

# Handwritten Text Recognition (HTR) for Irish-Language Folklore

**Brian Ó Raghallaigh, Andrea Palandri, Críostóir Mac Cárthaigh**

Dublin City University, University College Dublin

Ireland

{brian.oraghallaigh, andrea.palandri}@dcu.ie, criostoir.maccarthaigh@ucd.ie

## Abstract

In this paper we present our method for digitising a large collection of handwritten Irish-language texts as part of a project to mine information from a large corpus of Irish and Scottish Gaelic folktales. The handwritten texts form part of the Main Manuscript Collection of the National Folklore Collection of Ireland and contain handwritten transcriptions of oral folklore collected in Ireland in the 20th century. With the goal of creating a large text corpus of the Irish-language folktales contained within this collection, our method involves scanning the pages of the physical volumes and digitising the text on these pages using Transkribus, a platform for the recognition of historical documents. Given the nature of the collection, the approach we have taken involves the creation of individual text recognition models for multiple collectors' hands. Doing it this way was motivated by the fact that a relatively small number of collectors contributed the bulk of the material, while the differences between each collector in terms of style, layout and orthography were difficult to reconcile within a single handwriting model. We present our preliminary results along with a discussion on the viability of using crowdsourced correction to improve our HTR models.

**Keywords:** digital folkloristics, handwritten text recognition, Irish

## 1. Introduction

The research described here took place between Oct 2021 and Mar 2022 and was carried out as part of the AHRC/IRC-funded *Decoding Hidden Heritages in Gaelic Traditional Narrative with Text-Mining and Phylogenetics* project<sup>1</sup> being conducted jointly by researchers in the University of Edinburgh, Dublin City University, Durham University, University College Dublin and Indiana University. The overarching goal of the larger project is to collate and analyse a large number of the collected Gaelic folktales of Scotland and Ireland with a view to better understanding the joint cultural history of these two countries.

The Scottish component involves digitising material held in the School of Scottish Studies Archive in the University of Edinburgh and the Irish component involves digitising material held in the National Folklore Collection in University College Dublin. Once compiled, the Scottish and Irish corpora will be normalised and combined by the project team for analysis. While handwritten text recognition (HTR) for both the Scottish corpus and the Irish corpus is being carried out using Transkribus, a platform for recognising historical documents, this paper will focus only on the creation of the Irish corpus.

## 2. Irish-Language Folktale Corpus

The National Folklore Collection of Ireland is housed in University College Dublin and comprises several collections of material compiled by the Irish Folklore Commission and its successors during the 20th century (Almqvist 1977–9), for example the Schools' Collection and the Main Manuscript Collection (MMC). The *Dúchas* digitisation project,<sup>2</sup> which is running since 2012 (Ó Cleirín et al. 2014), has scanned and indexed the entire Schools' Collection (c.450k pages) and transcribed much of it via a crowdsourcing initiative. The *Dúchas* project has started digitising and indexing material from the MMC as well as the Audio Collection. The MMC is substantial and consists of 2,400 bound volumes comprising c.700k pages

of material. The Decoding Hidden Heritages (DHH) project will supplement the work of the *Dúchas* project by scanning and converting 100 volumes (c.40k pages) of the MMC to text, focusing on volumes containing folktales in Irish.

Despite the success of the crowdsourcing initiative to transcribe the Schools' Collection on a number of levels (e.g. c.400k pages transcribed, active learning resource, positive user engagement, etc.), it was obvious given the advancement of AI-powered transcription tools that it would be incumbent upon us to use semi-automatic techniques to transcribe the MMC to create our Irish-language folktale corpus for the DHH project.

## 3. Transkribus

The software being used to automate transcription of texts from the MMC is Transkribus (Sánchez et al. 2014), 'a comprehensive platform for the digitisation, AI-powered text recognition, transcription and searching of historical documents - from any place, any time, and in any language.'<sup>3</sup> The program allows users to train unique AI-powered text-recognition models that can quickly reach relatively-low character error rates (CER) that yield automated transcriptions from handwritten manuscripts. Transkribus also offers the function to create a language model based on your transcription data which can further reduce the CER, and especially the word error rate (WER), of the automated transcription.

Transkribus offers a number of tools and functions that can be used to transform images of handwritten documents into text, which include: tools for the manual and automatic segmentation of a document image, called 'Layout Analysis'; a console for manual transcription adjoining the image of the document; a tool to train new models based on your transcription data; a function to run your model to automatically transcribe any number of pages; the option of creating a Language Model (LM) based on your training data or to upload one from elsewhere, enhancing the performance of the HTR model; a function to mark which user has corrected any number of pages; tools to compare

<sup>1</sup> <https://www.gaois.ie/en/about/decoding-hidden-heritages>

<sup>2</sup> <https://www.duchas.ie/en>

<sup>3</sup> <https://readcoop.eu/transkribus/?sc=Transkribus>

and evaluate the efficiency of any number of HTR models on a given text; tools to search your document once it has been transcribed. All data is stored on Transkribus' cloud service and users can create 'Collections' in which large numbers of documents can be managed simultaneously.

Training models in Transkribus that produce a CER of ~5% is a relatively rapid and straightforward process. We found that a first rough model, giving CERs of ~10%, could be trained with only 50 pages of transcription data from the MMC. After this the law of diminishing marginal returns applied, whereby any additional production in data resulted in progressively smaller increases in output. Generally speaking, once a CER of ~5% was reached, the additional production in data necessary to further reduce the CER in a meaningful way began to reach unworkable levels for our small team (one full-time postdoctoral researcher and one full-time postgraduate research assistant), and other strategies were discussed in order to further improve the models in the future. For example, our most recent model for Seosamh Ó Dálaigh (one of the MMC collectors) was trained on 558 pages of manuscript, 69,457 words, and produced a CER of 4.39% and a WER of 12.43%. The model prior to this had been trained on 396 pages of manuscript, 49,078 words, and had yielded a CER of 4.69% and a WER of 13.61%. Therefore, this labour-intensive 41.5% increase in the training data only resulted in a 0.3% reduction of the CER and a 1% decrease of the WER.

On the other hand, more promising progress was made with other models that used less transcription data and much was found to depend on the general legibility and orderliness of each individual scribe.

## 4. Handwritten Text Recognition on the Main Manuscript Collection

The MMC presents two main challenges to HTR technology:

### 4.1 Dialect Variation

Most collectors involved in creating the MMC placed particular emphasis on remaining as close as possible to their informants' speech in their transcriptions. This approach was exhibited by Séamus Ó Duilearga himself, who founded the Irish Folklore Commission in 1927, in *Leabhar Sheáin Í Chonaille* (1948) and is described in the introduction to that work.

Ní raibh ionnam ach úirlis sgríte don tseanachaí: níor atharuíos siolla dá nduairt sé, ach gach aon ní a sgrí chò maith agus d'fhéadfainn é.<sup>4</sup>

Similarly, transcribers working for the Commission took great care to capture the dialects of their informants and in some cases we even find representations of pronunciation tendencies unique to individual speakers. For example, forms such as *do* replacing *go*, e.g. *dubhairt sí do raibh sí do maith*, appear in Seosamh Ó Dálaigh's transcriptions of a number of informants from West Kerry,<sup>5</sup> spellings such as *cén chaoi a ngohat sí* for *cén chaoi a ngabhfadh sí* are

used by Liam Mac Coisdealbha in Connemara,<sup>6</sup> and spellings such as *thenaic* for *tháinig* or *órc* for *amharc* are used by Liam Mac Meanman in transcribing speakers from West Donegal.<sup>7</sup>

While this feature of the MMC makes it a valuable resource for the study of twentieth century Irish dialects, this rich variation in linguistic forms makes the collection unsuitable to a general Irish Language Model (LM) that could assist the HTR. Indeed, an LM trained on the transcription data from the entire corpus would result in forms like *órc* or *ghohat sí* appearing in regions where those are not the pronunciations because of suggestions from the LM assisting the HTR. Similarly, given the uniqueness of spellings found throughout the MMC, the manuscripts' display of dialect variation would risk being lost to another Irish LM if this were to be uploaded from a dictionary or a corpus of printed texts.

Preliminary data compiled from Scottish Gaelic manuscripts and shared with us by researchers in the University of Edinburgh collaborating on the project showed that scribe-specific models that used an LM from the training data yielded the best results, i.e. produced the lowest CERs. For this reason and because of the linguistic nature of the MMC we decided to begin training a series of scribe-specific HTR and language models for the most prolific collectors involved in the gathering of *seanscéalta* 'folktales', the narrative form on which the project focuses, so that Transkribus could yield more accurate transcriptions that required less correction time.

### 4.2 Code and Script Switching

Another challenge the MMC presents to HTR technology is the switching between Irish and English in most manuscripts of the collection, which is also usually reflected in a change in script, i.e. whereas Irish is usually written in a form of Gaelic script, English words are usually written in a form of cursive. This practise of using a different script to write non-Irish words is old in Gaelic tradition and can be found in Early Modern manuscripts as well, therefore this is a broader challenge that will face the application of HTR on Irish manuscripts more generally in the future.

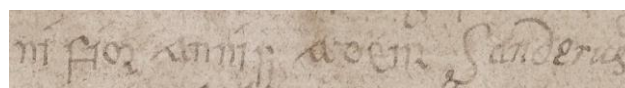


Figure 1: An example of script switching in a seventeenth-century copy of Keating's *Foras Feasa ar Éirinn* written by Iollan Ua Maolchonaire (RIA MS 23 O 19, fol. 90): *ní fíor an ní sin adeir Sanderus*. 'That is not true according to Sanderus'.

In the case of the MMC specifically, preliminary data suggests a correlation between collectors with high frequency in script switching and models with high CER levels, i.e. low accuracy. Manuscripts written by Seán Ó Flannagáin, which contain a number of macaronic texts, are a case in point, for whom we have struggled to reduce CER levels to below 10%. In the following example, a comparison of the capital *d* in *d'fhiarthuigh* and *dad*, the *f* in *fhiós* and *five*, the *s* in *sé* and *six*, the *r* in *dubhairt* and

<sup>4</sup> 'I was only a writing tool for the story teller: I didn't change a single syllable that he uttered, instead writing everything as accurately as I could.'

<sup>5</sup> MMC MS 242, p. 548.

<sup>6</sup> MMC MS 157, p. 29.

<sup>7</sup> MMC MS 168, p. 18.

or, the *g* in *geárr* and *night* and the *t* in *acht* and *night* gives some measure of what this scribe's HTR model is contending with.

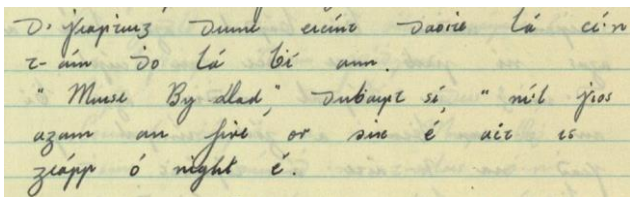


Figure 2: MMC MS 354, p. 207.

D'fhiarthuigh duine eicint daoithe lá cé'n t-ám dho lá bhí ann.  
 'Muise By Dad,' dubhairt sí, 'níl fios agam an five or six é acht is geárr ó night é.'<sup>8</sup>

One solution to this issue is to omit pages with large amounts of script switching, such as this one, from the training data in the hope of improving recognition of the Gaelic script. But in most cases script switching is confined to single words and is distributed so evenly throughout the pages of the MMC that large portions of data would end up being discarded only to filter out a handful of English words. This process would also produce a HTR model disproportionate to the language of the corpus, since many of these English words are integrated so seamlessly into the grammar of Irish that they form an integral part of its linguistic fabric, as shown by the following example from one of Tadhg Ó Murchadha's manuscripts where lenition (which occurs as a consonant mutation in Irish and is signified by a dot over a consonant letter in Gaelic script) is marked on the English word *practice* following the Irish word *aon*.

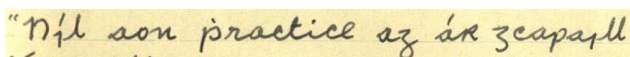


Figure 3: MMC MS 145, p. 14.

'Níl aon *phractice* ag ár gcapaill'

Keeping the English words in the training data remains the only option available for the moment and the hope is that the resulting AI-powered language and HTR models learn to cope with them. More often than not, however, these are mistranscribed, as shown by the following example from a Seosamh Ó Dálaigh manuscript containing the common Irish sentence *tá sé alright* 'it's alright', which was transcribed by a HTR model with a CER of 4.69% using a transcription-data LM as *tá sé aige*.

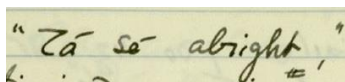


Figure 4: MMC MS 242, p. 548.

<sup>8</sup> Someone asked her what time of day it was. 'Well by dad,' she said 'I don't know whether it's five or six, but it won't be long till night.'

## 5. Method

### 5.1 Selecting Collectors

Since DHH is primarily concerned with the folktales in the MMC, the project also offers the Dúchas digitisation project a system for prioritising the transcription of material from the MMC, which consists of *c.*700k pages.<sup>9</sup> In conjunction with the team in the National Folklore Collection in UCD, a list was drawn up of the most prolific field workers involved in the collection of *seanscéalta*, the narrative form that is to be the core focus of the DHH project.

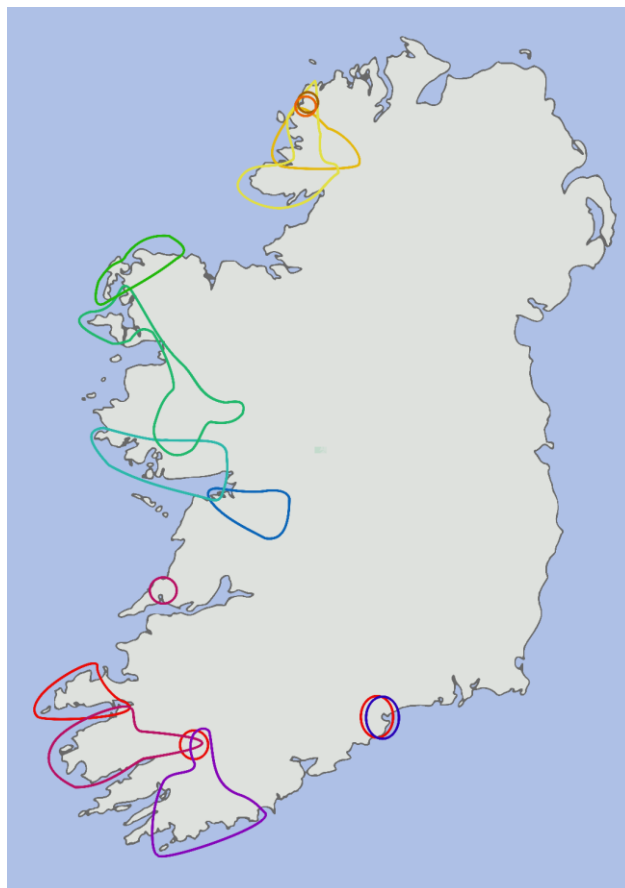


Figure 5: Core areas covered by the 12 Irish Folklore Commission field workers chosen by the project.

Aodh Ó Domhnaill		Liam Mac Coisdealbha	
Aodh Ó Duibheannaigh		Seán Ó Flannagáin	
Liam Mac Meanman		Seosamh Ó Dálaigh	
Seán Ó hEochaidh		Tadhg Ó Murchadha	
Pádraig Bairéad		Seán Ó Cróinín	
Proinsias de Búrca		Níoclás Breatnach	

In compiling this list special care was taken to ensure that as many areas of Ireland that were Gaelic speaking at the time the MMC was compiled were duly represented. The list was narrowed down to 12 full-time collectors who worked for the Irish Folklore Commission. These 12 collectors and their fieldwork areas are shown in Figure 5.

<sup>9</sup> <https://www.duchas.ie/en/info/cbe>.



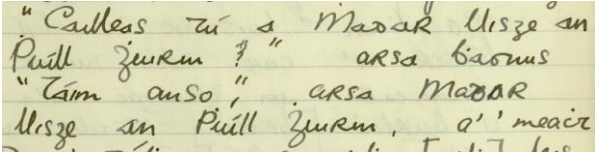
## 5.2 Digitisation and Transcription

As was described in the introduction to this paper, the digitisation and indexing of documents from the MMC had already begun under the Dúchas project and all manuscripts used to train the HTR models for the collectors listed above had already been digitised by the time the DHH project started in Oct 2021. Therefore, since the project already had access to a large number of digitised manuscripts from the MMC, the next steps of the methodology were all implemented using Transkribus, as follows.

1. Importing Document images into Transkribus.
2. Running the ‘Layout Analysis’ (LA), i.e. automatic segmentation of the Document images:
  - a. An automatic correction of the LA using the ‘merge small text lines’ widget was necessary in some cases.
3. Manual transcription of the Document, proofreading and marking of revised pages as ‘Ground Truth’ in the Document Manager.
4. Training a first model on c.50 transcribed pages, keeping 10 aside as a fixed validation set.
5. Evaluating the HTR model using the fixed validation set:
  - a. Running the HTR model on the fixed validation set produced in step 4. At this stage you can choose to use a LM from the training data.
  - b. Using the ‘Compare’ tool to produce accurate CERs and WERs.
6. Running the model on a set number of pages using the LM, about the same amount that was transcribed manually in Step 3.
7. Correction of the automated transcription produced in step 6, and marking of revised pages as Ground Truth in the Document Manager.
8. Training a new model on the increased data set.
9. Repetition of steps 5–8 until a model with satisfactory CERs and WERs is obtained.

## 6. Results

As of Mar 2022 eight scribe-specific HTR models have been trained using the method outlined above. The results of this work are presented in Table 1 which shows the size of the training data as a word count, beside the CERs and WERs of the latest model. A total of 234,693 words have been transcribed so far, the average CER produced by our models is 4.9% and the average WER is 12.5%.

Collector	#Words Transcribed	CER	WER
Seosamh Ó Dálaigh	69,457	4.39%	12.43%
 <p>UCD CBÉ MS 242 p. 30</p>			
Seán Ó hEochaidh	65,975	3.98%	6.1%

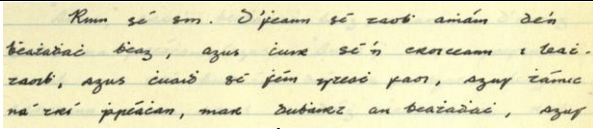
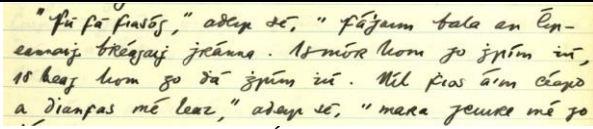
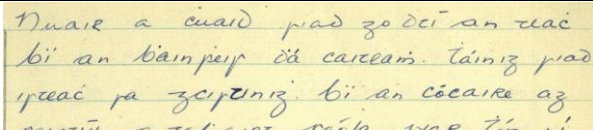
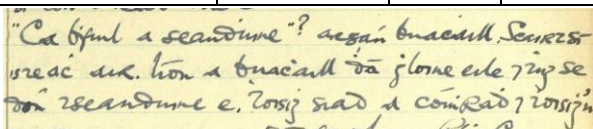
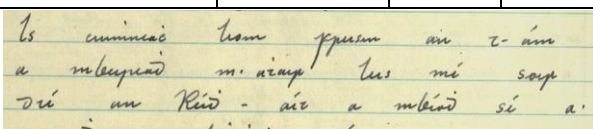
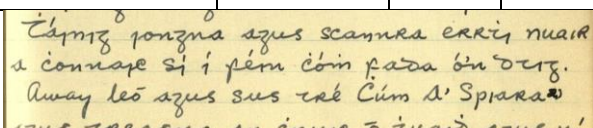
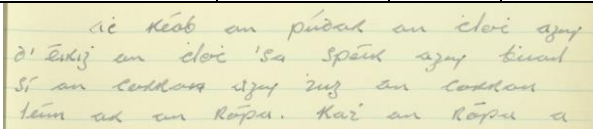
 <p>UCD CBÉ MS 139 p. 30</p>			
Liam Mac Coisdealbha	34,758	2.17%	2.66%
 <p>UCD CBÉ MS 157 p. 31</p>			
Proinsias de Búrca	24,654	4.49%	13.74%
 <p>UCD CBÉ MS 161 p. 31</p>			
Liam Mac Meanman	14,736	5.62%	17.97%
 <p>UCD CBÉ MS 168 p. 30</p>			
Seán Ó Flannagáin	9,378	10.28%	26.26%
 <p>UCD CBÉ MS 354 p. 200</p>			
Tadhg Ó Murchadha	8,102	3.59%	7.47%
 <p>UCD CBÉ MS 145 p. 30</p>			
Aodh Ó Duibheannaigh	7,633	4.28%	13.54%
 <p>UCD CBÉ MS 370 p. 29</p>			

Table 1: Results of the HTR models trained in Transkribus as of March 2022.

## 7. Discussion

Having successfully built models that give us a CER of <5% for 6 of our 12 collectors, we are satisfied that we will be able to do the same for the remaining 6. We are reasonably confident that we will be able to use the resulting textual representation of the manuscript writings to carry out digital folkloristic research on the folktales that occur in the dataset. We are satisfied that our approach of developing multiple HTR models (i.e. one for each collector's hand) was appropriate for obtaining a reasonably accurate transcription of a large quantity of data in multiple hands, within a short timeframe and with finite resources. In addition, as most subsequent processing can be automated, managing multiple HTR models will not be a burden. While a CER of <5% is satisfactory, a CER <2% is the ultimate goal, however, much of the errors we are seeing at *c.*5% are minor or are related to punctuation.

Other errors, such as the ones caused by the code and script switching described above, may continue to be a problem regardless. The law of diminishing returns means we are unlikely to reduce the CER much further with the resources and time we have, particularly for the more challenging handwriting styles. With this in mind, we are proposing to harness the resources of Meitheal Dúchas.ie,<sup>10</sup> a crowdsourcing initiative that was successfully utilised to transcribe the NFC Schools' Collection on Dúchas. We plan to carry out a pilot project where we will invite Meitheal members to correct MMC material which has been automatically transcribed using Transkribus. Researchers on the Transcribe Bentham project reported that volunteers were reluctant to switch from transcribing material from scratch to checking fellow volunteers' transcriptions (Causer et al. 2018). In our case, they will not be correcting human transcriptions, but we nonetheless expect less enthusiasm for correction over transcription. We want to test this hypothesis, and are also hopeful that enough volunteers will be sufficiently motivated to correct enough material to help us improve the HTR substantively. MMC material outside the scope of this project in both Irish and English will be made available to transcribe from scratch, so volunteers will have a choice.

Material being processed using Transkribus is stored on Transkribus Servers and is accessible via the Transkribus REST API. Dúchas material is stored on Dúchas Servers and is accessible via the Dúchas REST API. Dúchas images are stored in Azure Blog Storage in the Microsoft Cloud and are accessible via the Azure Blob service REST API. We plan to automate the steps below (if possible) with a Python script that will utilise the Transkribus REST API, the Dúchas REST API, and the Azure Blob service REST API, as well as other interfaces available to us as DHH and Dúchas administrators. API operations or endpoints are given in parentheses where possible. For each of up to 10 MMC volumes collected by each of the 12 collectors in Figure 5 (we have Transkribus credits available to us for HTR on *c.*40k pages and there are *c.*400 pages per volume) in which there is a substantial quantity of folktales, we will execute the following steps on each volume iteratively:

1. Create Document (/collection) in Transkribus within DHH Collection.

2. Upload (/uploads) volume pages (i.e. one image file per page) from Dúchas blob storage (Get Blob) to Transkribus Document.
3. Run Layout Analysis (/LA), Short Line Merge and HTR (/recognition) using scribe-specific model on Document.
4. Export Document to TXT format (i.e. one text file per page).
5. Import transcription text files into Dúchas and map to Dúchas metadata (which is being compiled by the DHH and Dúchas teams within the Dúchas system).
6. Make transcriptions available to the Meitheal Dúchas.ie crowd volunteers for correction.
7. Get corrected transcriptions from Dúchas (/api/{version}/cbe/?VolumeNumber={volume}) and import into Transkribus using TextToImage.
8. Retrain (/recognition) HTR model.

Given the full size of the dataset (12 collectors = 697 volumes, i.e. *c.*278,800 pages) and corresponding HTR cost implications, we plan to filter out volumes containing material other than folktales prior to recognition, and we will only process as many of the volumes containing folktales as is feasible within our timeframe and budget. We do not intend, however, to exclude individual pages within volumes from the recognition process. This is to simplify the administrative burden that partially transcribed volumes would create for the Dúchas team. This approach might be adapted should it become feasible to perform HTR on volume sections or even individual items (i.e. folktales) within volumes. Once the above steps are completed on 100–120 volumes, we will run An Caighdeánaitheoir<sup>11</sup> on the output and send the standardised texts along with associated metadata forward for text-mining and phylogenetic analysis.

## 8. Conclusion

In this paper we introduced the Decoding Hidden Heritages project which aims to carry out a deep analysis of the narrative traditions of Scotland and Ireland by analysing a large text corpus of Irish and Scottish Gaelic folktales using computational methodologies. This paper focused on the Irish component of the initial corpus creation phase of the project. We described how we are using the Transkribus software to carry out handwritten text recognition on a large number of scanned manuscript pages from the National Folklore Collection, and illustrated the difficulties we encountered in dealing with dialect variation, code switching and script switching that occur throughout the manuscript pages. We presented our methodology in which we are producing individual recognition models for each scribe. This was motivated by the significant interscribe variability in terms of style (e.g. letter size, angle), layout (e.g. spacing) and orthography (e.g. punctuation), but also by the fact that a manageable number of collectors (i.e. 12) would provide us with sufficient dialectal and folkloristic coverage for our study.

Our research so far indicates that Transkribus works extremely well at recognising historical documents, handwritten Irish-language texts in our case. A CER of <5% was achieved for six of eight different HTR models

<sup>10</sup> <https://www.duchas.ie/en/meitheal/>

125<sup>11</sup> <https://github.com/kscanne/caighdean>

trained so far by manually transcribing or correcting an average of 30,000 words per model. These 8 models would allow us to transcribe up to 568 MMC volumes, the volumes handwritten by these 8 full-time folklore collectors whose handwriting we have so far modelled individually, should we wish to do so. This would give us fulltext access to *c.*227,200 pages of folklore material, thus enabling us to carry out the next stage of our research where we will investigate convergence and divergence in the narrative traditions of Scotland and Ireland using text-mining and phylogenetics. Moreover, this work will also feed back into the Dúchas project in its efforts to fully digitise the collections of the NFC. The Dúchas project already provides fulltext search of much of the Schools' Collection but is yet to provide the same for the MMC. This research will lay down the foundation for this to be achieved.

## 9. Acknowledgements

This work was supported by the Arts and Humanities Research Council (Grant no. AH/W001934/1) and the Irish Research Council (Grant no. IRC/W001934/1). The Dúchas project is funded by the Department of the Gaeltacht (Government of Ireland).

## 10. References

- Almqvist, B. (1977–9). The Irish Folklore Commission: achievement and legacy. *Béaloideas* 45–7:6–26.
- Causer, T., Grint, K., Sichani, A. M., and Terras, M. (2018). Making such bargain: Transcribe Bentham and the quality and cost-effectiveness of crowdsourced transcription. *Digital Scholarship in the Humanities*, 33(3):467–87.
- Ó Duilearga, S. (1948). *Leabhar Sheáin Í Chonaill*, p. xxiv. Dublin: The Educational Company of Ireland Ltd.
- Ó Cleircín, G., Bale, A., and Ó Raghallaigh, B. (2014). Dúchas.ie: ré nua i stair Chnuasach Bhéaloideas Éireann. *Béaloideas*, 82:84–100.
- Sánchez, J. A., Romero, V., Toselli, A. H., and Vidal, E. (2014). ICFHR2014 Competition on Handwritten Text Recognition on Transcriptorium Datasets (HTRtS). In *14th International Conference on Frontiers in Handwriting Recognition*, pages 785–790, Crete, Greece, Sep, doi: 10.1109/ICFHR.2014.137.