Challenges in the Detection of Dialect for Historical Languages; the Case of Old Irish Text Resources

Adrian Doyle
Maynooth University
Department of Early Irish
School of Celtic Studies
adrian.odubhghaill@mu.ie

Abstract

Old Irish presents particular challenges for the study of automatic dialect detection. It is generally accepted that little trace of dialect can be found in Old Irish writings. Extant Old Irish text resources introduce a considerable amount of extra variation, which could impact dialect identification applications. While some scholarship has suggested that certain features may be indicative of dialect, such hypotheses are difficult to substantiate where authorship is anonymous, or where the text itself is not associated with a particular geographical region. This paper describes the application of stylometric dialect detection techniques to Old Irish texts, and discusses the features which emerge from this process as potential markers of dialect. The aim is not necessarily to identify Old Irish dialectal features with certainty, but rather to investigate the impact that Old Irish text resources could have on such applications. This paper does, nevertheless, add to the extant body of research by highlighting features which have been identified as stylistically distinct by stylometric techniques intended for dialect detection.

Keywords: Old Irish, Early Irish, Historical Languages, Dialect Detection

1. Introduction

For most of the 20th century, conventional wisdom in the field has been that Old Irish, the Gaelic language which survives from roughly the 7th to the 10th centuries, does not show significant signs of dialectal variation. Forms of variation certainly do occur in texts surviving from this period, and in 1985, McCone noted that "peculiarities" and "deviations" in the Old Irish glosses had "been known about for a long time". Such forms of variation as exist, however, have generally been attributed to reasons other than dialect. Thurneysen writes, for example, that "Linguistic differences in the Old Irish sources are almost all differences of period, and are the result of morphological development" (1946, 12). In other words, the language can be seen developing between the earliest and latest writings from the period, however, "Contemporary divergences, such as would point to dialectal peculiarities, are very rare..." (1946, 12). While affirming this position more recently, stating "The language encountered in Old Irish texts shows a surprisingly high degree of uniformity, with hardly any dialectal distinctions discernible," Stifter (2006, 10) goes on to assert that, nevertheless, "these certainly must have existed at the time." This echoes O'Rahilly's earlier claim that "it is not rash to assume that the Irish spoken in those times by the common people was not quite uniform in every part of the country" (1932, 248). Though O'Rahilly too admits that, "Of dialect differences in Old and Middle Irish we know nothing or next to nothing", earlier in the same work he makes the comparatively radical, and apparently contradictory statement, that "A perusal of the following pages will lead to the conclusion that historically there were but two main dialects in Irish, a Northern and a Southern..." (1932, 16).

Aside from the relatively small amount of Old Irish text surviving in contemporary sources, the problem of detecting Old Irish dialect is compounded by forms of text variation introduced at the editorial level. The orthographic representation of the language as it appears in manuscripts has been shown to be difficult to digitise accurately (Doyle et al., 2018), while the lack of a single digital editorial standard (Dereza et al., 2023), and of a single tokenisation or word separation standard (Doyle and McCrae, 2025a; Doyle et al., 2019), have been demonstrated to impact natural language processing (NLP) applications for Old Irish in other task areas. As has been discussed in Doyle and McCrae (2025a; 2025b), several extant repositories of Old Irish text follow their own discrete editorial and word separation guidelines (Griffith, 2013; Lash, 2014; Bauer et al., 2023; Stifter et al., 2021). This introduces variation between resources which is not indicative of original authorship in any way, and hence could likely impact stylometric analyses applied to texts drawn from more than a single text repository. Meanwhile, the quantity of text, and variety of sources, which have been edited in adherence to a single methodology (Doyle, 2018, 2023a,b) are too limited to be useful in a dialect detection experiment.

Section 2 of this paper will discuss the back-

Proceedings of Workshop on Dialects in NLP — A Resource Perspective (DialRes) , pages 33–47
16 May 2026.

ground to the problem of dialect in the field of Early Irish studies in more detail, while section 4 will discuss applications of stylometric techniques for the detection of dialect in other language areas. Section 3 will discuss existing text resources for Old Irish and their suitability to such stylometric applications. Finally, section 5 will describe an experiment carried out whereby stylometric analysis for the detection of dialects was applied to a selection of Old Irish texts drawn from the repository found to be most suitable for this type of application. The aim is not only to outline what the application of such techniques might suggest about authorship in the case of Old Irish, but to make the case that Old Irish text resources should be constructed in such a manner that they better facilitate the application of techniques such as these.

2. Background

It has been suggested that writings surviving in Old Irish may reflect a literary standard (Stifter, 2006, 10) rather than any particular dialect. Stifter (2009, 60) expands on this notion:

> There is little or no trace of synchronic variation in the Old Irish literary tradition, what variation exists being mostly stylistic rather than geographical (Kelly, 1982; McCone, 1989). This presupposes either the early adoption of a specific local variety as the basis for a standard, or the early codification of a standard grammar. The sporadic appearance already in the glosses of features of phonology, morphology and syntax which only become prominent in the Middle Irish period after the tenth century (McCone, 1985), suggests that the dominant register in these texts is a conservative literary standard at some remove from the spoken language, and perhaps one generation older than the earliest attested texts.

This description of Old Irish writings outlines two obstacles to the detection of dialect in the language. Firstly, little synchronic variation can be found. Secondly, even where synchronic variation may be apparent, it may not be indicative of dialect. Such variation could, for example, reflect a distinction in linguistic register, or scribal variation whereby a scribe copying an Old Irish text into a later manuscript may deliberately or inadvertently corrupt, alter or modernise some forms. Scribes may even introduce hypercorrections, attempting to reproduce unfamiliar, archaic forms, but introducing aberrations in the process. In order to be sure that any form of synchronic variation is indicative of dialect, therefore, it is necessary to be able to assign a geographical location to the variant in question. Murray (2005, 100) notes that this problem "is exacerbated by the fact that much of the material under investigation is anonymous."

Despite claims that "The language encountered in Old Irish texts shows a surprisingly high degree of uniformity..." (Stifter, 2006, 10), it should be noted that a significant amount of variation does nevertheless occur in Old Irish texts. Spelling variation, for example, is not uncommon in writings in historical languages, and the use of diacritics to mark long vowels and lenition in Old Irish specifically is inconsistent. *Corpus Palaeohibernicum* (CorPH) (Stifter et al., 2021), identifies 539 types of variation which occur within that text repository alone, and each of these is grouped into one of five distinct variation categories; phonological, orthographical, morphological, syntactical, and lexical[1]. The extent to which variation of any of these kinds might be useful in the identification of dialect requires further study, however, a handful of specific types of variation have been outlined in the literature as potential dialectal markers.

Kelly (1982) notes that, beside terms which are regularly used for certain animals, rarer words can also be found to refer to the same animals. Kelly posits that this is unlikely to be diachronic variation, as common terms, such as *bó* "ox/cow", *ech* "horse", *mucc* "pig", *cáera* "sheep", *gabor* "goat", and *sinnach* "fox", are used in the earliest Old Irish texts, and so cannot simply be later replacements for earlier terms which still appear, albeit rarely (1982, 86). Ahlqvist (1988, 34) later suggested 'it may be important to realise that the importance of the "prestigious forms of Old Irish" is largely a function of modern scholars' knowledge about the language that was codified in Thurneysen's (1946) Grammar', underscoring the potentially important role that non-standard forms, such as those outlined by Kelly, might play in identifying markers of Old Irish dialect. Accordingly, Ahlqvist points to variation in the form of the anaphoric pronoun (specifically *ón* versus *són*) in the Old Irish glosses as another potential marker of dialect.

An argument which may reasonably be levelled against either of these hypotheses is that, even if these forms of variation do represent genuine dialectal differences, it is nevertheless impossible to be sure of this, unless any of these forms can be reliably linked to a particular geographical region. In response to this, Murray (2005, 106) suggests that a study of "Placenames would seem to offer the most hope for geographically locating possible dialectal features in medieval Irish vocabulary."

[1] https://chronhib.maynoothuniversity.ie/chronhibWebsite/tables/variations?page=0&limit=100&fprop=&fval=&dtable=variations&ctable=&search=false

He argues that not only does variation in placenames allow for geographical localisation, as variants can be accurately pinpointed, but also that "placenames are liable to fossilise early forms of the language", thereby preserving quite old variants, without much fear they may have been standardised or modernised.

In one of the most comprehensive studies in this topic area to date, Malthaner (2022) claims to have achieved two aims. Firstly, "to seek out and examine synchronic variation within the Old Irish record" (2022, 273), and secondly, to identify "the existence of dialects within the material". Malthaner claims to have identified "a significant amount of variation", and hence, claims to have "dismantled the long-held academic belief that the Old Irish period was free from linguistic variation." Though she concedes that "... while a number of features were identified and investigated, the majority were found to not be conclusive evidence of diatopic variation" (2022, 274), Malthaner nevertheless suggests that "at least one feature [has] been determined to be a diatopic variant" (2022, 273). Malthaner also dismisses "the theory of Common Gaelic (Jackson, 1951) which has pervaded scholastic thought ... owing to the unlikelihood of an entirely invariant language being able to exist for such a significant period of time" and argues for the position that Old Irish was "a scholastic standard that was utilised by an educated elite ... deviations from which can be understood to be slips into more natural registers".

A significant amount of orthographic and script variation can occur in Old Irish manuscripts, even among the writings of individual scribes. When such writings are digitised, differing editorial standards can result in further variation being introduced, and several ways in which extant digital corpora for Old Irish are distinct from one another as a result of these factors are discussed in Doyle and McCrae (2025a, 1–4). While automatic dialect identification has been demonstrated to be effective in other languages (Alsuwaylimi, 2024; Bernier-Colborne et al., 2022; Imaizumi et al., 2022), the text resources available for Old Irish limit the types of techniques which are applicable, and it is conceivable that irrelevant forms of variation could impact applications which are otherwise feasible. As such, the following questions are worth asking:

1. How might the lack of editorial standardisation in Old Irish text resources impact such applications?

2. How might orthographic and script variation in manuscripts impact such applications?

3. What might be identified as potential markers of dialect by applicable NLP techniques?

4. How will any potential dialectal features which are automatically identified compare with extant research on Old Irish dialects?

3. Resources

Before question 1, posed in the preceding section, can be addressed, it is necessary to briefly discuss the types of text resources which are available for use in experiments like the one discussed in this paper. This will give an indication of the type and quantity of text data which is available from existing resources, as well as the state of non-standardisation between them. As has been noted by Dereza et al. (2023), no text editing standard exists for digital Old Irish editions. This results in editorial variation between texts which can impact downstream NLP applications. It has also been noted that word separation, and hence tokenisation, are not carried out in a uniform manner across different text resources for Old Irish either (Doyle and McCrae, 2025a). While it is not feasible to discuss the extent of this non-standardisation here, relevant examples will be provided below to demonstrate how this can be problematic.

The majority of text surviving in contemporary sources from the Old Irish period comes in the form of glosses. These typically take the form of short interlinear notes in manuscripts. Three large corpora exist comprising thousands of glosses, these are the Würzburg, Milan and St. Gall gloss collections. Otherwise, a small amount of prose and poetry also survives. Each of these three collections of glosses have been digitised (Griffith, 2013; Doyle, 2018; Bauer et al., 2023). Another resource, CorPH (Stifter et al., 2021), is the largest repository of annotated Early Irish texts, and incorporates two of these collections, Milan and St. Gall, along with many other Early Irish texts. All of these resources provide morphological annotation for their contents[2], however, no two of these resources make use of the same editorial standard or annotation style. Similarly, the first text repository containing Old Irish to be parsed, POMIC (Lash, 2014), uses a discrete POS tag-set (Doyle and McCrae, 2025a, 2). Finally, two Universal Dependencies (UD) treebanks exist, each containing a small number of glosses (Doyle, 2023a,b). As is discussed in Doyle and McCrae (2025b), each of these make use of the same editorial standard and POS tag-set as the digitised Würzburg collection (Doyle, 2018), however, with fewer than 100 parse trees each, these are two of the smallest treebanks in the UD collection. While all of the resources mentioned here provide some level of

[2]Though annotation is not complete for the Würzburg collection, it is stated in Doyle and McCrae (2025b, 395) that over 600 glosses have been annotated to date.

lexical annotation, it will be unsurprising given the current state of research into Old Irish dialects, that no existing corpus is annotated with labels to indicate dialectal differences.

While data annotation standards enforced by UD, alongside the diplomatic editing methodology employed during their creation (Doyle and McCrae, 2025b), might suggest that the Old Irish UD treebanks would be good candidates for use in an experiment like the one presented here, unfortunately, there simply isn't enough annotated data available in these resources. Similarly, while POMIC does contain texts from various sources, and there is somewhat more data than is available in UD treebanks, the total quantity is still quite small. Thus, though some research has suggested that POS-tags and syntactic data can have a predictive power in stylometric analyses (Eder and Górski, 2022), the only parsed corpora for Old Irish do not contain enough data to support meaningful analysis in this case.

Looking at the remaining annotated text data which is available for Old Irish, it becomes clear that any data used must be drawn from a single resource. Where different editorial standards are employed, it is shown in Doyle and McCrae (2025a, 2) that this can result in variation even between the same text content as presented by different resources. It is also demonstrated in Doyle and McCrae (2025a) that, even where raw text is initially the same between resources, when word separation is applied this can result in drastically different tokens being produced by different resources. This type of variation might easily be expected to impact stylometric analyses applied to such resources. While there may be a sufficient quantity of text data in any one of the three individual gloss collections (Griffith, 2013; Doyle, 2018; Bauer et al., 2023) to conduct stylometric analysis, if all text data is drawn from a single manuscript source the chances of finding meaningful, potentially dialectal distinctions would be significantly reduced. Thus, the best option is to utilise data from CorPH, the largest of the annotated resources, comprising texts drawn from multiple sources, all annotated using a single methodology.

While it is not possible to link the language used in any of these texts with any certainty to any particular geographical area, many of the manuscripts themselves have been linked to particular regions, or scriptoria. For example, the annals of Ulster and writings from the Book of Armagh, can reasonably be linked geographically with the general Ulster area. Similarly Hofman (1996, 19–23) argues that the manuscript containing the St. Gall glosses was created in either Nendrum or Bangor, hence, this can also be linked with Ulster. While manuscripts such as these may indeed represent a local variety of the language, this kind of geographical association alone does not necessarily ensure this, as scribes writing in these locations may have come to the area from elsewhere. Nevertheless, selecting texts linked with a specific region for use in the experiment described here remains the best available option for potentially isolating dialect markers.

4. Related Work

A variety of NLP approaches have demonstrated the potential for identifying dialect across a variety of languages. Alsuwaylimi (2024) discusses the use of two hybrid bidirectional long short-term memory models for dialect detection in Arabic, Bernier-Colborne et al. (2022) describe the use of multi-class support vector machine classifiers, as well as a probabilistic classifier for French, and Imaizumi et al. (2022) discuss a multi-task learning approach for Japanese dialect identification and multi-dialect automatic speech recognition using a transformer-based system. Several factors separate these approaches from the case of Old Irish, not least the quantity and standardisation of text data which is typically available for modern languages. The chief distinguishing factor, however, is that for each of these languages, it is already possible to identify specific dialects, and link them to geographical regions. As such, models can be trained and tested on dialect-labelled data. This is not possible for Old Irish, and so another approach is necessitated.

Several publications suggest that stylometric analyses can be employed reliably to identify features like dialect (Lahjouji-Seppälä et al., 2022), and register (Grieve, 2023). Lahjouji-Seppälä et al. (2022) demonstrate the utility of a stylometric approach to analyse the diachronic development of 20th century Standard Ukrainian, and Grieve (2023) suggests that stylometric analysis is consistent with standard theories of register variation. Stylometric methods are even demonstrated to be useful in separating AI generated Arabic text from that of humans, including dialectal social media posts, by identifying stylistic features consistent with large language model text generation (Al-Shaibani and Ahmed, 2025). Moreover, Goswami et al. (2020) describe a method for simultaneously learning sentence embeddings and clustering assignments from short texts for the purpose of dialect identification, advancing the state-of-the art in both supervised and unsupervised settings. This is particularly relevant for texts like Old Irish glosses, which are typically very short.

A stylometric approach engineered to separate texts based on dialectal features is the best available approach for an Old Irish application, as it does not necessitate the use of dialect-labelled

data for training or testing of models. While this means that the accuracy of such an approach cannot be tested for Old Irish, that is not the primary aim. Instead, the main focus is to determine what impact variation within extant text resources will have on the application of an experiment intended to detect dialect. A stylometric approach will allow the text resource to be treated as a variable, with the same approach being applied to raw text, and to word-separated text from the same repository. As has been discussed in Doyle and McCrae (2025a), in Old Irish orthography spacing is not inserted systematically between all words, and the application of word separation in CorPH results in the addition of many characters, and even duplication of many morphs[3], such that the resulting tokens cannot be directly aligned with the original text. As such, it is possible to apply a single stylometric technique to both the original text, and to the recombined tokens, to demonstrate the impact such editorial variation can have on such an application. This, in turn, can indicate how the standard of data available from text resources could impact other dialect detection applications for Old Irish.

5. Experiment

To date, traditional scholarship has been focussed primarily on the lexical level, however, dialectal difference can manifest in orthography, morphology, syntax and sub-word character sequences. Even looking at lexical choice, however, use of function words, particles and clitics can be telling, but these do not receive as much focus in the literature for Old Irish. The experiment described below applies a stylometric clustering approach intended to identify latent dialect structure in a collection of short texts, without relying on labelled training data or predefined dialect categories. This approach should be capable of identifying systematic differences in lexical choice, orthography, and morphosyntactic patterns which, among other things, may be indicative of dialect.

5.1. Methodology

Taking each sentence, gloss, or verse in the corpus to be a text instance, the aim is to find statistical regularities across these instances. Each instance is assumed to be capable of encoding diagnostic dialectal cues (such as variant spellings, function words, particles, clitic forms) alongside other, irrelevant forms of variation. Texts were drawn from CorPH only, and all text instances were extracted from only the following list of texts:

1. Annals of Ulster
2. Armagh Additamenta
3. Armagh Glosses Minor Glosses
4. Armagh Liber Angeli
5. Armargh Notulae [sic]
6. Milan Donatus Minor Glosses
7. Milan Glosses
8. Milan Priscian
9. Milan Sententia Sanctorum Doctorum Minor Glosses
10. Poems of Blathmac
11. St Gall Priscian
12. St Gall Prudentius Minor Glosses

These specific texts were chosen either because they can be somewhat reliably ascribed to a named geographical location, can be ascribed to a named author (hence, they represent a single dialect), or because they are one of the large collections of Old Irish glosses. All text instances were added to a list, and labelled with the name of the text from which they were taken. These labels were not used during clustering, and were retained for visualisation purposes only.

Sentence embedding was carried out using a pre-trained, multilingual sentence embedding model[4] from the `SentenceTransformers` framework. This pre-trained model was chosen as no labelled dialect data was available. Having been pre-trained on large multilingual corpora, the aim was to avail of the model's transfer learning capabilities. This should allow the model to apply its general understanding of semantic and syntactic structure to the task of encoding Old Irish text instances, even without having ever been trained on Old Irish. The use of sentence embeddings allows for syntactic structure to be taken into account as well as lexical features. Text instances are mapped to 768-dimensional vector space by this model, and cosine distance is used to compute similarity between instance vectors.

[3] CorPH refers to elements produced during word-separation in that corpus as "morphs" rather than "tokens". However, as many of these "morphs" represent entire words, including any inflections, this use of the term is atypical. The term "morph", where used in this paper, is intended in the more traditional sense of a word segment which represents the smallest unit of language that has meaning. To avoid confusion, the "morphs" of CorPH will be referred to here as tokens.

[4] `paraphrase-multilingual-mpnet-base-v2` https://huggingface.co/sentence-transformers/paraphrase-multilingual-mpnet-base-v2

Unsupervised clustering was carried out using Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN). HDBSCAN has a few benefits which make it more useful for this application than other clustering methods. Firstly, because the number of potential dialects is unknown, it was not desirable to predefine a number of clusters. Secondly, HDBSCAN labels text instances which do not clearly belong to any dense region as noise rather than assigning them to any cluster. With potentially thousands of noisy text instances, this allows for irrelevant instances to be removed so that only relevant clusters are represented when the points are plotted. The minimum cluster size was set to 15 and minimum samples were set to 5 to ensure a lower number of clusters, increasing the likelihood that clusters will represent something closer to the scale of dialect, rather than that of authorship or turns of phrase.

Dimensionality reduction was carried out using UMAP (McInnes et al., 2020). UMAP was chosen for its ability to preserve both local and global structures better than alternatives like t-SNE, while compressing noise and redundant dimensions. Moreover, because dialect continua may be less likely to form linearly separable clusters than other stylometric classifications, UMAP might be expected to better capture gradual transitions between varieties as it approximates the underlying manifold structure of the data. Embeddings are projected to two dimensions after dimensionality reduction, and plotted. Each point plotted represents an individual text instance which the HDBSCAN process deemed significant enough to be assigned to a specific cluster, i.e. not noise. Plotted instances are colour coded to identify them with the corpus they were originally drawn from.

5.2. Variants

Four further variants of this experiment were also carried out to see how altering certain variables would affect results. Each of these variants were set up as per the methodology outlined above, with the only difference being to the variable in question.

5.2.1. Variant 1: TF-IDF

Term Frequency-Inverse Document Frequency (TF-IDF) was used instead of a sentence embedding model to quantify the importance of individual terms in a document given the frequency with which they appear in that document. The TF-IDF for a given term, *t*, in a given document, *d*, taken from a collection of many documents, *D*, is calculated as follows:

$$tfidf(t, d, D) = tf(t, d) \times idf(t, D)$$

As the primary focus of this approach is on the lexical level, measuring the frequency with which specific strings of characters (i.e. terms) occur within a document, it may be more sensitive to spelling variation than an embedding model. For a historical language like Old Irish, the considerable amount of arbitrary variation typical of texts might therefore impact the usefulness of a TF-IDF approach in an application like this. On the other hand, the ability of embedding models to detect semantically similarity, and hence interpret two texts as saying the same thing even when written differently, may cause them to overlook genuinely meaningful forms of variation in a task such as this.

The minimum document frequency for a term was set to 2, and the n-gram range was set between 1 and 2 to allow the measure to take syntactic context into account also. Using the `TfidfVectorizer` from `scikit-learn` (Pedregosa et al., 2011), TF-IDF embeddings were produced. All other parameters remained the same as in the methodology described above.

5.2.2. Variant 2: Character-level TF-IDF

Because word-level TF-IDF only looks at the frequency with which word-tokens appear, it cannot account particularly well for features like spelling variation, morphology and punctuation, all of which might indicate dialectal variation. For this reason, character level TF-IDF was also applied. This looks at combinations of characters rather than entire words within a text instance. The minimum document frequency was again set to 2, but the n-gram range was set between 2 and 5. Otherwise parameters were as per the first variant, above.

5.2.3. Variant 3: t-SNE

Because some amount of separation seemed apparent between clusters for certain corpora, while other clusters appeared crowded together (see discussion below), one variant of this experiment was carried out using t-Distributed Stochastic Neighbor Embedding (t-SNE) for dimensionality reduction instead of UMAP. According to van der Maaten and Hinton (2008), t-SNE "produces significantly better visualizations by reducing the tendency to crowd points together in the center of the map." The hope was that this method might separate some of the more crowded clusters apart in the plot, making it easier to visualise potential points of dialectal difference. Perplexity was set to 30, and cosine similarity was used to measure high-dimensional similarities between data points. Otherwise, the embedding model, and all other variables, were as per the methodology described above.

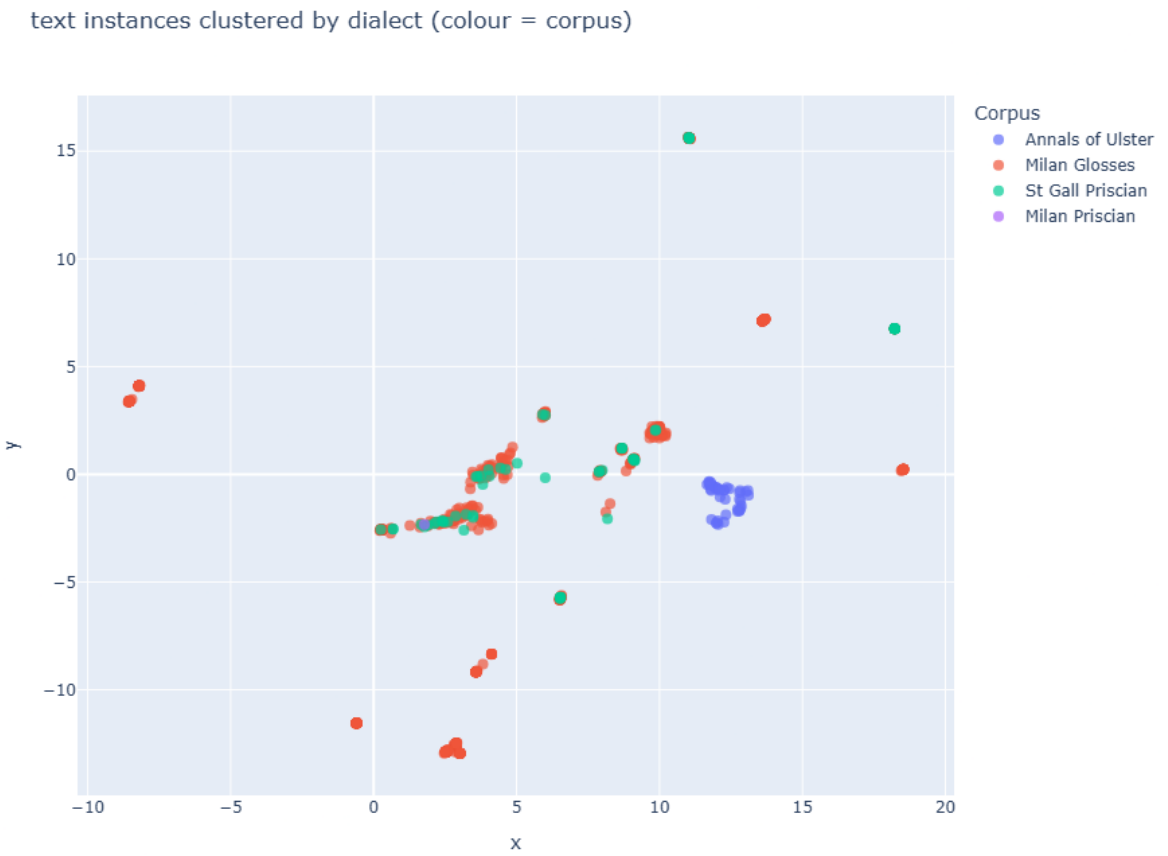

Figure 1: Plot of Dialect Clusters for Old Irish Using Sentence Embeddings.

5.2.4. Variant 4: Recombined Tokenised Text

As has been discussed in section 3, tokens resulting from CorPH's word-separation process do not necessarily match the original text exactly if recombined. A fourth variant of the experiment was set up to investigate how this kind of preprocessing would affect results. CorPH tokens for each text instance were concatenated with spaces between each token. While this may not result in new text instances which resemble actual Old Irish text very closely, words are nevertheless clearly separated, which might make it easier for the embedding model to identify differences at the lexical level which could indicate dialect. To exemplify the distinction, one text instance drawn from the Milan Glosses (Ml. 113a9), and used in all other experiment variants, reads *duróscai* in its original form. Recombining the tokens resulting from CorPH's word-separation as described here, however, results in the string, "`du· ·ró ós du·róscai`". Aside from the text input, all other variables and parameters for this experiment variant were as described in the methodology section above.

5.3. Results and Discussion

Figure 1 shows the results of the clustering process described in the methodology section above, after text instances were plotted. What is immediately clear is that a lack of data impacted the experiment. Once noisy data was removed, clusters were found to be comprised of data from only four collections; Annals of Ulster, Milan Glosses, St Gall Priscian, and Milan Priscian. As can be seen, contents from the Annals of Ulster form their own cluster to the centre-right of the plot, while another large cluster to the centre-left is comprised of data from the St. Gall Priscian, the Milan Glosses, and a single data point from the Milan Priscian. It must be kept in mind that these clusters could indicate stylistic features other than dialect, for example, linguistic register or text subject matter. Whatever the underlying factors though, it is significant that in a stylometric experiment geared towards dialect identification, these clusters remain after noise removal, and they are the largest two clusters. They account for more than simple similarity in lexical choice.

Several smaller clusters can be found around the edges of the plot, comprised of text instances from the Milan Glosses and St Gall Priscian. These are very telling of the impact that both manuscript and editorial variation can have on an experiment of this sort. Clusters 10 (-1, -12), 12 (-8.5, 4.5), 13 (3, -12.5) and 15 (1, -2.5) are primarily comprised of variant forms of the same word, with spelling variation. All instances in cluster 10 are either *.i. adǽ* or *ostú .i. adǽ*, and all instances in 12 are, *adæ*. Many instances of *.i. adæ* and *.i. adae* were lumped into the leftmost end of cluster 15, though the cluster itself continues as far to the right as (4, -1.5), and contains a wide variety of other text instances also. Finally, cluster 13 comprises the forms *adǽ, ádǽ, adé,* and *ádæ*. The fact that these forms are numerous enough to comprise four clusters, suggests that this term may be a significant feature of style in the Milan Glosses, however, the fact that they do not all appear in a single cluster suggests that irrelevant variation has impacted these results. It is possible that standardisation of manuscript spelling variation could result in clearer findings.

None of the features of style plotted in these clusters can be clearly likened to anything in the dialectal literature for Old Irish to date. Nevertheless, the considerable overlap between the St. Gall Priscian and Milan Glosses would seem to confirm arguments that little dialectal variation can be found between these Old Irish sources. While both are corpora of glosses, their subject matter is very different. St. Gall deals with the grammar of Latin, while Milan relates to the Psalms. As such, the overlap cannot be attributed simply to similarity in genre, and there does appear to be some stylistic similarity shared between these two collections which is not shared by the Annals of Ulster. While the separation of the Annals of Ulster from the gloss corpora in the plot might be significant stylistically, it may equally reflect the inherent distinction in style between glosses and annal entries. Alternatively, it could also reflect stylistic differences indicating diachronic language variation rather than dialectal.

Because of limited space, plots displaying the results for the experiment variants can be found in the Appendix.

5.3.1. Results for Variant 1: TF-IDF

The results from the first variant of the experiment can be seen in Appendix A. In this variant, TF-IDF was used to measure stylistic similarity rather than an embedding model. Comparing these results against those in Figure 1, this method appears to have had more difficulty separating text instances into meaningful clusters. The result is a single, large grouping of data points towards the right of the plot, containing text instances from all plotted corpora, including the Annals of Ulster. It may be noteworthy, however, that after noise reduction text instances from more corpora remained in this plot than that of the embedding model. As with Figure 1, the same outlying, small clusters can be found here, i.e. *adǽ*, *.i. adǽ*, *adæ*, etc.

5.3.2. Results for Variant 2: Character-level TF-IDF

Appendix B shows the results of the second variant of the experiment, which utilised character-level TF-IDF as a measure. Here, the Annals of Ulster are separated relatively clearly again into their own cluster, centred around (1, 1). Again, instances from most other corpora form into a relatively large grouping towards the bottom, centre of the plot, with small, outlying clusters surrounding. One notable feature of this plot is that content from the Poems of Blathmac is relatively tightly grouped around the (6.3, 8.6) mark, seemingly indicating either strong markers for the poet's own personal style, or simply a distinction in genre between poetry, glosses, and annal entries.

5.3.3. Results for Variant 3: t-SNE

As predicted, clusters are spread further apart in the plot resulting from dimensionality reduction using t-SNE (Figure 4). This results in several smaller clusters comprised primarily of content from both St. Gall Priscian and the Milan Glosses. Outliers from the other plots now form much tighter clusters at (50, -35), (85.5, 6.5), and (93,-13.5). Otherwise, the contents of the Annals of Ulster remain separated from those of other corpora, comprising a single cluster at the centre, left. This plot does make the visualisation of distinct clusters more intuitive, though the overlap of two or more corpora in several large clusters underscores the fact that the underlying embedding model still considers these corpora to be stylistically similar.

5.3.4. Results for Variant 4: Recombined Tokenised Text

The extra lexical information available to the model in the fourth variant of the experiment appears to have resulted in much tighter groupings, with clusters being more definitively separated from each other in the plot (Figure 5). Still, there remains significant crossover between the contents of the St. Gall Priscian and Milan Glosses corpora, while content from the Annals of Ulster again forms an entirely separate cluster in the bottom centre of the plot. Even the outliers remain the same, albeit, with spacing now introduced between elements. Cluster 22 (17.8, 1.5) is comprised of *a dǽ, a dé,* and *.i. a dé*, 11 (23.5, 6) is comprised of *a dæ* and

á dæ, while 12 (21.3, 2) is entirely comprised of *.i. a dæ*.

6. Future Work

A major limitation of the work presented here is that the accuracy of any of the models described cannot be quantified, as this would require existing Old Irish text resources to be already annotated with dialect labels. In lieu of this, a valuable avenue for future research would be to apply the same stylometric techniques described in this paper to the modern Gaelic languages. Because dialectal distinctions in Modern Irish are well understood, for example, it would be possible to assess the accuracy of stylometric techniques on such a corpus. Though modern and historical forms of Irish are distinct in many ways, and so it is not possible to extrapolate that any measured accuracy would hold between the two, this would nevertheless allow researchers a better understanding of the relationship between clusters and genuine dialects in a very closely related language. This knowledge could then be used to improve understanding of applications to Old Irish, informing the tuning of hyperparameters and reading of results.

Another potentially valuable area for future investigation into dialect detection for Old Irish would be training of models to identify texts known to have been produced in different geographical areas. This would allow the accuracy of the dialect detection tool to be assessed in a way which was not possible here. If sufficiently accurate, features deemed to be likely representative of dialect by such a model could then be examined. To create such a model, however, would first require the production of digital editions of texts like the West Munster Synod, which are known to have been produced in areas which are geographically distinct from those which were available for use in this research. More importantly, as this research has shown, it would be important that the production of such texts is standardised to minimise the impact of editorial variation on downstream NLP techniques like dialect detection.

7. Conclusion

This paper has conducted a stylometric dialect identification experiment on Old Irish text. The aim was not to identify specific features which can be certainly assigned to dialect in Old Irish, as separation of clusters alone cannot be taken as an indication of dialect. Rather, the aim here was to find features of style which do separate corpora, so that these features can be examined in more detail by experts to determine the reasons clusters were separated. Where corpora like the Annals of Ulster, or the Poems of Blathmac – which can be linked to either a geographical region or an individual author – form tight clusters in the plots, this may indicate the presence of some stylistic features indicative of dialect. This, however, will require further examination to determine. Nevertheless, the results presented here appear to show a clear distinction, in several experiment variants, between a single cluster containing text instances from the Annals of Ulster, and several other clusters in which text instances from the Milan Glosses and St Gall Priscian are regularly mixed. Unfortunately, too little data was available for other corpora for their contents to have been equally heavily represented in any of the plots after noise was removed, though some methods examined here clearly resulted in more aggressive noise removal than others.

Another pressing aim of this research was to determine the impact of editorial non-standardisation on the application of techniques such as this to a historical language, like Old Irish. It was clear that plots differed even where the same techniques were applied, and the only variable was the standard of the text. Word-separated text, which was recombined using spacing, produced clusters which were much more clearly defined than those produced using the original text. This underlines the impact that editorial decisions can have. A relatively major finding was that, even in the case of an apparently significant stylistic feature, editorial non-standardisation resulted in the generation of up to four spurious text clusters where, at most, one was required. As such, this paper reiterates the point made by Dereza et al. (2023), that it is important to to investigate the potential value of deciding a digital text editing standard for Old Irish to increase the efficacy of NLP applications to the language.

To summarise the value of the work presented here, the primary achievement has been in showing that the only currently viable techniques which might be applied to Old Irish texts in an effort to identify dialectal features can be heavily impacted by the editorial standards applied in the production of text resources. The paper has also shown a strong tendency for text instances from two large gloss corpora to be clustered together, while annal entries remained separated throughout most of the experimentation conducted. Similarly, the existence of several small clusters unique to only one of the large gloss collections suggests some stylistic features unique to that corpus, whether they be dialectal or not. All of these preliminary findings may be useful in informing future dialectal investigations by more traditional means, by drawing attention to stylistic features which have not been the focus of attention to date.

8. Acknowledgements

I would like to express my gratitude to Prof. David Stifter for kindly reviewing this paper prior to publication, and for providing very helpful insights.

This publication has emanated from research conducted with the financial support of Taighde Éireann – Research Ireland under Grant number IRCLA/2023/2124.

9. Limitations

As mentioned above, not enough data was available, edited and annotated to the same standard. The experiment would have benefitted from the inclusion of the third large corpus of Old Irish glosses, Würzburg, however, these are not yet available in CorPH. Other corpora which can be assigned to specific locations outside of Ulster would have also improved the outcome of this experiment, for example, if the West Munster Synod were available through CorPH.

10. Bibliographical References

Anders Ahlqvist. 1988. Remarks on the Question of Dialects in Old Irish. In Jacek Fisiak, editor, *Historical Dialectology: Regional and Social*, pages 23–38. De Gruyter Mouton, Berlin, New York.

Maged S. Al-Shaibani and Moataz Ahmed. 2025. The Arabic AI Fingerprint: Stylometric Analysis and Detection of Large Language Models Text.

Amjad A. Alsuwaylimi. 2024. Arabic Dialect Identification in Social Media: A Hybrid Model with Transformer Models and BiLSTM. *Heliyon*, 10(17).

Gabriel Bernier-Colborne, Serge Leger, and Cyril Goutte. 2022. Transfer Learning Improves French Cross-Domain Dialect Identification: NRC @ VarDial 2022. In *Proceedings of the Ninth Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 109–118, Gyeongju, Republic of Korea. Association for Computational Linguistics.

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

Adrian Doyle and John P. McCrae. 2025a. 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 and John P. McCrae. 2025b. Development of Old Irish Lexical Resources, and Two Universal Dependencies Treebanks for Diplomatically Edited Old Irish Text. In *Proceedings of the 5th International Conference on Natural Language Processing for Digital Humanities*, pages 393–402, Albuquerque, USA. Association for 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.

Maciej Eder and Rafał L. Górski. 2022. Stylistic Fingerprints, POS-tags, and Inflected Languages: A Case Study in Polish. *Journal of Quantitative Linguistics*, 30(1):86–103.

Koustava Goswami, Rajdeep Sarkar, Bharathi Raja Chakravarthi, Theodorus Fransen, and John P. McCrae. 2020. Unsupervised Deep Language and Dialect Identification for Short Texts. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1606–1617, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Jack Grieve. 2023. Register Variation Explains Stylometric Authorship Analysis. *Corpus Linguistics and Linguistic Theory*, 19(1):47–77.

Rijcklof Hofman. 1996. *The Sankt Gall Priscian Commentary. Part 1. Volume 1: Introduction; Book 1-5*. Nodus, Münster.

Ryo Imaizumi, Ryo Masumura, Sayaka Shiota, and Hitoshi Kiya. 2022. End-to-end Japanese Multi-dialect Speech Recognition and Dialect Identification with Multi-task Learning. *APSIPA*

Transactions on Signal and Information Processing, 11(1):1–23.

Kenneth Jackson. 1951. 'Common Gaelic': the Evolution of the Goidelic Languages. *Proceedings of the British Academy*, 37:71–97.

Patricia Kelly. 1982. Dialekte im Altirischen? In W. Meid, H. Ölberg, and H. Schmeja, editors, *Sprachwissenschaft in Innsbruck*, pages 85—89. Innsbrucker Beiträge zur Kulturwissenschaft, Innsbruck.

M.Z. Lahjouji-Seppälä, A. Rabus, and R. von Waldenfels. 2022. Ukrainian Standard Variants in the 20th Century: Stylometry to the Rescue. *Russian Linguistics*, 46(3):217–232.

Ariana Nicole Malthaner. 2022. *Synchronic Language Variation in the Old Irish Glosses*. Trinity College Dublin, Dublin. PhD Thesis.

Kim McCone. 1985. The Würzburg and Milan Glosses: Our Earliest Sources of 'Middle Irish'. *Ériu*, 36:85–106.

Kim McCone. 1989. Zur Frage der Register im frühen Irischen. In Stephen N. Tranter and Hildegard L. Tristram, editors, *Mündlichkeit und Schriftlichkeit in der frühen irischen Literatur*, pages 57–97. Gunter Narr Verlag, Tübingen.

Leland McInnes, John Healy, and James Melville. 2020. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction.

Kevin Murray. 2005. Dialect in Medieval Irish? Evidence from Placenames. *Studia Celtica Fennica*, 2:97–109.

Thomas F. O'Rahilly. 1932. *Irish Dialects Past and Present*. Dublin Institute for Advanced Studies, Dublin.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

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

David Stifter. 2009. Early Irish. In Martin J. Ball and Nicole Müller, editors, *The Celtic Languages*, chapter 4, pages 55–116. Routledge, Abingdon, New York.

Rudolf Thurneysen. 1946. *A Grammar of Old Irish*, 2 edition. The Dublin Institute for Advanced Studies, Dublin.

Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(86):2579–2605.

11. Language Resource References

Bauer, Bernhard and Hofman, Rijcklof and Moran, Pádraic. 2023. *St Gall Priscian Glosses, version 2.1*.

Doyle, Adrian. 2018. *Würzburg Irish Glosses*.

Doyle, Adrian. 2023a. *Diplomatic St. Gall Glosses Treebank*. Universal Dependencies.

Doyle, Adrian. 2023b. *Diplomatic Würzburg Glosses Treebank*. Universal Dependencies.

Griffith, Aaron. 2013. *A Dictionary of the Old-Irish Glosses*.

Lash, Elliott. 2014. *The Parsed Old and Middle Irish Corpus (POMIC). Version 0.1*. The Dublin Institute for Advanced Studies.

Stifter, David and Bauer, Bernhard and Lash, Elliott and Qiu, Fangzhe and White, Nora and Barrett, Siobhán and Griffith, Aaron and Bulatovas, Romanas and Felici, Francesco and Ganly, Ellen and Nguyen, Truc Ha and Nooij, Lars. 2021. *Corpus PalaeoHibernicum (CorPH) v1.0*. National University of Ireland, Maynooth.

Appendix

A. Experiment Variant 1: TF-IDF

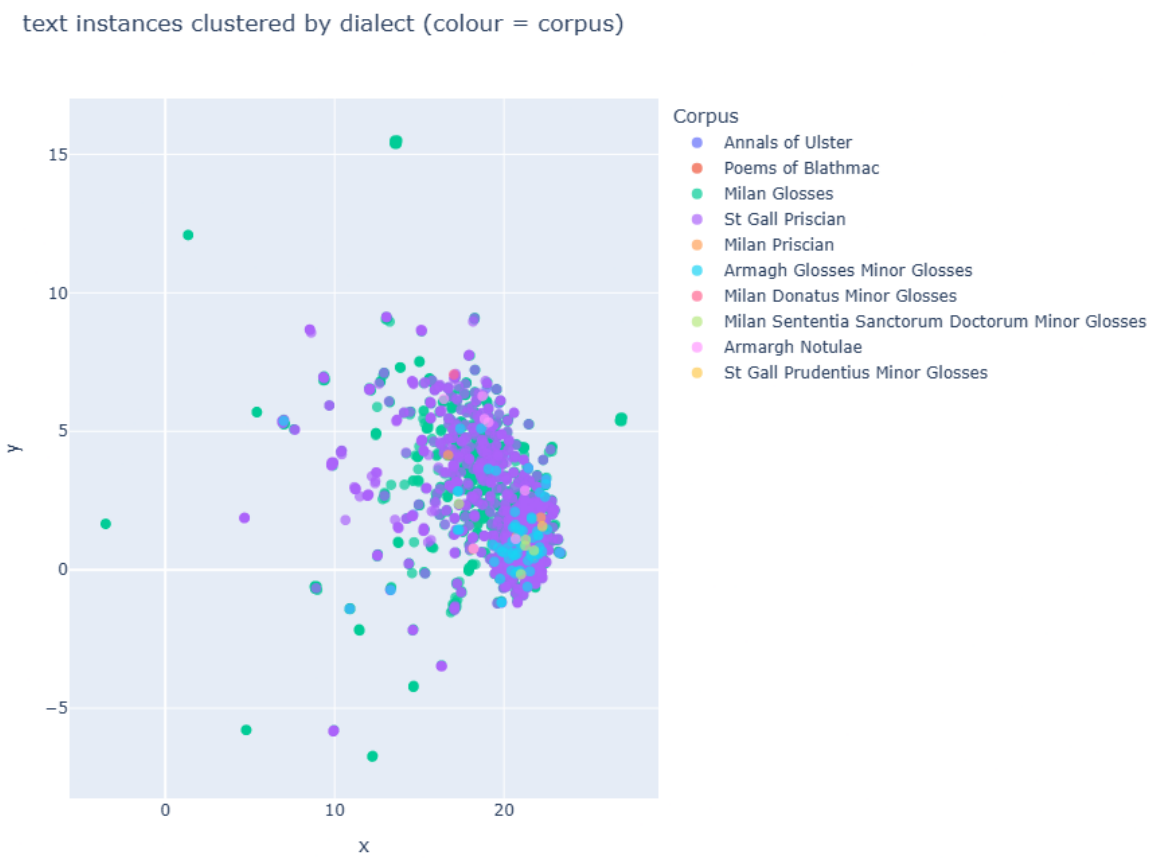

Figure 2: Plot of Dialect Clusters for Old Irish using TF-IDF.

B. Experiment Variant 2: Character-level TF-IDF

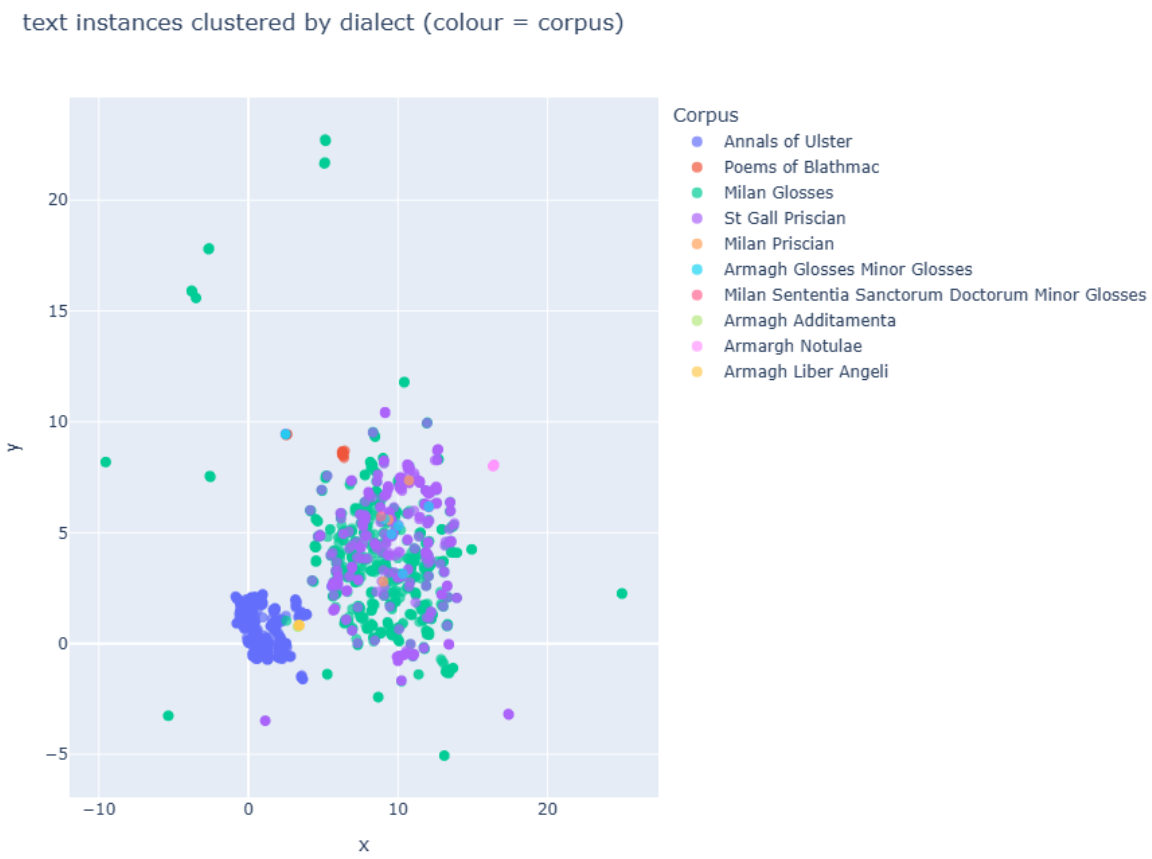

Figure 3: Plot of Dialect Clusters for Old Irish using Character-level TF-IDF.

C. Experiment Variant 3: t-SNE

Figure 4: Plot of Dialect Clusters for Old Irish using t-SNE.

D. Experiment Variant 4: Recombined Tokenised Text

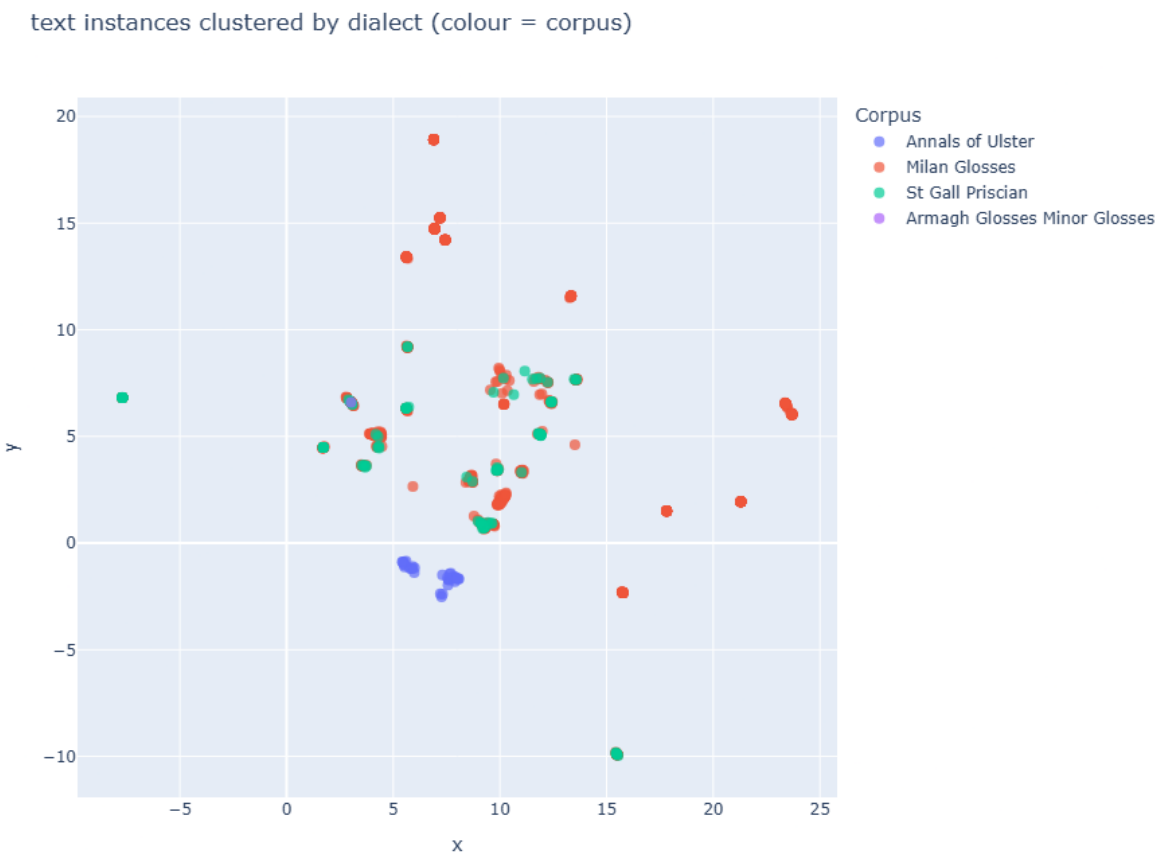

Figure 5: Plot of Dialect Clusters for Word-separated Old Irish Text.