Development of Old Irish Lexical Resources, and Two Universal Dependencies Treebanks for Diplomatically Edited Old Irish Text

Adrian Doyle
Department of Classics
School of Languages, Literatures, & Cultures
University of Galway
adrian.odubhghaill@universityofgalway.ie

John P. McCrae
Insight SFI Centre for Data Analytics
Data Science Institute
University of Galway
john@mccr.ae

Abstract

The quantity and variety of Old Irish text which survives in contemporary manuscripts, those dating from the Old Irish period, is quite small by comparison to what is available for Modern Irish, not to mention better-resourced modern languages. As no native speakers have existed for more than a millennium, no more text will ever be created by native speakers. For these reasons, text surviving in contemporary sources is particularly valuable. Ideally, all such text would be annotated using a single, common standard to ensure compatibility. At present, discrete Old Irish text repositories make use of incompatible annotation styles, few of which are utilised by text resources for other languages. This limits the potential for using text from more than any one resource simultaneously in NLP applications, or as a basis for creating further resources. This paper describes the production of the first Old Irish text resources to be designed specifically to ensure lexical compatibility and interoperability.

1 Introduction

While most Old Irish text surviving in contemporary manuscripts, those dated between roughly the seventh and tenth centuries, is accessible in discrete online repositories (Bauer et al., 2023; Griffith, 2013; Stifter et al., 2021a), a lack of standardisation between these resources as regards word separation, lexical annotation, and text normalisation has been well documented. Several studies have reported that attempts at applying natural language processing (NLP) techniques to Old Irish text have been impacted by this lack of standardisation (Doyle et al., 2019; Dereza et al., 2023a,b), and Old Irish had to be excluded from most subtasks undertaken as part of the SIGTYP 2024 Shared Task on Word Embedding Evaluation for Ancient and Historical Languages "as the quantity of historical Irish text data which has been tokenised and annotated to a single standard to date is insufficient for the purpose of training models to perform morphological analysis tasks" (Dereza et al., 2024, 162).

In the creation of new text resources for Old Irish, more consideration needs to be given from the outset to ensuring compatibility with existing resources. As for extant resources, concerns over their long-term sustainability are common, and anxieties often exist among those producing such resources regarding hosting costs, cyber attacks, and gradual obsolescence of technologies and frameworks over time. Stifter et al. (2021b, 8) identify *interoperability* and *sustainability* as key concerns, and claim that, during their workshops, "A recurring message was to keep things simple and stick to standard technologies".

This paper presents three Old Irish text resources which have been created with the express purpose of ensuring lexical compatibility between them. Word separation for Old Irish is not a trivial task, however, the recent development of a unified tokenisation method for Old Irish text (Doyle and McCrae, 2025) has made the prospect of lexically standardising Old Irish resources more attainable than before. The resources described in this paper were developed in tandem with that method, being kept up-to-date with all changes made to it throughout its development. Section 2 discusses the creation of the *Würzburg Irish Glosses* website (Doyle, 2018) which contains the text of the earliest large collection of glosses written in the Irish language. It goes on to describe some of the most substantial updates to the website's contents and functionality since its launch. Section 3 describes the tokenisation and annotation of the website's text, explaining how it conforms to Universal Dependencies

Proceedings of the 5th International Conference on Natural Language Processing for Digital Humanities, pages 393–402

(UD) guidelines (Zeman, 2016). Next, section 4 addresses the production of two UD treebanks containing diplomatically edited Old Irish text, each drawn from a different manuscript source. Finally, section 5 discusses the standardisation of lemmata across all of these resources.

2 The Würzburg Irish Glosses Website

Dating from about the 8th century, the Würzburg (Wb.) glosses on the Pauline epistles are the earliest of three large collections of glosses surviving from the Old Irish period, alongside the Milan and St. Gall (Sg.) glosses. As of 2017, however, they remained the last of the three to be made available online. The digitisation process of these glosses was described in Doyle et al. (2018), at which time it was reported that proofing of the digitised content was ongoing, alongside metadata annotation. It was claimed that "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" (2018, 70). In fact, the earliest version of a website hosting this data was live as early as October 2018, before proofing and annotation had even been completed, and the entirety of the Old Irish text contents were available on this resource by November 2018. The launch of this website made the digital text of the Würzburg glosses publicly available for the first time.

From the outset, the *Würzburg Irish Glosses* website (Doyle, 2018) utilised a JSON document to serve all gloss data to client-side machines upon loading the website. While bandwidth intensive, and perhaps slow to process on older hardware, this allows the website to be very responsive once initially loaded. This contrasts the operation of other online collections of Old Irish glosses based around a relational database back-end (Bauer et al., 2023; Griffith, 2013; Stifter et al., 2021a), which serve data for individual glosses on a request-by-request basis. While the contents of this JSON file have been updated several times, both to include more gloss information and to expand the metadata tag-set[1] from what was initially described in Doyle et al. (2018), the earliest version of the website was relatively rudimentary, and offered no interactivity to users. Updates in early 2019 allowed information about individual glosses from *Thesaurus Palaeohibernicus* (TPH; Stokes and Strachan, 1901, 499-712), to be displayed upon clicking on the text of a given gloss. This information included the Latin verse information and text which had been glossed, English translations of glosses, as well as footnotes and page numbers from the original print edition.

In early 2021 a new textual metadata field called `Site Notes` was introduced to provide new information and commentary for certain glosses, as well as to reference more recent scholarship than was available at the time TPH was published. Soon after, functionality was added to the site to display tokens, headwords, and part-of-speech (POS) tags beneath glosses, along with the already existing gloss information and `Site Notes`. Though only a small number of glosses had been experimentally tokenised or annotated by this stage, the number of tokenised and lexically annotated glosses would increase in stages over the following years. With this step, the contents of the *Würzburg Irish Glosses* website were brought into lexical alignment with those of the two UD treebanks described in section 4, though the site itself predated their creation.

The next major update did not take place until mid-2024, when two new metadata fields were introduced. The first, `New Reading`, allows for an updated transcription to be supplied for a gloss either where more recent scholarship has cast doubt on the transcription supplied in TPH, or where it has otherwise been impossible to tokenise the transcription supplied in TPH. The second field, `New Translation` allows for a new English translation to be supplied for a gloss, either where no translation was supplied in TPH, or where the supplied translation has been questioned in later scholarship. At the same time, a new lexicon feature containing all annotated tokens from the corpus was added to the website. Headwords are linked to entries in the *Electronic Dictionary of the Irish Language* (eDIL; Toner et al., 2019), currently the most complete digital lexicon to include Old Irish lexical information. Links were also added from folio numbers on the website to images of the facsimile available online

[1] For more details regarding this expanded tag-set see Doyle (2024, 48-54)

at TITUS (Stern, 1910).

All of the code required to generate the website is available on GitHub[2]. The `README` file for the GitHub repository explains how to download both the current and any historical versions of the website, and how to host any such version on a local machine. This provides a form of version control for the website. More importantly, though, if the JSON file were the only thing available, some programming knowledge would be necessary to extract any required information. The aim here is to ensure that not only will the text data remain available well into the future, even if the website itself should go offline for any reason, but that even the website's GUI will remain accessible for users who may have limited technical knowledge, or who simply do not wish to interact with the raw data. Meanwhile, interested parties with the required technical knowledge will be able to create a fork of the repository and adapt the website as their needs require, even long after support for the website ceases.

3 Tokenisation and POS-tagging of the Würzburg Irish Glosses Website

For the first two and a half years of its existence no lexical annotation was available on the *Würzburg Irish Glosses* website. At the time, all extant lexical resources for Old Irish made use of discrete word separation methods which are incompatible with one another, and which result in word forms that are not typical of word-level tokens used in lexical resources such as UD treebanks (Doyle and McCrae, 2025). As such, when the time came to apply lexical annotation to the contents of the *Würzburg Irish Glosses* website there was no clear preference as regards a method for applying word separation to the text.

In lieu of a generally agreed-upon method for separating Old Irish words, it was ultimately decided that a new approach should be utilised. While it was desirable to add lexical annotation to the website's contents, it was deemed unnecessary to produce the type of deep morphological analyses which were available in other gloss repositories (Griffith, 2013; Bauer et al., 2023; Stifter et al., 2021a), as a perfectly sufficient lexicon of this nature was already available in print for the Würzburg corpus (Kavanagh and Wodtko, 2001). Instead, with the aim of supporting downstream NLP applications, word separation and POS-tagging was carried out in a manner more closely resembling what is commonly applied to other European languages. Specifically, the decision was made to adhere to UD guidelines for tokenisation and POS-tagging (Zeman, 2016), as the popularity and widespread adoption of this format would likely provide the greatest level of future-proofing for the resulting annotated text. Tokenisation was applied manually in several stages, beginning as early as 2020, with headword annotation and POS-tagging (using the UD POS tag-set) being carried out in tandem. The tokenisation method, which would eventually be described in Doyle and McCrae (2025), was updated and refined regularly based on the emerging requirements of the text of both the Würzburg and the St. Gall glosses (see section 4) as the two corpora underwent the annotation process.

Of the 3,648 glosses which comprise the contents of the *Würzburg Irish Glosses* website, at the time of this writing 611 glosses (about 16.75% of the corpus) have already been tokenised and POS-tagged. This includes all of the glosses on the last three epistles (Titus, Philemon, and Hebrews), and of the three scribal hands which are evident in the manuscript, all glosses by the *prima manus* and the third scribal hand have already been tokenised and POS-tagged. Within the contents tagged to date, there are 1,890 unique Old Irish token types. Because code-switching is common in the glosses, 582 unique token types have also been identified as Latin.

4 Universal Dependencies Treebanks for Old Irish

As UD guidelines for tokenisation and POS-tagging were being applied to Old Irish text, the obvious next step was to produce an Old Irish UD treebank. In fact, two such treebanks were created at about the same time by different means. Syntactic parsing of Old Irish text had already been carried out at least once before, in the *Parsed Old and Middle Irish Corpus* (POMIC; Lash, 2014b), and POMIC

[2] https://github.com/AdeDoyle/WurzburgSiteCode

was also the first corpus containing Old Irish to make use of a widely adopted POS tag-set. Lash (2014a) notes that POMIC utilises a form of Penn-style POS-tags (Santorini, 1990), adapted for Old Irish from an earlier tag-set, itself having been developed for use with historical varieties of English (Santorini, 2016). As has been discussed in Doyle and McCrae (2025), the word separation used in POMIC necessitates alteration to the character content of the text, and therefore may not be adaptable to diplomatically edited text. As such, the treebanks described below represent a number of firsts for Old Irish. They are the first corpora of Old Irish to utilise a single, documented tokenisation method, the first diplomatically edited corpora of Old Irish to be lexically annotated, the first corpora of Old Irish to utilise a POS tag-set which is widely applied to other languages without adaptation, and the first dependency parsed corpora of Old Irish.

While the ubiquity of UD and of the CoNLL-U format will, hopefully, allow for the treebanks discussed here to be both easily accessible and interoperable, certain limitations should be referenced here also. The tokenisation method applied here adheres rigorously to UD guidelines (Zeman, 2016), and primarily the requirement that "the basic units of annotation are syntactic words (not phonological or orthographic words)". This necessarily dictates that certain lexical elements be separated, and hence annotated in ways which may not be familiar to Old Irish scholars. While much could be written about the implications of this on Old Irish morphology and syntax, it is not feasible to have this discussion in the space available here. Instead, the reader is directed to the sections on the verbal complex and miscellaneous tokens in Doyle and McCrae (2025, 5-7).

A distinction worth mentioning between most other UD treebanks, and those for Old Irish is that between "diplomatic" and "critical" editions. A diplomatic edition is typically one which reproduces text as closely as possible to how it appears in an original manuscript source. By contrast, a critical edition will generally contain a single version of a text, along with introductory matter, as well as explanatory, and textual notes. Features like spelling, word separation, and even vocabulary in a critical edition may be quite distinct from anything surviving in a manuscript source. Such alterations to texts may not be obvious in resources like UD treebanks which do not contain forewords explaining editorial decisions. As such, it may not be clear whether an Old Irish treebank contains text which remains very close to something preserved in a specific manuscript or whether it has been been altered to any extent by a modern editor. For this reason, all Old and Middle Irish treebanks are required to state in their `README` documentation which type of edition they represent by using either the "diplomatic" or "critical" designation. This information should also be included in the treebank name and URL using the abbreviations `Dip` and `Crit` (for example, the *Diplomatic St. Gall Glosses Treebank* URL ends: `…/UD_Old_Irish-DipSGG`). This should enable diplomatic and critical editions on UD to be automatically distinguished by web-scrapers. As it may be unclear for some treebanks which designation would be the most suitable, specific requirements and definitions are outlined in the language specific documentation for Old Irish on the UD website[3].

UD distinguishes between languages using ISO 639 codes. This has ramifications for many languages, but perhaps especially for historical language stages like Old Irish. While it is generally accepted that Primitive, Old, and Middle Irish are different historical stages of the same language, each has a distinct ISO 639 code. This means that, so far as UD is concerned, each is to be treated as a distinct language, and Old Irish text should be rigidly distinguished from either Primitive or Middle Irish text. In the case of historical language stages, however, such a distinction can be difficult to make. Old Irish could, for instance, be understood as a sort of linguistic standard, whereby if a particular set of grammatical and orthographic rules are followed, a text may be identified as "Old Irish" even if it is preserved in a manuscript dated to later than the Old Irish period, presumably having been copied from an earlier source. Alternatively, the case may be made that only text surviving in manuscripts dated to the Old Irish period itself, and not

[3] https://universaldependencies.org/sga/index.html

later, constitutes Old Irish. Good reasons exist for preferring either interpretation, for example, Stokes and Strachan note that "unfortunately the Middle-Irish transcribers have often modernised or corrupted these ancient documents. Therefore, in forming a collection of [Old Irish] texts on which scholars may rely with confidence the only safe rule is to exclude all matter not found in [manuscripts] anterior to the eleventh century" (1901, xi). On the other hand, McCone argues that "attempts at a more or less clear chronological definition of Old, Middle and Modern Irish along" temporal lines "are at best crude, particularly as regards the arbitrary transitional dates" (1997, 165). An attempt is made in UD treebanks to facilitate both interpretations as much as is possible while remaining consistent with the ISO 639 code scheme. Thus, if a treebank contains only text from a manuscript dated to the Old Irish period, it is considered an Old Irish treebank whether the edition is critical or diplomatic. If, however, the treebank contains text from a manuscript dated later than the Old Irish period, and it is a diplomatic treebank, it should use the appropriate ISO 639 code for the approximate date of the manuscript regardless of any linguistic dating of the text contents. Finally, if the editor of the text of a treebank has indicated that they have edited it such that it reflects the language of the Old Irish period, despite being drawn from one or more manuscripts dated later than the end of the Old Irish period, this may be identified as an Old Irish treebank but must also be designated a "critical edition".

4.1 The Diplomatic St. Gall Glosses Treebank (DipSGG)

The earliest attempt to create a tokeniser for Old Irish (Doyle et al., 2019) did not result in any particularly successful models, however, the paper concluded, "It may be possible to improve upon performance by training on a corpus of pre-processed glosses" (2019, 78). Of course, no sufficiently large collection of glosses existed at the time which had been either tokenised in a conventional manner at the word-level, or annotated using a common POS tagset. While it was feasible to manually tokenise and annotate a small number of glosses (see section 4.2), this would not be nearly enough to be useful for training LSTM-based tokeniser models like those described in Doyle et al. (2019). It was therefore necessary to produce a relatively large quantity of POS-tagged Old Irish text in a relatively short amount of time. For this reason an attempt was made to automate the process of transferring the annotations used in an existing corpus of Old Irish glosses over to UD annotations (Zeman, 2016).

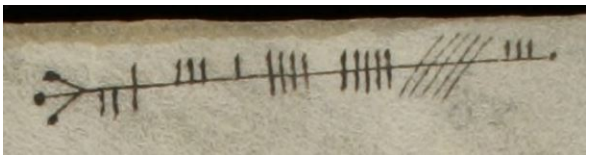

Figure 1: LATHEIRT in Ogam, from St. Gallen, Stiftsbibliothek, Cod. Sang. 904 f. 204 (`www.e-codices.ch`).

The *Diplomatic St. Gall Glosses Treebank* (Doyle, 2023a) was adapted from the contents of the *St Gall Priscian Glosses* database (Bauer et al., 2023), which were kindly made available by Bauer et al. for this purpose. The contents of the database were processed to generate a `CoNLL-U` file, with each gloss meeting UD requirements for tokenisation, headword assignment, POS-tagging and morphological feature annotation. First, however, certain grammatical and morphological features were re-analysed, new translations were provided, and the data was restructured. The St. Gall manuscript contains several glosses written in Ogam (or Ogham) script (see figure 1). These appear in the *St Gall Priscian Glosses* database transliterated into Roman Script, but were manually reverse-transliterated back into Ogam for the new treebank in the interest of producing the most diplomatic edition possible. Next, the text was automatically cleaned to remove HTML tags and ahistorical punctuation inserted by modern editors.

After cleaning, it was necessary in some cases to alter existing readings, or to provide new ones, which typically necessitated referencing the manuscript or other scholarly work. While it would be impossible to give an exhaustive list of examples in the space available here, the following few should be sufficient to demonstrate the kind of alterations made. In one case a personal name, written *donngvs* in the manuscript (see figure 2), is rendered *donn-*

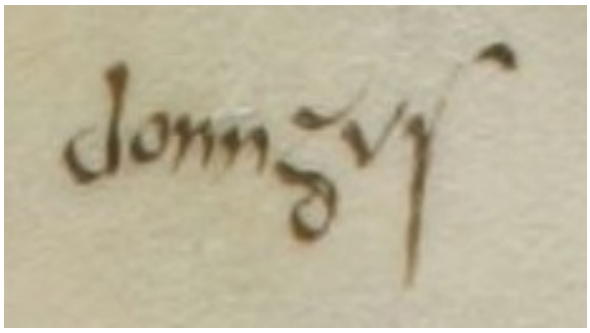

Figure 2: *donngvs* from St. Gallen, Stiftsbibliothek, Cod. Sang. 904 f. 194 (www.e-codices.ch).

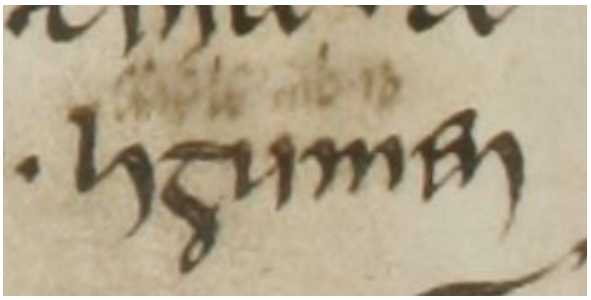

Figure 3: *cenéle ṁb[i]id*, glossing *ligumen* from St. Gallen, Stiftsbibliothek, Cod. Sang. 904 f. 113 (www.e-codices.ch).

gus in the database (see Bauer et al., 2023, `194b m.s.`). The manuscript spelling was restored for the new treebank in an attempt to be as diplomatic as possible. In another case, a gloss which reads *ruadri adest* "Rúadrí is here" (`159a m.s.`) was neither translated nor annotated in the database. A new translation and annotations were therefore supplied in the treebank. Finally, Bauer et al. include another gloss which does not appear in *Thesaurus Palaeohibernicus* (see Stokes and Strachan, 1903, 145). Bauer et al. provide this gloss with a new numbering of 113b32, and suggest the tentative reading *cenele*? ? *..b.so*?. No translation is provided for the entire gloss, however, the analysis correctly identifies the first word as *cenéle* "kind/sort". This is a gloss on the Latin *ligumen* "pod-vegetable". In fact, the Irish reads *cenéle ṁbiid* "a type of food", though the gloss is blurred and difficult to read in the manuscript (see figure 3). This new reading and translation were supplied in the new treebank, and an analysis was provided for the missing form, *ṁbiid*.

The next major undertaking was to manually transfer each of the 1,601 distinct morphological analyses are used by Bauer et al. (2023) to their equivalent UD POS-tags and morphological feature sets. A series of relatively complex regular expressions were used to parse analyses like "`3sg.pres.ind.pass. + infix.pron. class A 1sg.`" and extract necessary morphological information. This morphological information could then be mapped to the UD format for morphological features, like `Mood=Ind | Number=Sing | Person=3 | Tense=Pres | Voice=Pass`.

Once complete, the positions and placement of the annotated tokens had to be identified in the raw text. This step was necessary as only Irish words have been morphologically analysed by Bauer et al. (2023), and no "word forms" or analyses are provided for Latin text occurring in glosses. This meant that the only way to isolate the Latin text, so that it could be accurately annotated, was to first identify the Irish text. Thereafter, when the Irish text is removed, all remaining text can be assumed to be Latin. Matching each of the analysed Old Irish tokens to the correct substring within the raw text of the full gloss was often an extremely difficult task, however. In many cases, compound forms in the raw text are be split into multiple "word forms", each of which is morphologically analysed, and these "word forms" may not precisely reflect the exact character content of the raw text. As such, multiple analysed tokens or morphemes may need to be identified with a single word in the full gloss text (see figure 4).

To overcome this issue, a complicated method was devised which would be triggered when parts-of-speech which could potentially form compounds, like verbs and the copula, were found. Once such a POS was identified, the method would work backwards through the preceding tokens to determine if they were the types of POS which could potentially form a compound, and if so, they would be added to the token which had triggered the method if they did not already exist within it. Conversely, where preceding tokens were found to be doubled in a following compound, they had to be removed from it, taking care not to remove initial consonant mutations in doing so (see figure 5). This meant that a long list of all poten-

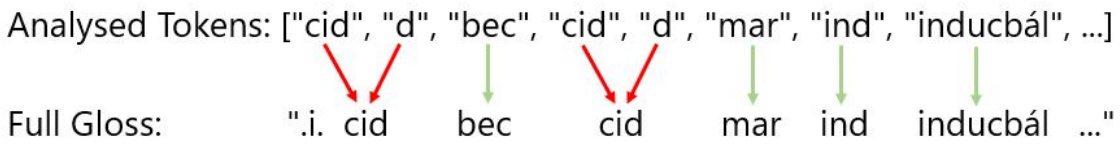

Figure 4: Example of repeated/doubled text characters from Sg. 2a7.

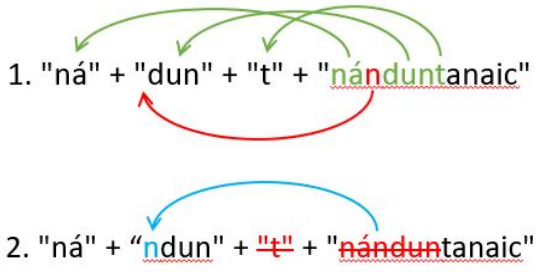

Figure 5: Example from Sg. 26b2 demonstrating how compounds are deconstructed. First elements which are doubled in the full verbal form are identified, second these are deleted from the verb form, and the nasal *n*, which was not doubled, is moved to the beginning of the separate token *dun*.

tially combining forms had to be painstakingly compiled by trial and error. Once complete, however, all Old Irish text adhered to UD tokenisation and annotation requirements. The remaining Latin content was then tokenised based on spacing, and the `X` POS-tag was applied to all Latin tokens.

The GitHub repository for the treebank was created in 2020, and the entirety of the St. Gall corpus was uploaded to the development branch in July of that year. Without dependency parsing, however, none of this content could be included in an official UD release at this time. Between February and April 2023, dependency annotation was manually added to sixty-three glosses from the St. Gall corpus. These included three poems, all of the Ogam glosses, and many of the more personal marginal notes. The contents of this treebank were officially included in UD version 2.12 at its release on May 15th 2023, and since then a further gloss has been added. The treebank is available under a `CC BY-NC-SA 4.0` license. Dependency parsing of the remaining glosses is ongoing, however, these can still be found in the development branch, fully POS-tagged and annotated with morphological features.

4.2 The Diplomatic Würzburg Glosses Treebank (DipWBG)

The experiment described in Doyle et al. (2019) utilised 41 glosses, specifically chosen from among the Würzburg corpus for their lexical features. To these was added another gloss (Wb. 19d29), and the resulting 42 glosses were set aside as gold standard to be used in experiments described in Doyle (2024). This required that they be tokenised, POS-tagged, and annotated with morphological features in accordance with UD guidelines (Zeman, 2016). This work was carried out manually using the `CoNLL-U` file format, and the gold standard test-set was first uploaded to GitHub in mid-2020[4].

In February 2023 the contents of this gold standard test-set were uploaded to the development branch of the *Diplomatic Würzburg Glosses Treebank* (Doyle, 2023b). Between February and May of that year dependency parsing annotation was manually added to a selection of the 42 glosses. By the time of the data-freeze at the beginning of May 2023, 34 of the 42 glosses had been fully annotated. These 34 glosses, along with the content of the the St. Gall glosses treebank, (Doyle, 2023a) marked the first inclusion of Old Irish in an official UD release as of version 2.12 on the 15th of May 2023. The contents of this treebank are available under a `CC BY-SA 4.0` license. The remaining eight glosses are still available in the development branch, and are intended to be included in a future release.

5 Lemmata and Lexicography

The `CoNLL-U` format used by UD treebanks requires that a lemma be provided for each token. As a historical language, the task of identifying

[4] https://github.com/AdeDoyle/Wb_POS-testfiles/blob/master/sga_wbgold-ud-test2.conllu

a lemma for a given word is deceptively difficult. On the one hand, spelling variation is common in manuscript sources, and so the form of a given word which might stand as a headword in a dictionary could be attested spelled several different ways. On the other hand, not all such forms are attested, and of those which are many are only attested in manuscripts dated much later than the Old Irish period. As such, it is not always clear what spelling should be used for a given lemma, and without a spelling standard for Old Irish the choice of one spelling over another is ultimately arbitrary.

A primary focus during the development of the resources discussed above was to ensure that there would be consistency of headwords across both treebanks and the *Würzburg Irish Glosses* website. This required a significant amount of manual annotation. An attempt was made to ensure that each lemma used, if not an attested form itself, was at least theoretically possible. Moreover, an effort was made to ensure no two distinct lexemes had both the same POS and the same spelling for their lemma. It should, therefore, be possible to distinguish between homonymous lemmata by looking at their POS-tags. It might have been preferable to use unique numerical IDs, particularly as these could be linked to the unique identifiers used to distinguish discrete entries on eDIL, however, numerical specifiers are not permitted in lemmata for UD treebanks, "because they are not part of the canonical surface form" (Zeman, 2016).

6 Future Work

Tokenisation and annotation of content on the *Würzburg Irish Glosses* website is currently ongoing, and in future it is expected that the site's functionality will be expanded, for example, by including or linking to HD images of the manuscript. It is intended that the UD treebanks will be expanded in the future also. Finally, the use of unique lemmata across these resources lays the groundwork for the future development of an Old Irish wordnet.

7 Conclusion

This paper has presented three new lexically annotated text resources for Old Irish, the *Würzburg Irish Glosses* website, and two UD treebanks. These are the first discrete corpora of Old Irish to use the same tokenisation method, POS tag-set, and headword annotation, making them the first distinct Old Irish resources to be lexically compatible with each other. Because the tokenisation method used was designed to allow for separation of words in diplomatically edited text, these are also the first diplomatically edited corpora of Old Irish to be lexically annotated. The *Würzburg Irish Glosses* website was the first resource to make the digital text of the Würzburg glosses available, which is noteworthy as these glosses constitute the earliest large collection of writings in the Irish language. It is expected that these resources will facilitate the application of NLP techniques to Old Irish which were not possible before, as well as the creation of further lexical resources like wordnets. It is expected that the use of a widely utilised framework, like that of UD, and hosting of website code on GitHub will assuage concerns about the accessibility of this data into the future.

Limitations

Tokenisation, headword annotation and POS-tagging are still ongoing for the *Würzburg Irish Glosses* website. While the entirety of the St. Gall glosses have been automatically tokenised, POS-tagged and annotated with headwords and morphological information, and all of this can be found in the development branch for that treebank (Doyle, 2023a), only a portion of this has been manually proofed, and errors may still exist. The size of the published UD treebanks remains quite small, and this has prevented them from being used in some data-intensive NLP tasks (Dereza et al., 2024). The use of ISO 639 codes by UD has implications for what can be said to constitute Old Irish (see discussion in section 4), and the definition of a word used by UD does not account for some features of Old Irish orthography, like the separation of nasals from the beginning of a word (see discussion in Doyle and McCrae, 2025, 6-7, and UD issue 927[5]).

[5] https://github.com/UniversalDependencies/docs/issues/927

Acknowledgements

I would like to express my immense gratitude to Bauer et al. (2023) for providing their data for use in this project, and for consenting to it being adapted in the manner described here. I would like also to express my sincere gratitude to my supervisor, Dr. Clodagh Downey, whose expertise has been invaluable during the course of this research. Any remaining errors and omissions are entirely my own.

This work has been possible thanks to the support of the Science Foundation Ireland (SFI) as part of Grant Number SFI/12/RC/2289_P2 Insight_2, Insight SFI Centre for Data Analytics. This work has also been funded by the University of Galway through the Digital Arts and Humanities Programme, and by the Irish Research Council through the Government of Ireland Postgraduate Scholarship Programme.

References

Bernhard Bauer, Rijcklof Hofman, and Pádraic Moran. 2023. St Gall Priscian Glosses, version 2.1. Accessed: February 3, 2025.

Oksana Dereza, Adrian Doyle, Priya Rani, Atul Kr. Ojha, Pádraic Moran, and John McCrae. 2024. Findings of the SIGTYP 2024 shared task on word embedding evaluation for ancient and historical languages. In *Proceedings of the 6th Workshop on Research in Computational Linguistic Typology and Multilingual NLP*, pages 160–172, St. Julian's, Malta. Association for Computational Linguistics.

Oksana Dereza, Theodorus Fransen, and John P. Mccrae. 2023a. Do Not Trust the Experts: How the Lack of Standard Complicates NLP for Historical Irish. In *The Fourth Workshop on Insights from Negative Results in NLP*, pages 82–87, Dubrovnik, Croatia. Association for Computational Linguistics.

Oksana Dereza, Theodorus Fransen, and John P. Mccrae. 2023b. Temporal Domain Adaptation for Historical Irish. In *Tenth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2023)*, Dubrovnik, Croatia. Association for Computational Linguistics.

Adrian Doyle. 2018. Würzburg Irish Glosses. Accessed: February 3, 2025.

Adrian Doyle. 2023a. Diplomatic St. Gall Glosses Treebank. Accessed: February 3, 2025.

Adrian Doyle. 2023b. Diplomatic Würzburg Glosses Treebank. Accessed: February 3, 2025.

Adrian Doyle. 2024. *Development of Natural Language Processing Techniques and Resources for Old Irish; with an Application for the Detection of Authors in the Würzburg Glosses*. University of Galway, Galway. PhD Thesis.

Adrian Doyle and John P. McCrae. 2025. An Assessment of Word Separation Practices in Old Irish Text Resources and a Universal Method for Tokenising Old Irish Text. In *Proceedings of the 5th Celtic Language Technology Workshop*, pages 1–11, Abu Dhabi [Virtual Workshop]. International Committee on Computational Linguistics.

Adrian Doyle, John P. McCrae, and Clodagh Downey. 2018. Preservation of Original Orthography in the Construction of an Old Irish Corpus. In *Proceedings of the LREC 2018 Workshop: "CCURL2018 – Sustaining Knowledge Diversity in the Digital Age"*, pages 67–70, Miyazaki, Japan.

Adrian Doyle, John P. McCrae, and Clodagh Downey. 2019. A Character-Level LSTM Network Model for Tokenizing the Old Irish text of the Würzburg Glosses on the Pauline Epistles. In *Proceedings of the Celtic Language Technology Workshop*, pages 70–79, Dublin, Ireland. European Association for Machine Translation.

Aaron Griffith. 2013. A Dictionary of the Old-Irish Glosses. Accessed: February 3, 2025.

Séamus Kavanagh and Dagmar S. Wodtko. 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.

Elliott Lash. 2014a. POMIC Annotation Manual. Manual, The Dublin Institute for Advanced Studies. Accessed: February 15, 2024.

Elliott Lash. 2014b. The Parsed Old and Middle Irish Corpus (POMIC). Version 0.1. Accessed: February 12, 2024.

Kim McCone. 1997. *The Early Irish Verb*, 2 edition. An Sagart, Maynooth.

Beatrice Santorini. 1990. Part-of-Speech Tagging Guidelines for the Penn Treebank Project (3rd Revision). Standard, Department of Computer and Information Science, University of Pennsylvania.

Beatrice Santorini. 2016. Annotation manual for the Penn Historical Corpora and the York-Helsinki Corpus of Early English Correspondence. Accessed: February 3, 2025.

Ludwig Christian Stern. 1910. Epistolae Beati Pauli Glosatae Glosa Interlineali: Irisch-lateinischer Codex der Würzburger Universitätsbibliothek in Lichtdruck herausgegeben und mit Einleitung und Inhaltsübersicht versehen von Ludw[ig] Chr.

Stern Halle 1910. online at TITUS: Thesaurus Indogermanischer Text- und Sprachmaterialien, Johann Wolfgang Goethe-Universität Frankfurt am Main, 2002; Accessed: February 3, 2025.

David Stifter, Bernhard Bauer, Elliott Lash, Fangzhe Qiu, Nora White, Siobhán Barrett, Aaron Griffith, Romanas Bulatovas, Francesco Felici, Ellen Ganly, Truc Ha Nguyen, and Lars Nooij. 2021a. Corpus PalaeoHibernicum (CorPH) v1.0. Accessed: February 3, 2025.

David Stifter, Nina Cnockaert-Guillou, Beatrix Färber, Deborah Hayden, Máire Ní Mhaonaigh, Joanna Tucker, and Christopher Guy Yocum. 2021b. Developing a Digital Framework for the Medieval Gaelic World; Project Report. Technical report, Developing a Digital Framework for the Medieval Gaelic World.

Whitley Stokes and John Strachan, editors. 1901. *Thesaurus Palaeohibernicus*, volume 1. The Dublin Institute for Advanced Studies, Dublin.

Whitley Stokes and John Strachan, editors. 1903. *Thesaurus Palaeohibernicus*, 2 edition, volume 2. The Dublin Institute for Advanced Studies, Dublin.

Gregory Toner, Sharon Arbuthnot, Máire Ní Mhaonaigh, Marie-Luise Theuerkauf, Dagmar Wodtko, Grigory Bondarenko, Maxim Fomin, Thomas Torma, Giuseppina Siriu, Caoimhín Ó Dónaill, and Hilary Lavelle. 2019. eDIL 2019: An Electronic Dictionary of the Irish Language, based on the Contributions to a Dictionary of the Irish Language (Dublin: Royal Irish Academy, 1913-1976). Accessed: February 3, 2025.

Dan Zeman. 2016. UD Guidelines V2. Accessed: February 3, 2025.