“To Have the ‘Million’ Readers Yet”: Building a Digitally Enhanced Edition of the Bilingual Irish-English Newspaper *An Gaodhal* (1881 – 1898)

Oksana Dereza[1,2], **Deirdre Ní Chonghaile**[3], **Nicholas Wolf**[3,4]
[1] University of Galway Library, Ireland
[2] Insight SFI Centre for Data Analytics, Data Science Institute, University of Galway, Ireland
[3] Glucksman Ireland House, New York University, USA
[4] Division of Libraries, New York University, USA
oksana.dereza@universityofgalway.ie, nichonghailed@ollscoilnagaillimhe.ie, nicholas.wolf@nyu.edu

Abstract

This paper introduces two new OCR models for the Irish language, a BART-based OCR post-correction model, and the core dataset on which they were trained: a monthly bilingual Irish-English newspaper named *An Gaodhal* that was produced from 1881 to 1898 by an Irishman living in Brooklyn, New York.

Keywords: Optical Character Recognition (OCR), OCR post-correction, Named Entity Recognition (NER), Irish, Gaeilge, bilingual data, code-mixing, digital edition, corpus creation, corpus annotation

1. Introduction

This paper introduces the *An Gaodhal* project, which aims to serve the historically under-resourced and endangered language of Irish[1] (known as Gaeilge) by providing new digital tools and resources.

The initial goal of the project was the extraction of full text of *An Gaodhal*, a monthly bilingual Irish-English newspaper produced from 1881 to 1898 by an Irishman living in Brooklyn, New York, to the highest possible degree of accuracy via Optical Character Recognition (OCR), with a view to making its printed content searchable. The methodology applied toward achieving this goal yielded additional digital outputs including:

- a new OCR model for the Irish language as printed in Cló Gaelach type;[2]
- a new OCR model for bilingual Irish-English content printed in Cló Gaelach and Roman types respectively;
- a BART-based OCR post-correction model for historical bilingual Irish-English data;
- a historical Irish training set for Named Entity Recognition (NER);

All but the first of these four additional outputs appear to be the first of their kind. Each of the project outputs is set for public release to enable open-access research.

This paper also identifies the challenges historical Irish data poses to Natural Language Processing (NLP) in general and OCR in particular, and reports on project results and outputs to date. Finally, it contextualises the project within the wider field of NLP and considers its potential impact on under-resourced languages worldwide.

2. Related Work

2.1. OCR

In December 2022, the Irish government launched a roadmap document titled *Digital Plan for the Irish Language: Speech and Language Technologies 2023-2027*, which “provides an overview of the research required to make Irish-language linguistic resources available in the coming years” (Government of Ireland, 2022). This digital plan acknowledges the need for a diverse ecosystem of Irish-language corpora and identifies a significant number already extant for Irish. To date, only one of these corpora — *Corpas Stairiúil na Gaeilge 1600 – 1926* (Acadamh Ríoga na hÉireann, 2017) — has been produced using OCR. To the best of our knowledge, the relevant OCR work was outsourced to a third-party, and any models deployed in that work were not available publicly.

At the outset of the present project in January 2023, there were no publicly available OCR models attuned to Cló Gaelach and pre-standardised spelling of the Irish language, in either monolingual or multilingual contexts.[3] The only related project in existence was a Cló Gaelach training dataset for Tesseract OCR software,[4] published by Scannell et al. (2020). In November 2023, an

[1] Moseley (2010) lists Irish as ‘definitely endangered’.

[2] Cló Gaelach – a typeface widely used for Irish until the 1960s when it was replaced by Roman type.

[3] In a bilingual / multilingual context, Irish appears most frequently alongside English, reflecting their co-existence in Ireland for centuries.

[4] `https://github.com/tesseract-ocr/tesseract`

Irish-only model for texts in either Cló Gaelach or Roman typefaces was made public on the Transkribus OCR platform (Farrell, 2023). The methodology that produced this model differs considerably from the approach discussed herein in respect of the treatment of different typefaces, the treatment of individual printed glyphs, and the broader span of centuries represented in the corpus upon which the model was trained.

OCR models vary enormously, ranging from bespoke monolingual models, some of them reading individual handwritten scripts, to large-scale models incorporating multiple languages. Transkribus Team (2021) created the multilingual multi-typeface print model *Transkribus Print M1*, "including antiqua and blackletter prints, typewriter, computer print outs and decorative fonts" and supporting Dutch, English, Finnish, French, German, Italian, Latin, Swedish, Portuguese, Spanish, Polish, Flemish, Czech, Slovak, Slovenian, and Castilian. Currently, Transkribus features 147 publicly available models for print and handwritten text recognition,[5] including Devanagari, Hebrew, Ethiopian script, 14^{th} and 15^{th} century Spanish Gothic script, 14^{th} century cursive Dutch charters, 16^{th} century Balinese palm-leaf manuscripts, Serbian and Russian Church Slavonic, Ottoman Turkish written in Arabic script, 19^{th} century Danish handwriting, and multiple varieties of Fraktur[6] to name but a few.

Some OCR models focus on individual ancient and historical languages (Furrer and Volk, 2011; Bukhari et al., 2017; Springmann et al., 2018; Reul, 2020; Reul et al., 2021; Martínek et al., 2020; Dölek and Kurt, 2022; Ma et al., 2024). Others address multilinguality in a historical context: a team at Cornell University developed a trilingual handwritten text recognition (HTR) model for Ancient Greek, Latin, and German (Rusten, 2020);[7] and Capurro et al. (2023) are testing the viability of different approaches to building multilingual OCR models for HTR. A significant share of research on pre-modern OCR draws on historical newspaper corpora (Drobac et al., 2017; Koistinen et al., 2020; Drobac, 2020; Kettunen et al., 2020), which are readily accessible thanks to trends in early institutional digitisation.

Predictably, OCR datasets that predate the emergence of more advanced technologies register higher error rates. Moreover, source images are not always retained. The resulting impossibility or cost of re-extracting text prompts researchers to explore OCR post-correction as a discrete task (Reynaert, 2008; Vobl et al., 2014; Reynaert, 2016; Afli et al., 2016; Schulz and Kuhn, 2017; Richter et al., 2018; Dong and Smith, 2018; Dannélls and Persson, 2020; Duong et al., 2021; Soper et al., 2021; Rijhwani et al., 2021; Besnier and Mattingly, 2021; Lyu et al., 2021; Suissa et al., 2022). OCR post-correction is also applied to critically endangered languages where a scarcity of data would otherwise impede the building of a targeted OCR model. In such cases, scholars train a correction model to transform outputs of an OCR model unfamiliar with the target language (Rijhwani et al., 2020), a method that ultimately aims to obtain the best OCR results for the target language.

The capacity of OCR to inspire "new kinds of research on previously inaccessible sources" in humanities and social sciences is driving unprecedented growth in this domain and scholars continue to explore ways of improving OCR outputs (Smith and Cordell, 2018).

2.2. NER

Like OCR, NER work around the globe reflects a wide variety of approaches, some of which are discussed in an extensive survey on NER in historical documents published last year (Ehrmann et al., 2023). To date, two shared tasks on "identifying historical people, places and other entities (HIPE)" have been organised (Ehrmann et al., 2020, 2022). Some NER work focuses on individual historical languages, including 19^{th} century French (Tual et al., 2023), 8^{th} century Armenian (Tambuscio and Andrews, 2021), and Ancient Greek (Yousef et al., 2023); some aims at developing multilingual NER models (Neudecker, 2016; Boros et al., 2020; Dekhili and Sadat, 2020; Provatorova et al., 2020; Schweter et al., 2022). Like the *An Gaodhal* project, some teams combine OCR, OCR post-correction, and NER in historical texts (Todorov and Colavizza, 2020). As with OCR, historical newspaper data is well-represented in NER research (Hubková, 2019; Schweter and Baiter, 2019; Hubková et al., 2020),

NER for the Irish language, whether modern or historical, represents uncharted territory. According to the Government of Ireland (2022): "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."

3. Historical Context

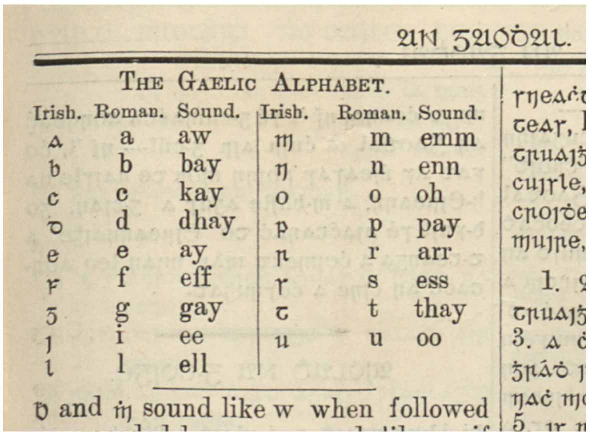
Historically, the Irish language has been printed in two different orthographies: Irish or Gaelic type,

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

[6] *Fraktur* denotes the German blackletter, or 'Gothic', fonts that derive from medieval handwriting.

[7] The authors could not locate this model on Transkribus, and assume it has not been made public.

known as Cló Gaelach, which originated in the scribal tradition (see Figure 1); and Roman type (McGuinne, 1992). Its corresponding Unicode characters draw on Roman (Latin) script. Cló Gaelach uses two kinds of diacritics: acute accents on vowels (ÁáÉéÍíÓóÚú); and dotted consonants (ḂḃĊċḊḋḞḟĠġṀṁṖṗṠṡṪṫ), the dots indicating a grammatical feature called lenition. Where Irish appears in Roman type, dotted consonants are replaced by Bh, Ch, dh, fh, etc.



An ᵹAoḋAl.

THE GAELIC ALPHABET.

Irish.	Roman.	Sound.	Irish.	Roman.	Sound.
a	a	aw	m	m	emm
b	b	bay	n	n	enn
c	c	kay	o	o	oh
ꝺ	d	dhay	p	p	pay
e	e	ay	ꞃ	r	arr
ꝼ	f	eff	ꞅ	s	ess
ᵹ	g	gay	ꞇ	t	thay
ı	i	ee	u	u	oo
l	l	ell			

ḃ and ṁ sound like w when followed

Figure 1: The Irish alphabet.

Although the quantity of printed material in Irish in the centuries prior to the appearance of *An Gaodhal* in October 1881 was small in comparison to many languages, a recent cataloguing of titles published in Irish between the 16^{th} and 19^{th} centuries (Sharpe and Hoyne, 2020) lists over a thousand entries, several of them with multiple editions. Prior to the 1880s, the most common genres for printing in Irish were religious texts (both Catholic and Protestant), academic texts, and so-called Gaelic columns in otherwise English-only newspapers in which a relatively small amount of content (usually letters, songs, or poetry) was printed in Irish. *An Gaodhal* thus appeared at a time when printing in Irish was taking place, but not on a mass scale, so the newspaper's production represented an energetic undertaking in the face of headwinds.

An Gaodhal was established and edited by Micheál Ó Lócháin (also known as Michael J. Logan).[8] It is regarded as the world's first serial dedicated to providing content to an Irish-language readership. The first four issues of the newspaper were printed commercially and at a loss. To save the enterprise, Logan took on the task of typesetting and printing the newspaper himself, most likely in his own home in Brooklyn. Over the next 17 years, Logan continued to issue the paper, supported by a transnational network of contributors. His commitment combined with the appetite among readers to achieve 1,200 subscriptions within the first year, growing to 3,000 at its peak, five times the number achieved by the contemporaneous Dublin-based *Irisleabhar na Gaedhilge*, also known as *The Gaelic Journal* (Uí Fhlannagáin, 1990).

As one might expect from an ethnic newspaper emerging in a diasporic setting, contributors to *An Gaodhal* and its readers welcomed the arrival of a new forum in which to identify their community of 'Éire Mhór' (Greater Ireland) and celebrate it (Knight, 2021). Nationalist politics at home in Ireland amplified that sense of pride, which extended to the use of Cló Gaelach throughout the newspaper to distinguish Irish expression from the English nation, its language and Roman type, and British imperialism. Indeed, the Irish type used in the newspaper, modelled on Watts type, was newly cast in the United States to avoid purchasing a set cast in a London foundry.

There is a palpable sense of energy and excitement in the newspaper as many of its contributors and readers were then gaining literacy in Irish for the first time. The standard of written Irish varied accordingly, as did the spelling, which had yet to be standardised. Add to this the use of three differing dialects and the emerging corpus of texts — however small at 1.86 million tokens — yields a welcome diversity in the prospective training data. To date, the adaptability of the OCR models developed by the project team supports this inference.

The challenges *An Gaodhal* faced were varied. The economics of audience size over printing costs, particularly for a newspaper printed for a transatlantic audience, drove its founder and editor to forgo any income from his work on the paper. The debate over the choice of type, whether Roman or Cló Gaelach, had long been a heated one; for those who insisted that Cló Gaelach was the only proper type for expressing Irish, there was the immediate challenge of procuring such a unique typeface — a matter of availability, not cost, as it could be purchased for the same price as Roman type. Even where Roman type was selected, any printer choosing to produce Irish texts in the nineteenth century or earlier faced difficulties in finding a sufficiently large, paying audience and, in a diasporic context, sufficiently fluent typesetters or compositors. The absence of mass literacy in Irish prior to the twentieth century combines with these challenges for printing to make the appearance of *An Gaodhal* and of similar undertakings in its aftermath[9] especially notable: they represent

[8] See `https://www.ainm.ie/Bio.aspx?ID=347` (Irish) and `https://www.dib.ie/biography/logan-michael-j-o-lochain-micheal-a4873` (English) for Michael J. Logan's biographies.

[9] See, for example, Knight (2021) on the Irish-language column in *New York Irish-American*, 1857 –

the first steps in creating a media landscape in the Irish language, an impact foretold in the ambition expressed by Logan in *An Gaodhal* “to have the ‘million’ readers yet.”[10]

4. Data

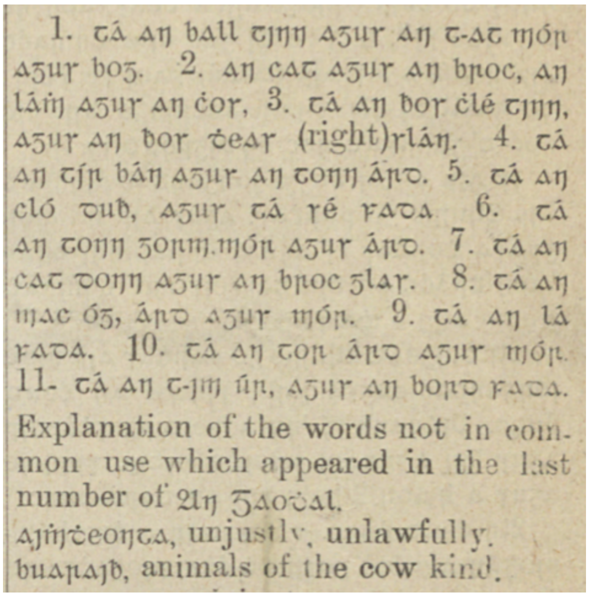
The only complete series of *An Gaodhal* spanning 1881 to 1898 survives in the James Hardiman Library at the University of Galway. This set was compiled, bound, and annotated by the Philadelphia-based scholar of Irish folklore and sean-nós song, Rev. Daniel J. Murphy, and forms part of his manuscript archive, which is also held in Galway (Ní Chonghaile, 2015). Since Rev. Murphy’s volumes of *An Gaodhal* were digitised in 2021 (University of Galway, 2021), the newspaper has been openly accessible as high resolution images via the University of Galway Library’s Digital Collections and Archives.[11] While the current interface provides searchable metadata, extending its functionality to include full-text searchability represents one of the ambitions of the present project, which aims to build a digitally enhanced edition of *An Gaodhal*.

As a monthly newspaper, *An Gaodhal* contained 12 numbers per volume. The corpus totals 147 issues from Vol. 1, No. 1, to Vol. 13, No. 3, and is complete and intact at 2,290 pages i.e. there are no missing pages. Most issues contain 16 pages; some contain 14, 12 or 8 pages. Page tears, ink spots, and blemishes are rare. Where such characteristics impair the legibility of text, human review relied on consulting the printed artefact or other extant samples of the relevant text.

The following list of key characteristics of the *An Gaodhal* corpus will help determine the relevance of the current project to the efforts of those seeking to apply OCR to other historical data:

- pages feature Irish mostly (381), English mostly (896), or both languages together (1,019);
- the use of two different typefaces throughout — Cló Gaelach and Roman — with infrequent changes of font and sometimes using Cló Gaelach for English content and Roman letters for Irish content (see Figures 2 and 3);
- the pre-standardised spelling of the Irish language in the late 19^{th} century;
- variations in spelling and vocabulary reflecting the three major dialects of Irish;
- variations in spelling reflecting the language aptitude of each contributor, many of whom were learners of the language or were gaining literacy in Irish for the first time;
- layout conventions reflecting the artisanal nature of the letterpress printing operation, which was small and domestic in scale and style, produced by the founder and editor Michael J. Logan entirely on a pro bono basis, and funded chiefly by subscriptions and advertisements.

4.1. Types, fonts, and marginalia



1. tá an ball tinn agus an t-at mór
agus bog. 2. an cat agus an broc, an
láṁ agus an ċos, 3. tá an bos ċlé tinn,
agus an bos ḋeas (right)slán. 4. tá
an tír bán agus an tonn árd. 5. tá an
cló duḃ, agus tá sé fada 6. tá
an tonn gorm.mór agus árd. 7. tá an
cat donn agus an broc glas. 8. tá an
mac óg, árd agus mór. 9. tá an lá
fada. 10. tá an tor árd agus mór.
11. tá an t-im úr, agus an bord fada.

Explanation of the words not in common use which appeared in the last number of An Gaoḋal.

aiṁḋeonta, unjustly, unlawfully.
buaraiḃ, animals of the cow kind.

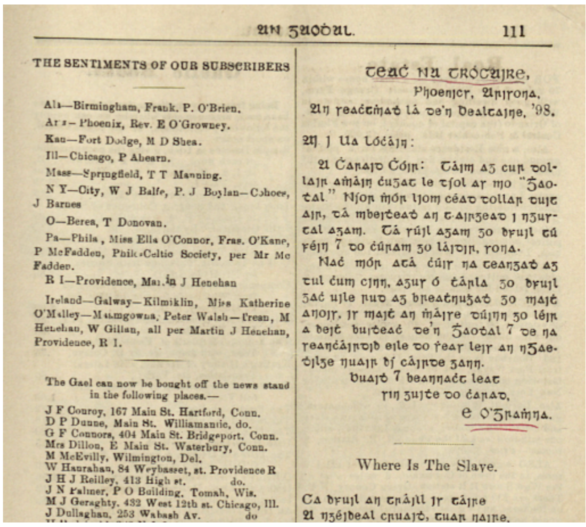
Figure 2: Example of mixed type usage in *An Gaodhal*.

The set of Cló Gaelach type used by Logan appears complete. A contemporary New York newspaper edited by Irish-born printer James Haltigan, *Celtic Monthly* (1879 – 1884), used a set of type that appears identical; however, the characters Ḃ, Ċ, Ḋ, Ḟ, Ġ, Ṁ, Ṗ, Ṡ, and Ṫ are applied variably therein (Knight, 2021). In lieu of dotted capital consonants, Haltigan and his colleagues sometimes rendered Ḃ, Ċ, and Ḋ as Bh, Ch, Dh, etc., a common substitution at this time and later where access to Cló Gaelach type was not guaranteed. To ensure that such nuances of contemporary typesetting and spelling conventions in a given printed artefact are preserved in the text extraction, the two new OCR models were trained to match a single Unicode character to each printed glyph; manually substituting Ḃ, Ċ, and Ḋ with Bh, Ch, and Dh, etc. was eschewed. Logan rarely adopted such substitutions and, in Irish-language texts, chose

1896, and Lyons (2021) on the Irish language revival, media and the transatlantic influence in 1857 – 1897.

[10] *An Gaodhal*, Vol. 9, No. 8 (January 1893): 236, accessible here: https://digital.library.universityofgalway.ie/p/ms/asset/16459

[11] https://digital.library.universityofgalway.ie/



AN GAOḊAL. 111

THE SENTIMENTS OF OUR SUBSCRIBERS

Ala—Birmingham, Frank. P. O'Brien.

Ariz—Phoenix, Rev. E O'Growney.

Kan—Fort Dodge, M D Shea.

Ill—Chicago, P Ahearn.

Mass—Springfield, T T Manning.

N Y—City, W J Balfe, P. J Boylan—Cohoes, J Barnes

O—Berea, T Donovan.

Pa—Phila , Miss Ella O'Connor, Fras. O'Kane, P McFadden, Phil.-Celtic Society, per Mr Mc Fadden.

R I—Providence, Martin J Henehan

Ireland—Galway—Kilmiklin, Miss Katherine O'Malley—Maumgowna, Peter Walsh—Crean, M Henehan, W Gillan, all per Martin J Henehan, Providence, R I.

The Gael can now be bought off the news stand in the following places.—

J F Conroy, 167 Main St. Hartford, Conn.
D P Dunne, Main St. Williamantic, do.
G F Connors, 404 Main St. Bridgeport. Conn.
Mrs Dillon, E Main St. Waterbury, Conn.
M McEvilly, Wilmington, Del.
W Hanrahan, 84 Weybasset, st. Providence R
J H J Reilley, 413 High st. do.
J N Palmer, P O Building, Tomah, Wis.
M J Geraghty. 432 West 12th st. Chicago, Ill.
J Dullaghan, 253 Wabash Av. do

TEAĊ NA TRÓĊAIRE,
Phoenix, Arizona.
An seaċtṁaḋ lá de'n Bealtaine, '98.

M. I Ua Lóċáin:

A Ċaraid Ċóir: Táim ag cur dollair aṁáin ċugat le síol ar mo "Gaoḋal." Níor ṁór liom céad dollar duit air, dá mbeiṫeaḋ an t-airgead i ngustal agam. Tá súil agam go bfuil tú féin 7 do ċúram go láidir, sona.

Naċ mór atá ċúis na teangaḋ ag dul ċum cinn, agus ó ṫárla go bfuil gaċ uile nuḋ ag breaṫnuġaḋ go maiṫ anois, is maiṫ an ṁáise dúinn go léir a beiṫ buiḋeaċ de'n Gaoḋal 7 de na seanċáirdiḃ eile do ṡeas leis an nGaeḋilge nuair bí cáirde gann.

Buaiḋ 7 beannaċt leat
sin guiḋe do ċarad,
E O'Graṁna.

Where Is The Slave.

Cá bfuil an tráill is táire
A nGéideal cruaiḋ, tuar naire.

Figure 3: Example of different Latin fonts and pica sizes in *An Gaodhal*.

to adhere to the relevant orthography, spelling some English words phonetically e.g. 'Nuaḋ Ġorc' for New York. In the present text extraction, the selected Unicode characters do not replicate exactly the design of the Cló Gaelach type such as Gaelchló[12] provides; rather, in deference to long-standing practice, Roman typeface characters — including those with diacritics (dots or accents) above the x-height or cap height e.g. ú or Ṁ — were chosen, thus ensuring interoperability between this dataset and others.

Printing errors are uncommon. Sometimes individual pieces of moveable type were placed in the printer's composing stick in the wrong order or upside-down, or supplies of particular letters e.g. a, á, e, é, ran short and were substituted with alternatives from either of the two orthographies. On occasion, insufficient ink or loose type rendered gaps. Corrections arising were tagged as 'supplied' or 'unclear' or 'gap' as appropriate to the word or line in question. Smaller pica sizes, which occurred only in the English-language fonts and most often in advertisements, proved challenging to the OCR software and thus prompted occasional manual text entry.

Handwritten marginalia corresponding to Rev. Murphy's handwriting occur on 495 of the 2,290 pages and were included in the OCR run. Appearing in black, blue, and red ink and in pencil, Rev. Murphy's annotations supply additional data including references to published books, journals, and newspapers; identify alternate song titles and associated song airs or melodies; and suggest corrections to the printed text content.

Abbreviations reflecting conventions of the period occur throughout, many of them serving to conserve space and type in printed matter, e.g. in English, 'Jas.' for 'James' and 'Patk.' for 'Patrick'. The names of American states are frequently abbreviated, with and without period marks and/or spacing, e.g. 'RI' and 'R. I.' for Rhode Island. In Irish, Logan frequently abbreviated 'agus' ('and') to the digit 7 in lieu of the Tironian symbol for the Latin 'et' (⁊). In correcting text extraction, human review reverted to the ampersand symbol (&) instead to avoid confusion with the digit 7.

5. OCR Workflow

The software selected for this process was READ-COOP's Transkribus (Kahle et al., 2017; Colutto et al., 2019), and the workflow included the following steps:

1. **Automating identification of predominantly Irish-language lines on pages.** This was done using Amazon's Textract software,[13] which could quickly and accurately produce token-detection and line segmentation regardless of language. The resulting OCR outputs were then categorised into Irish and non-Irish texts on a line-by-line basis by evaluating the dictionary-word accuracy of each line output. Pages scoring high as containing properly spelled English words were deemed 'non-Irish,' leaving a clear corpus of predominantly Irish-language pages to train an initial model. The team 'masked' English-language lines occurring in the pages of the selected corpus using overlaid opaque rectangles, enabling the creation of monolingual Irish-only page images.

2. **Training an OCR model for Irish-only pages.** From the masked Irish-only pages, the team selected 60 pages at random and, after excluding pages dominated by images or advertisements, a total of 57 proved viable for training. The team transcribed the texts on these pages manually and then used those transcriptions to create a model in Transkribus named *An Gaodhal Irish Model v. 1* for Cló Gaelach Irish-language detection (ID 50036), which incorporated 18,533 tokens.

3. **Training an OCR model on bilingual Irish-English pages.** The team selected 100 pages randomly from the entire collection, removing all masks to present fully bilingual texts. The language profile of each page determined which of the three selected OCR models ought to be applied. Pages predominantly in Irish were run through the Irish-only model (ID 50036); and pages predominantly

[12] https://www.gaelchlo.com/

[13] https://aws.amazon.com/textract

in English were run using *Transkribus Print M1* (ID 39995), which has been trained on over 5 million tokens and which also reflects the historical typographical conventions of the corpus. The resulting pages were then corrected manually, which provided the necessary content to train a bilingual OCR model. This bilingual model titled *An Gaodhal Irish / English Bilingual Model v. 1* (ID 51080) incorporated 54,406 input training tokens and achieved a character error rate (CER) close to 0 on the validation set.

4. **Correcting the outputs.** The team ran three different OCR models on the full 2,290 pages of the newspaper as appropriate to the language profile of each page: *Transkribus Print M1* on English-only or English-mostly pages; *An Gaodhal Irish Model v. 1* on the Irish-only or Irish-mostly pages; and *An Gaodhal Irish / English Bilingual Model v. 1* on bilingual pages. To date, half of these pages have been corrected by human review, including: all of the Irish-only pages; 41.9% of bilingual Irish-English pages; and 37.8% English-only pages (see Section 6.2 for more detail). With all 381 Irish-only or Irish-mostly pages corrected, a second Irish-only OCR model — *An Gaodhal Irish Model v. 2* (ID 61350) — was trained. It incorporated 164,015 words and achieved 1.4% CER on the validation set.

5. **Collecting supplementary page-level information.** The team reviewed and recorded key attributes of each page and presented the results of this review in the CSV file published together with the dataset. It lists: the presence of a table, advertisement, or image on each page; the language profile of the page — Irish, English, or bilingual; and the occurrence of verse (song or poem) or letters. This detail provides scope for further analysis of the content of the corpus (see Section 7 for the dataset description and reference).

6. OCR Post-Correction

6.1. Automatic correction

Whilst developing the bilingual OCR model, the team experimented with automatic OCR post-correction. The training set for OCR post-correction models included 103 pages from the 1 – 200 range; all of these pages had been manually corrected after the first application (as appropriate to the language profile of each page) of one of the project's chosen OCR models as outlined above. This dataset amounted to 9,994 lines of text and had 2.95% CER and 9.29% WER before manual correction. It was split into train and validation subsets with 0.9 : 0.1 ratio. The test set consisted of 235 lines from pages 10, 37 and 97 that were not used in the OCR model training. The test set CER and WER were 3.47% and 11.92% respectively.

The team decided to attempt fine-tuning state-of-the-art (SOTA) transformer models pre-trained for sequence-to-sequence tasks. In order to select the best transformer model for further experiments, we compared BART-base (Lewis et al., 2020), T5-base (Raffel et al., 2020), FLAN-T5-base (Chung et al., 2022), a BART-based English spellchecker (Guhr, 2023), and a T5-based spellchecker (Kundumani, 2022) by fine-tuning them with *An Gaodhal* data along with their default tokenisers. BART models performed significantly better than T5 models, as shown in Table 1.

Model	Test CER, %	Test WER, %
OCR output	3.47	11.92
BART-base	3.65	10.37
BART English spellchecker	3.40	10.50
T5-base	7.71	26.73
FLAN-T5-base	7.88	26.94
T5 English spellchecker	7.74	26.87

Table 1: Fine-tuning large language models on *An Gaodhal* data for OCR post-correction with default parameters.

The next step was to compare the performance of BART-base and BART-large models. Surprisingly, they demonstrated similar results: BART-base yielded 3.65% CER and 10.71% WER; and BART-large scores were 3.57% CER and 10.64% WER. As BART-large did not demonstrate a significant improvement compared to BART-base, the team proceeded with the smaller and less computationally-demanding BART-base model.

The team then measured how different tokenisers commonly used with transformer models[14] might influence performance. The standard tokeniser that comes with the BART-base model uses byte-level Byte-Pair Encoding, or BPE (Sennrich et al., 2016), and treats spaces like parts of the tokens. We trained three other tokenisers with slightly different architectures — SentencePiece (Kudo and Richardson, 2018), byte-level BPE, and character-level BPE — on bilingual Irish-English data from *An Gaodhal* and compared them to the standard BART tokeniser (see Table 2). BART-base performed best with our custom byte-level BPE tokeniser, achieving 3.33% CER and and 10.10% WER. This tokeniser was used in all subsequent experiments and is further referred to as 'custom tokeniser'.

[14] https://huggingface.co/docs/transformers/en/tokeniser_summary

Tokeniser	Test CER, %	Test WER, %
OCR output	3.47	11.92
Standard (BART-base)	3.65	10.71
Custom SentencePiece	3.63	10.37
Custom byte-level BPE	**3.33**	**10.10**
Custom char-level BPE	3.44	11.58

Table 2: The influence of different tokenisers on BART-base performance. The best result is marked in **bold**.

Finally, the team applied three data enhancement / fine-tuning techniques:

1. Masking. In large language models, it is common to mask 15% of tokens (Wettig et al., 2023) during pre-training to make the model more robust. The same strategy is recommended for BART fine-tuning.[15] Unfortunately, randomly masking 15% of words in our dataset at the pre-tokenisation stage did not yield better scores.

2. Balancing the dataset. The number of correct sentences in the training set for OCR post-correction outnumbered sentences containing OCR errors by a factor of 2.5. As such a class imbalance was likely to impede the efforts of a model to learn to correct errors, the team experimented with alternative ratios. We reduced the number of correct examples to a ratio of 1 : 1; and, in another set, to a slightly imbalanced ratio of 1 : 1.5, an adjustment that might mitigate the risk of over-correction. Though the model's performance improved in both settings, the difference between the 1 : 1 and 1 : 1.5 ratios was negligible.

3. Data augmentation. As the team aimed at training the model to correct very specific errors whilst also trying to avoid over-correction, it was decided to forgo introducing artificial noise. Instead, to augment the data, we elected to repeat every line in the dataset. However, the results revealed no improvement in the model's performance, either with 1 : 1 balancing or without.

The results are described in greater detail in Table 3. Analysing individual examples from the test set, we noted that models excel in correcting punctuation errors — such as an unnecessary or a missing space before/after a punctuation mark — or noise, usually in the form of dashes and square brackets. However, they are not as successful with incorrectly recognised letters, which is most likely due to the limited number of relevant examples in the training set.

All models were fine-tuned and tested with the help of the 'transformers' Python library (Wolf et al., 2020), and the best one is available on the HuggingFace model hub (Dereza, 2024) along with the corresponding dataset (Dereza et al., 2024).

6.2. Manual correction

The approach to correcting OCR output was curatorial, not editorial. As the newspaper was edited by the same individual from start to finish and printed under his guidance, there is a notable consistency of style throughout. Corrections were applied rarely and only then in the interests of ensuring discoverability. Non-standard forms of Irish-language spellings throughout prompted a strict adherence to the printed artefact as did printer's abbreviations — both conventional and idiosyncratic — that represent efforts to maximise space or optimise readability.

Punctuation and typographical conventions are generally preserved. However, some commas were inserted where printing rendered a period mark in the middle of a sentence; tilde marks ($\approx$) used in hyphenated compound words were replaced with a standard n-dash (–) to avoid confusion with the mathematical sign 'equal to' (=); and spaces were inserted on either side of m-dashes (—) to ensure that words on either side were recognised as separate entities. Some lines of text were justified from time to time but many more end with a word that is split between the end of that line and the start of the next, reflecting the physical restrictions of manual type-setting. In the printed artefact, the split is bridged by a n-dash (–). Excluding hyphenated compound words, we replaced such examples with the character ¬ (called a 'soft hyphen' or 'optional hyphen'). Such amendments aim to facilitate comprehension and deliver consistency for machine-reading tasks.

The bilingual Irish-English model (ID 51080) performed best when the content featured almost equal quantities of both languages and when the languages were confined to separate sections. Where the languages were intermixed in individual lines — in lists of translated Irish vocabulary or language instructional texts, for instance — the OCR output required more correction where the model failed to adjust to the rhythm of the orthographic exchanges on the page. Pages featuring a majority of English content required text entry for any Irish content therein where the English-only OCR model failed to render the Irish orthography. Likewise, pages featuring a majority of Irish content required text entry for any English content therein where the Irish-only OCR model failed to render the English orthography.

[15] https://huggingface.co/docs/transformers/model_doc/bart

Configuration	Train + valid data	Test CER, %	Test WER, %
OCR output	–	3.47	11.92
BART-base + standard tokeniser	9994 lines	3.65	**10.71**
BART-large + standard tokeniser	9994 lines	3.57	**10.64**
BART-base + custom tokeniser	9994 lines	**3.33**	**10.10**
BART spellchecker + custom tokeniser	9994 lines	**3.40**	**10.24**
BART-base + custom tokeniser + masking	9994 lines	3.60	**10.91**
BART-base + custom tokeniser + data balanced 50:50	5734 lines	**3.29**	**<u>9.83</u>**
BART-base + custom tokeniser + data balanced 40:60	7154 lines	**3.29**	**<u>9.83</u>**
BART-base + custom tokeniser + data augmented x2	19988 lines	**3.39**	**10.24**
BART-base + custom tokeniser + data augmented x2, balanced 50:50	11468 lines	**<u>3.27</u>**	**10.17**

Table 3: Comparison of different BART fine-tuning configurations. Improvements in CER / WER on the test set are marked in **bold**, and the best result is **<u>underlined</u>**.

Where OCR failed to render complete lines or word boxes, these were entered manually. Lines were sometimes joined or split to maximise comprehensibility of the extracted text. Corrections were provided at word-level, not simply at line-level, to enable future application of language-based technologies.

As is common in OCR workflows, layout detection was important to overall accuracy, especially given that columns and paragraph structures were used by the printers throughout. To yield workable baseline recognition, print block detection and layout analysis models offered by Transkribus were applied — at default settings — consecutively to each page. The occurrence of two columns on most pages, tables, advertisements, images, marginalia, and fine print demanded careful review of the page layout and sometimes required manual treatment including adjusting baselines and box boundaries and hand-drawing baselines for vertical text and marginalia.

7. Output

The work described above resulted in the machine-readable full text of *An Gaodhal* published on the NYU UltraViolet platform (Ní Chonghaile et al., 2023). The data constitute direct exports from Transkribus of the resulting full text. The files are presented in two forms:

1. Alto-format XML files that provide bounding box regions for text locations (at the individual token level) of separately tokenised pages,
2. Page-format XML files, which are comparable to Alto files but use a specific output format for Transkribus software.

XML files are internally self-describing, with tags providing names of fields. 'Page' Transkribus output format files are organised on a per-page basis into regions (<TextRegion>) or tables (<TableRegion>), lines (<TextLines>) or table cells (<TableCell>) respectively, and words (<Word>). Regions are also labeled according to a structure type: paragraph (<TextRegion type='paragraph'>), page-number (<TextRegion type='page-number'>), or marginalia (<TextRegion type='marginalia'>). These distinguish between the standard printed newspaper text, a page number printed on the page, and handwritten marginalia added to the original printed artefact.

Each structural element maps to the image uploaded to the software, reading each of the newspaper's two columns left to right from top to bottom. Exceptions arise where the usual layout deviates according to the printer's prerogative; for instance, when a reader's eye moves at intervals over and back between the two columns. In such rare cases, human review prompted the re-ordering of the sequence to ensure the extracted text output was as logical and comprehensible as the experience of reading the printed artefact.

Each region, line, and word has a unique identifier derived from its logical sequence on the page. Thus, for example, word id 'r5l1w2' refers to region 5, line 1, word 2. Tables, table cells, and words conform to the same style of sequencing e.g. region id 'tbl_4_4' refers to a table appearing between Regions 3 and 4 of standard text areas, and the relevant word id entries appear per line and word (left to right) as 'r_4_1_1' and 'r_4_1_2'.

Additional identifiers indicating separators (<Separator>) are retained in the data. The separator ID numbers do not conform to the sequence of identifiers mapping all other page elements; rather, they retain the identifiers generated automatically by the initial layout analysis. Hence, they appear somewhat random — two consecutive separators might appear as 'r_25' and 'r_39' — and are typically grouped together at the end of the page metadata. Each separator corresponds to a decorative hairline rule or border demarcating different elements of the printed page, separating articles or advertisements from each other. Such decorative elements aid the reader's navigation of a printed page. In a digital environment, an equivalent distinction is provided by the structural tags applied during text extraction. As such, separators were deemed surplus to the requirements of text extraction. In addition, such was the quantity of separators throughout, time did not allow for the re-sequencing of each individual separator between different text regions as they appear on each page.

Bounding coordinates for polygons and locations of points making up text baselines are oriented to an origin point (0,0) at the top left of the page, mapping each element to the image in question. X,Y coordinates are given as pairs in the form x1,y1; x2,y2; etc.

ALTO output format files follow the XML stylesheet maintained by the Library of Congress.[16] These files follow a similar region, line, string format, with the token provided at <string CONTENT>.

The accompanying CSV[17] provides additional metadata on a per-page basis that were recorded in the course of page layout review. Page metadata appear in rows with columns distinguishing between the following elements: page filename; the language profile of the page — Gaeilge (Irish), English, or Bilingual; presence of skew or tight gutters; and whether or not a page contains marginalia, images, advertisements, verse, or letters. Variables include:

- pageFilename: XML OCR output to which row data refer
- skew_gutter_fallaway: Yes/No on presence of a skew, gutter, or fallaway on digitised page that might affect OCR quality
- hasTable: Yes/No on presence of a table or table-like arrangement of tokens on page (includes list and list-like structures)
- language: Gaeilge/English/Mix, predominant language on page
- isCover: Yes/No on whether this page is the issue start (i.e. cover) page
- hasMarginalia: Yes/No on whether handwritten margin notes are present
- hasSong_Poem: Yes/No on whether a song or poem, or part thereof, is present
- hasAdvert: Yes/No on whether an advertisement is present
- hasLetter: Yes/No on whether a letter, or part thereof, is present
- hasImage: Yes/No on whether an image is present

8. Future Work

The team has begun working on Named Entity Recognition (NER) for Irish toward automatically extracting references to people, events, locations, dates, creative works, and more from the text. For this purpose, a dataset of 11,000 words was labeled manually according to IOBES annotation scheme to train / fine-tune a deep learning model for historical Irish NER in future. To the best of our knowledge, this attempt is the first of its kind for the Irish language. As for the English-language content from *An Gaodhal*, we applied NER to it separately using *en_core_web* models trained on large English-language corpora available through Python NLP framework spaCy (2023).

The team is also identifying corpora suitable for future applications of these new OCR and NER tools, e.g. the bilingual Irish-English newspaper *An Stoc*, 1917-1931 (University of Galway, 2022).

9. Conclusion

The project has presented its team with a unique set of challenges, some of which have been explored previously by only a handful of initiatives.

All project outputs will be made publicly accessible and available for further application in the field of computational linguistics. Creating an open interface enabling searches of the bilingual content of *An Gaodhal* will reveal to a wider public the vitality of Irish language practice in a diasporic context and reflect its co-existence alongside English. This new resource will enable historians to better contextualise the multilingual heritage of the Irish diaspora. The specificity of the newspaper's content and readership will be a particular boon to genealogists.

The OCR, OCR post-correction and NER tools produced by this project represent welcome additions to the digital tool-kit serving the Irish language into the future. Finally, the methodologies described here may come to inform and so serve other under-resourced and endangered languages worldwide.

10. Acknowledgements

This project has been supported to date by the Robert D. L. Gardiner Foundation, the Irish Institute of New York, New York University Glucksman Ireland House, and the University of Galway. We express our special gratitude to the James Hardiman Library at University of Galway for providing high-resolution images of *An Gaodhal*.

[16] Version 4.4 is the most current at the time of submission: https://www.loc.gov/standards/alto/v4/alto-4-4.xsd

[17] https://ultraviolet.library.nyu.edu/records/5ya5n-mc504/files/AnGaodhal_pageMetadata.csv

11. Authors' Contributions

Oksana Dereza focused on the NLP technologies, provided analysis of related work on OCR and NER, carried out computer-assisted OCR post-correction and NER experiments, and lead the writing of this paper.

Deirdre Ní Chonghaile provided historical and linguistic domain expertise, reviewed each corpus page toward generating the metadata CSV, conducted transcription work toward training OCR models on Transkribus, ran OCR, and performed manual OCR correction.

Nicholas Wolf acted as project PI, provided historical and linguistic domain expertise, conducted transcription work toward training OCR models on Transkribus and trained those models, reviewed metadata CSV, and prepared a NER training dataset that was reviewed by Deirdre.

All authors contributed to the project design and to the final manuscript.

12. Bibliographical References

Haithem Afli, Zhengwei Qiu, Andy Way, and Páraic Sheridan. 2016. Using SMT for OCR error correction of historical texts. In *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC'16)*, pages 962–966. European Language Resources Association (ELRA).

Clément Besnier and William Mattingly. 2021. Named-entity dataset for medieval Latin, Middle High German and Old Norse. *Journal of Open Humanities Data*, 7(0):23.

Tobias Blanke, Michael Bryant, and Mark Hedges. 2012. Open source optical character recognition for historical research. *Journal of Documentation*, 68(5):659–683.

Emanuela Boros, Elvys Linhares Pontes, Luis Adrián Cabrera-Diego, Ahmed Hamdi, José G Moreno, Nicolas Sidère, and Antoine Doucet. 2020. Robust named entity recognition and linking on historical multilingual documents. In *Conference and Labs of the Evaluation Forum (CLEF 2020)*, volume 2696 of *CEUR Workshop Proceedings*, pages 1–17. CEUR-WS.

Syed Saqib Bukhari, Ahmad Kadi, Mohammad Ayman Jouneh, Fahim Mahmood Mir, and Andreas Dengel. 2017. anyOCR: An open-source ocr system for historical archives. In *Proceedings of the 14th IAPR international conference on document analysis and recognition (ICDAR)*, volume 1, pages 305–310. IEEE.

Carlotta Capurro, Vera Provatorova, and Evangelos Kanoulas. 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.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022. Scaling instruction-finetuned language models. *arXiv:2210.11416*.

Sebastian Colutto, Philip Kahle, Hackl Guenter, and Günter Mühlberger. 2019. Transkribus: A platform for automated text recognition and searching of historical documents. In *Proceedings of the 15th International Conference on eScience (eScience)*, pages 463–466. IEEE.

Dana Dannélls and Simon Persson. 2020. Supervised OCR post-correction of historical Swedish texts: What role does the OCR system play? In *Proceedings of the Digital Humanities in the Nordic Countries, 5th Conference, Riga, Latvia, October 21-23, 2020*, volume 2612 of *CEUR Workshop Proceedings*, pages 24–37. CEUR-WS.

Ghaith Dekhili and Fatiha Sadat. 2020. Hybrid statistical and attentive deep neural approach for named entity recognition in historical newspapers. In *Working Notes of CLEF 2020 – Conference and Labs of the Evaluation Forum, Thessaloniki, Greece, September 22-25, 2020*, volume 2696 of *CEUR Workshop Proceedings*. CEUR-WS.

İshak Dölek and Atakan Kurt. 2022. A deep learning model for Ottoman OCR. *Concurrency and Computation: Practice and Experience*, 34(20):e6937.

Rui Dong and David A Smith. 2018. Multi-input attention for unsupervised OCR correction. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2363–2372. Association for Computational Linguistics.

Senka Drobac. 2020. *OCR and post-correction of historical newspapers and journals*. Ph.D. thesis, University of Helsinki.

Senka Drobac, Pekka Kauppinen, and Krister Lindén. 2017. OCR and post-correction of historical Finnish texts. In *Proceedings of the 21st Nordic Conference on Computational Linguistics*, pages 70–76. Association for Computational Linguistics.

Quan Duong, Mika Hämäläinen, and Simon Hengchen. 2021. An unsupervised method for OCR post-correction and spelling normalisation for Finnish. In *Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 240–248. Linköping University Electronic Press, Sweden.

Maud Ehrmann, Ahmed Hamdi, Elvys Linhares Pontes, Matteo Romanello, and Antoine Doucet. 2023. Named entity recognition and classification in historical documents: A survey. *ACM Computing Surveys*, 56(2).

Maud Ehrmann, Matteo Romanello, Alex Flückiger, and Simon Clematide. 2020. Extended overview of CLEF HIPE 2020: Named entity processing on historical newspapers. In *CLEF 2020 Working Notes. Conference and Labs of the Evaluation Forum*, volume 2696 of *CEUR Workshop Proceedings*. CEUR-WS.

Maud Ehrmann, Matteo Romanello, Sven Najem-Meyer, Antoine Doucet, and Simon Clematide. 2022. Extended overview of HIPE-2022: Named entity recognition and linking in multilingual historical documents. In *Proceedings of the Working Notes of CLEF 2022 – Conference and Labs of the Evaluation Forum*, volume 3180 of *CEUR Workshop Proceedings*. CEUR-WS.

Florian Fink, Klaus U Schulz, and Uwe Springmann. 2017. Profiling of OCR'ed historical texts revisited. In *Proceedings of the 2nd International Conference on Digital Access to Textual Cultural Heritage*, DATeCH 2017, page 61–66. Association for Computing Machinery.

Lenz Furrer and Martin Volk. 2011. Reducing OCR errors in Gothic-script documents. In *Proceedings of the Workshop on Language Technologies for Digital Humanities and Cultural Heritage*, pages 97–103. Association for Computational Linguistics.

Government of Ireland. 2022. Digital plan for the Irish language. Speech and language technologies 2023-2027. Accessed: 29 March 2024.

Ivan Gruber, Marek Hrúz, Pavel Ircing, Petr Neduchal, Tomáš Zítka, Miroslav Hlaváč, Zbyněk Zajíc, Jan Švec, and Martin Bulín. 2021. OCR improvements for images of multi-page historical documents. In *International Conference on Speech and Computer (SPECOM 2021)*, volume 12997 of *Lecture Notes in Computer Science*, pages 226–237. Springer.

Helena Hubková. 2019. Named-entity recognition in Czech historical texts: Using a CNN-BiLSTM neural network model. Master's Thesis in Language Technology, Uppsala University.

Helena Hubková, Pavel Král, and Eva Pettersson. 2020. Czech historical named entity corpus v. 1.0. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 4458–4465. European Language Resources Association.

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

Kimmo Kettunen, Mika Koistinen, and Jukka Kervinen. 2020. Ground truth OCR sample data of Finnish historical newspapers and journals in data improvement validation of a re-ocring process. *LIBER Quarterly: The Journal of the Association of European Research Libraries*, 30(1):1–20.

Matthew Knight. 2021. *"Our Gaelic Department": The Irish-Language Column in the New York Irish-American, 1857-1896*. Ph.D. thesis, Harvard University Graduate School of Arts and Sciences.

Mika Koistinen, Kimmo Kettunen, and Jukka Kervinen. 2020. How to improve optical character recognition of historical Finnish newspapers using open source Tesseract OCR engine – final notes on development and evaluation. *Human Language Technology. Challenges for Computer Science and Linguistics*, pages 17–30.

Taku Kudo and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71. Association for Computational Linguistics.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880. Association for Computational Linguistics.

Fiona Lyons. 2021. *Thall is abhus: Irish language revival, media and the transatlantic influence 1857-1897*. Ph.D. thesis, University College Dublin.

Lijun Lyu, Maria Koutraki, Martin Krickl, and Besnik Fetahu. 2021. Neural OCR post-hoc correction of historical corpora. *Transactions of the Association for Computational Linguistics*, 9:479–493.

Hsing-Yuan Ma, Hen-Hsen Huang, and Chao-Lin Liu. 2024. Reading between the lines: Image-based order detection in OCR for Chinese historical documents. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38:21, pages 23808–23810. The Association for the Advancement of Artificial Intelligence.

Jiří Martínek, Ladislav Lenc, and Pavel Král. 2020. Building an efficient OCR system for historical documents with little training data. *Neural Computing and Applications*, 32:17209–17227.

Dermot McGuinne. 1992. *Irish type design: A history of printing types in the Irish character*. Art and Architecture Series. Irish Academic Press.

Christopher Moseley, editor. 2010. *Atlas of the World's Languages in Danger*, 3rd edition. Memory of Peoples. UNESCO Publishing, Paris.

Clemens Neudecker. 2016. An open corpus for named entity recognition in historic newspapers. In *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC'16)*, pages 4348–4352. European Language Resources Association (ELRA).

Deirdre Ní Chonghaile. 2015. "Sagart gan iomrádh": An tAthair Domhnall Ó Morchadha (1858-1935) agus amhráin Philadelphia. In Máirín Nic Eoin, Ríona Nic Congáil, Pádraig Ó Liatháin, Meidhbhín Ní Úrdail, and Regina Uí Chollatáin, editors, *Litríocht na Gaeilge ar fud an Domhain*, pages 191–214. *Leabhair*COMHAR, Baile Átha Cliath.

Vera Provatorova, Svitlana Vakulenko, Evangelos Kanoulas, Koen Dercksen, and Johannes M. van Hulst. 2020. Named entity recognition and linking on historical newspapers: UvA. ILPS & REL at CLEF HIPE 2020. In *CLEF 2020: CLEF 2020 Working Notes: Working Notes of CLEF 2020 – Conference and Labs of the Evaluation Forum*, volume 2696 of *CEUR Workshop Proceedings*, pages 1–8, Thessaloniki, Greece. CEUR-WS.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551.

Christian Reul. 2020. *An Intelligent Semi-Automatic Workflow for Optical Character Recognition of Historical Printings*. Ph.D. thesis, Bayerische Julius-Maximilians-Universität Würzburg (Germany).

Christian Reul, Christoph Wick, Maximilian Nöth, Andreas Büttner, Maximilian Wehner, and Uwe Springmann. 2021. Mixed model OCR training on historical Latin script for out-of-the-box recognition and finetuning. In *The 6th International Workshop on Historical Document Imaging and Processing*, HIP'21, pages 7–12. Association for Computing Machinery.

Martin Reynaert. 2008. Non-interactive OCR post-correction for giga-scale digitization projects. In *International Conference on Intelligent Text Processing and Computational Linguistics*, pages 617–630. Springer.

Martin Reynaert. 2016. OCR post-correction evaluation of Early Dutch books online – revisited. In *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC'16)*, pages 967–974. European Language Resources Association (ELRA).

Caitlin Richter, Matthew Wickes, Deniz Beser, and Mitch Marcus. 2018. Low-resource post processing of noisy OCR output for historical corpus digitisation. In *Proceedings of the 11th International Conference on Language Resources and Evaluation (LREC 2018)*. European Language Resources Association (ELRA).

Shruti Rijhwani, Antonios Anastasopoulos, and Graham Neubig. 2020. OCR Post Correction for Endangered Language Texts. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5931–5942. Association for Computational Linguistics.

Shruti Rijhwani, Daisy Rosenblum, Antonios Anastasopoulos, and Graham Neubig. 2021. Lexically aware semi-supervised learning for OCR post-correction. *Transactions of the Association for Computational Linguistics*, 9:1285–1302.

Jeff Rusten. 2020. Training a multilingual model in Transkribus. Transkribus Blog. Accessed: 29 March 2024.

Sarah Schulz and Jonas Kuhn. 2017. Multimodular domain-tailored OCR post-correction. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2716–2726, Copenhagen, Denmark. Association for Computational Linguistics.

Stefan Schweter and Johannes Baiter. 2019. Towards robust named entity recognition for historic German. In *Proceedings of the 4th Workshop on Representation Learning for NLP (RepL4NLP-2019)*, pages 96–103. Association for Computational Linguistics.

Stefan Schweter, Luisa März, Katharina Schmid, and Erion Çano. 2022. hmBERT: Historical multilingual language models for named entity recognition. In *Proceedings of the Working Notes of CLEF 2022 – Conference and Labs of the Evaluation Forum*, volume 3180 of *CEUR Workshop Proceedings*. CEUR-WS.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725. Association for Computational Linguistics.

Prawaal Sharma, Poonam Goyal, Vidisha Sharma, and Navneet Goyal. 2024. VOLTAGE: A versatile contrastive learning based OCR methodology for ultra low-resource scripts through auto glyph feature extraction. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 881–899. Association for Computational Linguistics.

Richard Sharpe and Micheál Hoyne. 2020. *Clóliosta: Printing in the Irish Language, 1571-1871*. Dublin Institute for Advanced Studies, Dublin.

David A. Smith and Ryan Cordell. 2018. A research agenda for historical and multilingual optical character recognition. Accessed: 29 March 2024.

Elizabeth Soper, Stanley Fujimoto, and Yen-Yun Yu. 2021. BART for post-correction of OCR newspaper text. In *Proceedings of the 7th Workshop on Noisy User-generated Text (W-NUT 2021)*, pages 284–290. Association for Computational Linguistics.

Uwe Springmann, Christian Reul, Stefanie Dipper, and Johannes Baiter. 2018. Ground truth for training OCR engines on historical documents in German Fraktur and Early Modern Latin. *Journal for Language Technology and Computational Linguistics*, 33(1):97–114.

Omri Suissa, Maayan Zhitomirsky-Geffet, and Avshalom Elmalech. 2022. Toward a period-specific optimized neural network for OCR error correction of historical Hebrew texts. *ACM Journal on Computing and Cultural Heritage (JOCCH)*, 15(2):1–20.

Marcella Tambuscio and Tara Lee Andrews. 2021. Geolocation and named entity recognition in ancient texts: A case study about Ghewond's Armenian history. In *Proceedings of the Conference on Computational Humanities Research 2021*, volume 2989 of *CEUR Workshop Proceedings*, pages 136–148. CEUR-WS.

Konstantin Todorov and Giovanni Colavizza. 2020. Transfer learning for named entity recognition in historical corpora. In *Working Notes of CLEF 2020 – Conference and Labs of the Evaluation Forum, Thessaloniki, Greece, September 22-25, 2020*, volume 2696 of *CEUR Workshop Proceedings*. CEUR-WS.

Solenn Tual, Nathalie Abadie, Joseph Chazalon, Bertrand Duménieu, and Edwin Carlinet. 2023. A benchmark of nested named entity recognition approaches in historical structured documents. In *Document Analysis and Recognition – ICDAR 2023*, pages 115–131. Springer Nature Switzerland.

Fionnuala Uí Fhlannagáin. 1990. *Mícheál Ó Lócháin agus An Gaodhal*. An Clóchomhar Tta., Baile Átha Cliath.

Thorsten Vobl, Annette Gotscharek, Uli Reffle, Christoph Ringlstetter, and Klaus U. Schulz. 2014. PoCoTo – an open source system for efficient interactive postcorrection of OCRed historical texts. In *Proceedings of the 1st International Conference on Digital Access to Textual Cultural Heritage (DATeCH 2014), Madrid, Spain, May 19-20, 2014)*, pages 57–61. Association for Computing Machinery.

Alexander Wettig, Tianyu Gao, Zexuan Zhong, and Danqi Chen. 2023. Should you mask 15% in masked language modeling? In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2985–3000. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45. Association for Computational Linguistics.

Tariq Yousef, Chiara Palladino, and Stefan Jänicke. 2023. Transformer-based named entity recognition for Ancient Greek. In *Digital Humanities 2023: Book of Abstracts*, pages 194–195. Zenodo.

13. Language Resource References

Acadamh Ríoga na hÉireann. 2017. *Corpas Stairiúil na Gaeilge 1600-1926*. Royal Irish Academy. Accessed: 29 March 2024.

Dereza, Oksana. 2024. *BART-base fine-tuned for OCR post-correction of historical bilingual Irish-English data*. HuggingFace. Accessed: 29 March 2024.

Dereza, Oksana and Ní Chonghaile, Deirdre and Wolf, Nicholas. 2024. *Historical bilingual Irish-English dataset for OCR post-correction*. HuggingFace. Accessed: 29 March 2024.

Facebook. 2022. *BART-base*. HuggingFace. Accessed: 29 March 2024.

Farrell, Gerard. 2023. *Irish, Gaelic and Roman type (Seanchló agus Cló Rómhánach) v.3*. Transkribus. Accessed: 29 March 2024.

Dublin Institute for Advanced Studies. 2019. Irish script on screen. Website. Accessed: 29 March 2024.

Google. 2023a. *FLAN-T5-base*. HuggingFace. Accessed: 29 March 2024.

Google. 2023b. *T5-base*. HuggingFace. Accessed: 29 March 2024.

Guhr, Oliver. 2023. *Spelling Correction English Base*. HuggingFace. Accessed: 29 March 2024.

Kundumani, Bhuvana. 2022. *T5 Base Spellchecker*. HuggingFace. Accessed: 29 March 2024.

Ní Chonghaile, Deirdre and Dereza, Oksana and Wolf, Nicholas. 2023. *An Gaodhal Newspaper (1881-1898): Full-Text OCR Output Files (Version 1)*. New York University. Accessed: 29 March 2024.

Scannell, Kevin and Regan, Jim and Damazyn, Kevin. 2020. *Tesseract Irish Uncial Training Data*. GitHub. Accessed: 29 March 2024.

Schweter, Stefan. 2020. *Europeana BERT and ELECTRA models (1.0.0)*. Zenodo. Accessed: 29 March 2024.

spaCy. 2023. English language models. Website. Accessed: 29 March 2024.

Transkribus Team. 2021. *Transkribus print M1*. Transkribus. Accessed: 29 March 2024.

University of Galway. 2021. *An Gaodhal Newspaper*. University of Galway. Accessed: 29 March 2024.

University of Galway. 2022. *An Stoc Newspaper*. University of Galway. Accessed: 29 March 2024.