An Assessment of Word Separation Practices in Old Irish Text Resources and a Universal Method for Tokenising Old Irish Text

Adrian Doyle and John P. McCrae
Insight SFI Centre for Data Analytics
Data Science Institute
University of Galway
adrian.odubhghaill@universityofgalway.ie and john@mccr.ae

Abstract

The quantity of Old Irish text which survives in contemporary manuscripts is relatively small by comparison to what is available for well-resourced modern languages. Moreover, as it is a historical language, no more text will ever be generated by native speakers of Old Irish. This makes the text which has survived particularly valuable, and ideally, all of it would be annotated using a single, common annotation standard, thereby ensuring compatibility between text resources. At present, Old Irish text repositories separate words or sub-word morphemes in accordance with different methodologies, and each uses a different style of lexical annotation. This makes it difficult to utilise content from more than any one repository in NLP applications. This paper provides an assessment of distinctions between existing annotated corpora, showing that the primary point of divergence is at the token level. For this reason, this paper also describes a new method for tokenising Old Irish text. This method can be applied even to diplomatic editions, and has already been utilised in various text resources.

1 Introduction

The majority of text which survives in contemporary Old Irish manuscripts has already been digitised and lexically annotated. This content is available online from various text repositories. Methods used for separating and annotating words and morphemes differ between repositories, however, with the result that data is incompatible between existing repositories. As interest in the application of various NLP techniques to historical Irish texts increases, several sources have reported that experiments were impacted by the lack of standardisation between text resources such as these (Doyle et al., 2019; Doyle and McCrae, 2024; Dereza et al., 2023a,b). Regarding digital resources for Gaelic languages, Stifter et al. found that "The most pressing issues include lack of standardisation and agreement of norms ... and inconsistency as far as tokenisation and use of unique identifiers across various Gaelic resources" (2021b, 8), which they suggest "can cause confusion and hinders linkage and interoperability." Moreover, Dereza et al. concluded that "the necessity of a text editing standard, especially for NLP applications, has not been properly debated and investigated by the historical Irish academic community" (2023a, 86).

This paper addresses the lack of standardisation among Old Irish text resources. It will demonstrate some of the main ways that text data and lexical annotations differ between existing resources in section 2, and will discuss some of the grammatical and orthographic reasons such distinctions exist. It will be shown that diplomatic editions, those in which editors attempt to faithfully reproduce text as it appeared in an original manuscript, can cause particular difficulty for Old Irish word separation. A novel method for tokenising diplomatically edited Old Irish text, which can prevent lexical variation between tokenised corpora, will be presented in section 3. It will be demonstrated that this method can also be applied to normalised, or otherwise altered text. Finally, section 4 will discuss how this tokenisation method has allowed for the consistent annotation of distinct Old Irish text resources, ensuring compatibility between them.

2 Currently Available Corpora

The historical stage of the Irish language as it was written between roughly the 7th and 9th centuries is termed Old Irish. Many texts which may be described linguistically as Old Irish can be found in manuscripts which date from later than the 9th century, having been copied from earlier sources. As Stokes and Strachan note, however, "Middle-Irish transcribers have often modernised or corrupted these ancient documents" (1901, xi). For this reason, the corpus of Old Irish text which survives in

Proceedings of the 5th Celtic Language Technology Workshop (CLTW 5), pages 1–11
Abu Dhabi, UAE (Online), January 20, 2025

Examples	Source	Text Ref.	Raw Text	Words
1a	**SGP**	Sg. 1b1	“.i. ci insamlar”	“ci”, “in”, “in·samlar”
1b	**CorPH**	Sg. 1b1	“.i. ci in·samlar”	“.i.”, “ci’”, “in·”, “in·samlar”
2a	**SGP**	Sg. 7b8	“do·furgabtais”	“do”, “fur”, “-”, “do·furgabtais”
2b	**CorPH**	Sg. 7b8	“do·furgabtais”	“do·”, “·fur”, “∅”, “do·furgabtais”
3a	**MlDB**	Ml. 2b3	“.i. dintsruth”	“di”, “int”, “sruth”
3b	**CorPH**	Ml. 2b3	“.i. dintsruth”	“di”, “int”, “sruth”
4	**POMIC**	Arm. 64	–	“d-a-beir”, “side”, “0”

Table 1: Comparison of Old Irish raw text and word separation between various text repositories: **CorPH** (Stifter et al., 2021a), **MlDB** (Griffith, 2013), **POMIC** (Lash, 2014b), **SGP** (Bauer et al., 2023)

manuscripts dated to the Old Irish period itself is of particular value.

Compared to the total quantity of existing text which may be described as Old Irish, the contemporary Old Irish corpus is relatively small, and the types of texts which comprise it are more limited. A small amount of Old Irish prose and poetry survives in contemporary manuscripts, though the majority of the contemporary Old Irish corpus is comprised of glosses. These glosses can vary in length from a single word to several sentences, though the majority are quite short. Three large collections exist, the Würzburg (Wb.) glosses, the Milan (Ml.) glosses, and the St. Gall (Sg.) glosses. A significant amount of code-switching occurs between Old Irish and Latin in each of these collections, however, Ml. contains the largest quantity of Old Irish text with 8,443 glosses being collected for that corpus by Stifter et al. (2021a). Sg. has the least Irish content with 3,478 glosses according to e-codices (2005), meanwhile there are 3,501 Irish glosses in Wb. (Doyle, 2018).

Separate projects have been undertaken to digitise and annotate the three corpora of glosses (Griffith, 2013; Bauer, 2015; Bauer et al., 2023; Doyle, 2018). Two Universal Dependencies (UD) treebanks have since been created (Doyle, 2023a,b), each containing a small selection of these glosses. Otherwise, the *Parsed Old and Middle Irish Corpus* (POMIC; Lash, 2014b) contains some Old Irish prose text, and a variety of content has been collected and annotated in *Corpus PalaeoHibernicum* (CorPH; Stifter et al., 2021a). The resources discussed in section 4, which make use of the tokenisation method described here in section 3, use UD style part-of-speech (POS) tags (Zeman, 2016). Aside from these, though each of the remaining resources provide lexical annotation, only POMIC makes use of an established POS tag-set. According to Lash (2014a), POMIC uses a variety of Penn-style POS-tags (Santorini, 1990) which were originally adapted for use with Old English (Santorini, 2016). Each of the other resources utilise discrete lexical annotations.

The more noteworthy distinction between resources than lexical annotation, however, is that each separates words in accordance with different methods. Separating words is a deceptively difficult task for Old Irish (Doyle et al., 2019). While the orthographies of many modern European languages require spacing to occur between most words, for Old Irish “... words which are grouped round a single chief stress and have a close syntactic connexion with each other are written as one in the manuscripts” (Thurneysen, 1946, 24). Often this can result in word clusters which are difficult to separate. For example, where the words “*is*” and “*samlid*” come together, sometimes a letter is elided, forming a compound which is difficult to separate, “*isamlid*”, ‘it is thus’, (as in Wb. 4a4 and 5b36). Occurrences of such clusters can result in different words being separated and annotated in different ways by different resources, even where they represent the same manuscript text.

A handful of examples of Old Irish text from various repositories can be found in Table 1. While an exhaustive list of distinctions between existing text repositories is not possible, these examples are sufficient to demonstrate some of the major differences between editorial standards and word separation methods used by each repository. The “Raw Text” column displays how each repository represents the text of the manuscript before applying word separation. POMIC (Lash, 2014b) is an exception as it does not contain pre-separation text data. The “Words” column displays the words iden-

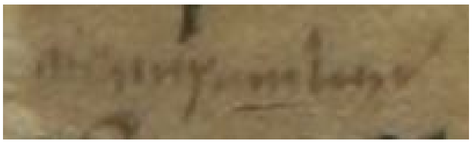

Figure 1: *.i.ciinsamlar* (1b1) from St. Gallen, Stiftsbibliothek, Cod. Sang. 904 (www.e-codices.ch).

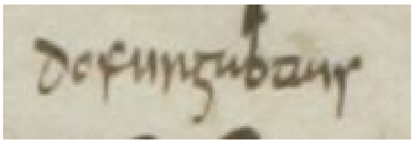

Figure 2: *dofurgabtais* (7b8) from St. Gallen, Stiftsbibliothek, Cod. Sang. 904 (www.e-codices.ch).

tified by each repository after separation.

Examples 1a and 1b demonstrate that the raw text can differ between repositories based on editorial decisions. In the case of 1b the editors have supplied punctuation in "*ci in·samlar*" ('if I should imitate') which was not supplied by the editors of the exact same text, "*ci insamlar*", in example 1a. Though faded, it can just about be seen in Figure 1 that no punctuation occurs in the original manuscript either. Similarly, while the editors of both 2a and 2b supply punctuation in the raw text, "*do·furgabtais*" ('they should enunciate'), it can be seen in Figure 2 that no such punctuation appears in the manuscript. Because of editorial distinctions such as this, a tokenisation method for Old Irish will need to be capable of handling text both with and without this manner of punctuation. For this same reason it is currently a requirement that Old and Middle Irish treebanks added to UD must be identified as either "diplomatic" or "critical", where "diplomatic" treebanks cannot include punctuation, capitalisation or other text characters inserted by editors (with the exception of expanded abbreviations), unless they appear in the manuscript[1].

Further distinctions between resources become apparent when examining how words are separated. Even where text has been drawn from a single source, and the raw text is identical, different repositories will often separate different words. For example, 2a has "*do*", "*fur*" and "-" equating to "*do·*", "*·fur*" and "∅" in 2b. There is also a tendency among resources for separated words not to reflect the raw text character-for-character, making it impossible to reproduce the raw text by simply concatenating the separated words. In 3a and 3b, for example, only a single *i* occurs in the raw text, "*d**i̲**ntsruth*" ('from the torrent'), however, concatenating the words identified by each resource, "*di*" ('from') and "*int*" ('the'), would result in "*d**i̲i̲**ntsruth*" with two *i*s. More egregiously, in 2b where the raw text reads "*do·furgabtais*", concatenating the words identified by the resource would result in the gibberish string "*do··fur∅do·furgabtais*". In three examples, 2a, 2b, and 4, an "empty" word is supplied to represent a semantic element which is understood to occur in that position, but not represented in the raw text. This duplication and addition of characters is not typical of word-level tokenisation but is common in Old Irish resources, particularly where an attempt is made separate the verbal complex into its various components, while also portraying it as a single word. In stark contrast, example 4 presents the entire verbal complex, "*d-a-beir*" ('he gives it'), as a single word only. While this is more representative of typical tokenisation practice, hyphenation which would not have occurred in the manuscript was introduced to identify the infixed pronoun, "*-a-*" ('it') from the rest of the verb. As such, this word separation method necessitates altering the original text for clarity.

As a comparison is being drawn here between the separation of words in various Old Irish text repositories and what might be typically expected of tokenisation, it must be noted that only Lash (2014a) actually uses the term "tokens" in the annotation manual for *POMIC*, and only once. Otherwise, he generally refers to "words" and "word-division", while other resources use the terms "phonolog[ical] word" (Griffith, 2013), "word form" (Bauer et al., 2023), and "morph" (Stifter et al., 2021a). This reflects the fact that these resources were not necessarily developed to be used in NLP applications, but as aides to linguistic research. Griffith (2013), for example, describes the Ml. database as a "dictionary" and a "lexicon" rather than as an annotated digital text. It would therefore be unreasonable to expect word division in these repositories to reflect tokenisation in a traditional sense. Indeed, the methods used by each

[1] Conversely, any treebank containing editorial alterations to the text such as these must be identified as "critical", though this definition does not align perfectly with the common use of the term "critical edition". For more information see discussion of Treebank Classification at https://universaldependencies.org/sga/index.html.

resource for separating words, and sometimes also smaller morphemes, are perfectly valid from a linguistic perspective, even though resources may differ from one another. If facilitating downstream NLP applications is to be treated as a realistic objective in the future development of Old Irish text resources, however, compatibility between these resources at the word level must be afforded more consideration than it has been to date. Identifying a single, universally applicable method for tokenising Old Irish text is clearly the first step which must be taken in this direction, as tokenisation necessarily impacts following steps like POS-tagging and dependency parsing. Such a tokenisation method will need to satisfy the requirements of both diplomatically edited manuscript text, and text which has been normalised or otherwise altered.

3 Tokenisation Method

The purpose of this section is to present a new tokenisation method which can be universally applied to all Old Irish text, be it diplomatically edited or altered by modern editors in any of a variety of ways (including silent word separation, expanding manuscript contractions and abbreviations, supplying capitalisation or punctuation, etc.). The main principles of this tokenisation method are as follows:

1. The character content of the raw text should not be altered by the tokenisation process, other than by the removal of whitespace characters between words.
2. Tokens (other than punctuation and symbols) resulting from the process should represent lexical words, not orthographic combinations made up of multiple parts-of-speech.
3. Tokens should represent synchronically Old Irish words, regardless of how such words may have developed diachronically.
4. No “empty/zero” characters should be introduced to represent lexemes which are not already represented in the raw text.
5. Resulting tokens should conform to the expectations of widely used text-data frameworks and POS-tagging schemes, such as UD.

For reasons of space, it would be impossible to provide a comprehensive discussion of every type of word here, however, detailed examples of the suggested tokenisation of various parts-of-speech can be found in Tables 2, 3, 4 and 5[2]. These can be found in Appendices A, B, C and D respectively.

3.1 Unproblematic Parts-of-speech

Many parts-of-speech are relatively unproblematic insofar as tokenisation is concerned, and can be separated relatively intuitively. Nouns like “*fer*” (‘man’), “*ben*” (‘woman’) or “*guide*” (‘prayer’), adjectives like “*becc*” (‘small’), “*már*” (‘large’) or “*maith*” (‘good’), and numerals like “*óen*” (‘one’), “*cethir*” (‘four’) or “*secht*” (‘seven’), are generally separated from surrounding words in modern editions and learning material using spacing, and this can be applied consistently with no further alteration typically occurring in the text as a result. Such parts-of-speech will always form discrete tokens of their own. A more complete list of parts-of-speech which can be separated into discrete tokens with relative ease can be found in Table 2.

While the parts-of-speech represented in Table 2 can be tokenised in a manner similar to most other languages, without any substantial linguistic disagreement, a few points should be noted about particular examples. Firstly, *olchena*, though it has a discrete entry in the *Electronic Dictionary of the Irish Language* (eDIL; Toner et al., 2019), is not considered an adverb in its own right, but a combination of *ol* and *cene*. This is necessary as the form occasionally occurs with spacing between these components in manuscript sources. In all other cases, adverbs form discrete tokens. Secondly, conjugating prepositions are treated as individual tokens in Old Irish treebanks. This is in line with Stifter’s claims that these constitute “a single entity” (2006, 87) and that “It is not possible to separate one element from the other”. It is also in line with the example of UD treebanks for Modern Irish, however, it should be noted that Scottish Gaelic and Manx Gaelic treebanks currently treat these as compounds of prepositions with pronouns.

3.2 Problematic Parts-of-speech

Consistent separation of words other than those in Table 2 can pose more difficulty, particularly where phenomena like syncope and apocope affect

[2] Discrete examples are separated by commas in these tables. Where a single example includes more than one token, the relevant token appears in bold and underlined. For example, where “***a** sind*” and “***do** nd*” appear as examples in Table 3 in the “Prepositions” row, the prepositions in these examples are “***a***” and “***do***” respectively. In such examples, spacing is used to separate all tokens, even where spacing may not have occurred in the raw text.

compounds of multiple words, but also in many other cases where shifting stress patterns affect the orthographic representation of clitics. Thurneysen claims "The absence of stress is most complete in (1) the article or a possessive pronoun standing between a preposition and the word it governs, (2) infixed pronouns and (sometimes) **ro** between preverbs and verbs, and (3) the copula between conjunctions and the predicate" (1946, 31). Indeed, the verbal complex, the article, the copula, and other words with which they can combine, are responsible for most of the difficulty in tokenising Old Irish. Table 3 demonstrates the suggested tokenisation for some of the more problematic parts-of-speech in Old Irish, other than those directly related to the verbal complex. Copula and Verb tokens, being some of the most problematic, are presented in Table 4, while other parts-of-speech which make up the verbal complex can be found in Table 5. Each of these tables demonstrate how tokens should be separated when they occur in compounds.

For many word-types represented in Table 3, separation is only problematic where they combine with other words. Independent personal pronouns like "*mé̄*" ('me'), and possessive pronouns like "*mo*" ('my'), for example, are not problematic to tokenise. Where they are compounded, however, producing forms like the "*mei-*" of "*meisse*" ('me!'), or the "*m-*" of "*móinur*" ('I alone'), knowing whether these should be separated can be less intuitive. Nevertheless, to enable the production of text resources in widely adopted formats, such as UD treebanks, a single, consistent tokenisation method must be applied in cases like these. It is the suggestion of this paper that all of the word types identified in Table 3 should be separated such that they form discrete tokens.

Certain conjunctions can be particularly problematic, especially in cases where what might be considered individual conjunctions can be found with spacing between their component morphemes in both manuscripts and learning material. Stifter, for example, lists "*in tain*" ('when'), "*íarsindí*" ('after'), "*fo bíth*" ('because'), "*in chruth*" ('so/as') and "*is cumme*" ('it is the same as if') as conjunctions (2006, 248–249), though it is suggested here that they be interpreted instead as multi-word expressions, and tokenised accordingly. To these, Stifter adds discrete negative forms of conjunctions like "*an(n)a*" ('while not'), "*arná*" ('so that not'), and the space-separated "*ol ní*" ('because not'). In accordance with this tokenisation method, these too should be separated to form discrete tokens. Conversely, certain items which should probably be considered discrete lexical conjunctions by the Old Irish period, like "*cenmitha*" ('aside from/in addition to'), can nevertheless be found written graphically as two words in manuscripts, "*cen mitha*" (see Sg. 150b3). Such cases present some difficulty for tokenisation as they require either that a lexical word be separated into subword morphemes, or that a space character can occur within a token, which is exceptional in UD treebanks. Nevertheless, the suggestion here is that conjunctions like "*cen mitha*" should be represented by a single token, even if that token contains a space character.

3.3 The Copula

The Copula deserves particular attention. The basic, non-combining forms (*"am", "at", "is",* etc.) can be tokenised relatively easily. It becomes difficult, however, to systematically separate copula forms from certain other morphs which may be considered parts-of-speech in their own right. It is tempting, for example, to separate negative particles from what may be seen as copula endings ("*níta*" = "*ní*" + "*ta*"). As the third singular negative form "*ní*" contains no such ending in the orthography, however, no distinct copula token could result. For this reason, discrete negative copula forms should be retained as tokens for all persons and numbers (see Table 4).

3.4 The Verbal Complex

As the size of Table 5 might indicate, the verbal complex is the single feature in Old Irish orthography which creates the most difficulty for tokenisation. This can be ascribed to the sheer number of distinct types of words which can be compounded within it, as well as to the effects of syncope and apocope on the resulting compounds. It is not possible in this paper to discuss the various elements which make up the verbal complex in detail, however, it is necessary to note the following qualities. Firstly, verbs have dependent and independent forms, with dependent forms being used following conjunct particles, including the negative, interrogative and relative particles, the semantically empty verbal particle, "*no*", as well as certain conjunctions. Secondly, Old Irish has compound verbs, comprised of one or more "preverbs" followed by a verbal root. McCone maps how up to five preverbs can precede a verbal root (1997, 90). Thirdly, where the object of the verb is expressed by a pronoun, this pronoun is generally “infixed” between either the initial preverb, or a conjunct particle, and the remainder of the verb, though in certain situations suffixed pronouns are used instead.

The dependent (or “prototonic”) forms of compound verbs often look quite different from the independent (or “deuterotonic”) forms, as the use of a conjunct particle shifts stress from the second element in the compound to the initial preverb. Hence, negating the compound verb “*dobeir*” (‘he gives’), which contains the initial preverb “*do*”, results in the prototonic form “*nítabair*” (‘he does not give’), where the preverb has become “*ta*”. Where a pronoun is infixed into the deuterotonic form it follows the initial preverb, “*do**m**beir*” (‘he gives me’), but where it is infixed into the prototonic form it precedes it, “*ní**m**tabair*” (‘he does not give me’). This creates a systematic difficulty for tokenisation. If it is desirable to separate the pronoun from the remainder of the verb during tokenisation, this can be achieved in prototonic verb forms without affecting the initial preverb, (“*ní*” + “*m*” + “*tabair*”), but in deuterotonic forms would necessitate separating the initial preverb also, (“*do*” + “*m*” + “*beir*”). The alternative would be to retain “*dombeir*” in its entirety as a single token, and treat the pronoun as if it were verbal morphology. This is the approach taken by POMIC, (see example 4 in Table 1), though hyphenation is used to identify the pronoun. In a more diplomatic edition it would be much more difficult to identify which part of the verb constituted inflection for the verbal object[3].

As can be seen in Table 5, this tokenisation method requires that initial preverbs be separated from the remainders of compound verbs in deuterotonic form, but not in prototonic form. Initial preverbs, therefore, will stand as discrete tokens where verbs occur in deuterotonic form, but will form the stressed anlaut of the verb token itself in prototonic form. Infixed pronouns will always form standalone tokens, as will suffixed pronouns, and all conjunct particles.

The augment, “*ro*”/“*ru*”, creates further difficulty. In most cases, it will act as a non-initial preverb, either standing in stressed position, as in “*as**ru**bart*” (‘he has said’), or later within the compound, as in “*nito**r**gaítha*” (‘he should not defraud him’). In these situations it should be treated as part of the verb token. In rare situations, however, it stands in pretonic position, sometimes even standing between an initial preverb and infixed pronoun, as in “*for**ru**mchennadsa*” (‘I have been destroyed’, see Thurneysen, 1946, 256). In such cases, it should form its own separate token in the same manner as initial preverbs in deuterotonic forms of verbs (“*for*” + “*ru*” + “*m*” + “*chennad*” + “*sa*”).

This tokenisation method is also capable of handling instances of tmesis, where any POS other than an infixed pronoun separates an initial preverb or conjunct particle from the remainder of the verb. A good example of this is “*ad cruth cáin cichither*” (‘a beautiful form will be seen’), where both “*cruth*” (‘form’) and “*cáin*” (‘fair/beautiful’) are infixed between the preverb, “*ad*”, and remainder of the verb, “*cichither*” (‘will be seen’). As is demonstrated in Table 4, where tmesis occurs, the initial preverb or conjunct particle, any infixed pronouns, other parts-of-speech preceding the remainder of the verb (such as adjectives and nouns), and the remainder of the verb itself, each form separate tokens from one another.

3.5 Miscellaneous Tokens

Moving away from the verbal complex, a few further tokenisation issues remain. The first regards nasalisation markers (“*m*”/“*ṁ*” and “*n*”/“*ṅ*”), which indicate a phonetic change to the anlaut of a following word. They are generally written as a part of that following word, as in “*is inse **ṅ**duit*” (‘it is impossible for you’, Wb. 5b28), or “*isdered **ṁ**betho*” (‘it is the end of the world’, Wb. 10b3), but are also frequently separated from it by spacing, and even enclosed by punctuation (see Bronner, 2016), as in “*a**ṅ** grammatice*” (‘the *grammatice*’, Sg. 204a8), “*laa **ṁ** brátha*” (‘doomsday’, Wb. 26a1), and “*lae **.m.** brátho*” (Thurneysen, 1946, 147). In these situations, tokens with internal space characters are permissible, and indeed required by UD treebanks[4]. Therefore, forms like “*ṅ grammatice*”, “*ṁ brátha*”, and “*.m. brátho*” should be treated as single tokens which contain a space.

Ambiguity may still arise regarding word boundaries where letters have been elided in combinations between clitics and stressed words such as “*isamlid*” (for “*is*” + “*samlid*”, ‘it is thus’),

[3] Fransen (2020) has demonstrated it may be possible to parse this kind of complex Old Irish verbal morphology using finite state technology, however, no such morphological analyser has yet been made available for general use.

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

“*hit̲uilsiu*” (for “*hit*” + “*tuilsiu*”, ‘in your will’), “*oc̲umtuch*” (for “*oc*” + “*cumtuch*”, ‘building’), etc. (see Thurneysen, 1946, 91). The rule of thumb adhered to here is that extra letters, which did not occur in the original orthography, should never be supplied during tokenisation. Instead, in accordance with this tokenisation method, the clitic should always lose the letter when separating words, hence, “*isamlid*” = “*i*” + “*samlid*”, “*hituilsiu*” = “*hi*” + “*tuilsiu*”, and “*ocumtuch*” = “*o*” + “*cumtuch*”.

3.6 Abbreviations, Contractions, Symbols and Punctuation

The tokenisation of abbreviations and contractions (where these are not expanded by editors) remains an issue. UD guidelines (Zeman, 2016) suggest that “abbreviations for single words ... are assigned the part of speech of the full form”. This is possible for abbreviations like the Tironian *et*, “⁊”, which can be simply annotated as a conjunction, as would the full form, “*ocus*” (‘and’). It is not possible, however, for abbreviations like “*.i.*” which represent multiple words in Irish, “*ed ón*” (‘*id est*’). Instead, such abbreviations should be maintained as discrete tokens, inclusive of any punctuation characters they may have. These can then be POS-tagged as appropriate, for example, “*.i.*” is POS-tagged `ADV` in Old Irish UD treebanks, which matches its treatment in Modern Irish treebanks.

Where a marking or grapheme is used to abbreviate a specific character sequence (such as where “ↄ” stands for “*con*”), these should be treated as if they were letter characters. Where the abbreviated sequence constitutes only a portion of an abbreviated word, the grapheme or marking should form a part of the whole word token. A diplomatic edition may retain the abbreviated token, “*ↄall*”, for example, which is equivalent to the normalised form “*Conall*”. Similarly, where markings with no set phonetic value, such as suspension strokes, are used to abbreviate some portion of a word, these should form part of the same token as the rest of the word they abbreviate. Again, for example, an abbreviated token like “*ↄchob̄*”, with a suspension stroke above the final letter, *b*, might occur in a diplomatic edition representing the normalised form “*Conchobar*”.

The rules outlined in the preceding two paragraphs hold for markings intended to denote abbreviations, even where they include non-letter characters. If, however, a sequence of one or more non-letter characters (such as ∴ or .,.,.,) is used in an edition to approximate a manuscript marking which does not denote either an abbreviation, or one or more words (see Groenewegen, 2011), this entire sequence should form a single, discrete token. This token may then be POS-tagged as appropriate. Depending on how it is used, it may be a form of punctuation, or it may be treated as a symbol as in the case of a *signe de renvoi*.

3.7 Applicability to Different Types of Text

While this tokenisation method was designed to be utilised for diplomatically edited Old Irish text it is easily adaptable to texts which have been normalised or otherwise altered by modern editors. For example, in a diplomatic edition “*dombeir*” should be split into three tokens (“*do*” + “*m*” + “*beir*”), however, in another edition an editor may mark the stressed part of the verb using punctuation (hyphenation or a mid-height dot). This should then form its own token and be POS-tagged as punctuation. Hence “*dom·beir*” would be tokenised “*do*” + “*m*” + “·” + “*beir*”. As such, this tokenisation method can be applied to any Old or Middle Irish corpus, whether or not it is edited diplomatically. It therefore has the potential to ensure syntactic compatibility between Early Irish text resources in a manner which has not been possible to date.

4 Applications to Old Irish Text

To date the tokenisation method described in this paper has been employed by the online text repository of the Würzburg glosses (Doyle, 2018), as well as by two UD treebanks (Doyle, 2023a,b). In fact, the tokenisation method was developed in tandem with the *Diplomatic St. Gall Glosses* treebank (Doyle, 2023a) and with the *Würzburg Irish Glosses* website (Doyle, 2018) to ensure that it could fulfil the various tokenisation requirements of each corpus. As annotation of these corpora progressed, the tokenisation method was periodically reevaluated and updated as necessary to account for the wide variety of lexical features which occur in these texts.

As the present focus is on tokenising Old Irish text, any more comprehensive discussion of these text resources falls outside the scope of this paper. It is notable, however, that at the time of this writing the entirety of the St. Gall glosses have already been tokenised using the method set out here, including those glosses written in the Ogham script. Therefore, the tokenisation method described in this paper has already been proven to successfully support the consistent separation of word-level tokens throughout a relatively large portion of the surviving body of Old Irish text, and across two writing systems.

5 Future Work

A significant obstacle to the production of large amounts of annotated Old Irish text remains the lack of an automatic tokeniser for the language. The earliest investigation into the viability of such a resource not only demonstrated the considerable difficulty involved in tokenising Old Irish, but also noted that the lack of standardisation between Early Irish text repositories in terms of word separation led to a lack of consistent data with which to train such a model (Doyle et al., 2019). The tokenisation method presented above aims to address this data sparsity by providing a blueprint which could potentially be used to bring discrete text repositories into alignment regarding word boundaries, without needing to alter their raw text content in any way. It is hoped that as more Old Irish text becomes available, which has been tokenised in accordance with the method describe here, it will be possible to train an automatic tokeniser model, and thereby further increase the speed with which Old Irish text can be tokenised and annotated.

6 Conclusion

This paper has demonstrated that the methods by which words are separated in various Old Irish text repositories are inconsistent, making their lexical contents incompatible with one another for the purpose of downstream NLP applications. To address this, a novel tokenisation method has been presented here which can be applied even to diplomatically edited Old Irish text. This removes the impetus to alter the character content of tokens when separating words, a practice which is common in Old Irish text resources.

Before a suitable tokenisation method had been identified for Old Irish, it had not been practicable to standardise the separation of words between Old Irish text resources. The tokenisation method described in this paper has allowed lexical uniformity to exist between resources for the first time. The corpora which have already made use of this tokenisation method are not only the first diplomatically edited Old Irish corpora to have been tokenised, but also the first discrete corpora of Old Irish to share a common word separation method. That it has already been successfully applied to text in three Old Irish resources, including the entirety of the relatively large St. Gall collection of glosses, demonstrates that this new tokenisation method enables consistent tokenisation across a selection of the most challenging scenarios which can result from Old Irish grammar and manuscript orthography.

The importance of word-level compatibility between annotated text resources cannot be understated, though it may be taken for granted in the case of many European languages with more settled spelling and word separation. Particularly where word-level tokens play a role in the application of downstream NLP tasks, any variability between corpora regarding what constitutes a word could potentially skew results. As such, it is envisioned that the tokenisation method presented here will allow for a wider variety of NLP techniques to be applied across the Old Irish texts which already utilise it than would have been possible before. The intention for this paper is that it can act as a reference for those who may wish to tokenise corpora of Early Irish text in the future, and thereby contribute to the lexical standardisation of Early Irish text resources. Ultimately, if this or a comparable tokenisation standard were to become widely adopted by Old Irish text repositories, it is expected that this would not only bolster ongoing linguistic research, but that it could also support new areas of investigation which require more standardised datasets.

Acknowledgements

I would like to express my sincere gratitude to my supervisor, Dr. Clodagh Downey, whose expertise has been invaluable during the course of this research. Without her guidance and keen attention to detail the current work would not be what it is. 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 Postgradu-

ate Scholarship Programme.

References

Bernhard Bauer. 2015. A Dictionary of the Old Irish Priscian Glosses. Accessed: September 12, 2024.

Bernhard Bauer, Rijcklof Hofman, and Pádraic Moran. 2023. St Gall Priscian Glosses, version 2.1. Accessed: September 12, 2024.

Dagmar Bronner. 2016. Nasalierung im Buch von Armagh: Überlegungen zu altirischen Schreibkonventionen. *Zeitschrift für celtische Philologie*, 63(1):29–48.

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: September 12, 2024.

Adrian Doyle. 2023a. Diplomatic St. Gall Glosses Treebank. Accessed: September 19, 2024.

Adrian Doyle. 2023b. Diplomatic Würzburg Glosses Treebank. Accessed: September 19, 2024.

Adrian Doyle and John P. McCrae. 2024. Developing a Part-of-speech Tagger for Diplomatically Edited Old Irish Text. In *Proceedings of the Third Workshop on Language Technologies for Historical and Ancient Languages (LT4HALA) @ LREC-COLING-2024*, pages 11–21, Torino, Italia. ELRA and ICCL.

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.

e-codices. 2005. e-codices - Virtual Manuscript Library of Switzerland. Accessed: September 12, 2024.

Theodorus Fransen. 2020. Automatic Morphological Analysis and Interlinking of Historical Irish Cognate Verb Forms. In Elliott Lash, Fangzhe Qiu, and David Stifter, editors, *Morphosyntactic Variation in Medieval Celtic Languages. Corpus-Based Approaches*, pages 49–84. De Gruyter Mouton, Berlin.

Aaron Griffith. 2013. A Dictionary of the Old-Irish Glosses. Accessed: September 12, 2024.

Dennis Groenewegen. 2011. Tionscadal na Nod. Accessed: September 21, 2024.

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

Elliott Lash. 2014b. The Parsed Old and Middle Irish Corpus (POMIC). Version 0.1. Accessed: September 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. Accessed: September 19, 2024.

Beatrice Santorini. 2016. Annotation manual for the Penn Historical Corpora and the York-Helsinki Corpus of Early English Correspondence. Accessed: September 19, 2024.

David Stifter. 2006. *Sengoidelc*. Syracuse University Press, New York.

David Stifter, Bernhard Bauer, Elliott Lash, Fangzhe Qiu, Nora White, Siobhán Barrett, Aaron Griffith, Romanas Bulatovas, Ellen Felici, Francesco abd Ganly, Truc Ha Nguyen, and Lars Nooij. 2021a. Corpus PalaeoHibernicum (CorPH) v1.0. Accessed: September 12, 2024.

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. Accessed: September 23, 2024.

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

Rudolf Thurneysen. 1946. *A Grammar of Old Irish*, 2 edition. 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: September 19, 2024.

Dan Zeman. 2016. UD Guidelines V2. Accessed: September 19, 2024.

Appendix

A Unproblematic Parts-of-speech for Tokenisation of Old Irish

Word Type	Examples	UD POS
Adjectives	*becc, beccaib,* *lugu, lugimen,* *dían, dénithir, déniu,* ***sen*** *tintúd, is* ***siniu***	ADJ
Adverbs	*trá, nammá, íarum*	ADV
Anaphoric Pronouns	*do* ***ṡuidiu****, ol* ***suide****,* *amal* ***ṡodain****,* *as beir* ***side***	PRON
Conjugating Prepositions	*ass, dam, lemm, occaib*	ADP
Deictic Particle	*int* ***í****, forsna* ***hí****,* *inna* ***hí****, a* ***ní*** *siu*	PART
Demonstrative Particles	*so, sin*	PART
Nouns	*fer, fir, feraib*	NOUN
Numerals	*tri, téoraib*	NUM
Numeric Particle	***a*** *óen,* ***a*** *cethir,* ***a*** *secht*	PART
Vocative Particle	***á*** *ḟir,* ***á*** *chéiliu,* ***a*** *rómanu*	PART

Table 2

B Problematic Parts-of-speech for Tokenisation of Old Irish

Word Type	Examples	UD POS
The Article	*in, ind, inna, a,* *la* ***sin****, la* ***ssa****, co* ***ssind****,* *do* ***nd****, do* ***naib***	DET
Conjunctions	*ocus, acht, cía, má, ara,* *"cen mitha",* ***ar*** *ná,* ***a*** *nna,* ***ol*** *ni,* ***ma*** *nip,* ***ce*** *ni d ḟil,* ***dia*** *cairigther,* ***co*** *naccae,* ***co*** *ndom accae,* ***co*** *ndid tuctis,* ***ci*** *d,* ***ci*** *so,* ***ma*** *d,* ***ma*** *so*	CCONJ **OR** SCONJ
Emphatic Suffixes	*sa, siu, som,* *mei* ***sse****, a thu* ***su****, hé* ***som***	PRON
Independent Personal Pronouns	*mé, hé, ed, sní,* ***mei*** *sse, a* ***thu*** *su,* ***hé*** *som*	PRON
Interrogative Pronouns	*cía, cid, cesí,* ***ci*** *de,* ***c*** *indas on,* ***ci*** *pad,* ***cía*** *bed*	PRON
Possessive Pronouns	*mo, do, a,* ***m*** *óinur, i* ***mm*** *eícndarcus,* *i* ***t*** *chóimthecht*	PRON
Prepositions	*a, do, la, oc,* ***a*** *sind,* ***do*** *nd,* ***la*** *sin,* ***oc*** *ind*	ADP

Table 3

C Tokenisation of the Old Irish Verb and Copula

Word Type	Examples	UD POS
The Verb	*gaibid, biru, caraimm,* *at **tá**, fo **gaib**, as **biur**,* *ní **gaib**, ní **biur**, ní **caraimm**,* *ní **fil**, ní **fagaib**, ní **epur**,* *ní m **fil**, f a **ngaib**, a t **biur**,* *ní s **ngaib**, no b **caraimm**,* *in dam **biur**,* *ad cruth cáin **cichither**,* *no m choimmdiu **cóima***	VERB
The Copula	*am, at, is, ammi, adi, it,* *bid, as, ata,* *níta, ní, nítad,* *nacham, nách, nachib,* *ce **so**, cia **so**, ma **so**,* *ma **d**, ci **d**, co **ndid**,* *a **mtar**, cía **bed**, ci **pad**,* *rop, robbu,* *amal **nonda**, amal **nondad**,* *amal **nondan**, ce **notad***	AUX

Table 4

D Tokenisation of Elements of the Old Irish Verbal Complex

Word Type	Examples	UD POS
The Augment (*ro, ad, com*)	*for **ru** m chennad sa,* *amal **ro** n gab,* *rosechestar, rotoltanaigestar,* *as **rubart**, im **ruidbet**,* *do **rochuirsemmar**,* *ní **roimdibed**, ní **roscríbad**,* *ni **torgaítha**, in **ruchumsan**,* *fo da **rorcenn**,* *ni m **thorgaíth**,* *con **acab**, con **abbong*** *con **ascar**, fris **comorg**,* *do **comrig***	PART **OR** VERB
Conjunct Particles	***ní*** *léici,* ***ní*** *tuit,* ***ní*** *fúasna,* ***in*** *foircnea,* ***in*** *naccai,* ***in nád*** *fail,* ***ní*** *m léici,* ***ní*** *t accai,* ***ní*** *tuit,* ***ní*** *s fúasna,* ***ní*** *b ben,* ***nach*** *am dermainte,* *ar **nach** it rindarpither,* ***in*** *ndom léici,* ***in*** *ndot accai,* ***in*** *ndid tuit,* ***in*** *nda fúasna* ***in*** *ndob ben*	PART
Infixed Pronouns	*a **tom** chí, a **tot** beir,* *d **a** mbeir, fo **s** ngaib,* *a **t** chí, a **tonn** beir,* *do **b** beir, fo **s** ngaib,* *no **m** chara, no **t** ben,* *n **a** cúalae, ní **s** naccai,* *ní chara, ní **n** ben,* *nách **ib** cúalae, in **da** accai*	PRON
Initial Preverbs	**DEUTEROTONIC** ***ad*** *cí,* ***do*** *beir,* ***fo*** *gaib,* ***du*** *airṅgir,* ***a*** *tot chí,* ***d*** *a beir,* ***fo*** *m gaib,* ***do*** *b airṅgir,* **PROTOTONIC** *ní **accai**, ní **tabair**, ní **fagaib**,* *ní **tairngir**,* *ní t **accai**, ní **tabair**,* *ní m **fagaib**, ní b **tairngir***	PART **OR** VERB
Relative Particle	*ar **a**, di **a**, hu **a**, la **sa**, oc **a***	PART
Suffixed Pronouns	*beirth **i**, léicsi **us**, guidm **it***	PRON
Verbal Particle (*no/nu*)	***no*** *bed,* ***no*** *berinn,* ***no*** *léicthea,* ***no*** *marbthae,* ***no*** *m chara,* ***n*** *a cara,* ***no*** *b cara,* ***no*** *da deligedar,* ***no*** *nda failsigetar,* *ce **nu** d sluindi*	PART

Table 5