Preservation of Original Orthography in the Construction of an Old Irish Corpus.

Adrian Doyle, John P. McCrae, Clodagh Downey
National University of Ireland Galway
a.doyle35@nuigalway.ie, john.mccrae@insight-centre.org, clodagh.downey@nuigalway.ie

Abstract

Irish was one of the earliest vernacular European languages to have been written using the Latin alphabet. Furthermore, there exists a relatively large corpus of Irish language text dating to this Old Irish period (c. 700 – c 950). Beginning around the turn of the twentieth century, a large amount of study into Old Irish revealed a highly standardised language with a rich morphology, and often creative orthography. While Modern Irish enjoys recognition from the Irish state as the first official language, and from the EU as a full official and working language, Old Irish is almost incomprehensible to most modern speakers, and remains extremely under-resourced. This paper will examine considerations which must be given to aspects of orthography and palaeography before the text of a historical manuscript can be represented in digital format. Based on these considerations the argument will be presented that digitising the text of the Würzburg glosses as it appears in *Thesaurus Palaeohibernicus* will enable the use of computational analysis to aid in current areas of linguistic research by preserving original orthographical information. The process of compiling the digital corpus, including considerations given to preservation of orthographic information during this process, will then be detailed.

Keywords: manuscripts, palaeography, orthography, digitisation, optical character recognition, Python, Unicode, morphology, Old Irish, historical languages

1. Introduction

Encoded within the original, handwritten text of the Würzburg glosses, the earliest large collection of glosses written in the Irish language, is a wealth of scribal knowledge. This paper argues for the preservation of this knowledge in the creation of a digital corpus of Old Irish text by faithfully representing the original orthographic features of the glosses.

An argument will be made in favour of focusing on material found only in manuscripts contemporary with the Old Irish period of c. 700 – c. 950. This paper will next outline why the text of the Würzburg glosses as it appears in *Thesaurus Palaeohibernicus* (Stokes and Strachan, 1901) is the best candidate for digitisation. Finally, the automated digitisation and proofing process of the corpus will be detailed.

2. Irish

Irish was one of the earliest vernacular European languages to have been written using the Latin alphabet. Thurneysen describes the language as "the earliest form of a Celtic language which can be more or less completely reconstructed from extant sources" (1946, p.1). There are many accepted stages in the development of the language from its earliest attested form to that which is spoken today. Of these stages, the earliest three, Primitive, Old and Middle Irish, are collectively referred to as Early Irish. Consisting, for the most part, of personal names, generally in the genitive case, engraved on standing stones utilising the Ogham alphabet, Primitive Irish is poorly attested compared to Old and Middle Irish.

2.1 Old Irish

A decent number of Old Irish texts survive into the modern period, among these the Old Irish Glosses; Würzburg, Milan and St. Gall. Despite the availability of textual source material, however, it is not necessarily useful to treat all texts written in Old Irish as equal. Stifter distinguishes, for example, between *Classical Old Irish* and *Late Old Irish* based on "linguistic variation [within the Old Irish glosses]" (2006, p.10). Before any text can be deemed suitable for inclusion in a digital corpus of Old Irish, it is first necessary that the term, Old Irish, be examined.

McCone notes that "some scholars have been wont to recognise four main phases [in the evolution of Irish]," (1997, p.163), whereby Old Irish can be understood as the language attested "from roughly the beginning of the 8th century to the middle of the 10th century A.D." Material preserved in manuscripts dated within this Old Irish period "is inevitably the corpus from which the norms of Old Irish grammar have been established in the first instance by modern scholarship" (McCone, 1997, p.164), and it is upon such material that Rudolf Thurneysen based his seminal work, *A Grammar of Old Irish* (1946). The surprisingly high degree of uniformity apparent in these texts suggests that Old Irish must have existed as "a literary language whose standard was taught to the Irish 'men of writing' in school, much as standardised Latin was taught to Continental pupils as a language of literary communication, long after Classical Latin had ceased to be a spoken language of the people" (Stifter, 2006, p.10). McCone asserts that "Old Irish can be defined linguistically in terms of a wide range of specific grammatical traits that together constitute a distinctive system" (1997, p.165).

If "essential conformity to the appropriate criteria… constitute grounds for describing … a text as Old Irish, regardless of the date of the manuscript in which it is preserved" (McCone, 1997, p.165), however, it follows that texts such as *Críth Gablach*, surviving in manuscripts dated later than the end of the Old Irish period, and even some written in the modern day, could be described as Old Irish provided they fall within the prescribed linguistic parameters. This notion, raises an issue regarding the potential inclusion of such texts in a digital corpus. A text composed later than the Old Irish period will be more reflective of a scribe's own understanding of an already archaic literary standard than it will be of the standard itself. Even text copied from earlier sources may be unreliable as "Middle-Irish transcribers have often modernised or corrupted these ancient documents" (Stokes and Strachan, 1901). While McCone lambasts "attempts at a more or less clear chronological definition of Old, Middle

and Modern Irish" (1997, p.165) citing "arbitrary transitional dates" as cause for concern, he concedes that material dating to within the Old Irish period "alone can be safely assumed to be free of the possible distortions of significantly later recopying" (McCone, 1997, p.164). A further issue with the inclusion of texts in a digital corpus of Old Irish based on their conformity to outlined linguistic criteria is that it begs the question, how much deviation from these criteria is too much? Despite the high degree of linguistic uniformity apparent in texts preserved in manuscripts dated earlier than the 10th century, McCone (1985) outlines several examples of deviation from the Old Irish norm already apparent in some of the earliest textual sources of Old Irish, including the Würzburg glosses. These deviations, McCone argues, are more consistent with linguistic developments associated with the subsequent Middle Irish period. Outlining hard linguistic criteria with which to justify a given text's inclusion in, or exclusion from a digital corpus is beyond the scope of this paper, and in any case, this practice may limit the utility of the corpus to researchers. Current research projects such as Chronologicon Hibernicum (Stifter, 2015), and LexiChron (Toner and Han, 2018), focus on linguistic features of select texts in order to establish reasonable means by which to linguistically date others. For these reasons this paper will focus on Old Irish text which is preserved only in manuscripts contemporary with the Old Irish period of the 8th to the middle of the 10th century, and will not exclude any such material based on linguistic criteria.

2.2 Old Irish Text and Resources

There are three large sources of Old Irish text which survive in manuscripts dated to within the Old Irish period. These are collectively known as the Old Irish Glosses. These consist of three large collections of interlinear and marginal glosses on Latin texts. The earliest of these, dated to the middle of the 8th century (Stifter, 2006), are the Würzburg glosses on the Pauline epistles. From the early 9th century come the Milan glosses on the psalms, and from the middle of the 9th century come the St. Gall glosses on the Priscian grammar of Latin. Projects undertaken by Dr. Aaron Griffith (2013) of the University of Vienna, and Dr. Pádraic Moran (2014) of the National University of Ireland, Galway have already collected, and published in digital format, the text of the Milan and St. Gall glosses respectively. While Kavanagh and Wodtko (2001) have produced a lexicon based on the Würzburg glosses, no collection has been published in digital format to date. For this reason, this project has been focused on the process of digitising the text of the Würzburg glosses.

Of the glosses which have been digitised, Moran (2014) suggests that St. Gall contains about 9,400 glosses, over a third of which are written in Old Irish. These do not equate to full sentences, as many glosses are fragmentary, or contain single words or phrases. Nonetheless, assuming a similar number are present in the Milan glosses, that brings the extant digital corpus of Old Irish to only about 6000 glosses. There currently exists no part-of-speech (POS) tagged corpus for any complete set of glosses. In fact, POMIC (Lash, 2014), a collection of fourteen Old and Middle Irish texts, contains the only currently available POS tagged text in Old Irish. While this provides an excellent resource for computer-based Early Irish research, texts which match this paper's definition of Old Irish are few, and those which have been POS tagged are fewer. As such, Old Irish remains highly under-resourced.

3. Old Irish Orthography and Palaeography

Having settled upon an appropriate Old Irish corpus, namely the Würzburg glosses, consideration must next be given to the source of text which will be drawn upon. As will be demonstrated in this section, drawing upon the original text as it appears on the folio would present many technical issues resulting from the original orthographic stylings of Old Irish scribes. In many cases, modern editors cannot preserve characters of the original manuscript script, and hence, must make emendations which alter the orthography of resultant modern editions.

The insular script employed by Irish scribes utilises a selection of variations on Latin alphabetical symbols. Many of its distinctive letters, diacritics and symbols, such as "ɔ", "Ṡ", and "Ḟ", are supported by Unicode. As such, much of the orthography of Old Irish text can be represented digitally. Nevertheless, a variety of contraction markers which remain unsupported by Unicode, and which are used throughout the Old Irish glosses, prevent them from being perfectly represented by Unicode characters alone. These abbreviating contractions come in many forms, and are used in place of the plene spelling of a word. One common example is the suspension stroke which can be used in combination with a variety of different letters to produce various differing sounds. Combination with the letter "b", for example, could produce the sound "bar". Hence, words like "Conchobar", could be written out in full, or contracted, "ɔchob" with a suspension stroke over the "b". As such, the use of contractions in Old Irish text saves valuable space on vellum. Editors compiling modern editions may opt to represent such contractions by supplying the full plene spelling in their place. Such is the case with the two-volume collection of Early Irish texts, *Thesaurus Palaeohibernicus* (TPH) (Stokes and Strachan, 1901; Stokes and Strachan, 1903) Importantly for the purpose of this project, the editors of TPH retain many diacritics and symbols such as those outlined earlier. Moreover, where orthographic features could not be retained, the editors identify plene text which they have supplied. Therefore, by drawing upon the text as it appears in TPH, it is possible to digitise the contents of the Würzburg glosses extremely faithfully, without sacrificing important elements of the source material's orthography. This, in turn, will allow for statistical linguistic analysis to be carried out on a digital corpus of Old Irish text which represents the language in as close a manner as possible to its original format.

4. Digitisation of *Thesaurus Palaeohibernicus*

Both volumes of TPH were initially captured using a Kirtas (Kirtas, 2015) scanner with APT manager software, and edited with Book Scan Editor software. At this point, ABBYY FineReader (ABBYY, 2018) OCR software was utilised to recognise the text in the captured image files. The output of this process was a machine-readable PDF file containing both the image, and digital text of the entire two volumes of TPH.

The character recognition, while generally successful on the English language content, apparently had difficulty

with the Latin, and particularly with the Irish text. Footnote markers were regularly mistaken for a variety of characters not present in the hard copy, including "®", "*", and "^". Diacritic markers present in the Irish text posed a similar problem. Often, acute accents were represented mistakenly as umlauts, for example "domsa höre" for "domsa hóre". Even in the English text where character recognition had been generally better, the regularity with which characters were mistaken warrants strict proofing of each line. Examples of such mistakes include, "…because 1 believe…", and "he vvho shall believe". Information originally appearing in marginalia, such as folio and line numbers, were often combined erroneously with linguistic text.

4.1 Automated Analysis of OCR Success

A number of Python scripts were written to measure the general success of the OCR process. Initial efforts were focussed on measuring the success of character recognition in page headers, as these contain TPH page numbers. The page range of the Würzburg glosses in TPH spans from 499 to 712. The first script written checked to see if all of these page numbers exist, in sequence, within the digital text. This found that 28 of 214 page numbers, roughly 13%, had been incorrectly digitised. Once identified, these missing page numbers were manually corrected.

The next concern was to discover how many page headers had their textual content correctly digitised. Another Python script identified 27 page headers, about 12.6%, which had been incorrectly digitised. Again, these were manually corrected.

4.2 Automatic Approach to Proofing

With all page headers and numbers now correctly in place, a script was written to count each line of text per page and represent it as a sequentially increasing decimal following the relevant page number. For example, the title line on the first page of the Würzburg glosses would be indexed at 499.1. This new indexing system allows for quick comparison between the often difficult to recognise digitised text and its equivalent text in TPH. This has increased the speed with which the digital text can be proofed by eliminating excess time spent attempting to recognise a given line of text.

At this point the text was still interspersed with arbitrary characters where diacritics and footnotes had been incorrectly assigned. These characters made manual proofing a cumbersome task. A script was written to replace any such unexpected characters with a single underscore, this would serve as a clear signal to a proof reader that something had been removed, and hence, a given section of text would require particular attention.

4.3 Preservation of Information

TPH contains a variety of metadata related to the text on a given page. Page headers, mentioned earlier, contain not only TPH page numbers, but also information pertaining to the content of the text on the page. Even numbered page headers read "Biblical Glosses and Scholia", a common theme throughout the first volume of TPH. Between pages 499 and 712, odd numbered page headers read "Glosses on the Pauline Epistles." followed by information on the specific letters referenced on the page, beginning "Rom. I." on page 499 and working through to "Heb. V, VI." on page 711. It is a simple matter, therefore, to create a Python dictionary into which page number and content data can be automatically collected as keys and values respectively. Similar information is contained in secondary titles on pages where a new set of letters begin. A primary section title, "CODEX PAULINUS WIRZIBURGENSIS", is given on page 499. Such titles, when encountered during proofing, are surrounded by square bracket tags, [H2][/H2] and [H1][/H1] respectively, in order to enable automatic identification of them at a later stage.

The text presented on a typical TPH page is split into three sections. The first, presented at the top of each page, is the Latin text of the Pauline Epistles. Only lines containing glosses are included, and the point within a line of text to which a gloss corresponds is marked with a superscript number. These footnote-style numbers caused particular trouble during the OCR process. No instance of these was correctly digitised. In proofing these are replaced with the same number, enclosed within square brackets. The section itself is also enclosed within tags which identify it as the original Latin text, [Lat][/Lat].

The second section, positioned in the middle of each page, contains the text of the glosses which relate to the Latin text above. Each gloss is numbered in accordance with the superscript numbers of the above Latin section. These glosses are written in a combination of Latin and Old Irish, with code switching occurring regularly. The editors of TPH distinguish between the two languages by printing Irish content in italics, while leaving Latin text unaltered. Where part of an Irish word has been supplied by editors in place of a manuscript contraction, this is identified by the editors by returning to roman type. Such supplements are surrounded by contraction tags, [Con][/Con], during the proofing process to preserve metadata relating to breaks from original orthography. Similarly, letters supplied by the editors but omitted in the manuscript are identified in TPH by square brackets. In proofing these are replaced by supplement tags, [Sup][/Sup].

Like the Latin section above, the glosses are surrounded with [SG][/SG] tags, identifying the section as a whole. However, Latin content within the section is separated from Irish content by means of separate Latin tags, [GLat][/GLat], which surround uninterrupted strings of Latin text, as well as individual instances of Latin abbreviations such as ".i." and "ł" which frequently appear in TPH. The Latin tags used within the glosses' section are distinct from those used earlier to ensure that the Latin content of each section can be automatically identified as separate. Within this section footnotes are marked out by means of superscript alphabetical letters. These are matched in a footnote section at the bottom of each page. As with the superscript numbers of the Latin section, these markers caused difficulty for the OCR software and none were correctly identified. In proofing these are replaced by the same letter enclosed within square brackets. In instances where the footnote suggests that the editors have emended a manuscript form, or supplied a form not present in the manuscript, the word is surrounded by opening and closing tags bearing the letter of the relevant footnote, for example, [a][/a]. This will allow the original manuscript orthography to be automatically restored. The tag-set utilises square brackets so that single-letter tags such as these will not be mistaken for html tags identifying elements such as hyperlinks, bold text, or paragraphs.

The third section, towards the bottom of a page, placed just above the footnotes, provides an English translation of the

Irish gloss content. Where present the Latin gloss content is left untranslated, however, much of it is simply removed. Footnotes continue into this section from the preceding section of glosses, and are treated in the same manner. The section is enclosed within tags, [Eng][/Eng], which identify its content as the translation of the glosses above.

Information regarding the location of a given page's text within the original manuscript is given in the outer marginalia of each page, to the left or right of the block of text to which they refer. Information supplied here includes the folio number, and a letter corresponding to the column on that folio, from which the text was taken. This folio information, regularly combined mistakenly with the main body of text during the OCR process, is removed during proofing and replaced with folio tags which surround the relevant blocks of text, for example, [f. 1a][/f. 1a].

The preservation of this metadata by means of a specialised tag-set creates a number of possibilities for researchers (Petrova, et al., 2009). The original text can be drawn upon as easily as the text which appears within the pages of TPH. Moreover, the identification of original orthographic details by means of tags allows for statistical analysis of variant spellings and word choices which may be useful to researchers in the identification, by computational means, of different scribal hands, linguistic registers, and dialect within the glosses.

5. Further Use of the Digital Corpus

As this paper is being written, proofing of the text content is ongoing. Once this process has been completed, focus will shift to POS and dependency tagging of the glosses, after which the corpus will be made available online. Ultimately, it is expected that this corpus will aid researchers in the field of Early Irish by allowing automation of a variety of research tasks, a possibility first proposed by Teresa Lynn (2012).

6. Conclusion

In creating a digital corpus for a historic language, preservation of the original orthographical content enables significant forms of text analysis to be performed on the resultant digital corpus. This paper advocates careful selection of source material, such as *Thesaurus Palaeohibernicus* (Stokes and Strachan, 1901; Stokes and Strachan, 1903), which, where possible, carefully preserves distinct orthographic diacritics and symbols where present in the original manuscript. A method is outlined for the preservation of metadata relating to original orthographical features of manuscripts where editors have been unable to preserve the features themselves in their edition.

In the case of Old Irish, it is envisioned that the production of this digital corpus will aid in research tasks which rely on the study of orthographical features by allowing the automation of tasks dependent on these features. Such tasks may include identification of different scribal hands, identification of linguistic register or dialect, and linguistic dating, where such tasks may be based on the frequency or location of orthographical features within a text.

This paper has shown that the speed with which Old Irish text can be digitised can be significantly increased by the combined use of OCR software with a variety of techniques intended to improve the proofing process.

A tag-set has been created which will be used to identify features within the digital corpus including original folio information, points of scribal contractions, text supplied by editors, code switching between Irish and Latin, editorial emendations of provided manuscript forms, as well as headers, sections and footnotes present in the source material. As the text requires proof reading, implementation of this new tag-set will be carried out in tandem with this process. Therefore, time taken to produce this digital corpus will not be significantly increased by its introduction.

7. Acknowledgements

This research is supported by the National University of Ireland, Galway's DAH (Digital Arts and Humanities) scholarship and by Science Foundation Ireland under Grant Number SFI/12/RC/2289 (Insight).

8. Bibliographical References

Griffith, A. (2013). A Dictionary of the Old-Irish Glosses. http://www.univie.ac.at/indogermanistik/milan_glosses.htm (Accessed: 10/01/2018).

Kavanagh, S. & Wodtko, D.S. (2001). A Lexicon of the Old Irish Glosses in the Würzburg Manuscript of the Epistles of St. Paul. Verlag der Österreichischen Akademie der Wissenschaften, Vienna.

Lash, E. (2014). The Parsed Old and Middle Irish Corpus (POMIC). Version 0.1. https://www.dias.ie/celt/celt-publications-2/celt-the-parsed-old-and-middle-irish-corpus-pomic/ (Accessed: 10/01/2018).

Lynn, T. (2012). Medieval Irish and Computational Linguistics. *Australian Celtic Journal*, 10:13-28.

McCone, K. (1985). The Würzburg and Milan Glosses: Our Earliest Sources of 'Middle Irish'. *Ériu*, 36:85-106.

McCone, K. (1997). The Early Irish Verb. An Sagart, Maynooth, 2nd edition.

Moran, P. (2014). St Gall Priscian Glosses. http://www.stgallpriscian.ie/ (Accessed: 10/01/2018).

Petrova, S., Solf, M., Ritz, J., Chiarcos, C. & Zeldes, A. (2009). Building and using a Richly Annotated Interlinear Diachronic Corpus: The Case of Old High German Tatian. Traitement automatique des langues, 50(2), 47–71.

Stifter, D. (2006). Sengoidelc. Syracuse University Press, New York.

Stokes, W. & Strachan, J. (Eds.). (1901). Thesaurus Palaeohibernicus Volume I. The Dublin Institute for Advanced Studies, Dublin, 3rd edition.

Stokes, W. & Strachan, J. (Eds.). (1903). Thesaurus Palaeohibernicus Volume II. The Dublin Institute for Advanced Studies, Dublin, 3rd edition.

Thurneysen, Rudolf. (1946). A Grammar of Old Irish. The Dublin Institute for Advanced Studies, Dublin.

9. Language Resource References

ABBYY. (2018). FineReader. https://www.abbyy.com/en-eu/finereader/ (Accessed: 10/01/2018).

Kirtas. (2015). https://www.kirtas.com/ (Accessed: 10/01/2018).

Stifter, D. (2015). Chronologicon Hibernicum. http://dhprojects.maynoothuniversity.ie/chronhib/ (Accessed: 10/01/2018).

Toner, G. & Han, X. (2018). LexiChron. https://www.qub.ac.uk/schools/ael/Research/Languages/LexiChronProject/ (Accessed: 10/01/2018).