Journal of Open Humanities Data

Applying OCR to Bilingual Irish-English Texts in *An Gaodhal* (1881–98)

COLLECTION: DIGITAL RESOURCES FOR THE LANGUAGES IN IRELAND AND BRITAIN

DISCUSSION PAPER

DEIRDRE NÍ CHONGHAILE
OKSANA DEREZA
NICHOLAS M. WOLF
*Author affiliations can be found in the back matter of this article

ubiquity press

ABSTRACT

Conducting computerised text extraction for the Irish language (Gaeilge) faces a number of challenges, the most significant of which is the machine-readability of Cló Gaelach, the typeface most commonly used in handwritten and printed Irish-language material before the 1960s. Further difficulties include: variable spelling and grammar in pre-standardised Irish-language texts; the idiosyncrasies of artisanal letterpress printing; and a diversity of genres, contributors, dialects, and vocabulary found in historical texts of the time. This paper discusses how a team at New York University and the University of Galway used the text-recognition software Transkribus to develop two new optical character recognition (OCR) models: a Gaeilge-only model and a bilingual Gaeilge-English model. The core dataset for this OCR training exercise is the Brooklyn-based bilingual monthly newspaper *An Gaodhal* (1881–98), the first serial dedicated to providing content to an Irish-language readership, which was established, edited, and printed by Micheál Ó Lócháin. The *An Gaodhal* project produced a machine-readable dataset and, with that, full-text searchability of a free and open digital edition of the newspaper, hosted by University of Galway Library. This paper considers the technical requirements and labour that ensure effective, efficient applications of OCR technology to historical Irish-language texts. It also explores the prospects for developing future OCR models for languages printed in historical, multilingual, and multiscript settings.

CORRESPONDING AUTHOR:
Nicholas M. Wolf
New York University Libraries, New York University, New York, USA
nicholas.wolf@nyu.edu

KEYWORDS:
optical character recognition (OCR); Cló Gaelach; Irish; Gaeilge; bilingual data

TO CITE THIS ARTICLE:
Ní Chonghaile, D., Dereza O., & Wolf, N. M. (2026). Applying OCR to Bilingual Irish-English Texts in *An Gaodhal* (1881–98). *Journal of Open Humanities Data*, 12: 145. DOI: https://doi.org/10.5334/johd.621

1 CONTEXT AND MOTIVATION

1.1 OCR AND IRISH HISTORICAL TEXTS

The Irish language experienced a significant change in 1963–64 when the national education system in Ireland jettisoned its native Cló Gaelach script in favour of roman script (Ó Murchú, 1977). Since then, historical texts handwritten or printed solely in Cló Gaelach – in different typefaces e.g. Petrie, Newman (McGuinne, 1992) – remain inaccessible to those who cannot read it. The advent of machine learning presents an opportunity to bridge the gap of comprehension that has grown in the decades since. Applying optical character recognition (OCR) to Irish-language texts printed in Cló Gaelach promises to enable a necessary transliteration at scale and with greater efficiency, thus ensuring older texts are not ignored. It also enables the migration of datasets captured in print to digital environments in which textual and corpus analysis tasks can be carried out more easily and the potential of linked data can be explored.

In the present case, the application of this particular digital approach to historical texts in the Irish language combines with the study of the history of journalism and media, specifically the story of the world's first Irish-language newspaper, the bilingual Irish-English monthly *An Gaodhal* (1881–98). Although it was published in Brooklyn, New York, the only surviving complete set of *An Gaodhal* is held in the University of Galway Library. Consequently, a research partnership was formed between New York University and the University of Galway to produce a fully searchable digital edition of the newspaper, the first of its kind for any newspaper printed in the Cló Gaelach script. This edition both supplements and expands on previous projects focused on historical Irish corpora. These include, most notably, the extraction of Cló Gaelach text from printed artefacts that has been presented independently of images of those artefacts as a digital corpus in *Corpas Stairiúil na Gaeilge 1600–1926* (Acadamh Ríoga na hÉireann, 2017)[1] and monolingual English-language OCR models that have been applied to printed bilingual Irish-English works featuring both Cló Gaelach and roman scripts (e.g. national, provincial, and radical newspapers in Irish Newspaper Archives[2]). The latter strategy tends to produce unintelligible outputs for the printed Irish-language content therein, an outcome that is symptomatic of the persistent under-resourcing and minoritisation of this endangered language.

1.1.1 Publicly Available OCR Models

Although pre-existing training datasets for Irish-language Cló Gaelach characters were extant at the outset of the *An Gaodhal* project in January 2023 (Scannell et al., 2020), these had not been trained with bilingual texts in mind. With *An Gaodhal* presenting material in both Irish and English, often on the same page and sometimes in the same line, there was concern that a previous set of training data focused solely on Irish would prove unsuitable for the proposed bilingual corpus. The team thus elected to build OCR models from scratch – one monolingual and single-script (Transkribus model #61350), and one bilingual and multiscript (Transkribus model #501397) – and release them publicly for others to use. These additional outputs would enable others to adopt digital approaches to other Irish-language datasets.

The team began by building a model focusing on the Irish-language content, which was printed in Cló Gaelach and used pre-standardised spelling. That model was then used to develop a second, bilingual and multiscript model that could recognise both Irish and English in their respective typefaces, Cló Gaelach and roman letters. For that purpose, training data from Irish-only pages and from the newspaper's English-language content were combined. Separately, in November 2023 Gerard Farrell (2023) released a monolingual multiscript OCR model for Irish-language texts in either Cló Gaelach or roman scripts (Transkribus model #56262).[3] As of 30 January 2026, there are three OCR models that can read Cló Gaelach available to the public via Transkribus: two for monolingual settings; and one for bilingual Irish-English settings. Each of these models differs in significant ways (see Dereza et al., 2024, for more detail).

1 https://corpas.ria.ie, last accessed 13 October 2026.

2 https://irishnewsarchive.com, last accessed 13 October 2026.

3 https://www.transkribus.org/models/irish-gaelic-and-roman-type-seanchl-agus-cl-rmhnach-v3, last accessed 13 October 2026.

1.1.2 Choosing the Dataset

An Gaodhal was an appropriate choice of dataset for the present project, not least for embodying its own pioneering emergence in a post-Famine, diasporic context, but also for its idiosyncratic form and content, which together presented a welcome diversity of challenges for the development of robust OCR models for Cló Gaelach. Additional motives for choosing *An Gaodhal* included:

1. at 2,290 pages typically featuring up to 120 page-width lines (depending on pica size) set in two columns and up to 800 words on each page, the 147 issues of the newspaper presented a workable quantity with which to begin training OCR models;
2. the newspaper was out of copyright;
3. the rights to the surrogate digital images supplied for the OCR training were held by the University of Galway, which kindly granted permission for their use;
4. *An Gaodhal* can be regarded as a foundation stone for the modern media landscape of the Irish language and, as such, merits further exploration; and
5. the Galway run of *An Gaodhal* is especially valuable because it forms part of a wider collection of internationally significant material in the Rev. Daniel J. Murphy Collection at the University of Galway Library,[4] the majority of which consists of handwritten and typed manuscripts documenting more than 1,200 sean-nós songs transcribed in Pennsylvania that, in the course of his life-long song scholarship, Rev. Murphy indexed and referenced in marginalia that occurs throughout his copies of *An Gaodhal* (Ní Chonghaile, 2015). The team's wish to capture Rev. Murphy's handwritten marginalia whilst applying OCR to printed text presented an additional layer of complexity to the project.

1.1.3 Choosing Transkribus

READ-COOP's Transkribus[5] supports historical text recognition for printed or handwritten material by providing custom model training along with a robust editing interface. The ease with which this technology could be learned and applied, combined with a welcome range of features, made Transkribus the optimum choice, especially as the resources of time and labour available to the project were limited. Moreover, Transkribus has a proven track record in recent years in enabling collaborative OCR training and corrections for humanities projects (Muehlberger et al., 2019; Nockels et al., 2022), for multilingual and multiscript data in particular (Capurro et al., 2023; Miyagawa, 2023).

1.1.4 Purpose of Discussion Paper

This paper discusses the manner in which the newspaper's character – including its content and printed layout – determined the project methodology, the challenges of training OCR models, how the digital infrastructure on which the respective institutional partners rely impacted the project design, and how these factors influenced the resulting dataset. Its purpose is to share some project learnings and so help guide future work in this field.

1.2 CHARACTERISTICS OF *AN GAODHAL*

The *An Gaodhal* dataset amounts to 1.8 million words. The distribution of the two languages across its 2,290 pages presents as follows: 381 Irish mostly, 896 English mostly, and 1,019 with both languages together on the same page. The pages were typically set in two columns. In terms of linguistic features, spellings varied considerably according to the author of each contribution, reflecting Ó Lócháin's commitment to acknowledging dialectal differences and variable vocabulary and to supporting the efforts of contributors seeking to improve their command of the written language. This might create difficulty for some users of the dataset seeking a relatively standardised version of the historical language: compared with, for example, a contemporary printed book in Irish, Ó Lócháin's renderings exhibit greater diversity. As

4 https://archivesearch.library.universityofgalway.ie/index.php/rev-murphy-collection, last accessed 13 October 2026.

5 https://www.transkribus.org, last accessed 13 October 2026.

for content, the newspaper featured articles by the editor and others on current affairs in Ireland, the United States, and elsewhere; letters to the editor; poetry; songs and folkloric material collected and submitted by readers; and texts reprinted from published works or other newspapers including the *Brooklyn Daily Eagle*. It also included lists of subscribers' names and addresses at intervals and text-based and image-based advertisements in almost every issue.

2 DATASET DESCRIPTION

REPOSITORY LOCATION

https://doi.org/10.58153/7s7n3-58a49

REPOSITORY NAME

New York University UltraViolet Research Repository

OBJECT NAME

An Gaodhal Newspaper (1881–98) Full-Text OCR Output Files

FORMAT NAMES AND VERSIONS

ALTO XML, PAGE XML, TXT

CREATION DATES

2023-01-03/2023-12-14

DATASET CREATORS

Deirdre Ní Chonghaile (New York University, University of Galway): data curation, funding acquisition, methodology, resources, validation; Oksana Dereza (University of Galway): data curation, methodology, resources, validation; Nicholas Wolf (New York University): data curation, funding acquisition, methodology, resources, validation.

LANGUAGE

Irish (Gaeilge), English, Old Irish, Scottish Gaelic (Gàidhlig), Latin

LICENCE

Creative Commons Attribution Share Alike 4.0 International

PUBLICATION DATE

v.1: 2023-12-14; v.2: 2026-04-23

3 METHODS

3.1 TEXT EXTRACTION DECISION POINTS

Early on, a key curatorial decision rested on the question of how the extracted text might represent the printed Cló Gaelach characters. Equivalent Unicode characters were selected to represent a characteristic typographical feature of Cló Gaelach, that is, dotted consonants (e.g. Ḃ, Ċ, Ḋ, ḃ, ċ, ḋ), which indicate the linguistic feature known as lenition that is deployed in the Irish language. The team felt there was an opportunity to convey the contentious history of the typeface itself; the choice between Cló Gaelach characters and roman characters was so ideologically charged among Irish speakers in the nineteenth and early twentieth centuries, imbued with questions of cultural and political authenticity, that it merited representation. In the case of *An Gaodhal*, capturing Ó Lócháin's choice of characters would support future analysis, so the team chose to produce as diplomatic a rendering as possible.

Other curatorial decisions followed more practical needs, such as substituting the ampersand symbol (&) for the use of the number 7 (similar in shape to the intended Tironian symbol ⁊ for the Latin "et" used commonly in Irish-language texts) to avoid ambiguity. As with all transcription projects, the team settled on conventions for how to represent hyphens, dashes, and line-break continuation indicators, aiming to minimise complexity for users of the extracted text and smooth the usual inconsistencies in representing such characters in nineteenth-century newspaper typesetting.

3.2 TEXT EXTRACTION PROCESS

3.2.1 Digital Image Resolution

As OCR requires a minimum 15 px per letter for optimum output, the number of lines and dimensions of a page of text will dictate the necessary resolution. For its source images, the project used digital surrogates of *An Gaodhal*, specifically JPG images scanned at 600 dpi at a resolution of approximately 4,000 × 6,000 px. These were provided to the team by University of Galway Library.

3.2.2 Preparing Images for OCR Monolingual Model Training

To train the Irish-only model, the team sought an efficient means of filtering out text in English, whether on entire pages or in a single line. This was done by using pre-existing and widely available tools that recognise English-language tokens (e.g. Transkribus, Amazon Textract[6]), tools providing fast and accurate text detection and line segmentation across both languages. A selection of 100 pages were subjected to this first step of line segmentation and token-location identification, and then an English-language recognition model was deployed on those pages. The output was then processed to classify each line by language based on dictionary-word accuracy of the extracted tokens. Pages scoring high as containing properly spelled English words were deemed to be in the English language and temporarily removed, yielding a corpus of predominantly Irish-language pages with which to train an initial model. Elsewhere, on pages that provided a mix of English and Irish lines, the English lines were masked using a shape overlay annotation process, applying computer-drawn opaque rectangles that matched the yellowed newspaper colour of the digitised page. This measure provided additional clean training data, increasing the number of Irish-language tokens for OCR training purposes.

From the filtered and masked Irish-only pages, an initial model trained on 36 pages (Transkribus model #49592, not made public) was created. This bootstrap approach enabled first-pass transcription of pages, helping to reduce the number of corrections necessary for the next stage. The team then produced a second set of transcription pages – a selection of 57 from an initial 60, which excluded images or advertisements – with the first model helping to reduce correction time. This second model served as the basis for a successful monolingual Irish-language Cló Gaelach recognition model (Transkribus #50036) that would, in conjunction with outputs produced by applying English-language models, enable the creation of a viable set of transcribed bilingual pages for further training. Throughout the process of producing the dataset, corrections were completed via single-person review by a native Irish speaker able to read Cló Gaelach script, an approach that ensured curatorial consistency while maximising budgetary efficiency.

3.2.3 Preparing Images for OCR Bilingual Model Training

Once training on this initial Irish-only model was complete and its error rate deemed low enough to deploy it more widely on the corpus, the team turned to training a model capable of handling bilingual pages. A new subset of 100 pages from the full collection was randomly selected, restoring all masked regions to present fully bilingual texts. The language profile of each of the 2,290 pages of the newspaper was identified as predominantly English, predominantly Irish, or bilingual. Pages predominantly in Irish were then processed using the intermediate Irish-only model (Transkribus #50036), while those mainly in English were processed using *Transkribus Print M1* (#39995), a pre-trained model that is based on over 5 million tokens and is capable

6 https://aws.amazon.com/textract, last accessed 13 October 2026.

of handling the historical typographical conventions of the English-language portions of the *An Gaodhal* corpus. As before, the outputs were then corrected manually.

These corrected transcriptions could, at last, be used to train a preliminary bilingual model, *An Gaodhal Irish/English Bilingual Model v.1* (#51080). Trained on 54,406 tokens, this model was then applied to the full corpus, enabling a passable set of outputs. These were corrected manually to provide training data for two final public models. It should be noted that, even with capable language models in hand, the process of extraction was still fundamentally beholden to the quality of layout analysis. The newspaper posed some layout difficulties including: two-column layouts throughout; gridded advertisements; and the start of each issue frequently featured a multi-column table presenting the Cló Gaelach alphabet to facilitate those learning to read Irish. Page by page, the associated layout analysis output had to be regularly adjusted to produce effective results. Bilingual pages predictably required the most correction, while monolingual pages required minimal correction. Cognisant of the paucity of corpora in the Irish language, however, and hoping to release models of the highest-possible accuracy, the team dedicated many months to carefully reviewing each page and making necessary corrections.

With the reviewed and corrected corpus in hand, the team could proceed to producing OCR models via Transkribus. The output from this final phase provided the necessary content to train a successful bilingual model (#501397) and a second, improved Irish-only OCR model (#61350).

4 MODEL AND DATA RELEASES

4.1 DATA

The publicly available outputs associated with this project include:

- Extracted text is available on the New York University UltraViolet repository as a series of PAGE XML, ALTO XML, and plain-text files along with page-level metadata identifying features such as advertisements, images, marginalia, and predominant language present.[7] The data are presented as a series of files, one per page, containing word- and line-level information that identifies each individual word and its coordinate location on the page. An initial set of files representing the outputs with human correction of all Irish-language pages was released as version 1 in 2023. Fully corrected outputs of all pages are released as version 2 as of 2026 under a CC BY-SA 4.0 International licence.
- For those who wish to search pages or to compare text outputs to the original digitised pages, the fully corrected transcriptions were converted from PAGE XML into the ALTO XML format used by the University of Galway Library and uploaded to its Digital Collections and Archives, together with plain-text transcripts. This enables both collection-level and item-level full-text search of the *An Gaodhal* content.[8] To support this process, a custom converter was developed, which is openly available on GitHub.[9] The search functionality is part of the Digital Asset Management System used by the University of Galway Library and is provided by the company Aetopia; thus, the team was responsible for data and testing but took no part in implementing search functionality. High-resolution images of *An Gaodhal* on University of Galway Digital Collections and Archives are offered under a CC BY-NC 4.0 International licence.
- A detailed project page on the University of Galway website,[10] providing an overview of *An Gaodhal*'s history and highlighting its significance.

Figure 1 presents an extract of text from the digitised pages, followed by a portion of the equivalent PAGE XML rendering the first two words and their coordinates.

7 https://doi.org/10.58153/7s7n3-58a49, last accessed 13 October 2026.

8 https://digital.library.universityofgalway.ie/p/ms/categories/an-gaodhal, last accessed 13 October 2026.

9 https://github.com/universityofgalway-library/page-to-alto, last accessed 13 October 2026.

10 https://universityofgalway.ie/angaodhal/, last accessed 13 October 2026.

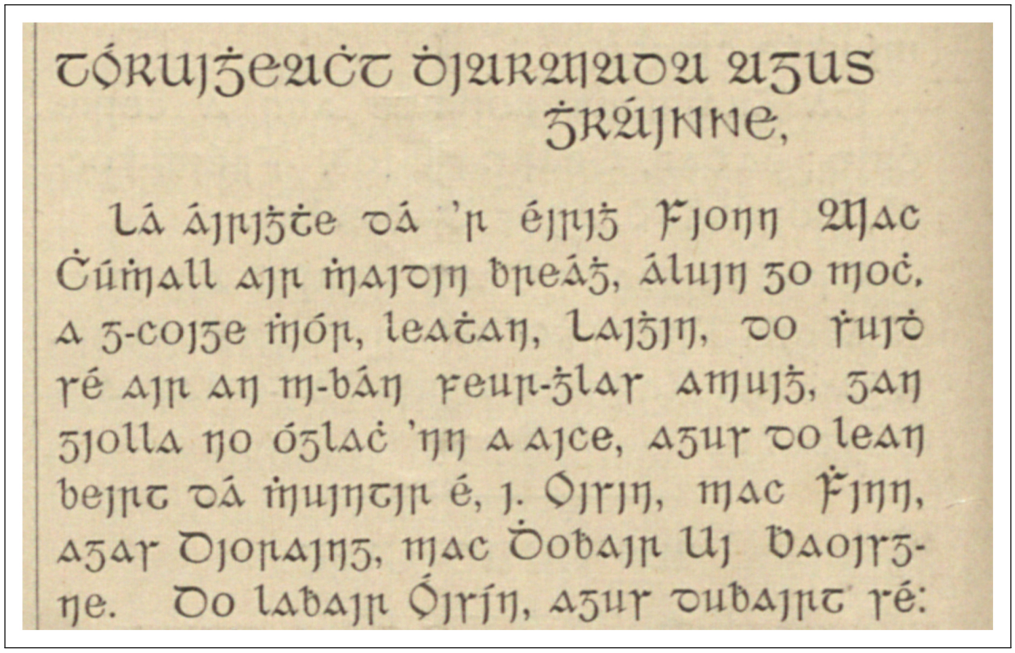


Figure 1 A snippet showing the prose tale *Tóruigheacht Dhiarmada agus Grainne* and below, an XML rendering of part of that same text. Image courtesy of the University of Galway Library.

```
<Word id="r6l1w1" custom="readingOrder {index:0;}">
    <Coords points="2159,1095 2159,1155 2756,1155 2756,1095 2159,1095"/>
    <TextEquiv>
        <Unicode>TÓRUIĠEAĊT</Unicode>
    </TextEquiv>
</Word>
<Word id="r6l1w2" custom="readingOrder {index:1;}">
    <Coords points="2826,1094 2826,1154 3344,1154 3344,1094 2826,1094"/>
    <TextEquiv>
        <Unicode>ḊIARMADA</Unicode>
    </TextEquiv>
</Word>
```

4.2 MODELS

Two finalised models resulting from the process of creating the datasets are available on the Transkribus AI Model Hub to support future work on historical Irish and bilingual texts as *An Gaodhal Gaeilge/Irish Monolingual Model* (#61350)[11] and *An Gaodhal Irish-English Bilingual Model* (#501397).[12] The final training set for the bilingual model comprised 2,067 pages, with 229 pages reserved for validation. The publicly released bilingual model, trained on 1,412,355 tokens, achieved a character error rate (CER) of 1.27%. The released Irish monolingual model was trained on 344 pages consisting of 164,015 tokens, and achieved a CER of 1.4%.[13]

5 RESULTS AND DISCUSSION

5.1 BALANCING PROJECT OBJECTIVES

5.1.1 Training OCR versus Accurate Text Extraction

In combining the two key objectives of creating OCR models from scratch and generating accurate text extraction of an idiosyncratic source, the project encountered a methodological tension between, on the one hand, producing reliable OCR models and, on the other, rendering faithful transcriptions of the printed artefacts that preserve printing errors and thus ensure

11 https://www.transkribus.org/models/an-gaodhal-gaeilge-irish-monolingual-model, last accessed 13 October 2026.

12 https://www.transkribus.org/models/an-gaodhal-irish-english-bilingual-model, last accessed 13 October 2026.

13 Because not every page consists of Irish, the number of Irish-language tokens available for model training is lower than the 1.8 million total tokens in the corpus.

historical accuracy. In such cases, the team came to rely on the positive factors that would shape the project outcomes. These included:

- the entirety of the editorial content of *An Gaodhal* and all but four of the printed issues themselves, the physical artefacts, were produced by the same person, Micheál Ó Lócháin, who showed a remarkable consistency of approach to spelling, writing, editing, typography, and typesetting over nearly 18 years of work;
- the wide variety of contributors ensured a broad range of dialectal vocabulary, spellings, and phraseology, thus increasing the adaptability and applicability of the OCR models to more localised texts; and
- the relatively small number of print anomalies was unlikely to impact the accuracy rate of a newly trained OCR model; they would, in fact, be displaced over time should the *An Gaodhal* model enable further generation of high-quality Irish-language extracted texts from other corpora.

In short, as scalability offsets the risks associated with relying on more localised texts for training OCR models, the team chose to adhere as closely as possible to the printed artefact in order to produce a dataset faithful to its source.

Progress was slower in the beginning while the team became familiar with the Transkribus technology and workflows. Some pages needed to be processed a second time. Errors were unavoidable given the short time allocated to processing 2,298 images (eight of which were relevant items inserted into the bound volumes containing a total of 2,290 newspaper pages) and manual correction of OCR text output.

5.1.2 Word-Level Correction versus Line-Level Correction

The team elected to produce word-level corrections in order to facilitate future natural language processing (NLP) work and thus maximise potential research impact. Unfortunately, it failed to anticipate the time-cost associated with that choice. Word-level corrections are slower than line-level corrections, especially where the page contains a high volume of words. In word-level correction mode, each keystroke often took nearly one second to load. Consequently, the most efficient way to make adjustments was to avoid using the “delete” key and, instead, to select the characters that needed replacing before typing the correct text. This time-cost is greater for languages that rely on virtual keyboards where physical keyboards do not provide all of the relevant characters via typed shortcuts. Such time-costs prompted careful layout analysis prior to running OCR to minimise the burden of correcting text output. Throughout the project, word-level corrections could only be applied in Transkribus’s Expert desktop client (see the spanner icon for line-level and word-level options). It remains the case that the desktop client offers more functionality than the web version of Transkribus.

5.2 KEY LEARNINGS

5.2.1 Significance of Layout Detection

Regardless of the accuracy rate of a given text model, precise layout detection proved crucial to the quality of text output and, with that, the success of the overall extraction task. This was particularly evident on pages featuring text columns, structured paragraph layouts, tables, advertisements, images, marginalia, and fine print. Consequently, the project workflow dedicated a considerable percentage of time per page toward reviewing the output of the layout analysis tools provided by Transkribus, specifically its Print Block Detection and Layout Analysis models, which were applied using default settings. Corrective work included adjusting text baselines or paragraph or region bounding boxes. For vertical text or handwritten marginalia, baselines were drawn manually to ensure accurate text line segmentation.

5.2.2 Success of Bootstrap Approach to Building Models

This process confirmed that, with a small team and limited budgetary resources, a bootstrap approach can create new OCR models from scratch. The initial application of a substandard model produces a workable, if basic, transcription of a selection of pages that serves as training data for subsequent phases of model development, thereby reducing the correction burden over time. For this project, three iterations of running a model, correcting, and then creating a new model from the improved outputs produced a successful bilingual model for Irish and English.

6 IMPLICATIONS/APPLICATIONS

6.1 USES OF THE CORPUS

Recognising recent work countering the historically "cavalier attitude towards periodical sources in Irish" (O'Leary, 2010, 152), the provision of the *An Gaodhal* dataset has the potential to enrich scholarship further and across a broad range of fields from Irish-language media studies, history, and ethnomusicology to historical linguistics, lexicography, dialectology, and corpus analytics (Uí Chollatáin, 2008; Ní Loingsigh et al., 2022). A clear sense of the linguistic make-up of the corpus will require further exploration, with challenges such as how to lemmatise pre-standard Irish spellings, distinguish Irish words on multilingual pages (and especially without confusing words such as the English indefinite article "an" with the Irish definite article "an" or the Irish interrogative particle "an"), and assembling stop-word lists for historical forms of the language. Nevertheless, a preliminary sense of the distribution of words in either language throughout the corpus (Table 1) and the frequency of non-trivial Irish words on the Irish-only pages (Table 2) invites future linguistic analysis.

PREDOMINANT LANGUAGE ON PAGE	NUMBER OF PAGES	WORD COUNT
Irish	381	205,419
English	896	872,341
Mixed	1,019	730,537

Table 1 Counts of Non-Lemmatised Words in *An Gaodhal* by Page Type.

WORD	COUNT
féin	293
lá/lae	191
duine/daoine	168
fear/fir	145
maiṫ	111
ċéile	90

Table 2 Frequency Counts of Most Common Irish-Language Words in *An Gaodhal* Found on Irish-Only Pages (per 100,000 words).

Beyond linguistic analysis, as a newspaper source rooted in the nationalist revival period of Irish history, searchable access to the contents of *An Gaodhal* enables more scholars to retrieve individual points of interest from its pages, such as contributor views on contemporary events in the 1880s and 1890s in Ireland and the United States, and thus arrive at a more holistic sense of the newspaper's achievement. It also facilitates direct access to individual poems, songs, prose, speeches, letters, and other Irish-language material that further elucidate historical points of view from the period, all the more interesting because of their appearance in a U.S. newspaper, and document the popularity of certain compositions that sustained beyond Ireland. The corpus itself, with its spelling variations and presentation of historical forms of the language, also presents relevant historical material for linguistic analysis of the evolution of Irish as well as teaching texts for language instruction focused on older forms of Irish.

6.2 POTENTIAL APPLICATIONS OF OCR FOR GAEILGE

With the creation of three OCR models for Irish-language text printed in Cló Gaelach script, either in monolingual or bilingual settings, the stage is set for further applications of these new tools. Each of the models is publicly available via Transkribus, a platform providing a range of access options from free accounts to premium subscriptions. Publishers might apply the models toward re-issuing Irish-language works that are out of print. Institutions providing digital access to written artefacts – in this case, printed Irish texts found in books, newspapers, journals, or manuscripts – might embark on wholesale OCR-led text extraction where appropriate.

Although the cost of conducting OCR work is an important factor, the question of sourcing language expertise while the language itself remains in a vulnerable position is all the more vital.

One alternative approach to text extraction – crowdsourcing volunteer manual transcription work – is often applied to manuscript material featuring variable handwriting and scripts generated by multiple scribes (for example, the project *Meitheal Dúchas*[14]). OCR work conducted by smaller groups of people is often carried out in institutional settings with team leaders overseeing the work of paid researchers or student labour.[15] More recently, some projects are combining OCR with crowd-sourcing.[16] Applying OCR at scale to Irish-language texts – for instance, to newspapers printed wholly or in part in Cló Gaelach – will require commitment and investment from institutions and funders.

6.2.1 *An Stoc* Newspaper

University of Galway Library holds a second Irish-language newspaper that both merits text extraction via OCR and resembles *An Gaodhal* in ways that would help maximise the return on such an investment. The monolingual Irish-language monthly *An Stoc* (1917–31)[17] was edited from start to finish by the university's first ever Professor of Irish, Tomás Ó Máille (1880–1938). He was just as consistent as his fellow Galwayman Micheál Ó Lócháin in his writing and editing work. Both men included folkloric material in their respective newspapers: while Ó Lócháin encouraged readers to collect and submit material (Uí Fhlannagáin, 1990), Ó Máille recruited his students to collect and submit material to *An Stoc* and to help him distribute it (Mac Peaircín, 2020). Contributors to both papers came from the same source communities in the West of Ireland and sometimes from the same families (de Hae & Ní Dhonnchadha, 1938–1940; Mac Peaircín, 2027), thus promising an unusual opportunity for comparative research on networks of correspondence, scholarship, and song transmission. While *An Gaodhal* featured a variety of dialects and variable spelling, *An Stoc* prioritised Ó Máille's native Connaught Irish in more standardised spellings in order to offer a counterbalance to the proliferation of Munster Irish found in contemporaneous publications.

An Stoc was printed commercially in the Cló Newman script, which resembles the typeface used in *An Gaodhal*. As such, the *An Gaodhal Gaeilge/Irish Monolingual Model* (#61350) is well placed to adapt to the challenge it presents. The form and layout of *An Stoc* are also comparable. Finally, as the digital infrastructure upon which the University of Galway Library relies supports the inclusion of multiple datasets in a single search, users would have the opportunity of searching both *An Gaodhal* and *An Stoc* simultaneously, thus rendering both newspapers equally accessible and gifting scholars the potential for new learnings that can come from digital approaches to research questions.

6.2.2 Conradh na Gaeilge Archive

The Conradh na Gaeilge archive at University of Galway Library[18] is another candidate for the potential application of OCR models for Gaeilge. The collection contains a good deal of material in handwritten and printed Cló Gaelach and is contemporaneous with both *An Gaodhal* and *An Stoc*. The diversity of material would likely present considerable challenges to layout analysis prior to OCR work.

6.3 DEVELOPING NER FOR GAEILGE

While research on named entity recognition (NER) for historical texts is widespread and varied (Ehrmann et al., 2023), NER for the Irish language, whether modern or historical, is relatively new. In 2022, the Irish government observed: "To date, there is no named entity recognition system available for Irish. There are some basic resources (lists of named entities) available through the part-of-speech tagger technology, and place names at logainm.ie but much work is required to extend this research into a comprehensive NER tool" (Government of Ireland, 2022). In recent years, such governmental encouragement of research into NER has yielded concrete results in model development for Irish (Adkins et al., 2025), albeit relying on modern corpora

14 https://www.duchas.ie/en/meitheal, last accessed 13 October 2026.

15 https://www.transkribus.org/blog/success-stories, last accessed 13 October 2026.

16 https://www.transkribus.org/crowdsourcing-transcription-platform, last accessed 13 October 2026.

17 https://digital.library.universityofgalway.ie/p/ms/categories/an-stoc, last accessed 13 October 2026.

18 https://digital.library.universityofgalway.ie/p/ms/categories/conradh-na-gaeilge, last accessed 13 October 2026.

such as governmental proceedings. The *An Gaodhal* corpus could help address this lacuna by providing a base corpus to which NER models can be applied or with which NER can be newly developed, as described in Dereza et al. (2024). Any future application of NER to the *An Gaodhal* corpus would benefit from relying on well-developed pre-existing linked-data models in the form of logainm.ie (for names and creative works) and duchas.ie (for folklore content) to demonstrate the interoperability of the data.

AI DECLARATION

No GenAI tools were used in writing or editing this paper. Claude Code was used for programming assistance during the process of creating the PAGE XML to ALTO XML converter mentioned in this paper, and the documentation for that tool.

DATA ACCESSIBILITY STATEMENT

REPOSITORY LOCATION

https://doi.org/10.58153/7s7n3-58a49

OBJECT NAME

An Gaodhal Newspaper (1881–98) Full-Text OCR Output Files

FORMAT NAMES AND VERSIONS

ALTO XML, PAGE XML, TXT

LICENCE

Creative Commons Attribution Share Alike 4.0 International

ACKNOWLEDGEMENTS

We express our special gratitude to the University of Galway Library for providing high-resolution images of *An Gaodhal* and archival support and also to the Data Science Institute (now part of Rinn AI) at University of Galway for providing research support.

FUNDING STATEMENT

This project was supported by the Robert D. L. Gardiner Foundation, the Irish Institute of New York, New York University Glucksman Ireland House, the University of Galway, Foras na Gaeilge, and the Department of Rural and Community Development and the Gaeltacht, Government of Ireland.

COMPETING INTERESTS

The authors have no competing interests to declare.

AUTHOR CONTRIBUTIONS

Deirdre Ní Chonghaile: conceptualisation, data curation, funding acquisition, investigation, methodology, project administration, validation, writing – original draft.

Oksana Dereza: conceptualisation, data curation, formal analysis, investigation, methodology, resources, software, writing – original draft.

Nicholas Wolf: conceptualisation, data curation, funding acquisition, investigation, methodology, project administration, resources, supervision, validation, writing – original draft.

AUTHOR AFFILIATIONS

Deirdre Ní Chonghaile orcid.org/0000-0001-8147-9874
Gaois, Fiontar & Scoil na Gaeilge, Dublin City University, Dublin, Ireland

Oksana Dereza orcid.org/0000-0003-2896-6037
University of Galway Library, University of Galway, Galway, Ireland

Nicholas M. Wolf orcid.org/0000-0001-5512-6151
New York University Libraries, New York University, New York, USA

REFERENCES

Acadamh Ríoga na hÉireann (2017). *Corpas Stairiúil na Gaeilge 1600–1926: Historical Irish corpus*. https://corpas.ria.ie/ (Accessed: 16 July 2026).

Adkins, J., Collins, H., Wagner, J., Walsh, A., & **Davis, B.** (2025). Named Entity Recognition for the Irish Language. In A. K. Ojha, V. Giouli, V. B. Mititelu, M. Constant, G. Korvel, A. S. Doğruöz, & A. Rademaker (Eds.), *Proceedings of the 21st Workshop on Multiword Expressions (MWE 2025)* (pp. 82–96). Association for Computational Linguistics. https://doi.org/10.18653/v1/2025.mwe-1.9

Capurro, C., Provatorova, V., & **Kanoulas, E.** (2023). Experimenting with Training a Neural Network in Transkribus to Recognise Text in a Multilingual and Multi-Authored Manuscript Collection. *Heritage*, *6*(12), 7482–7494. https://doi.org/10.3390/heritage6120392

Dereza, O., Ní Chonghaile, D., & **Wolf, N.** (2024). "To have the 'million' readers yet": Building a digitally enhanced edition of the bilingual Irish-English newspaper *An Gaodhal* (1881–1898). In R. Sprugnoli & M. Passarotti (Eds.), *Proceedings of the Third Workshop on Language Technologies for Historical and Ancient Languages (LT4HALA) @ LREC-COLING 2024* (pp. 65–78). ELRA and ICCL. https://doi.org/10.63317/5oc72of6o6ph

Ehrmann, M., Hamdi, A., Pontes, E. L., Romanello, M., & **Doucet, A.** (2023). Named entity recognition and classification in historical documents: A survey. *ACM Computing Surveys*, *56*(2), 1–47. https://doi.org/10.1145/3604931

Farrell, G. (2023). *Irish, Gaelic and Roman type (Seanchló agus Cló Rómhánach) v.3*. Transkribus public model. https://www.transkribus.org/models/irish-gaelic-and-roman-type-seanchl-agus-cl-rmhnach-v3 (Accessed: 1 July 2024).

Government of Ireland (2022). *Digital Plan for the Irish Language: Speech and Language Technologies*, 2023–2027. https://assets.gov.ie/241755/e82c256a-6f47-4ddb-8ce6-ff81df208bb1.pdf (Accessed: 16 July 2026).

de Hae, R., & **Ní Dhonnchadha, B.** (1938–1940). *Clár Litridheacht na Nua-Ghaedhilge 1850–1936*. Baile Átha Cliath: Foillseacháin Rialtais.

Mac Peaircín, L. (2020). *Breac sa bhFarraige: An Stoc agus An Chearnóg (1917–1931)*. An Drom Bán, Durlas, Co. Thiobraid Árann: An Droichead Miotail.

Mac Peaircín, L. (2027). *An Stoc agus An Chearnóg (1917–1931): Clár*. An Drom Bán, Durlas, Co. Thiobraid Árann: An Droichead Miotail.

McGuinne, D. (1992). *Irish Type Design: A history of printing types in the Irish character*. Irish Academic Press.

Miyagawa, S. (2023). Digitizing Old Nubian dictionary: Optical Character Recognition for multi-lingual and multi-script text from medieval Africa. In *2023 7th IEEE Congress on Information Science and Technology (CiSt)*, (pp. 587–592). Essaouira. https://doi.org/10.1109/CiSt56084.2023.10409998

Muehlberger, G., Seaward, L., Terras, M., Ares Oliveira, S., Bosch, V., Bryan, M., Colutto, S., Déjean, H., Diem, M., Fiel, S., Gatos, B., Greinoecker, A., Grüning, T., 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., & **Zagoris, K.** (2019). Transforming scholarship in the archives through handwritten text recognition: Transkribus as a case study. *Journal of Documentation*, *75*(5), 954–976. https://doi.org/10.1108/JD-07-2018-0114

Ní Loingsigh, K., Ó Meachair, M. J., Ó Cleircín, G., Bhreathnach, Ú., & **Ó Raghallaigh, B.** (2022). Anailís dhiacronach chorpasbhunaithe ar cheantair theangeolaíocha na hÉireann. *Teanga: Iris Chumann na Teangeolaíochta Feidhmí in Éirinn*, *29*, 169–193. https://doi.org/10.35903/teanga.v29i.2564

Nockels, J., Gooding, P., Ames, S., & **Terras, M.** (2022). Understanding the application of handwritten text recognition technology in heritage contexts: A systematic review of Transkribus in published research. *Archival Science*, *22*(3), 367–392. https://doi.org/10.1007/s10502-022-09397-0

Ní Chonghaile, D. (2015). "Sagart gan iomrádh": An tAthair Domhnall Ó Morchadha (1858–1935) agus amhráin Philadelphia. In R. Nic Congáil, M. Nic Eoin, M. Ní Úrdail, P. Ó Liatháin, & R. Uí Chollatáin (Eds.), *Litríocht na Gaeilge ar fud an domhain, imleabhar 1: Cruthú, caomhnú agus athbheochan* (pp. 191–214). Baile Átha Cliath: LeabhairCOMHAR.

O'Leary, P. (2010). Yesterday's Papers: Review of *Iriseoirí Pinn na Gaeilge: An Cholúnaíocht Liteartha: Critic Iriseoireachta*. *The Irish Review* (Cork), *42*, 152–154.

Scannell, K., Regan, J., & **Damazyn, K.** (2020). Tesseract Uncial Training Data. GitHub. https://github.com/kscanne/tesseract-gle-uncial

Uí Chollatáin, R. (2008). *Iriseoirí Pinn na Gaeilge: An cholúnaíocht liteartha, critic iriseoireachta*. Baile Átha Cliath: Cois Life.

Uí Fhlannagáin, F. (1990). *Mícheál Ó Lócháin agus An Gaodhal*. Baile Átha Cliath: An Clóchomhar.

Ó Murchú, M. (1977). Successes and failures in the modernization of Irish spelling. In J. A. Fishman (Ed.), *Advances in the Creation and Revision of Writing Systems* (pp. 267–289). Mouton. https://doi.org/10.1515/9783110807097-012

TO CITE THIS ARTICLE:
Ní Chonghaile, D., Dereza O., & Wolf, N. M. (2026). Applying OCR to Bilingual Irish-English Texts in *An Gaodhal* (1881–98). *Journal of Open Humanities Data*, 12: 145. DOI: https://doi.org/10.5334/johd.621

Submitted: 17 July 2026
Accepted: 18 September 2026
Published: 04 November 2026

COPYRIGHT:
© 2026 The Author(s). This is an open-access article distributed under the terms of the Creative Commons Attribution 4.0 International License (CC-BY 4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited. See https://creativecommons.org/licenses/by/4.0/.

Journal of Open Humanities Data is a peer-reviewed open access journal published by Ubiquity Press.