figure 1: Distribution of complete training corpus data over year of collection.



(a) Segmentation



(b) Transcription

Figure 2: Examples of the Segmentation and Transcription tasks for HWR

provide a Gaelic model, it would have to be trained. This training process is described in Section 6 below.
In Figures 4a and 4b we provide examples of how the output quality of a Scottish Gaelic HWR model can vary depending on whether it is applied to the writing of the principal hand, or one with little or no training data. While the model performs fairly well on the principal hand it does extremely badly on the other hand. We think that this is mainly due to an unseen writing style, especially the way some of capital letters are curved, as well as the spaced out writing in this case.[13] Transkribus seems to fail to recognise that this is a sequence of text and only recognises a few, individual words. The latter example is one of the worst outputs we have encountered and we include it here to illustrate that HWR is not a solved problem. However, our evaluation results presented in Section 7 show that our models can yield promising results on unseen test data, especially when using larger training datasets and a language model.

6. Model Training Workflow

The workflow of the project comprised an iterative process of manual and automatic tasks. A systematic representation of the workflow is shown in Figure 5.
The sequence of tasks are as follows:

- A large quantity (1000s) of documents are scanned or photographed[14]
- Documents are loaded into Transkribus and are automatically segmented
- A Gaelic Domain Expert transcribes a portion of the documents (100s; using the Transkribus interface)

[13] See Lamb (2012, 121, fn 24) for more information on this transcriber.

[14] Transkribus provides an Android app to facilitate document photography.

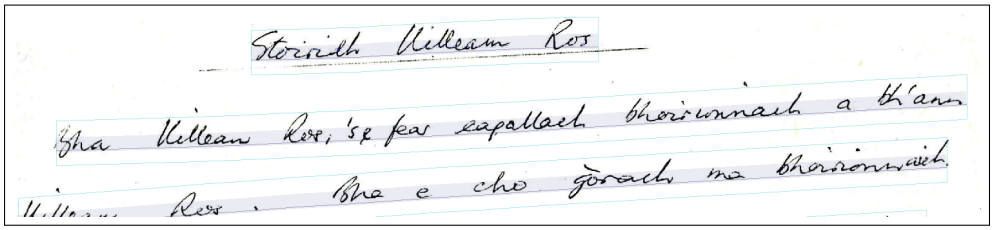
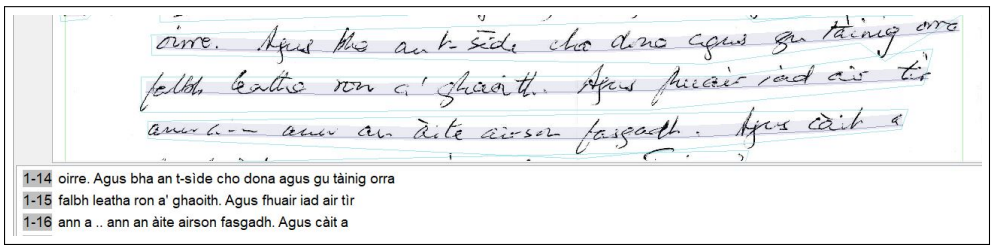
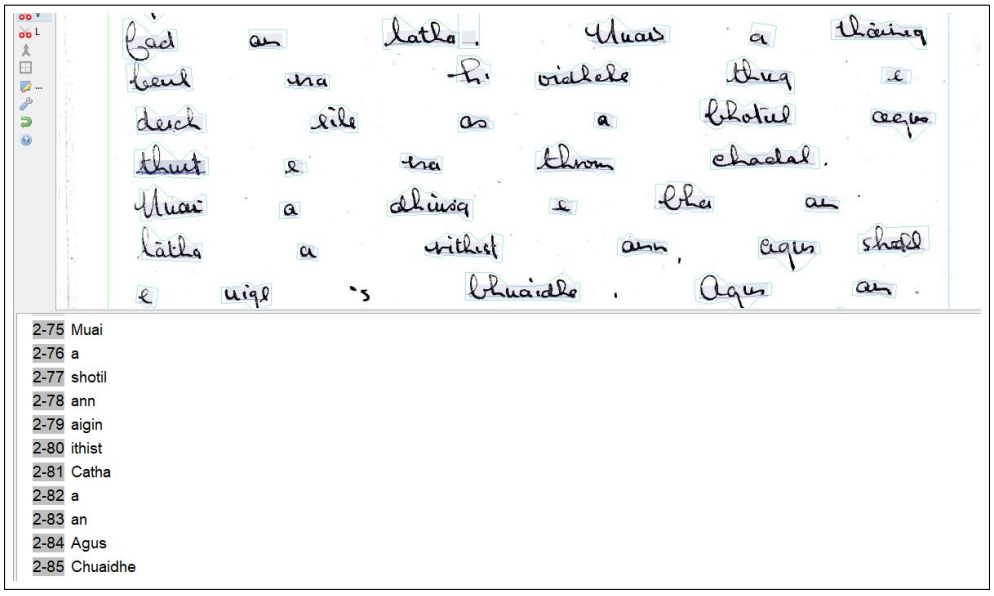


Figure 3: An example of automatic segmentation of Scottish Gaelic from the Transkribus software platform



(a) Good quality output for the principal hand



(b) Bad quality output for a hand with little training data

Figure 4: HWR output examples for the Scottish Gaelic Transkribus model

- Transcribed documents are divided into a training set (90%) and an evaluation set (10%)
- The training set is used to train a Transkribus neural network model for Scottish Gaelic
- The first (seed) model is used to transcribe the evaluation set
- Transkribus hypothesis on the evaluation set is scored against the manual transcription, i.e. Word Error Rate (WER) and Character Error Rate (CER) are computed
- If the error is unacceptable (above some defined threshold)
  - Auto-transcribe more training data (100s of pages) from the scanned documents that have not already been transcribed
  - The Gaelic Domain Expert corrects errors of Transkribus transcriptions
  - The corrected data augments the existing training set
  - Training and evaluation are repeated
- Else, if error is acceptable (below some defined threshold)
  - Auto transcribe all remaining scanned documents

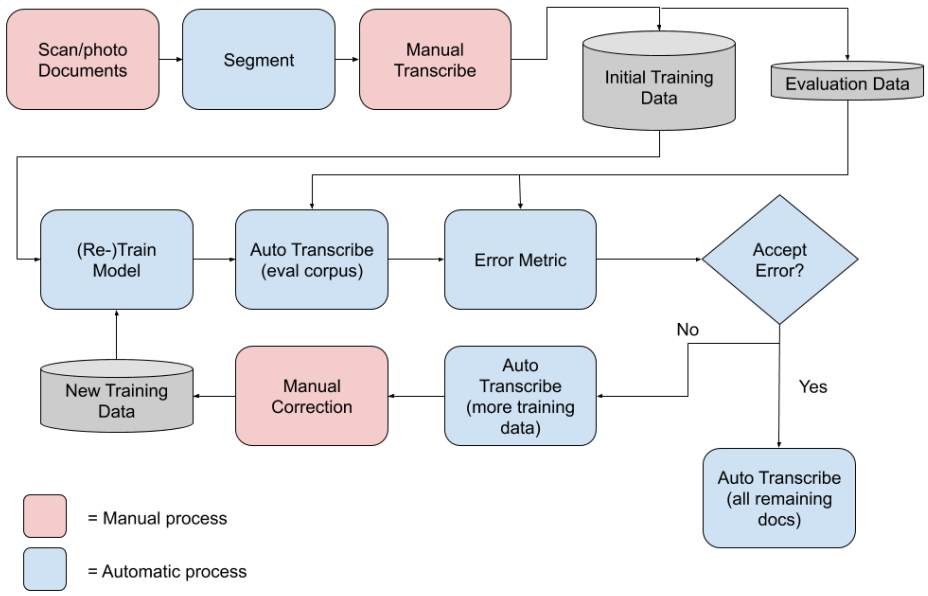


Figure 5: An overview of the machine-assisted transcription workflow. Red components are manual processes and blue components are automated. The system is initialised with the first manual transcription process, then enters the iterative feedback cycle where new data is automatically transcribed, manually corrected and fed back into training a new model.

The whole process begins with a small manual investment. The principle is that the manual correction phase gets easier at each iteration, because the automated system is improving its hypothesis. That is, there should be fewer mistakes to correct and, over time, more transcription data can be brought up to a manually-corrected standard with the same effort. The nature of neural network training methods suggests that we should expect an exponential decay in error until the limits of the model are reached. This means that there will likely be diminishing returns and a natural point will be found where the value of further transcribing/correcting data for the purpose of training the model is no longer economical. At this point, if the model performance is sufficiently acceptable, it can be used to automatically transcribe any remaining documents without the need for manual correction.

7. Experimental Results

7.1. Machine Assisted Transcription

In total, we completed three iterations through the workflow: a 75-hour initialisation iteration involving fully-manual transcription, followed by two further iterations with 75 and 380 hours of manual effort respectively, where HWR models were trained and used as the basis for manual correction. Table 1 shows the resulting transcription yield from each of the three stages. The first 75 hours of manual effort produced 18,158 words, making a yield of 242.11 words per hour. This initial tranche of data was used to train our first HWR model (**118_P_LM**), which was then used to generate an automatic transcription. The next 75 hours of manual effort in the second iteration were used to correct the output.

The second iteration produced an additional 18,397 words, making a yield of 245.29 words per hour: this was only slightly higher than the first. This suggested that, with the initial model, the machine assisted transcription had a very similar yield to a fully-manual transcription approach, i.e. it was taking just as long to correct the errors as it would have taken to transcribe from scratch, unassisted.

Combining all of the data from the first and second iterations, making a total of 36,555 words, we trained a second model (**221_P_LM**). It was clear at this point that the second model had performed much better than the first and was providing substantially greater assistance to the manual transcription process. For this reason, we decided to perform automated recognition on all remaining documents and focus the remaining manual transcription time budget on correction of the output. An additional 340,237 words were transcribed during this iteration, over 380 hours, making a yield of 895.36 words per hour. This means that with a modest investment of 150 hours of manual effort, we increased our transcription yield by a factor of over 3.5 times.

Table 1: Yield from manual transcription effort across 3 iterations of the workflow shown in Figure 5. The *Segmentation Only* row is the initial fully-manual transcription and subsequent rows were seeded with an automated transcription hypothesis using a model trained on the data from the previous row. The model name represents some information about the model separated by underscores: the number of pages used in training the model, that the data came from the (P)rincipal hand, and a Language Model (LM) was used – see subsequent sections for more detail.

HWR Seed Model	Manual Hours	Words Transcribed	Words per Hour
Segmentation Only	75	18,158	242.11
118_P_LM	75	18,397	245.29
221_P_LM	380	340,237	895.36
Total/Average	530	376,792	710.93

Table 2: Experimental results for HWR models with different quantities of training data (118, 221, 1678 or 1917 pages), Principal or Mixed hands (P or M), and with Lexical support from a Language Model (LM), (150k) word vocabulary dictionary or None. Results show Character Error Rate (CER) and Word Error Rate (WER) for Principal (P) and Other (O) hands evaluation data. The best results for each evaluation condition are highlighted in bold.

Model Code	**#Pages**	**#Words**	**Mixed**	**Lex.**	**CER(P)**	**WER(P)**	**CER(O)**	**WER(O)**
1917_M_None	1,917	376,792	TRUE	None	2.19	6.75	5.89	17.65
1917_M_150k	1,917	376,792	TRUE	150k	4.5	12.88	8.18	24.02
1917_M_LM	1,917	376,792	TRUE	LM	1.7	5.04	**5.01**	**14.86**
1678_P_None	1,678	318,967	FALSE	None	2.07	6.38	25.76	49.59
1678_P_150k	1,678	318,967	FALSE	150k	4.4	12.56	25.62	47.69
1678_P_LM	1,678	318,967	FALSE	LM	**1.67**	**4.94**	23.06	43.54
221_P_None	221	36,555	FALSE	None	2.58	7.53	25.07	49.68
221_P_150k	221	36,555	FALSE	150k	3.68	9.2	25.19	47.76
221_P_LM	221	36,555	FALSE	LM	2.53	7.28	24.14	47.34
118_P_None	118	18,158	FALSE	None	4.97	14.44	30.05	57.08
118_P_150k	118	18,158	FALSE	150k	5.62	14.84	29.78	54.28
118_P_LM	118	18,158	FALSE	LM	4.75	13.76	29.16	54.95

Ultimately, after 530 hours of manual effort we managed to achieve a total of 376,792 transcribed words. Assuming our initial fully-manual yield of 242.11 words per hour, the same quantity of transcription would have otherwise taken around 1,556 hours of manual effort. This means that the machine-assisted approach presents a significant reduction to costs, vis-à-vis manual handwriting transcription.

7.2. Lexical Support for HWR Models

The HWR models learnt to predict the most likely characters of the texts, given observation features derived from their images. HWR models are typically purely optical models that have no specific knowledge of the language they are transcribing, other than its character set. However, it is also possible to supplement models with information from additional lexical models in order to support, and potentially improve, the hypothesis. In particular, the Transkribus platform allows the provision of a lexicon or language model during the recognition inference.

The lexicon essentially provides an *allow*-list of tokens that can be permitted in the hypothesis. If an HWR hypothesis predicts a character sequence that does not correspond to an entry in the lexicon, then it can be rejected in favour of an another hypothesis that is represented. Each token is also weighted according to its prior probability, meaning that in cases of ambiguity, tokens that are more common are more likely to be selected. This can help to remove illegitimate character sequences (non-valid tokens) from the hypothesis but, conversely, any legitimate tokens that happen to be out-of-vocabulary in the lexicon may never be predicted. Therefore, it is important that the lexicon is comprehensive.

The language model differs in comparison to the lexicon in that it is not simply a model of tokens in isolation, but predicts the most likely sequences of tokens. This means that if there is ambiguity in a hypothesis, or noise in the input features, the lexical context can help to inform the most likely token that would have come next. By modelling more intrinsic information about the structure of a language in this way, we typically have more powerful lexical support than the basic lexicon. Each time Transkribus is used to train a HWR model, it also trains a language model using the same reference text as training data. These language and HWR models are tied in a way that they cannot

be mixed and matched between different training runs.

8. Discussion

Our results show that increasing the quantity of data helped to improve recognition performance. For example, going from our **118_P_LM** through **221_P_LM** to **1678_P_LM**, we see a reduction in WER from 13.76% through 7.28% to 4.94% for the principal hand evaluation case. This suggests a non-linear relationship between data quantity and error reduction, i.e. reducing the error rate by a constant factor would require increasing the data quantity factor. However, we do not have enough data points to estimate the true nature of this relationship.

The lexicon does not seem to help to improve recognition accuracy. However, we believe this is because our lexicon contains mostly base dictionary form words, i.e. it does not contain a lot of morphological permutations. For that reason, restricting the output to the lexicon entries is likely to create a lot of out-of-vocabulary (OOV) issues where words that are not present are assigned another word that has a similar character sequence. This is supported by the fact that the WER degraded more significantly than the CER when introducing the lexicon, i.e. the OOV word substitution can still result in getting most of the characters correct even if the word is incorrect.

Introducing the Language Model (LM) as a lexical support does help to improve recognition accuracy for both CER and WER. While the LM always improved accuracy, it demonstrated a more substantial improvement for the HWR models trained on more data. The LM can be particularly useful when the HWR hypothesis has fewer alternatives to choose from. As the HWR model improves, it is more likely to correctly recognise sub-word units of words (e.g. stems and affixes) that were previously poorly recognised. This can narrow the hypothesis and make it more likely for the LM to select the correct result. The introduction of a portion of mixed hand data resulted in substantial improvements on mixed hand evaluation data with only a negligible reduction in performance on the principal hand evaluation data: e.g. the WER reduced from 43.54% to 14.86% on mixed hands between **1678_P_LM** and **1917_M_LM** respectively, while only increasing from 4.94% to 5.04% for the same models.

9. Conclusion

We have shown that the use of machine-assisted handwriting recognition can significantly improve transcription efficiency with a modest manual effort investment. The data that has been digitised is now available to be easily searched and archived for humanities research. It can also be used as a data resource for other NLP tasks such as Automatic Speech Recognition (ASR), language modelling and entity extraction.

We believe that the iterative framework that we employed for this task could be re-purposed for other low-resource languages where the lack of an initial HWR requires such a bootstrapping approach. The acceleration in yield could have been improved further by re-training the model more often so as to gain the benefits at a more frequent cadence. The framework also supports the possibility of multiple manual transcribers and an asynchronous approach to model updates and manual effort, i.e. training a new model is not blocked by waiting for all transcribers to finish their current tasks.

10. Future Work

While the models developed for this project proved valuable for improving the efficiency of transcription on our target corpus, we would like to investigate how well the approach would generalise to corpora in other domains. In particular, we would like to create a general Scottish Gaelic HWR model than can be used as a reliable resource for digitising handwritten documents. This work would involve acquiring new datasets both to evaluate our existing models against and develop contrasting systems.

We were able to demonstrate that increasing data quantity improved model performance, but we did not have enough data to accurately estimate the trend. As with many machine learning tasks, it is likely that there will be an issue with diminishing gains where equivalent performance improvements may require exponential increases in data. Having enough data and examples of models trained with different quantities to estimate this would be useful when designing future experiments.

Another interesting approach is to consider the use of multi-lingual training data. Handwriting corpora for the related Goidelic language, Irish, could be used to supplement our training data; their character sets and many aspects of their grapheme distribution are similar. This kind of data could also help to act as a kind of natural regularisation for our models and prevent over-fitting to certain hands that are over-represented in our data. The combined data could be used to develop a multi-lingual model that can handle more general Gaelic-language handwriting.

11. Acknowledgements

We gratefully acknowledge funding received from the University of Edinburgh's Challenge Investment Fund. We would like to thank Prof James Loxley (University of Edinburgh) for his contributions to the early stages

of the project, Prof Melissa Terras (University of Edinburgh) for her advice and putting us in touch with Transkribus, and Michael Bauer, who carried out the recognition and editing of the Gaelic text. Finally, our sincere thanks to the staff at Transkribus and the School of Scottish Studies Archives, for their excellent support and assistance.

12. Bibliographical References

Akhmetov, I., Pak, A., Ualiyeva, I., and Gelbukh, A. (2020). Highly language-independent word lemmatization using a machine-learning classifier. *Computación y Sistemas*, 24(3):1353–1364.

Al-Salman, A. S. and Alyahya, H. (2017). Arabic online handwriting recognition: a survey. *Proceedings of the 1st International Conference on Internet of Things and Machine Learning*.

Alipour, S. A., Tabatabaey-Mashadi, N., and Abbassi, H. (2016). Recent approaches in online handwriting recognition for persian and arabic right-to-left languages. *2016 2nd IEEE International Conference on Computer and Communications (ICCC)*, pages 358–364.

Altwaijry, N. and Al-Turaiki, I. M. (2020). Arabic handwriting recognition system using convolutional neural network. *Neural Computing and Applications*, pages 1–13.

Arkhipov, A., Barinskaya, A., and Shtefura, R. (2021). Using handwritten text recognition on bilingual evenki-russian manuscripts of konstantin rychkov1. *Scripta & E-Scripta*, 21.

Batchelor, C. (2016). Automatic derivation of categorial grammar from a part-of-speech-tagged corpus in scottish gaelic. *PARIS Inalco du 4 au 8 juillet 2016*, page 1.

Batchelor, C. (2019). Universal dependencies for scottish gaelic: syntax. In *Proceedings of the Celtic Language Technology Workshop*, pages 7–15.

Bauer, M. and MacDhonnchaidh, U. (????). Am faclair beag. Online; accessed 19-February-2022.

Bella, G., McNeill, F., Gorman, R., O Donnaile, C., MacDonald, K., Chandrashekar, Y., Freihat, A. A., and Giunchiglia, F. (2020). A major Wordnet for a minority language: Scottish Gaelic. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 2812–2818, Marseille, France, May. European Language Resources Association.

Bhunia, A. K., Das, A., Bhunia, A. K., Kishore, P. S. R., and Roy, P. P. (2019). Handwriting recognition in low-resource scripts using adversarial learning. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4762–4771.

Blokland, R., Partanen, N., Rießler, M., and Wilbur, J. (2019). Using computational approaches to integrate endangered language legacy data into documentation corpora: Past experiences and challenges ahead. In *Workshop on Computational Methods for Endangered Languages, Honolulu, Hawai'i, USA*, volume 2, pages 24–30.

Bluche, T. (2015). *Deep neural networks for large vocabulary handwritten text recognition*. Ph.D. thesis, Paris 11.

Boizou, L. and Lamb, W. (2020). An online linguistic analyser for scottish gaelic. In *Human Language Technologies–The Baltic Perspective: Proceedings of the Ninth International Conference Baltic HLT 2020*, volume 328, page 119. IOS Press.

Boquera, S. E., Bleda, M. J. C., Gorbe-Moya, J., and Zamora-Martínez, F. (2011). Improving offline handwritten text recognition with hybrid hmm/ann models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33:767–779.

Burlacu, C. and Rabus, A. (2021). Digitising (romanian) cyrillic using transkribus: new perspectives. *Diacronia*, 2021(14):A196–A196.

Carbune, V., Gonnet, P., Deselaers, T., Rowley, H., Daryin, A. N., Lafarga, M. C., Wang, L.-L., Keysers, D., Feuz, S., and Gervais, P. (2020). Fast multilanguage lstm-based online handwriting recognition. *International Journal on Document Analysis and Recognition (IJDAR)*, 23:89–102.

Chammas, E., Mokbel, C., and Likforman-Sulem, L. (2018). Handwriting recognition of historical documents with few labeled data. *2018 13th IAPR International Workshop on Document Analysis Systems (DAS)*, pages 43–48.

Chen, X., Jin, L., Zhu, Y., Luo, C., and Wang, T. (2021). Text recognition in the wild: A survey. *ACM Computing Surveys (CSUR)*, 54(2):1–35.

Choudhary, U., Bhosale, S., Bhise, S., and Chilveri, P. G. (2017). A survey: Cursive handwriting recognition techniques. *2017 2nd IEEE International Conference on Recent Trends in Electronics, Information & Communication Technology (RTEICT)*, pages 1712–1716.

Chowdhury, A. and Vig, L. (2018). An efficient end-to-end neural model for handwritten text recognition. In *BMVC*.

Das, Y. K., Jain, P., and Sreekumar, K. G. (2018). Comprehensive survey on machine learning application for handwriting recognition. *International Journal of Applied Engineering Research*, 13(8):5823–5830.

Dimauro, G., Impedovo, S., Pirlo, G., and Salzo, A. (1997). Automatic bankcheck processing: A new engineered system. *Int. J. Pattern Recognit. Artif. Intell.*, 11:467–504.

Dimond, T. (1957). Devices for reading handwritten characters. In *IRE-ACM-AIEE '57 (Eastern)*.

Firmani, D., Maiorino, M., Merialdo, P., and Nieddu, E. (2018). Towards knowledge discovery from the vatican secret archives. in codice ratio - episode 1: Machine transcription of the manuscripts. *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*.

Fischer, A., Wuthrich, M., Liwicki, M., Frinken, V., Bunke, H., Viehhauser, G., and Stolz, M. (2009). Automatic transcription of handwritten medieval documents. *2009 15th International Conference on Virtual Systems and Multimedia*, pages 137–142.

Fischer, A., Baechler, M., Garz, A., Liwicki, M., and Ingold, R. (2014). A combined system for text line extraction and handwriting recognition in historical documents. *2014 11th IAPR International Workshop on Document Analysis Systems*, pages 71–75.

Graves, A., Liwicki, M., Fernández, S., Bertolami, R., Bunke, H., and Schmidhuber, J. (2009). A novel connectionist system for unconstrained handwriting recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31:855–868.

Hafemann, L. G., Sabourin, R., and Oliveira, L. (2017). Offline handwritten signature verification — literature review. *2017 Seventh International Conference on Image Processing Theory, Tools and Applications (IPTA)*, pages 1–8.

Hull, J. J. (1994). A database for handwritten text recognition research. *IEEE Trans. Pattern Anal. Mach. Intell.*, 16:550–554.

Kahle, P., Colutto, S., Hackl, G., and Mühlberger, G. (2017). Transkribus-a service platform for transcription, recognition and retrieval of historical documents. In *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, volume 4, pages 19–24. IEEE.

Keysers, D., Deselaers, T., Rowley, H., Wang, L.-L., and Carbune, V. (2017). Multi-language online handwriting recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39:1180–1194.

König, A., Lyding, V., Gorgaini, E., Grote, G., and Pretti, M. (2020). Community involvement for transcribing historical correspondences of south tyrolean interest: A di-öss use case. Technical report, -.

Kumbhar, O. and Kunjir, A. (2017). A survey on optical handwriting recognition system using machine learning algorithms. *International Journal of Computer Applications*, 175:28–31.

Lamb, W. and Danso, S. (2014). Developing an automatic part-of-speech tagger for scottish gaelic. In *Proceedings of the First Celtic Language Technology Workshop*, pages 1–5.

Lamb, W. and Sinclair, M. (2016). Developing word embedding models for scottish gaelic. In *Actes de la conférence conjointe JEP-TALN-RECITAL*, volume 6, pages 31–41.

Lamb, W. (2012). The storyteller, the scribe, and a missing man: Hidden influences from printed sources in the gaelic tales of duncan and neil macdonald. *Oral Tradition*, 27(1):109–160.

Manoj, A., Borate, P., Jain, P., Sanas, V., and Pashte, R. (2016). A survey on offline handwriting recognition systems. *International journal of scientific research in science, engineering and technology*, 2:253–257.

Marti, U.-V. and Bunke, H. (2001). Using a statistical language model to improve the performance of an hmm-based cursive handwriting recognition system. *Int. J. Pattern Recognit. Artif. Intell.*, 15:65–90.

Menasri, F., Louradour, J., Bianne-Bernard, A.-L., and Kermorvant, C. (2011). The a2ia french handwriting recognition system at the rimes-icdar2011 competition. In *Electronic Imaging*.

Mondal, T., Bhattacharya, U., Parui, S. K., Das, K., and Mandalapu, D. (2010). On-line handwriting recognition of indian scripts - the first benchmark. *2010 12th International Conference on Frontiers in Handwriting Recognition*, pages 200–205.

Muehlberger, G., Seaward, L., Terras, M., Oliveira, S. A., Bosch, V., Bryan, M., Colutto, S., Déjean, H., Diem, M., Fiel, S., Gatos, B., Greinoecker, A., Grüning, Tobias, Hackl, G., Haukkovaara, V., Heyer, G., Hirvonen, L., Hodel, T., Jokinen, M., Kahle, P., Kallio, M., Kaplan, F., Kleber, F., Labahn, R., Lang, E. M., Laube, S., Leifert, G., Louloudis, G., McNicholl, R., Meunier, J.-L., Michael, J., Mühlbauer, E., Philipp, N., Pratikakis, I., Puigcerver Pérez, J., Putz, H., Retsinas, G., Romero, V., Sablatnig, R., Sánchez, J. A., Schofield, P., Sfikas, G., Sieber, C., Stamatopoulos, N., Strauss, T., Terbul, T., Toselli, A. H., Ulreich, B., Villegas, M., Vidal, E., Walcher, J., Weidemann, M., Wurster, H., and Zagoris, K. (2019). Transforming scholarship in the archives through handwritten text recognition: Transkribus as a case study. *Journal of documentation*.

Murchú, E. P. Ó. (2019). Using intergaelic to pretranslate and subsequently post-edit a sci-fi novel from scottish gaelic to irish. In *Proceedings of the Qualities of Literary Machine Translation*, pages 20–25.

National Records of Scotland. (2015). Scotland's census 2011: Gaelic report (part 1). Technical report, National Records of Scotland, Edinburgh.

Naz, S., Umar, A. I., Ahmad, R., Ahmed, S. B., Shirazi, S. H., and Razzak, M. I. (2015). Urdu nasta'liq text recognition system based on multi-dimensional recurrent neural network and statistical features. *Neural Computing and Applications*, 28:219–231.

Nguyen, H. T., Nguyen, C. T., and Nakagawa, M. (2018). Icfhr 2018 – competition on vietnamese online handwritten text recognition using hands-vnondb (vohtr2018). *2018 16th International Conference on Frontiers in Handwriting Recognition (ICFHR)*, pages 494–499.

O Maolalaigh, R. (2016). Dasg: Digital archive of scottish gaelic/dachaigh airson stòras na gàidhlig. *Scottish Gaelic Studies*, 30:242–262.

Pal, U., Jayadevan, R., and Sharma, N. (2012). Handwriting recognition in indian regional scripts: A survey of offline techniques. *ACM Trans. Asian Lang. Inf. Process.*, 11:1:1–1:35.

Parvez, M. T. and Mahmoud, S. A. (2013). Offline

arabic handwritten text recognition: A survey. *ACM Comput. Surv.*, 45:23:1–23:35.

Pham, V., Kermorvant, C., and Louradour, J. (2014). Dropout improves recurrent neural networks for handwriting recognition. *2014 14th International Conference on Frontiers in Handwriting Recognition*, pages 285–290.

Plamondon, R. and Srihari, S. N. (2000). On-line and off-line handwriting recognition: A comprehensive survey. *IEEE Trans. Pattern Anal. Mach. Intell.*, 22:63–84.

Potanin, M. B., Dimitrov, D., Shonenkov, A., Bataev, V., Karachev, D., and Novopoltsev, M. (2021). Digital peter: Dataset, competition and handwriting recognition methods. In *HIP@ICDAR*.

Rabus, A. (2022). Handwritten text recognition for croatian glagolitic. *Slovo: časopis Staroslavenskoga instituta u Zagrebu*, 72(1):181–192.

Ramzan, M., Khan, H. U., Akhtar, W., Zamir, A., Awan, S. M., Ilyas, M., and Mahmood, A. (2018). A survey on using neural network based algorithms for hand written digit recognition. *environment*, 9(9).

Rasipuram, R., Bell, P., and Doss, M. M. (2013). Grapheme and multilingual posterior features for under-resourced speech recognition: a study on scottish gaelic. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 7334–7338. IEEE.

Sánchez, J.-A., Romero, V., Toselli, A. H., and Vidal, E. (2014). Icfhr2014 competition on handwritten text recognition on transcriptorium datasets (htrts). *2014 14th International Conference on Frontiers in Handwriting Recognition*, pages 785–790.

Sánchez, J.-A., Romero, V., Toselli, A. H., Villegas, M., and Vidal, E. (2017). Icdar2017 competition on handwritten text recognition on the read dataset. *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, 01:1383–1388.

Santosh, K. C. and Nattee, C. (2010). A comprehensive survey on on-line handwriting recognition technology and its real application to the nepalese natural handwriting. *Kathmandu University Journal of Science, Engineering and Technology*, 5:31–55.

Siewert, J., Scherrer, Y., and Tiedemann, J. (2021). Towards a balanced annotated low saxon dataset for diachronic investigation of dialectal variation. In *Proceedings of the 17th Conference on Natural Language Processing (KONVENS 2021)*, pages 242–246.

Tagougui, N., Kherallah, M., and Alimi, A. M. (2012). Online arabic handwriting recognition: a survey. *International Journal on Document Analysis and Recognition (IJDAR)*, 16:209–226.

Terras, M. (2006). *Image to interpretation: an intelligent system to aid historians in reading the Vindolanda texts*. OUP Oxford.

Terras, M. (2022). Inviting AI into the archives: The reception of handwritten recognition technology into historical manuscript transcription. transcript Verlag.

Thompson, W. et al. (2021). Using handwritten text recognition (HTR) tools to transcribe historical multilingual lexica. *Scripta & e-Scripta*, 2021(21):217–231.

Voigtlaender, P., Doetsch, P., and Ney, H. (2016). Handwriting recognition with large multidimensional long short-term memory recurrent neural networks. *2016 15th International Conference on Frontiers in Handwriting Recognition (ICFHR)*, pages 228–233.

Wang, Y., Xiao, W., and Li, S. (2021). Offline handwritten text recognition using deep learning: A review. *Journal of Physics: Conference Series*, 1848.

Wu, Y.-C., Yin, F., and Liu, C.-L. (2017). Improving handwritten chinese text recognition using neural network language models and convolutional neural network shape models. *Pattern Recognit.*, 65:251–264.

Xu, L., Krzyżak, A., and Suen, C. Y. (1992). Methods of combining multiple classifiers and their applications to handwriting recognition. *IEEE Trans. Syst. Man Cybern.*, 22:418–435.

Yin, F., Wang, Q.-F., Zhang, X.-Y., and Liu, C.-L. (2013). Icdar 2013 chinese handwriting recognition competition. *2013 12th International Conference on Document Analysis and Recognition*, pages 1464–1470.

Zhang, X.-Y., Yin, F., Zhang, Y., Liu, C.-L., and Bengio, Y. (2018). Drawing and recognizing chinese characters with recurrent neural network. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40:849–862.